

**RNA Topology as a Framework for Structure, Dynamics and  
Adaptation**

By

Maximillian Honorio Bailor

A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
(Chemistry)  
in The University of Michigan  
2010

Doctoral Committee:

Professor Hashim M. Al-Hashimi, Chair  
Professor Ayyalusamy Ramamoorthy  
Associate Professor Heather Carlson  
Associate Professor Anna K. Mapp

“Genius begins great works; labor alone finishes them.”

–Joseph Joubert

“Here is a test to find whether your mission on earth is finished: If you are alive, it isn't.”

–Richard Bach

© 2010 by Maximillian Honorio Bailor

For my namesakes Alfred Maximillian Bailor and Madam Honoria Caulker-Bailor.

## Acknowledgements

First and foremost, I would like to thank my advisor, Hashim Al-Hashimi, for all the time, support and advise that he has provided over the years. I would also like to thank my committee members: Professors Heather Carlson, Anna Mapp and Ayyalusamy Ramamoorthy. Additionally, many thanks to all members of the Al-Hashimi group, past and present. They are a great group of people and I have been very fortunate to know and work with all of you. In particular, thanks to Qi Zhang who got me started in the lab, and Xioayan Sun and Anette Casiano-Negroni who have both been the older sisters I never had... or in Anette's case, younger.

Most importantly, I want to thank and recognize the contributions of my family. My parents, Karen and Hilton, and my Brother, Remy who have all been so supportive and wonderful, I can't thank you enough for your love. Thanks to the Murch-clan (Bruce, Nancy, Sam, Cecilia and Will), who have been a home away from home. Additionally, Ann Arbor has been a wonderful place to live and work over the last six years. I am very grateful for, and will miss immensely, all of the people – chemistry and non-chemistry – who I have had the fortune to know, and in whose lives I have shared. Above all, I would like to thank Josh Cope, Marie Tillema Cope, Steve Dumais, Theresa Finney-Dumais and Amy Payeur.

Last, and certainly not least, I want to recognize my Boo-Bear, Diedre, who has been a constant source of love, support, understanding and encouragement over this journey.

## Table of Contents

Dedication .....	ii
Acknowledgements .....	iii
List of Figures .....	vii
List of Tables .....	ix
List of Appendices .....	x
Abstract .....	xi
<b>Chapter 1 Introduction .....</b>	<b>1</b>
1.1 RNA structure determination .....	1
1.2 Local RNA structure .....	5
1.3 RNA secondary structure and nomenclature .....	7
1.4 Prediction of secondary structure .....	8
1.4.1 Turner rules .....	9
1.4.2 Comparative sequence analysis and covariation algorithms .....	10
1.5 Prediction of three-dimensional RNA conformations from secondary structure .....	11
1.5.1 RNA Fragment-based comparative modeling .....	12
1.5.2... Restraint-based comparative modeling .....	14
1.6 Dynamics and Adaptation .....	15
1.7 Global RNA topology .....	17
1.8 References .....	27
<b>Chapter 2 Characterizing the relative orientation and dynamics of RNA A-form helices .....</b>	<b>38</b>
2.1 Introduction .....	38
2.2 Long-range orientational information from residual dipolar couplings .....	40
2.3 Limits of applicability and practical considerations .....	43

2.4	Experimental determination and modeling of RNA helices .....	46
2.4.1	Experimental validation of predicted Watson-Crick base pairs.....	46
2.4.2	Measurement of RDCs .....	46
2.4.3	Order tensor analysis .....	49
2.4.4	Determining the average inter-helical alignment.....	50
2.4.5	Characterizing inter-helix motions .....	52
2.5	Structure and dynamics of the HIV-1 transactivation response element .....	52
2.6	Conclusions.....	53
2.7	References .....	64
<b>Chapter 3 Linking RNA Topology to Structure and Dynamics .....</b>		<b>72</b>
3.1	Introduction .....	72
3.2	Defining two-way junctions .....	73
3.3	PDB search of RNA two-way junctions .....	74
3.3.1	Computing inter-helical angles for arbitrary RNA two-way junctions .....	75
3.4	Computing topologically allowed inter-helical orientations .....	78
3.5	Distribution of bulge-type two-way junctions.....	80
3.6	Prediction of internal loop conformations.....	81
3.7	Conclusions.....	82
3.8	References .....	90
<b>Chapter 4 RNA Topology and Adaptation: The Rules Behind RNA-Ligand</b>		
<b>Selection.....</b>		<b>92</b>
4.1	Introduction .....	92
4.2	Preparation and purification of <sup>13</sup> C/ <sup>15</sup> N labeled HIV-1 TAR RNA.....	94
4.3	Chemical shift perturbation mapping and titrations .....	95
4.4	Chemical shift intensities .....	97
4.5	Measurement of residual dipolar couplings and order-tensor analysis .....	99
4.6	Determination of inter-helical Euler angles.....	101
4.7	TAR-Ligand relationships.....	102
4.8	Universality of size-encoded conformational selection .....	103
4.9	Conclusions.....	104
4.10	References .....	120
<b>Chapter 5 Conclusions and Future Directions.....</b>		<b>124</b>

**5.1 Conclusions and Future.....124**  
**5.2 References .....127**



## List of Figures

<b>Figure 1.1: Important achievements and milestones in the history of RNA structure determination.....</b>	<b>21</b>
<b>Figure 2.1: The field dependant orientation and distance dependence of RDCs. ....</b>	<b>57</b>
<b>Figure 2.2: Determining the relative orientation and dynamics of A-form helices using an order tensor analysis of RDCs. ....</b>	<b>58</b>
<b>Figure 2.3: Typical RDCs measured in base and sugar moieties of RNA using the pulse sequences listed in Table 2.1. ....</b>	<b>59</b>
<b>Figure 2.4: Implementation of strategy to determine the relative orientation and dynamics of two helices for the free state of HIV-1 TAR.....</b>	<b>60</b>
<b>Figure 2.5: The relative orientation (interhelical bend, <math>\beta_h</math>, and twist angle, <math>\zeta</math>) and dynamics (<math>\vartheta_{int}</math>) of RNA helices obtained from order tensor analysis of RDCs in various contexts. ....</b>	<b>61</b>
<b>Figure 3.1: Topological confinement of RNA inter-helical orientations across <math>H_{.3}S_XH_{.3}S_0</math> bulges. ....</b>	<b>85</b>
<b>Figure 3.2: Effects of topological constraints on helix length and relative orientation of sequentially arranged two-way junctions.....</b>	<b>86</b>
<b>Figure 3.3: Topological confinement of RNA inter-helical orientations across <math>H_{.3}S_XH_{.3}S_Y</math> internal loops. ....</b>	<b>87</b>
<b>Figure 4.1: HIV-1 transactivation element (TAR) structure adaptation pinwheel.....</b>	<b>106</b>

<b>Figure 4.2: Size-encoded selection of TAR inter-helical orientations using aminoglycosides.....</b>	<b>107</b>
<b>Figure 4.3: Titration curves as a function of total aminoglycoside concentration.....</b>	<b>108</b>
<b>Figure 4.4: Chemical shift perturbations, intensities and residual dipolar couplings and derived models for aminoglycoside-bound TAR complexes.....</b>	<b>109</b>
<b>Figure 4.5: Chemical structures of ligands from complexes of bound two-way junction RNA elements.....</b>	<b>111</b>
<b>Figure 4.6: Generalized size-encoded RNA conformational selection using small molecules.....</b>	<b>112</b>
<b>Figure 4.7: Control examples of generalized size-encoded RNA conformational selection using small molecules.....</b>	<b>114</b>
<b>Figure 4.8: Inter-helical Euler angles of A-site in the context of the 16S ribosome bound to a variety of aminoglycosides.....</b>	<b>115</b>
<b>Figure 4.9: Generalized electrostatic RNA conformational selection using small molecules.....</b>	<b>116</b>

## List of Tables

<b>Table 1.1: Secondary structure prediction from primary sequence.</b> .....	<b>22</b>
<b>Table 1.2: Secondary structure prediction algorithms.</b> .....	<b>23</b>
<b>Table 1.3: Comparative modeling programs.</b> .....	<b>23</b>
<b>Table 1.4: Three-dimensional structure prediction.</b> .....	<b>26</b>
<b>Table 2.1: Pulse sequences implemented in the measurement of scalar and residual dipolar couplings in nucleic acids.</b> .....	<b>62</b>
<b>Table 2.2: Established media used to align nucleic acids in structural studies.</b> .....	<b>63</b>
<b>Table 2.3: Chemical shift changes for protonated carbons and nitrogens due to RCSA contributions from alignment assuming <math>1.0 \times 10^{-3}</math> degree of order. Corresponding shifts for attached protons are <math>\pm 8</math> ppb.</b> .....	<b>63</b>
<b>Table 3.1: PDB accession numbers used to compute inter-helical angles for <math>H_{\geq 3}SXH_{\geq 3}SY</math> junctions. <sup>a</sup>PDBs that were excluded in the <math>H_{\geq 4}SXH_{\geq 4}SY</math> series.</b> .....	<b>88</b>
<b>Table 4.1: One bond CH and NH RDCs (in Hz) measured in TAR bound to aminoglycosides.</b> .....	<b>117</b>
<b>Table 4.2: Statistics for order tensor analysis of RDCs measured in HIV-1 TAR bound to aminoglycosides.</b> .....	<b>119</b>

## List of Appendices

<b>Appendix A: EULER-RNA-AFORM.pl .....</b>	<b>129</b>
<b>Appendix B: newmax_basics.pm .....</b>	<b>137</b>
<b>Appendix C: newmax_basics2.pm .....</b>	<b>154</b>
<b>Appendix D: newlsf4.pl .....</b>	<b>164</b>
<b>Appendix E: abgconNEWFRAME.pl .....</b>	<b>170</b>
<b>Appendix F: abgconverter.pl .....</b>	<b>176</b>
<b>Appendix G: maxpdbfitgenerator2.pl .....</b>	<b>182</b>
<b>Appendix H: maxdistgenerator.pl .....</b>	<b>184</b>
<b>Appendix I: maxtoranggenerator2.pl .....</b>	<b>186</b>
<b>Appendix J: rnaconformations2.pl .....</b>	<b>188</b>

## Abstract

The thermodynamic rules that link RNA sequences to secondary structure are well established while the link between its secondary structure and three-dimensional global conformation remains poorly understood. A-form helices – which represent ~50% of all RNA secondary structural elements in the PDB – are the most common secondary structural element of RNA. In fact, the global conformation and shape of RNA is largely defined by the relative arrangement of A-form helices linked by flexible pivot points. Of these flexible pivots, approximately 70% consist of two-way junctions such as bulges and internal loops. We have constructed comprehensive three-dimensional cubic maps depicting the relative orientation of A-form helices across RNA junctions as collected from the Protein Data Bank, and rationalized the findings using modeling and NMR spectroscopy. First, we derive a new angular nomenclature that utilized Euler angles to measure and describe the relative orientation of two helices. Next, we show that the secondary structure of junctions code for readily computable topological constraints that accurately predict the three-dimensional orientation of helices across all two-way junctions. And Finally, we rationalize adaptive changes that occur upon ligand binding in terms of the conformational orientations available to different junctions. The results herein suggest that the global conformation and orientation of helices is largely defined by topological constraints encoded within RNA's secondary

structure, and that tertiary contacts, along with other intermolecular interactions, act to stabilize specific structures from within a larger continuum of available topologically allowed conformers. Results from this work have the potential to enhance the development of applications in biotechnology and the targeting and development of therapeutics, as well as expand our understanding of RNA folding, and disordered RNA conformers in general (e.g. riboswitches). Additionally, there are a number of direct applications of this work that will readily impact the field of RNA structure prediction and modeling. This research will have a direct impact on our understanding of the role of structural dynamics and conformational adaptation within biology at large.

# Chapter 1

## Introduction

### 1.1 RNA structure determination

*“... while we have devoted a long time to the study of nucleic acid structure, the examination of topological properties was not attempted until recently. However, the interest of describing them first is that they require a minimum of detailed knowledge, and that the conclusions reached present a wide range of validity.”*

–Jacques Ninio

“Properties of nucleic acid representations I. Topology”  
Biochimie, 1971, 53, 485-494.

Following the initial discovery of nucleic acids in 1868 by Friedrich Miescher<sup>1</sup>, it was 50 years before any type of basic chemical structure for any of the nucleotide subunits was known<sup>2</sup>. Phoebus Levene finally demonstrated that each nucleotide contained a base, sugar and phosphate group<sup>2</sup>. Another 50 years passed before the first significant physical characterization of RNA was reported. The discovery entailed the determination of the first RNA sequence. The RNA in question was tRNA-ala and its primary sequence and secondary structure was determined in 1965 by Robert W. Holley<sup>3,4</sup>, an achievement that later garnered the 1968 Nobel Prize in Physiology or Medicine. Ten years later, Walter Fiers extended upon those findings to successfully determine, for the first time, the sequence and secondary structure of an entire viral genome: bacteriophage MS2<sup>5</sup>. The sequence was determined from enzyme digests, and the secondary structure was inferred based

on simplified thermodynamic principles, as set forth by Ignacio Tinoco Jr and coworkers<sup>6,7</sup>. Bacteriophage MS2 has a genome of 3,569 nucleotides of which “10.2% constitute untranslated segments,” and it marked the first “living organism” whose entire primary chemical structure had been determined<sup>5</sup>. Surprisingly, it wasn’t until 1978 that the first high-resolution three-dimensional structure of a RNA was obtained. All prior efforts had resulted in diffraction patterns too low-resolution to produce a reliable structure<sup>8-10</sup>. However the breakthrough that led to the first atomic resolution structure of an RNA was the result of three scientists – Hingerty, Brenin and Jack – who determined the structure of tRNA-phe at 2.5 Å resolution<sup>11</sup> (Figure 1.1a). Following this initial result a ~15 year span followed in which the only RNA structures determined were of tRNA, its variants and an array of assorted helical fragments. Although these first structures were largely homologous in sequence and structure, they provided the first glimpse of the architectural organization that had long eluded structural biologists. Additionally, these initial structures yielded the first reported values for two RNA A-form helical parameters: rise (2.5 Å) and twist (32° per base pair). Moreover, these structures provided the first observations of tertiary interactions that would be later characterized in more depth.

In 1994, Peter Moore’s group determined the first RNA structure by nuclear magnetic resonance (NMR) spectroscopy, which was Helix I from the 5S RNA of *Escherichia coli*<sup>12</sup>. The achievement marked the beginning of structure determination for a slew of new and unique RNAs. Most notably, the structure of a group I intron was determined in 1996 by Jennifer Doudna<sup>13</sup>(Figure 1.1b). Tom



Cech would follow with a larger group I intron structure (~250 nucleotides) of his own – albeit of lower resolution than the first<sup>14</sup>. The determination of the group I intron crystal structure was notable for a number of reasons. First, group I introns were known to undergo self-splicing in vivo. It was the discovery of this function that led to Cech being awarded the Nobel Prize in 1989, along with Sidney Altman who had discovered a similar catalytic ability for RNase P. Most importantly, it demonstrated that larger RNA structures could be experimentally determined given current structure determination techniques, especially given that the majority of structures at the time had been smaller – between 30 and 40 nucleotides – and predominately helical in nature.

Structure determination of the large ribosomal subunit from *Haloarcula marismortui* in 2000 by Thomas Steitz and Peter Moore was a major break-through in RNA structure determination and a huge accomplishment for molecular biology<sup>15</sup>. It was the first atomic resolution model of rRNA and it resolved more than 90% of nucleotides and 27 of 32 proteins (Figure 1.1c). The structure definitively rationalized the role of proteins in the rRNA complex, and revealed that the ribosome was indeed a ribozyme. Since that time a number of structures, both large and small, have demonstrated the immense organizational and structural complexity that exists in RNA. Moreover, it has become increasingly more apparent why structure determination of RNA has been far more difficult than the progresses achieved for other biological macromolecules, namely proteins. In part, this is a product of the intrinsically dynamic conformational behavior of helices in solutions, as has been highlighted by NMR spectroscopy and FRET studies. While there is

much data to suggest that this conformational heterogeneity is important for biological function, it also makes the determination of structures by x-ray crystallography all the more difficult. Still, the structural data gained thus far from the current database of RNA structures, large and small, has been tremendously insightful with respect to the structure and dynamics of RNA junctions, and arrangement and organization of base pairs.

It is interesting to note that the determination of large RNA structures is just out of its infancy. It's been 10 years since the ribosome structure determined by Steitz and Moore was produced, and it was another 3 years before another RNA of a comparable or larger size was determined. The vast majority of RNA structures determined today are rather small compared to the ribosome, further highlighting some of the intrinsic difficulty that remains for RNA structure determination. Going forward, techniques to advance methods for RNA structure determination will be critical to advance our knowledge of function and RNA architecture.

A majority of the solved RNA structures are known to participate in or regulate critical cellular functions. For instance, the ribosome, group I intron and riboswitches are all examples of RNAs that are central to processes that affect translation or transcription (Figure 1.1b-c). However, of the solved structures, all are bound to their cognate ligands, often in high-salt conditions. Although, we are limited to particular structures at specific points in the functional life of these RNAs, they have revealed an invaluable amount of information concerning the local geometries of base pairs and the global arrangement of helices. One area of interest

going forward will be the determination and modeling of the structural and dynamic characteristics of conformational adaptive RNA.

## 1.2 Local RNA structure

The local structure and configuration of base pairs is a well-characterized structural feature of RNA. For instance, idealized bond lengths, torsion angles, hydrogen bond distances and stacking energies are known for each Watson-Crick base pair and a number of non-canonical pairings<sup>16</sup>. Moreover, base step parameters (buckle ( $\kappa$ ), propeller ( $\omega$ ), opening ( $\sigma$ ), incline ( $\eta$ ), tip ( $\theta$ ), and twist ( $\Omega$ )) and sugar torsions ( $\nu_0$ - $\nu_4$ )<sup>17</sup> are well established and readily obtained using a variety of software: Curves 5.1<sup>18</sup>, FreeHelix98<sup>19</sup>, 3DNA<sup>20</sup>, SCHNAaP<sup>21</sup>, NUPARM and NUCGEN<sup>22</sup>. Contiguous stretches of RNA helices are known to adopt an A-form helical geometry<sup>17</sup>, and these geometrical parameters have been used extensively in analysis of local base pair configurations. In addition, the establishment of a common reference frame<sup>23</sup> has unified computational base step parameters, and permitted the determination of global helical parameters for stretches of Watson-Crick base pairs<sup>20,21</sup>. Experimentally, nucleobase and base pair parameters have been incorporated into structure determination protocols, such as NMR, where idealized geometries with variable error integrate needed physical constraints for known A-form helical regions. Similarly, idealized nucleobase parameters are integral components of molecular dynamics force fields. Thus, the parameterization of local RNA structural features, such as individual nucleobases and Watson-Crick base pairs, has been instrumental to approaches for structure determination, modeling and RNA dynamics simulations.

While parameters for A-form helices and Watson-Crick base pairs are readily accessible, parameters describing their counterpart non-canonical base pairs, as well as more exotic base associations (e.g. base triples and quadruples) are less studied. However, databases, such as the Non-canonical Base Pair Database<sup>24,25</sup>, are beginning to highlight and direct attention to the local structural diversity that exists within RNA. In part, this is due to the discrete number of known non-canonical base pair structures, such as Hoogsteen GA base pairs, which are relatively common among non-canonical base pairs. However, shortcomings such as these should resolve themselves as more RNA structures are determined and published.

Even the limited amount of data currently available has afforded valuable detailed information that has served to expand our understanding of the inclusion and incorporations of these local structural motifs into large networks of A-form helices<sup>26-29</sup>. Furthermore, RNA backbone interactions, such as ribose zippers<sup>30</sup> and A-minor<sup>31</sup> motifs illustrate the wide variety of possible intermolecular interactions. Moreover, recurrent themes have been observed for many higher-order junctions which allow the RNA to take advantage of these unique interactions to form highly ordered structures<sup>32-34</sup>. For example, it was long thought that group II introns formed complex arrays of backbone interactions from biochemical experiments<sup>35</sup> that were later confirmed from the structure determined in 2008<sup>36</sup>. Despite the inherently rich information garnered from base pair geometries and nucleobase parameters, there is an inherent disconnect between local and global RNA Structure. Thus, local RNA structure fails to adequately describe the “higher-scale geometry”

that is an intrinsic component of RNA global shape with a set of guiding principles that underlie its basic structural organization<sup>37</sup>.

### **1.3 RNA secondary structure and nomenclature**

Secondary structure determination of RNA has become commonplace since its feasibility was first demonstrated by Holley in 1965<sup>3,4</sup>. Advances of secondary structure determination have spanned biochemical and computational methodologies and techniques. Hydroxyl footprinting<sup>38</sup>, SHAPE chemistry<sup>39,40</sup> and other chemical modification experiments<sup>41,42</sup> have made secondary structure determination almost trivial. Moreover, many of these experiments have time resolved variants<sup>43,44</sup>. Software that predicts the secondary structure from a sequence, or a number of related sequences, has made significant contributions to our understanding of RNA. For instance, hydroxyl footprinting has provided informative data concerning the organization of folding pathways<sup>45-47</sup>, as well as transitions among RNA structural switches<sup>48,49</sup>. Additionally, conserved sequences have been used to identify phylogenic relationships among group II introns<sup>50,51</sup>.

Furthermore, the realization RNA is capable of forming non-canonical base pairs has increased the accuracy of secondary structure prediction, which, in turn, has greatly increased the accuracy and usefulness of secondary structure prediction and characterization. In part, the expansion of RNA secondary structure nomenclature<sup>52</sup> has developed hand-in-hand with the determination of more and more three-dimensional structures. The wealth of information from base pair interactions – both canonical and non – and tertiary interactions has had a tremendous effect on

our ability to perceive and rationalize results obtained from biochemical experiments. Programs, such as RNAVIEW<sup>53</sup> and FR3D<sup>54</sup> have been instrumental in aiding in the characterization of three-dimensional structures into two-dimension representations. As a result, RNA secondary structure represents a common ground between three-dimensional RNA conformations and their sequence. Its emergence as a readily available and coarse grain descriptor of RNA has meant that we can more easily marry structural data with its biochemical counterpart – function.

#### **1.4 Prediction of secondary structure**

Methods to predict RNA secondary structures from known sequences were first developed during the early part of the 1980s. The first published algorithm attempted to optimize predicted structures based on minimal free energy (MFE) constraints<sup>55</sup>. Free energy parameters were obtained from thermodynamic data of small RNA elements, often bulges, asymmetric and symmetric internal loops encapsulated within stretches of Watson-Crick base pairs, as well as variable length apical loops that capped the ends of small stretches of nucleotides. These early algorithms were successful, in part, because RNA secondary structures are more stable than in other biomolecules. For example, protein secondary structure formation is often driven by stabilization from tertiary interactions, whereas the opposite is generally true for RNA. Moreover, RNA secondary structures are amenable to mathematical analyses using simple combinatorial rules. Over time and in concert with the development of computational resources more advanced structure prediction algorithms have been developed. For example, comparative sequence analysis of covariation pattern analysis<sup>56</sup>, as discussed below, utilizes

comparisons among a number of RNA sequences to determine a common secondary structure. The Sankoff algorithm, first published in 1985, is a dynamic programming algorithm capable of determining the appropriate sequence alignment and fold for multiple sequences<sup>57</sup>. Of course there are many algorithms<sup>58-60</sup> that exist, each with its own strengths and weaknesses (Table 2). Discussed below are two of the more common secondary structure prediction algorithms.

#### **1.4.1 Turner rules**

Work in the 1980's from Douglass Turner's group resulted in the publication of free-energy parameters for base pairs, bulges and internal and apical loops base predominantly on melting curves. Using these parameters it is possible to predict secondary structures for target RNA sequences with ~73% accuracy. At the time, secondary structure prediction was heavily reliant on comparative computational methods, which were themselves heavily reliant on sequences with a high-degree of homology. However, knowing the energy contributions of base pair formation and structural motifs made secondary structural comparisons among a group of candidate solutions trivial. For proposed secondary structures, a free energy could readily be computed and scored. These configuration energy scores were used to rank and relate structures whose organization may have varied radically. However, the approach was derived predominately from energies of structural motifs sequestered within smaller RNAs. Thus, energetic contributions from tertiary interactions, which could have stabilized certain motifs in larger RNAs, were neglected. Such tertiary interactions have been shown to make critical contributions to the structural integrity of larger RNAs, like the group I introns<sup>61</sup>. Moreover, cases

with energies derived from calculations that optimized parameters for certain RNA classes (e.g. tRNA or 5S rRNA), biases these calculations towards tertiary interactions from those particular RNAs. Thus, a trade-off in the accuracy occurs at the expense of other RNA classes. Nonetheless, this approach to RNA secondary structure prediction has been largely successful. In particular, updated energy parameters have increased accuracy, as well as development of new computational algorithms<sup>62-66</sup>.

#### **1.4.2 Comparative sequence analysis and covariation algorithms**

Comparative sequence analysis (CSA) is a method used to determine the unifying structural characteristics of related RNAs. Given a large number of sequences with near identical function, comparative analysis produces reliable secondary structural characterization. Covariation algorithms are used to determine the likelihood that two sites are able to base pair. By analyzing a number of different sequences, one can determine if base pair formation is preferred at a specific site by analyzing if mutation of a particular nucleotide  $i$  leads to concomitant mutations of nucleotide  $j$ . Using this method structures of very large, complex RNAs have been predicted. For example, Gutell and coworkers implemented CSA to determine rRNA secondary structure and correctly predicted 97% to 98% of observed base pairs for the 30S<sup>67</sup> and 50S<sup>15</sup> ribosome. The analysis requires large data sets, but given ample data it produces very reliable results. In some cases, largely heterogeneous sequences can be problematic depending on the covariation algorithm used. However, CSA has been very useful at predicting base pairs, even non-canonical. Additionally, CSA has been utilized to identify and determine phylogenetic relationships among a number



of important RNAs. For instance, CSA has been used to determine the evolutionary and structural relationships among group I<sup>68-70</sup> and II<sup>71,72</sup> introns, RNase P<sup>73-75</sup>, telomerase<sup>76,77</sup>, and SRP<sup>78</sup> RNAs. Relevant programs to carry out CSA and covariation analysis are given in Table 3.

### **1.5 Prediction of three-dimensional RNA conformations from secondary structure**

Principles of three-dimensional RNA modeling were born in some of the early efforts to determine x-ray crystal structures. For example, the program NUCLIN-NUCLSQ – discussed below – was used in the structure determination of one of the first tRNA molecules. This necessity was derived from difficulty of obtaining high-resolution crystals. In fact, some of the first crystal experiments of RNA were unable to obtain structures due to insufficient structural resolution<sup>8-10</sup>. Therefore, early modeling efforts were a necessity of early RNA structure determination process. Early efforts in this area materialized in the form of NUCLIN-NUCLSQ, a restrained least-squares program used primarily for refinement<sup>79</sup>, and later on as the basis for modeling approaches of nucleic acid structures. The first program, NUCLIN, is used to initialize restraints that were first based on idealized bond lengths and angles from x-ray crystallography of small-molecules<sup>79</sup>. The second program, NUCLSQ, performs the restrained least-squares simulation. Since its creation, NUCLIN-NUCLSQ has been incorporated into MANIP<sup>80</sup> – a nucleic acid homology-modeling program – as a component refinement module. As experimentally determined RNA structures have steadily increased in number, so too have the number and sophistication of RNA modeling approaches. Many recent three-dimensional structure determination protocols have since employed strategies that build on

known RNA geometries, and often incorporate information from previously determined structures.

The majority of modern modeling strategies currently employed for RNA were first utilized in proteins. Some of the challenges faced for RNA, which are less obtrusive for proteins, stem from the dearth of known structures. Of the methods currently employed for structure modeling, two types of strategies dominate: fragment-based comparative modeling<sup>81,82</sup> and restrained-based modeling<sup>83</sup>. As discussed in more detail below, the major difference between fragment-based and restrained-based modeling approaches lie primarily in the manner in which each technique generates three-dimensional structures. Both strategies rely on a sequence alignment or some type of secondary structure comparison in order to determine general structural-likeness. The process of comparison largely guides the modeling, and determines how successful target RNA structures are predicted. However, a major hindrance to these modeling approaches is the dearth of structural information currently available. Other approaches utilize strategies analogous to the two presented here and are listed in Table 4.

### **1.5.1 RNA Fragment-based comparative modeling**

Fragment-based comparative modeling is a database method that utilizes previously determined RNA structures. The database is searched for short fragments of similar sequence. Once all candidate fragments for a target structure are found, they are then stitched together into larger three-dimensional models, with each structure typically evaluated on a number of criteria (e.g. Van der Waals contacts,

electrostatics and hydrogen bonds). The first step for most fragment-based comparative methods involves the alignment of the target sequence or comparison of secondary structure with known structures. Homology modeling is the specific case in which the sequences of known structure and target sequence are compared among one another. A potential pitfall of such a strategy is when the sequence of interest is unrelated to anything that is present within the structural database. Further complications arise if structures within the database fail to fully sample the available physical space or do not contain the appropriate structural motifs. Regardless, many successful approaches have implemented some type of fragment-based approach to their algorithm.

One of the first structure prediction programs MANIP<sup>80</sup> used a fragment-based method approach. The program utilized a database of RNA junctions and an algorithm to build RNA helices from canonical and non-canonical base pairs. The first step to constructing models involved building helical fragments and choosing appropriate junction fragments from a database of structures. Next, fragments were stitched together with the assistance of the user. Once a suitable structure had been built, it was subject to refinement using NUCLIN-NUCLSQ<sup>79</sup>. One disadvantage of this approach is that modeled structures are dependent upon the conformational variety available within the database. Also, substantial user interference was necessary, meaning that the individual constructing the models required a certain level of expertise.

Another program, MC-sym<sup>84</sup>, also implements a fragment-based comparative method. In MC-sym, nucleotide cyclic motifs (NCMs)<sup>85</sup> are used to identify and represent a minimal RNA structural motif. The NCM motifs are defined as the shortest physical distance necessary to arrive at the start without crossing paths. The program takes a target secondary structure and reduces it into sets of NCMs. Then using a database of RNA structure fragments, it constructs models by weaving together NCMs and evaluating energies. Once a structure has been completed the program cycles through a prescribed number of iterations. Ultimately, one “best” structure is chosen from amongst the many cycles. Again, the difficulty here is that structures are limited to what is “known” by the database, however, less user knowledge is required than a program like MANIP.

### **1.5.2 Restraint-based comparative modeling**

Many of the steps in restraint-based modeling are analogous, if not identical, to the strategy employed for fragment-based comparative modeling. The initial steps are to determine the best template for a target sequence from a database of known sequences. In the case of RNA, this may also include the incorporation of known secondary sequence relationships like Watson-Crick base pairs. Having found the best match constraints are created. In proteins, restraints typically involve the heavy atoms about Ramachandran phi/psi and functional group dihedral angles specific to residue types and known secondary and tertiary structural configurations. However, other properties of the molecule can also be utilized (e.g. hydrogen bonding, torsion angles, etc.). Having parameterized the target structure with values from a known structure, a simulated annealing protocol is, most often,

used to generate an ensemble target structures with accurate global folds. In the final step, structures are energy minimized and geometries optimized.

For RNA there is one program commonly used to employ this type of modeling procedure: fragment assembly of RNA (FARNA)<sup>86</sup>. FARNA utilizes known structural RNA fragments in order to derive a knowledge-based potential energy function that accounts for the preferences of backbone configurations and base orientations observed in experimental structures. Having determined the appropriate constraints to apply for a particular RNA system, a Monte Carlo simulation is employed to generate candidate structures, which are evaluated based on the knowledge-based energy function. However, this strategy still suffers from many of the same challenges that hinder other approaches. Namely, sufficient conformational sampling is difficult for moderate to large RNA (> 40 nucleotides), and potential conformational constraints are limited to what is contained within the structural database. Thus it is beyond the scope of the program to potentially predict novel interactions or structural organization. However the program is automated, just like MC-sym (see section 1.5.1), which makes it widely accessible to a variety of researchers.

## **1.6 Dynamics and Adaptation**

It is becoming increasingly apparent that dynamics and adaptation are central features of many cellular processes. RNA, in particular, given its simple and homogeneous structural composition has been shown to display a wide range of very dynamic behavior. For example the Varkud satellite (VS) ribozyme, the largest known nucleolytic ribozyme, has been shown to undergo extensive inter-helical

domain motions<sup>87-91</sup>. These large domain-domain motions have been shown to be critical to the function of the VS ribozyme<sup>90,91</sup>. The determination of conformational dynamics in systems like the VS ribozyme has been predominately studied using techniques like gel electrophoresis and Förster resonance energy transfer (FRET). While informative, these techniques lack the site-specific resolution provided by x-ray and NMR. However, NMR often faces many technical challenges involving sample preparation and data interpretation, such that it is often difficult to create accurate three-dimensional models of the dynamic conformations that exist in solution.

Residual dipolar couplings (RDCs) have been monumentally important in helping to define the structure and dynamics of RNA. The most direct application of RDCs has been their incorporation into more traditional NMR structure determination procedures, most of which rely on the measurement of NOEs. Traditional NMR structure determination techniques often take a significant amount of time to derive structures, and most studies are often limited in size. Using an experimental protocol that incorporates residual dipolar coupling (RDC) measurements has contributed greatly toward characterizing the average arrangement of A-form helices, as well as providing a means from which we can probe and describe their dynamics.

The RDC-derived modeling strategy described within builds directly off of information gained from an inspection of RNA secondary structure and incorporates well known global structural elements in such a way that descriptive models of the

relative orientation and helical configuration of RNA can be constructed. To date the protocol has been implemented in the study of the structure and dynamics for a number of RNAs<sup>92-95</sup>. A detailed description of the implementation of this protocol and the general conclusions can be found in chapter 2. Most remarkable, however, has been the integration of RDC measurements and domain-elongation, which yielded the first experimentally derived movie of RNA inter-helical motions<sup>96</sup>. Moreover, the work marked the first time anyone had constructed a physical model of large domain dynamics exclusively from experimental data. This result has set the stage to extend experimental methods geared toward elucidating conformational dynamics for any number of RNA.

### **1.7 Global RNA topology**

Far less information is available to characterize and parameterize properties that pertain to the description of RNA's global shape. Currently, only two parameters exist that are commonly used to describe RNA's global topology: size and shape. The size of RNA is often parameterized as the radius of gyration,  $R_g$ , which is generally understood to be the measure of compactness.  $R_g$  is readily determined from small-angle scattering (SAS) measurements and fluorescence anisotropy (FA). In time-resolved versions of SAS and FA experiments, the time-dependence of  $R_g$  can be determined, which can provide insights into events involving structural compactness of RNA during folding<sup>97-100</sup>. RNA shape is typically characterized by a density function,  $P(r)$ , which relays information on the average relative distances of atoms.  $P(r)$  is readily determined in SAS experiments, and the distribution of distances it yields is regularly used to rationalize global conformational changes of

RNA. In particular, the longest distance,  $D_{\max}$ , is typically indicative of the longest possible distance between two ends of the RNA. In a few reported cases,  $D_{\max}$  has been able to resolve and rationalize large conformational changes that occur during RNA folding or in response to ligand binding<sup>48,87,101,102</sup>. However, neither of these variables is able to relay specific information regarding the detailed nature of the helical arrangements, or the orientational changes of helices. These types of detailed insights are often rationalized in terms of an existing structure, or in some cases models<sup>87,101-103</sup>. Hence, supplementary structural data is required in order to maximize data analysis.

The angular characterization of relative helical orientations has also been used, however, much less than  $R_g$  or  $P(r)$ . Experiments employing gel mobility and transient electric birefringence<sup>104,105</sup>, as well as electron microscopy<sup>106,107</sup> (EM) have been employed in the determination of the global orientation of helices, specifically for tRNA, a hammerhead RNA and a group I intron. However, many RNAs are not amenable to helix elongation methodology. For instance, it has been shown that elongated helices abolished, or severely reduced, catalytic activity in some RNAs<sup>78,106,107</sup>. Hence, the most notable shortcoming of these techniques is interference from chemical modification with native features and RNA function. As discussed in chapter 2, a strength of the described RDC derived modeling strategy to deduce inter-helical angles is that chemical modification of the RNA is unnecessary. Moreover, inter-helical twist angles are readily derived from experimental measurements or the resultant models.

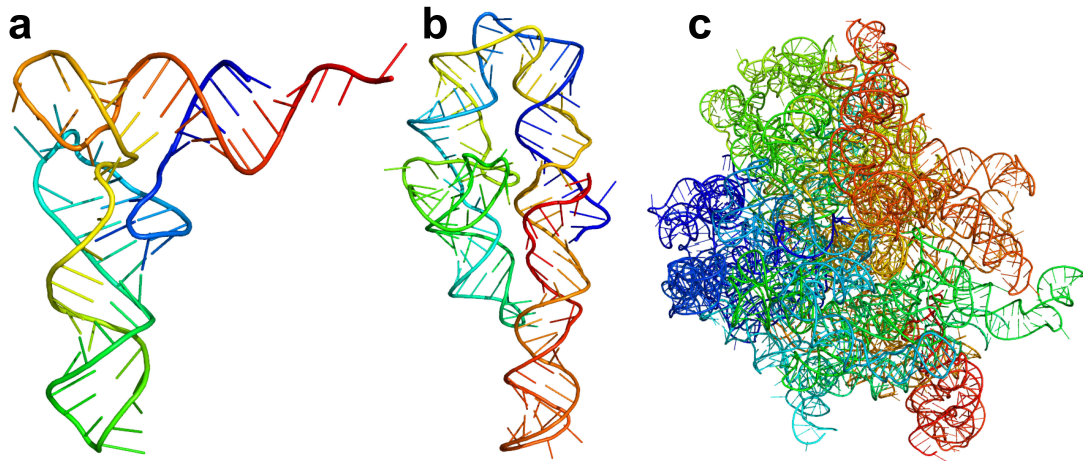


Determination of the relative angular orientation of helices is not a trivial endeavor. In fact, even determining how to experimentally measure and obtain inter-helical orientations is quite difficult. Before 2006, characterization of inter-helical angles was reliable only for inter-helical bending. Numerous procedures have been used to obtain this parameter. For instance inter-helical bend angles in the RNA Junction database<sup>108</sup> were calculated by superimposing idealized helices on to C4' atoms, and deriving the helical axis by averaging over all atomic coordinates. The bend angle was then calculated by taking the dot product for the resulting vectors for each helix. Prior experimental determination of an inter-helical bend angle using NMR data from this lab also involved a strategy utilizing idealized elongated A-form helices<sup>109</sup>. Having modeled the conformation of the RNA, the chi-tensor alignment of a reference helix was determined and the model was superimposed to this frame. An idealized A-form helix was superimposed onto the alternate helix. From this configuration the rotation required to orient the second helix within the reference frame generated the bend angle.

The decision to employ an Euler angle nomenclature to describe the relative inter-helical orientation of two helices arose, in part, due to their prior use in the singular value decomposition (SVD) analysis of RDC data. In a typical SVD analysis, the orientation of the structural fragment under consideration is arbitrarily oriented within the molecular frame, and the Euler angles derived from determination of the order tensor describe the rotation necessary to correctly place the structural fragment in its principle axis system. A trivial extension of this method is to coaxially orient both helices in their initial configuration, with coincident helical and

molecular z-axes. Euler angles derived in this manner produce a complete description of the inter-helical conformation for any two helices: two angles to describe the twist of each helix and one angle to describe the inter-helical bend. The broader applicability of the nomenclature lies in the ability to accurately and completely describe the conformation of two RNA helices.

The goal of the work described within has two objectives. One is to standardize and develop a high-throughput method for the characterization of global features of RNA. Second is to accurately describe the derived conformations. Aside from this work, there currently does not exist a means to uniquely describe the orientation and relative arrangement of RNA helices. Additionally, we have successfully applied the characterization of the relative orientation of two helices to RNA structures determined by other spectroscopic means. The results of this data, as discussed in the later chapters, has uncovered a previously unrecognized relationship between allowed orientations of RNA helices, and permitted us to rationalize the underlying principles in terms of the structure, dynamics and adaptation of RNA.



**Figure 1.1: Important achievements and milestones in the history of RNA structure determination.** (a-c) Shown are x-ray crystal structures of tRNA (2.5 Å), group I intron (2.8 Å) and 50S ribosome (2.5 Å), respectively, with resolution indicated in parentheses.

**Table 1.1: Secondary structure prediction from primary sequence.**

<b>Name</b>	<b>Comments</b>	<b>Pseudo knots</b>	<b>Ref</b>
CentroidFold	Secondary structure prediction based on generalized centroid estimator.	no	110
CONTRAFold	Secondary structure prediction method utilizing conditional log-linear models (CLLMs), which are a flexible class of generalized probabilistic models from Self-consistent mean field.	no	111
KineFold	Folding kinetics of RNA sequences including pseudoknots by including an implementation of the partition function for knots.	yes	112
Mfold	Minimum free energy (MFE) RNA structure prediction algorithm.	no	113
Pknots	Dynamic programming algorithm optimized for RNA pseudoknot prediction using nearest neighbor energy model.	yes	114
PknotsRG	Dynamic programming algorithm for prediction of restricted RNA pseudoknots class.	yes	115
RNAfold	Minimum free energy (MFE) RNA structure prediction algorithm. Includes implementation of partition function for computing base pair probabilities.	no	116
RNAshapes	Will predict minimum free energy (MFE) secondary structure of single sequences, or can be used as a dynamic programming algorithm.	no	117
RNAstructure	Predicts lowest free energy structures and base pair probabilities for RNA and DNA. Structure prediction can be constrained using variety of experimental data. Graphical user interfaces are available for Windows and Mac OS-X/Linux. Programs available for use with Unix-style interfaces, and C++ class library is available as stand alone.	no	118,119
Sfold	Program performs a statistical sampling of all possible structures using a weighted partition function of probabilities.	no	120,121
UNAFold	Integrated package of software that simulates folding, hybridization, and melting pathways for one or two single-stranded nucleic acid sequences.	no	122

<b>Name</b>	<b>Comments</b>	<b>Ref</b>
McCaskill-MEA	Computes a matrix of base pairing probability for each input sequence within an alignment. It then obtains the alignment base pairing probability matrix by averaging over each individual matrix. A consensus secondary structure is predicted from consensus matrix to maximized the expected prediction accuracy.	58
Self-consistent mean field (SCMF)	Initially treats all base pairs as having an equal probability of forming, however with each iteration the more energetically favored base pairs show increasing probability of forming, as long as they are consistent with their neighbors.	59
Sankoff	Incorporates folding and alignment into a single algorithm. By optimizing a linear combination of the objective functions, an N-dimensional generalization of the alignment method to the problem of reconstruction allows objectives – folding and alignment – to be carried out simultaneously.	57
3D Triangular Lattice	Simulates folding dynamics of the RNA sequences on a 3D triangular lattice. Base pairs from the best lattice conformation are identified by the folding simulation.	60

<b>Name</b>	<b>Comment</b>	<b>Num Seq</b>	<b>Align</b>	<b>Struct</b>	<b>Pseudo knots</b>	<b>Ref</b>
Carnac	Combination of comparative analysis combined and minimum free energy folding.	any	no	yes	no	123 124
CMfinder	Algorithm based on maximizing expectation via covariance models. Searches for RNA motifs using both folding energy and covariation.		yes	yes	no	125
CONSAN	Implementation of variation on pinned Sankoff algorithm for simultaneous pairwise RNA alignment and consensus prediction.	2	yes	yes	no	126
Dynalign	Algorithm that combines free energy minimization with comparative sequence analysis to find common low free energy structure for two sequences independent of sequence identity.	2	yes	yes	no	127-129
FoldalignM	Sankoff based approach, where partition function calculates base pair probability matrices. Matrices are aligned to produce a multiple alignments and predict consensus structure.	any	yes	yes	no	125,130, 131
KNetFold	Computes consensus RNA secondary structure from machine learned RNA sequence alignment.	any	input	yes	yes	132

LARA	A lagrangian transformation relaxation, using a numerical optimization approach.	any	yes	yes	no	125
LocaRNA	Successor of PMcomp with improved time complexity. A variant of Sankoff's algorithm for simultaneous folding and alignment, with input from McCaskill's algorithm. Method also means to compare base pair probability matrices.	any	yes	yes	no	133
MASTR	Using Markov chain Monte Carlo simulation to iteratively improve sequence alignment and structure prediction input RNA sequence set.	any	yes	yes	no	125,134
Murlet	Efficient algorithm for the multiple alignment of structural RNA sequences. Variant of Sankoff algorithm, uses scoring system to reduce time and space requirements.	any	yes	yes	no	135
MXSCARNA	Multiple alignment tool for RNA sequences using progressive alignment based SCARNA's on pairwise structural alignment algorithm.	any	yes	yes	no	136
PARTS	A method for joint prediction of alignment and common secondary structures of two RNA sequences using a probabilistic model based on pseudo free energies obtained from precomputed base pairing and alignment probabilities.	2	yes	yes	no	137
Pfold	Folds alignments using a stochastic context-free grammar (SCFG) trained on rRNA alignments. "It assumes an alignment and gives one common structural prediction for all the sequences."		input	yes	no	138
PMcomp/PMulti	Variant of Sankoff's algorithm and takes input base pair probability matrices from McCaskill's algorithm. PMulti is wrapper program to do repeated progressive multiple alignments.		yes	yes	no	139
R-COFFEE	RNAfold computes secondary structure of input sequences. Modified version of T-Coffee compute multiple sequence alignments. Can be combined with existing sequence alignment methods.	any	yes	yes	no	140,141
RNAalifold	Predicts structure given an alignment using both free energy and a covariation measure for base pair regions.	any	input	yes	no	125
RNAcast	Predicts common shapes for all sequences and their energetic best structure.	any	no	yes	no	125

RNAforester	Performs multiple alignments based on input sequence set.		yes	input	no	125
RNAmine	Software tool to extract the structural motifs from a set of RNA sequences.	any	no	yes	no	142
RNASampler	Stems are aligned by comparing all base pairs sequences. A conservation score measures the quality, and a structural alignment is built.	any	yes	yes	yes	125
SCARNA	Fast, convenient tool for structural alignment of RNA sequence pair. Aligns two RNA sequences and calculates similarities based on common secondary structures.	2	yes	yes	no	143
SimulFold	Employs Bayesian Markov chain Monte Carlo method to sample joint posterior distribution of RNA structures, alignments, and trees.	any	yes	yes	yes	144
Stemloc	Program for pairwise RNA structural alignment based on probabilistic models of structures from known Pair stochastic context-free grammars.	any	yes	yes	no	145
StrAl	Heuristic method for alignment of ncRNA; reduces sequence-structure alignment to two-dimensional problem similar to standard multiple sequence alignment.		yes	no	no	146
WAR	Webserver that simultaneously use a number of methods for performing multiple alignment and secondary structure prediction for noncoding RNAs		yes	yes	no	125
Xrate	Program for analysis of multiple sequence alignments using phylogenetic grammars.	any	yes	yes	no	147
RNASoft	RNASoft program suite provides tools for predicting secondary structure of DNA or RNA. The tools are based on standard thermodynamic models of RNA secondary structure formation.					148
RNA-DECODER	Program explicitly considers known protein-coding context of an RNA-sequence alignment when predict evolutionarily conserved secondary structure.					149

<b>Name</b>	<b>Comments</b>	<b>Ref</b>
JUMNA	Does energy optimizes nucleic acids and nucleic acid-ligand complexes. The force field used, input and output data, various options for symmetry, conformational constraints and energy mapping are discussed as well as recent combinatorial search techniques.	150
MANIP	Program does rapid assembly of RNA fragment motifs into a three-dimensional architecture. Assembly is performed in real time with buttons and dials that rotate and translate any specific fragment. Program is interfaced with NUCLIN-NUCLSQ for rapid, online automated refinement of partial or full structures.	80
NUCLIN-NUCLSQ	NUCLIN-NUCLSQ is a version of Wayne Hendrickson's "PROLSQ," specifically modified for nucleic acids.	79,151
S2S	Sequence to Structure (S2S) proposes a framework in which an user can easily display, manipulate and interconnect heterogeneous RNA data, such as multiple sequence alignments, secondary and tertiary structures.	152
RNA2D3D	Using a RNA sequence and secondary structure, the program produces automatic and rapid first-order approximations of 3-dimensional conformations.	153
ERNA-3D	A molecular modeling system specially focused on the creation of large-scale RNA models.	154
YUP	Yammp Under Python (YUP), aka Yammp 2, is a molecular modeling program, focused on carrying out molecular simulations (mechanics), reduced representations and multiscale modeling.	155
NAB	Program provides a programming environment for geometric and force-field manipulations of nucleic acids. Part of the AmberTools.	156
BARNACLE	A probabilistic sampling Python library for RNA structures compatible with a given nucleotide sequence.	157
FARNA	Automated strategy for de novo prediction of RNA tertiary structures.	86
iFoldRNA	Three-dimensional RNA structure prediction and folding program.	158
MC-Fold, MC-Sym	Program built around nucleotide cyclic motifs (NCM), which employs a fragment-based approach to RNA structure prediction.	84
NAST	Large RNA molecule coarse grain modeling approach, which utilizes knowledge-based potentials and structure filters	159



## 1.8 References

1. Dahm, R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Hum Genet* **122**, 565-581 (2008).
2. Levene, P. A. The Structure of Yeast Nucleic Acid. *Studies from the Rockefeller Institute for Medical Research* 105 (1918).
3. Holley, R. W. Structure of an alanine transfer ribonucleic acid. *JAMA* **194**, 868-871 (1965).
4. Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., *et al.* Structure of a Ribonucleic Acid. *Science* **147**, 1462-1465 (1965).
5. Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500-507 (1976).
6. Borer, P. N., Dengler, B. & Tinoco Jr, I. O. C. Uhlenbeck. 1974. Stability of ribonucleic acid double-stranded helices. *J Mol Biol* **86**, 843-853 (1974).
7. Tinoco, I., Borer, P. N., Dengler, B., Levin, M. D., *et al.* Improved estimation of secondary structure in ribonucleic acids. *Nat New Biol* **246**, 40-41 (1973).
8. Kim, S. H., Quigley, G. J., Suddath, F. L., McPherson, A., *et al.* Three-dimensional structure of yeast phenylalanine transfer RNA: folding of the polynucleotide chain. *Science* **179**, 285 (1973).
9. Kim, S. H., Quigley, G., Suddath, F. L. & Rich, A. High-resolution x-ray diffraction patterns of crystalline transfer RNA that show helical regions. *Proc Natl Acad Sci U S A* **68**, 841 (1971).
10. Kim, S. H. & Rich, A. Single crystals of transfer RNA: An X-ray diffraction study. *Science* **162**, 1381-1384 (1968).
11. Hingerty, B., Brown, R. S. & Jack, A. Further refinement of the structure of yeast tRNAPhe. *J Mol Biol* **124**, 523-534 (1978).
12. White, S. A., Nilges, M., Huang, A., Brunger, A. T. & Moore, P. B. NMR analysis of helix I from the 5S RNA of Escherichia coli. *Biochemistry* **31**, 1610-1621 (1992).
13. Cate, J. H., Gooding, A. R., Podell, E., Zhou, K., *et al.* Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* **273**, 1678 (1996).
14. Golden, B. L., Gooding, A. R., Podell, E. R. & Cech, T. R. A preorganized active site in the crystal structure of the Tetrahymena ribozyme. *Science* **282**, 259 (1998).
15. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905 (2000).
16. Turner, D. H. & Mathews, D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res* **38**, D280-D282 (2010).

17. Neidle, S. *Oxford Handbook of Nucleic Acid Structure*. 1999 (New York: Oxford University Press, .
18. Ravishanker, G., Swaminathan, S., Beveridge, D. L., Lavery, R. & Sklenar, H. Conformational and helicoidal analysis of 30 PS of molecular dynamics on the d (CGCGAATTCGCG) double helix: "curves", dials and windows. *Journal of biomolecular structure & dynamics* **6**, 669 (1989).
19. Dickerson, R. E. DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic acids research* **26**, 1906 (1998).
20. Lu, X. J. & Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res* **31**, 5108 (2003).
21. Lu, X. J., El Hassan, M. A. & Hunter, C. A. Structure and conformation of helical nucleic acids: analysis program (SCHNAaP). *J Mol Biol* **273**, 668-680 (1997).
22. Bansal, M., Bhattacharyya, D. & Ravi, B. NUPARM and NUCGEN: software for analysis and generation of sequence dependent nucleic acid structures. *Comput Appl Biosci* **11**, 281-287 (1995).
23. Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., *et al.* A standard reference frame for the description of nucleic acid base pair geometry. *J Mol Biol* **313**, 229-237 (2001).
24. Nagaswamy, U., Voss, N., Zhang, Z. & Fox, G. E. Database of non-canonical base pairs found in known RNA structures. *Nucleic Acids Res* **28**, 375-376 (2000).
25. Nagaswamy, U., Larios-Sanz, M., Hury, J., Collins, S., *et al.* NCIR: a database of non-canonical interactions in known RNA structures. *Nucleic Acids Res* **30**, 395-397 (2002).
26. Stombaugh, J., Zirbel, C. L., Westhof, E. & Leontis, N. B. Frequency and isostericity of RNA base pairs. *Nucleic Acids Res* **37**, 2294-2312 (2009).
27. Leontis, N. B., Stombaugh, J. & Westhof, E. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res* **30**, 3497 (2002).
28. Lescoute, A., Leontis, N. B., Massire, C. & Westhof, E. Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic Acids Res* **33**, 2395-2409 (2005).
29. Lee, J. C. & Gutell, R. R. Diversity of base pair conformations and their occurrence in rRNA structure and RNA structural motifs. *J Mol Biol* **344**, 1225-1249 (2004).
30. Tamura, M. & Holbrook, S. R. Sequence and structural conservation in RNA ribose zippers. *J Mol Biol* **320**, 455-474 (2002).
31. Nissen, P., Ippolito, J. A., Ban, N., Moore, P. B. & Steitz, T. A. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc Natl Acad Sci U S A* **98**, 4899-4903 (2001).

32. Holbrook, S. R. Structural principles from large RNAs. *Annu Rev Biophys* **37**, 445-464 (2008).
33. Laing, C., Jung, S., Iqbal, A. & Schlick, T. Tertiary motifs revealed in analyses of higher-order RNA junctions. *J Mol Biol* **393**, 67-82 (2009).
34. de la Peña, M., Dufour, D. & Gallego, J. Three-way RNA junctions with remote tertiary contacts: a recurrent and highly versatile fold. *RNA* **15**, 1949-1964 (2009).
35. Michel, F., Umesono, K. & Ozeki, H. Comparative and functional anatomy of group II catalytic introns--a review. *Gene* **82**, 5-30 (1989).
36. Toor, N., Keating, K. S., Fedorova, O., Rajashankar, K., *et al.* Tertiary architecture of the *Oceanobacillus iheyensis* group II intron. *RNA* (2009).
37. Ninio, J. Properties of nucleic acid representations. I. Topology. *Biochimie* **53**, 485-494 (1971).
38. Wang, X. D. & Padgett, R. A. Hydroxyl radical "footprinting" of RNA: application to pre-mRNA splicing complexes. *Proc Natl Acad Sci U S A* **86**, 7795-7799 (1989).
39. Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J Am Chem Soc* **127**, 4223-4231 (2005).
40. Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci U S A* **106**, 97-102 (2009).
41. Konforti, B. B., Liu, Q. & Pyle, A. M. A map of the binding site for catalytic domain 5 in the core of a group II intron ribozyme. *EMBO J* **17**, 7105-7117 (1998).
42. Tijerina, P., Mohr, S. & Russell, R. DMS footprinting of structured RNAs and RNA-protein complexes. *Nat Protoc* **2**, 2608-2623 (2007).
43. Hampel, K. J. & Burke, J. M. Time-resolved hydroxyl-radical footprinting of RNA using Fe(II)-EDTA. *Methods* **23**, 233-239 (2001).
44. Mortimer, S. A. & Weeks, K. M. Time-resolved RNA SHAPE chemistry. *J Am Chem Soc* **130**, 16178-16180 (2008).
45. Uchida, T., Takamoto, K., He, Q., Chance, M. R. & Brenowitz, M. Multiple monovalent ion-dependent pathways for the folding of the L-21 Tetrahymena thermophila ribozyme. *J Mol Biol* **328**, 463-478 (2003).
46. Laederach, A., Shcherbakova, I., Liang, M. P., Brenowitz, M. & Altman, R. B. Local kinetic measures of macromolecular structure reveal partitioning among multiple parallel pathways from the earliest steps in the folding of a large RNA molecule. *J Mol Biol* **358**, 1179-1190 (2006).
47. Laederach, A., Shcherbakova, I., Jonikas, M. A., Altman, R. B. & Brenowitz, M. Distinct contribution of electrostatics, initial conformational ensemble, and macromolecular stability in RNA folding. *Proc Natl Acad Sci U S A* **104**, 7045-7050 (2007).
48. Lipfert, J., Das, R., Chu, V. B., Kudravalli, M., *et al.* Structural transitions and

- thermodynamics of a glycine-dependent riboswitch from *Vibrio cholerae*. *J Mol Biol* **365**, 1393-1406 (2007).
49. Brooks, K. M. & Hampel, K. J. A rate-limiting conformational step in the catalytic pathway of the glmS ribozyme. *Biochemistry* **48**, 5669-5678 (2009).
  50. Toor, N., Hausner, G. & Zimmerly, S. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA* **7**, 1142-1152 (2001).
  51. Simon, D. M., Kelchner, S. A. & Zimmerly, S. A Broadscale Phylogenetic Analysis of Group II Intron RNAs and Intron-Encoded Reverse Transcriptases. *Molecular Biology and Evolution* **26**, 2795-2808 (2009).
  52. Leontis, N. B. & Westhof, E. Geometric nomenclature and classification of RNA base pairs. *RNA* **7**, 499-512 (2001).
  53. Yang, H., Jossinet, F., Leontis, N., Chen, L., *et al.* Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res* **31**, 3450-3460 (2003).
  54. Sarver, M., Zirbel, C. L., Stombaugh, J., Mokdad, A. & Leontis, N. B. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J Math Biol* **56**, 215-252 (2008).
  55. Zuker, M. & Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* **9**, 133 (1981).
  56. Parsch, J., Braverman, J. M. & Stephan, W. Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics* **154**, 909-921 (2000).
  57. Sankoff, D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J Appl Math* **45**, 810-825 (1985).
  58. Kiryu, H., Kin, T. & Asai, K. Robust prediction of consensus secondary structures using averaged base pairing probability matrices. *Bioinformatics* **23**, 434-441 (2007).
  59. Kleesiek, J. & Torda, A. E. RNA secondary structure prediction using a self-consistent mean field approach. *J Comput Chem* (2009).
  60. Gillespie, J., Mayne, M. & Jiang, M. RNA folding on the 3D triangular lattice. *BMC Bioinformatics* **10**, 369 (2009).
  61. Silverman, S. K., Zheng, M., Wu, M., Tinoco, I. & Cech, T. R. Quantifying the energetic interplay of RNA tertiary and secondary structure interactions. *RNA* **5**, 1665-1674 (1999).
  62. Kim, N., Shiffeldrim, N., Gan, H. H. & Schlick, T. Candidates for novel RNA topologies. *J Mol Biol* **341**, 1129-1144 (2004).
  63. Matsui, H., Sato, K. & Sakakibara, Y. Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures. *Proc IEEE Comput Syst Bioinform Conf* 290-299 (2004).

64. Dirks, R. M. & Pierce, N. A. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J Comput Chem* **24**, 1664-1677 (2003).
65. Zhao, J., Malmberg, R. L. & Cai, L. Rapid ab initio prediction of RNA pseudoknots via graph tree decomposition. *J Math Biol* **56**, 145-159 (2008).
66. Roovers, M., Oudjama, Y., Kaminska, K. H., Purta, E., *et al.* Sequence-structure-function analysis of the bifunctional enzyme MnmC that catalyses the last two steps in the biosynthesis of hypermodified nucleoside mnm5s2U in tRNA. *Proteins* **71**, 2076-2085 (2008).
67. Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Morgan-Warren, R. J., *et al.* Structure of the 30S ribosomal subunit. *Nature* **407**, 327-339 (2000).
68. Damberger, S. H. & Gutell, R. R. A comparative database of group I intron structures. *Nucleic Acids Res* **22**, 3508 (1994).
69. Michel, F. & Westhof, E. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J Molr Biol* **216**, 585-610 (1990).
70. Cech, T. R. Conserved sequences and structures of group I introns: building an active site for RNA catalysis--a review. *Gene* **73**, 259-271 (1988).
71. Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., *et al.* The Comparative RNA Web(CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC bioinformatics* **3**, 2 (2002).
72. Michel, F., Kazuhiko, U. & Haruo, O. Comparative and functional anatomy of group II catalytic introns--a review. *Gene* **82**, 5-30 (1989).
73. Harris, J. K., Haas, E. S., Williams, D., Frank, D. N. & Brown, J. W. New insight into RNase P RNA structure from comparative analysis of the archaeal RNA. *RNA* **7**, 220-232 (2001).
74. Brown, J. W., Haas, E. S., James, B. D., Hunt, D. A., *et al.* Phylogenetic analysis and evolution of RNase P RNA in proteobacteria. *J Bact* **173**, 3855 (1991).
75. James, B. D., Olsen, G. J., Liu, J. & Pace, N. R. The secondary structure of ribonuclease P RNA, the catalytic element of a ribonucleoprotein enzyme. *Cell* **52**, 19-26 (1988).
76. Chen, J. L., Blasco, M. A. & Greider, C. W. Secondary structure of vertebrate telomerase RNA. *Cell* **100**, 503-514 (2000).
77. Romero, D. P. & Blackburn, E. H. A conserved secondary structure for telomerase RNA. *Cell* **67**, 343-353 (1981).
78. Walter, P. & Blobel, G. Signal recognition particle contains a 7 S RNA essential for protein translocation across the endoplasmic reticulum. *Nature* **299**, 691-698 (1982).
79. Westhof, E., Dumas, P. & Moras, D. Crystallographic refinement of yeast aspartic

- acid transfer RNA. *J Mol Biol* **184**, 119-145 (1985).
80. Massire, C. & Westhof, E. MANIP: an interactive tool for modelling RNA. *J Mol Graph Model* **16**, 197-205, 255-7 (1998).
81. Blundell, T., Carney, D., Gardner, S., Hayes, F., *et al.* 18th Sir Hans Krebs lecture. Knowledge-based protein modelling and design. *Eur J Bioch/FEBS* **172**, 513 (1988).
82. Blundell, T. L., Sibanda, B. L., Sternberg, M. J. & Thornton, J. M. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **326**, 347-352 (1987).
83. Sali, A. & Blundell, T. L. Comparative protein modeling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815 (1993).
84. Parisien, M. & Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**, 51-55 (2008).
85. Lemieux, S. & Major, F. Automated extraction and classification of RNA tertiary structure cyclic motifs. *Nucleic Acids Res* **34**, 2340-2346 (2006).
86. Das, R. & Baker, D. Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences* **104**, 14664 (2007).
87. Lipfert, J., Ouellet, J., Norman, D. G., Doniach, S. & Lilley, D. M. The complete VS ribozyme in solution studied by small-angle X-ray scattering. *Structure* **16**, 1357-1367 (2008).
88. Pereira, M. J., Nikolova, E. N., Hiley, S. L., Jaikaran, D., *et al.* Single VS ribozyme molecules reveal dynamic and hierarchical folding toward catalysis. *J Mol Biol* **382**, 496-509 (2008).
89. Bouchard, P., Lacroix-Labonté, J., Desjardins, G., Lampron, P., *et al.* Role of SLV in SLI substrate recognition by the *Neurospora* VS ribozyme. *RNA* **14**, 736-748 (2008).
90. Walter, N. G., Hampel, K. J., Brown, K. M. & Burke, J. M. Tertiary structure formation in the hairpin ribozyme monitored by fluorescence resonance energy transfer. *EMBO J* **17**, 2378-2391 (1998).
91. Pereira, M. J., Harris, D. A., Rueda, D. & Walter, N. G. Reaction pathway of the trans-acting hepatitis delta virus ribozyme: a conformational change accompanies catalysis. *Biochemistry* **41**, 730-740 (2002).
92. Mollova, E. T., Hansen, M. R. & Pardi, A. Global structure of RNA determined with residual dipolar couplings. *J Am Chem Soc* **122**, 11561-11562 (2000).
93. Sun, X., Zhang, Q. & Al-Hashimi, H. M. Resolving fast and slow motions in the internal loop containing stem-loop 1 of HIV-1 that are modulated by Mg<sup>2+</sup> binding: role in the kissing-duplex structural transition. *Nucleic Acids Res* **35**, 1698-1713 (2007).
94. Getz, M., Sun, X., Casiano-Negroni, A., Zhang, Q. & Al-Hashimi, H. M. NMR studies of RNA dynamics and structural plasticity using NMR residual dipolar couplings. *Biopolymers* **86**, 384-402 (2007).

95. Casiano-Negroni, A., Sun, X. & Al-Hashimi, H. M. Probing Na(+)-induced changes in the HIV-1 TAR conformational dynamics using NMR residual dipolar couplings: new insights into the role of counterions and electrostatic interactions in adaptive recognition. *Biochemistry* **46**, 6525-6535 (2007).
96. Zhang, Q., Stelzer, A. C., Fisher, C. K. & Al-Hashimi, H. M. Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature* **450**, 1263-1267 (2007).
97. Koch, M. H., Vachette, P. & Svergun, D. I. Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Q Rev Biophys* **36**, 147-227 (2003).
98. Kwok, L. W., Shcherbakova, I., Lamb, J. S., Park, H. Y., *et al.* Concordant exploration of the kinetics of RNA folding from global and local perspectives. *J Mol Biol* **355**, 282-293 (2006).
99. Lakowicz, J. R. in *Principles of Fluorescence Spectroscopy* 63-95 (Springer US, 2006).
100. Anunciado, D., Agumeh, M., Kormos, B. L., Beveridge, D. L., *et al.* Characterization of the dynamics of an essential helix in the U1A protein by time-resolved fluorescence measurements. *J Phys Chem B* **112**, 6122-6130 (2008).
101. Baird, N. J., Westhof, E., Qin, H., Pan, T. & Sosnick, T. R. Structure of a folding intermediate reveals the interplay between core and peripheral elements in RNA folding. *J Mol Biol* **352**, 712-722 (2005).
102. Lipfert, J., Chu, V. B., Bai, Y., Herschlag, D. & Doniach, S. Low-resolution models for nucleic acids from small-angle X-ray scattering with applications to electrostatic modeling. *J Appl Cryst* **40**, s229-s234 (2007).
103. Ali, M., Lipfert, J., Seifert, S., Herschlag, D. & Doniach, S. The Ligand-Free State of the TPP Riboswitch: A Partially Folded RNA Structure. *J Mol Biol* (2009).
104. Zacharias, M. & Hagerman, P. J. Bulge-induced bends in RNA: quantification by transient electric birefringence. *J Mol Biol* **247**, 486-500 (1995).
105. Zacharias, M. & Hagerman, P. J. The influence of symmetric internal loops on the flexibility of RNA. *J Mol Biol* **257**, 276-289 (1996).
106. Nakamura, T. M., Wang, Y. H., Zaug, A. J., Griffith, J. D. & Cech, T. R. Relative orientation of RNA helices in a group 1 ribozyme determined by helix extension electron microscopy. *EMBO* **14**, 4849 (1995).
107. Amiri, K. M. A. & Hagerman, P. J. Global conformation of a self-cleaving hammerhead RNA. *Biochemistry* **33**, 13172-13177 (1994).
108. Bindewald, E., Hayes, R., Yingling, Y. G., Kasprzak, W. & Shapiro, B. A. RNAJunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign. *Nucleic Acids Res* **36**, D392-D397 (2008).
109. Pitt, S. W., Zhang, Q., Patel, D. J. & Al-Hashimi, H. M. Evidence that electrostatic interactions dictate the ligand-induced arrest of RNA global flexibility. *Angew Chem*

*Int Ed Engl* **44**, 3412-3415 (2005).

110. Sato, K., Hamada, M., Asai, K. & Mituyama, T. CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Res* (2009).

111. Do, C. B., Woods, D. A. & Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**, e90 (2006).

112. Xayaphoummine, A., Bucher, T. & Isambert, H. Kinifold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res* **33**, W605 (2005).

113. Zuker, M. *mfold-2.3* (1996).

114. Rivas, E. & Eddy, S. R. A dynamic programming algorithm for RNA structure prediction including pseudoknots1. *Journal of Molecular Biology* **285**, 2053-2068 (1999).

115. Reeder, J., Steffen, P. & Giegerich, R. pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows. *Nucleic Acids Res* (2007).

116. Denman, R. B. Using RNAFOLD to predict the activity of small catalytic RNAs. *Biotechniques* **15**, 1090 (1993).

117. Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J. & Giegerich, R. RNASHAPES: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* **22**, 500 (2006).

118. Mathews, D. H. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* **10**, 1178 (2004).

119. Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., *et al.* Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A* **101**, 7287 (2004).

120. Ding, Y., Chan, C. Y. & Lawrence, C. E. Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic acids research* **32**, W135 (2004).

121. Ding, Y. & Lawrence, C. E. 8 Rational design of siRNAs with the Sfold software. *RNA interference technology: From basic science to drug development* 129 (2005).

122. Markham, N. R. & Zuker, M. UNAFold: software for nucleic acid folding and hybridization. *Meth Mol Biol (Clifton, NJ)* **453**, 3 (2008).

123. Perriquet, O., Touzet, H. & Dauchet, M. Finding the common structure shared by two homologous RNAs. *Bioinformatics* **19**, 108 (2003).

124. Touzet, H. & Perriquet, O. CARNAC: folding families of related RNAs. *Nucleic Acids Res* **32**, W142 (2004).

125. Torarinsson, E. & Lindgreen, S. WAR: Webserver for aligning structural RNAs. *Nucleic Acids Res* **36**, W79-W84 (2008).

126. Dowell, R. D. & Eddy, S. R. Efficient pairwise RNA structure prediction and



- alignment using sequence alignment constraints. *BMC Bioinformatics* **7**, 400 (2006).
127. Harmanci, A. O., Sharma, G. & Mathews, D. H. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics* **8**, 130 (2007).
128. Mathews, D. H. Predicting a set of minimal free energy RNA secondary structures common to two sequences. *Bioinformatics* **21**, 2246 (2005).
129. Mathews, D. H. & Turner, D. H. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences1. *J Mol Biol* **317**, 191-203 (2002).
130. Torarinsson, E., Havgaard, J. H. & Gorodkin, J. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics* **23**, 926-932 (2007).
131. Havgaard, J. H., Lyngsø, R. B. & Gorodkin, J. The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. *Nucleic Acids Res* **33**, W650-W653 (2005).
132. Bindewald, E. & Shapiro, B. A. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA* **12**, 342 (2006).
133. Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F. & Backofen, R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol* **3**, e65 (2007).
134. Lindgreen, S., Gardner, P. P. & Krogh, A. MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics* **23**, 3304-3311 (2007).
135. Kiryu, H., Tabei, Y., Kin, T. & Asai, K. Murlet: a practical multiple alignment tool for structural RNA sequences. *Bioinformatics* **23**, 1588-1598 (2007).
136. Tabei, Y., Kiryu, H., Kin, T. & Asai, K. A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics* **9**, 33 (2008).
137. Harmanci, A. O., Sharma, G. & Mathews, D. H. PARTS: Probabilistic Alignment for RNA joint Secondary structure prediction. *Nucleic Acids Res* **36**, 2406 (2008).
138. Knudsen, B. & Hein, J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res* **31**, 3423-3428 (2003).
139. Hofacker, I. L., Bernhart, S. H. F. & Stadler, P. F. Alignment of RNA base pairing probability matrices. *Bioinformatics* (2004).
140. Moretti, S., Wilm, A., Higgins, D. G., Xenarios, I. & Notredame, C. R-Coffee: a web server for accurately aligning noncoding RNA sequences. *Nucleic Acids Res* (2008).
141. Wilm, A., Higgins, D. G. & Notredame, C. R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res* (2008).
142. Hamada, M., Tsuda, K., Kudo, T., Kin, T. & Asai, K. Mining frequent stem patterns from unaligned RNA sequences. *Bioinformatics* **22**, 2480 (2006).

143. Tabei, Y., Tsuda, K., Kin, T. & Asai, K. SCARNA: fast and accurate structural alignment of RNA sequences by matching fixed-length stem fragments. *Bioinformatics* **22**, 1723 (2006).
144. Meyer, I. M. & Miklós, I. SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput Biol* **3**, e149 (2007).
145. Holmes, I. Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics* **6**, 73 (2005).
146. Dalli, D., Wilm, A., Mainz, I. & Steger, G. STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics* **22**, 1593-1599 (2006).
147. Klosterman, P. S., Uzilov, A. V., Bendaña, Y. R., Bradley, R. K., *et al.* XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics* **7**, 428 (2006).
148. Andronescu, M., Aguirre-Hernández, R., Condon, A. & Hoos, H. H. RNAsoft: A suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res* **31**, 3416-3422 (2003).
149. Pedersen, J. S., Meyer, I. M., Forsberg, R., Simmonds, P. & Hein, J. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res* **32**, 4925-4936 (2004).
150. Harvey, S. C., Wang, C., Teletchea, S. & Lavery, R. Motifs in nucleic acids: molecular mechanics restraints for base pairing and base stacking. *J Comput Chem* **24**, 1-9 (2003).
151. Sheriff, S. Addition of symmetry-related contact restraints to PROTON and PROLSQ. *J Appl Cryst* **20**, 55-57 (1987).
152. Jossinet, F. & Westhof, E. Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics* **21**, 3320-3321 (2005).
153. Martinez, H. M., Maizel, J. V. & Shapiro, B. A. RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA. *J Biomol Struct Dyn* **25**, 669-683 (2008).
154. Mueller, F. & Brimacombe, R. A new model for the three-dimensional folding of Escherichia coli 16 S ribosomal RNA. I. fitting the RNA to a 3D electron microscopic map at 20 Å. *J Mol Biol* **271**, 524-544 (1997).
155. Robert, K. Z. T. & Harvey, S. C. Yammp: Development of a molecular mechanics program using the modular programming method. *J Comp Chem* **14**, 455-470 (2004).
156. Macke, T. J. & Case, D. A. Modeling unusual nucleic acid structures. *ACS Symposium Series* **682**, 379-393 (1998).
157. Frelsen, J., Moltke, I., Thiim, M., Mardia, K. V., *et al.* A probabilistic model of

RNA conformational space. *PLoS Comput Biol* **5**, e1000406 (2009).

158. Sharma, S., Ding, F. & Dokholyan, N. V. iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics* **24**, 1951 (2008).

159. Jonikas, M. A., Radmer, R. J., Laederach, A., Das, R., *et al.* Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA* **15**, 189-199 (2009).

## Chapter 2

# Characterizing the relative orientation and dynamics of RNA A-form helices

### 2.1 Introduction

The functions of many regulatory non-coding RNAs (ncRNAs) entail large conformational changes that occur in response to a variety of cellular signals, ranging from the recognition of proteins and ligands, metal binding, changes in temperature, to RNA synthesis itself<sup>1-4</sup>. Conformational transitions are a means for RNA molecules to carry out many biochemical transactions. For example, the RNA conformation required for the assembly of a complex ribonucleoprotein (RNP) may differ from that required for executing the RNP function<sup>5,6</sup>. Additionally, conformational changes provide a basis for sensing signals and transmitting regulatory responses. As an example, a large class of mRNA riboswitches regulate gene expression by changing conformation in response to recognition of small metabolite molecules or changes in temperature<sup>7,8</sup>.

The above examples, and many others, highlight the broad and rugged structural landscape that characterizes the structural properties of ncRNAs. The vast nature of this landscape is especially daunting considering the near infinite number of possible conformations that may exist; some of which may be suitable therapeutic

targets for infectious diseases<sup>9-11</sup>. Moreover, the characterization of complex structural landscapes is a challenge that must be conquered to acquire a full understanding of the structural organization and characteristics that give rise to the biological function of ncRNAs. High-throughput characterization of RNA conformations at atomic resolution is a challenge unequaled by current techniques. Despite significant advances, RNA structure determination by X-ray crystallography and NMR spectroscopy still requires several months. X-ray crystallography, in particular, is limited to conditions that yield well diffracting crystals and this can preclude insight into less ordered RNA conformers. And while NMR spectroscopy can be applied under a variety of conditions, it is currently limited to RNA molecules the size of  $\sim 100$  nt<sup>12</sup>. Moreover, heavy reliance on short-range distance constraints makes it difficult to reliably define global aspects of RNA architecture<sup>13-15</sup>.

Described herein is a strategy that incorporates the measurement and analysis of residual dipolar couplings (RDCs) with modeled idealized A-form helices. This approach allows for the rapid and reliable characterization of the relative orientation and dynamics of RNA A-form helices from partially aligned samples (Figure 2.1)<sup>16,17</sup>. The relative orientation and dynamics of A-form helical domains is a salient feature of global RNA conformation as changes in the relative orientations is mechanistically important to RNA folding, recognition and catalysis<sup>2,3,18,19</sup>. Compared with other techniques capable of determining the inter-helical bend angle between RNA helices, such as gel mobility and transient electric birefringence<sup>20,21</sup>, this strategy allows one to directly distinguish between rigid and flexible inter-

helical bends. Additionally, one can gain direct information and insights into inter-helical twist angles.

Models generated from the analysis described herein sacrifice structural resolution for greater efficiency and breadth of application. By forgoing the higher structural resolution of more complete NMR structure determination or x-ray crystallography, one is able to gain insight into potentially broader applications and larger RNA systems. Thus, while it is incapable of yielding complete highly resolved structures, it provides a new opportunity to explore and directly probe the conformational and dynamic nature of helical domains for a variety of conditions. Furthermore, these simple helix-systems are ideal model templates for further structural refinement, more sophisticated modeling, as well as synergistic integration with other spectroscopic techniques and experiments.

## **2.2 Long-range orientational information from residual dipolar couplings**

The origin of residual dipolar couplings (RDCs) arises from incomplete averaging of the dipolar interaction in partially aligned molecules<sup>16,17,22</sup>. They report on the orientation of bond vectors relative to an applied magnetic field, and specifically the time-averaged function,  $\left\langle \frac{3\cos^2\theta - 1}{2} \right\rangle$ , where  $\theta$  is the angle between the bond vector and the magnetic field (Figure 2.1). Additionally, RDCs provide a clear-cut method to obtain information concerning the relative orientation and dynamics of molecular fragments particularly when their local conformation, and specifically the orientation of RDC targeted bond vectors, is known a priori (Figure 2.2a)<sup>13,23,24</sup>.

In this work, RNA fragments consist of two or more non-terminal contiguous hydrogen-bonded Watson-Crick base pairs (Figure 2.2a). Recently, a statistical survey<sup>25</sup> was conducted of 421 such Watson-Crick base pairs derived from 40 unbound and bound RNA X-ray structures (solved with  $< 3\text{\AA}$  resolution) and the  $2.4\text{\AA}$  X-ray structure of the ribosome<sup>26</sup>. Results show that the local conformation of Watson-Crick base pairs can accurately be modeled a priori using a standard idealized A-form helix geometry<sup>27,28</sup> (Figure 2.2a). These Watson-Crick base pairs are experimentally identified by trans-hydrogen bond  $J_{\text{NN}}$ -COSY type NMR experiments, which are capable of direct detection of N-H--N hydrogen bonds<sup>29,30</sup> (Figure 2.2a). For Watson-Crick base pairs flanked by G-U pairs or non-canonical motifs, their geometries can also be incorporated into this analysis, however, a higher degree of structural noise will need to be considered<sup>31</sup>.

With the ability to model the local conformation and geometries of Watson-Crick base pairs, the measurement of more than five independent RDCs per helix is required to determine the five order tensor elements<sup>16,32</sup>. The order tensor describes the average alignment of each helix relative to the applied magnetic field. Additionally, three Euler angles specify the order tensor frame for a fixed helix ( $S_{xx}$ ,  $S_{yy}$ ,  $S_{zz}$ ), which describes its average orientation relative to the magnetic field. The average orientation of fragments can be obtained through superimposition of sequential order tensor frames (Figure 2.2c)<sup>33-35</sup>. This involves the assumption that helical fragments share, on average, a common view of the magnetic field direction when assembled into a proper structure. Two additional principal order tensor parameters exist, one to describe the degree of helix alignment (

$\vartheta = \sqrt{\frac{2}{3}(S_{xx}^2 + S_{yy}^2 + S_{zz}^2)}$  and the other its asymmetry ( $\eta = \frac{|S_{yy} - S_{xx}|}{S_{zz}}$ )<sup>33,35</sup>. Both of

these parameters can be directly compared with values obtained from other helices in the same structure to determine the existence of relative inter-domain motions occurring over sub-millisecond time scales<sup>35</sup> (Figure 2.2c).

While helices will report identical parameters when rigid with respect to one another, inter-helical motions can lead to differences. Specifically, the  $\vartheta$  value for a given helix will be attenuated relative to the observed value of a helix that dominates total alignment, with the degree of attenuation generally increasing with motional amplitudes. Although often difficult to determine reliably, the asymmetry parameter ( $\eta$ ) can provide insight into the directionality of inter-helix motions with spatially isotropic (directionless) motions having a smaller effect on the relative helix  $\eta$  values compared to anisotropic (directional) motions<sup>33,35</sup>.

Although an idealized A-form geometry is assumed for the Watson-Crick base pairs, structural deviations can, and in certain instances do, arise, creating uncertainty (e.g. “structural noise”<sup>36</sup>) that must be accounted for, and ultimately propagated throughout the RDC-derived order tensor parameterization, as well as the conformational and dynamical analysis of helices. To this end, a statistical survey of RNA X-ray structures was used to parameterize standard angular deviations in base pair and base pair step angles (buckle ( $\kappa$ ), propeller ( $\omega$ ), opening ( $\sigma$ ), incline ( $\eta$ ), tip ( $\theta$ ), and twist ( $\Omega$ )) and sugar torsions ( $\nu_0$ - $\nu_4$ ) relevant to the analysis of one bond C-H and N-H RDCs (Figure 2.2a)<sup>25</sup>. The effects of structural noise in RNA helices and



the uncertainty arising from RDC measurements were accounted for in the order tensor determination using the program AFORM-RDC<sup>25</sup>. Other more general approaches exist for dealing with the structural noise from RNA helices in the determination and analysis of alignment tensors as described elsewhere<sup>36</sup>.

### **2.3 Limits of applicability and practical considerations**

There are three main considerations that arise when implementing the strategy described here. First, how many RDCs are required for each helical fragment in order to perform the order tensor analysis? To accurately determine an RNA helix's order tensor a minimum of five spatially independent (e.g. nonparallel) RDCs are necessary to solve for each of the five order tensor parameters. For practical considerations, satisfying this condition almost always requires the measurement of more RDCs – at least 8 one bond C-H and N-H RDCs for both sugar and base moieties. In many cases, 8 RDCs will yield the necessary spatial distribution of bond vectors, as defined by a condition number (CN),  $< 5$ <sup>25,35</sup>. For  $\geq 11$  RDCs with a CN  $< 5$ , A-form structural noise and typical experimental uncertainty (ca. 1.5 Hz) is expected to yield average errors in the magnitude and orientation of the principal axis on the order of  $< 9\%$  and  $< 4^\circ$  respectively<sup>25</sup>(Figure 2.3). Errors will decrease to  $< 5\%$  and  $< 4^\circ$  for  $\geq 17$  RDCs<sup>25</sup>.

The choice of which RDCs measurements to target is guided by the desire to maximize the magnitude, precision of measurement ratio and spatial distribution of the targeted vectors. The most optimum and commonly targeted RDCs are those for directly bonded C-H and N-H nuclei as they yield the largest RDC magnitudes

(Figure 2.3a). Additionally, other one, two, and three bond RDCs can be measured from the pulse sequences listed in Table 2.1 (Figure 2.3b) to provide further modeling restraints. These latter RDCs are generally much smaller, and in some cases difficult to accurately measure, especially for larger RNAs (>60 nt). Although not discussed in depth here, it is possible to include nucleobase residual chemical shift anisotropies (RCSA) in this analysis, which can easily be measured in larger RNAs due to favorable TROSY effects<sup>37-40</sup>. However, as long as more than 8 RDC measurements with a CN < 5 are obtained the experimental strategy described herein using AFORM-RDC will yield faithful estimates of the order tensor error due to structural noise and uncertainty in the RDC measurement<sup>25</sup>.

The order tensor analysis of RDCs relies on the assumption that one of the helical fragments dominates the overall molecular alignment of the RNA<sup>33,35,41,42</sup>. This “decoupling limit” (i.e. the condition that helical motions do not influence the global alignment of RNA) is readily satisfied when one helix dominates alignment or when helices are held rigid relative to one another. There are two additional regimes to consider when helices are flexible. In the extreme coupling limit, helices are of similar size and shape, and can contribute equally to global alignment of the molecule. If the motions of helices – one relative to the other - result in equivalent changes in total alignment, then a similar degree of helical order will be observed regardless of motions<sup>42</sup>. For circumstances like these, the conclusion that  $\vartheta_{\text{int}}$  equals one implies the absence of inter-helix motions fails. However, depending on the motional trajectory of helices, nonequivalent  $\vartheta$ s can be observed for helices even if they have equivalent size and shape, thus placing them under the extreme motional

coupling limit. For example, twisting around the axis of a given helix will reduce its own  $\vartheta$  without observing a reciprocal effect on the  $\vartheta$  value of the other helix.

In the intermediate coupling limit, which is most common, one helix partially, but not completely, dominates total alignment. Here, the derived motional amplitudes will underestimate the actual motional amplitudes<sup>33</sup>. Simulations<sup>43</sup> and experimental results have shown that differences on the order of three base pairs are, in some cases, sufficient to take an RNA system from outside the extreme coupling limit and into an intermediate regime<sup>25,33,42,44-46</sup>. As a solution, elongation of helices using isotopic labeling strategies that render residues of the extended helix invisible can move the RNA into the decoupling limit<sup>47</sup>.

Finally, the order tensor analysis assumes that local fluctuations are similar in magnitude for Watson-Crick base pairs in a variety of helical contexts. A survey of NMR relaxation studies and molecular dynamics simulation analyses support this assumption for Watson-Crick base pairs that flank other canonical base pairs, which suggests that local fluctuations uniformly reduce  $\vartheta$  values, and thus do not affect derived  $\vartheta_{\text{int}}$  values by more than  $\sim 5\%$ <sup>25</sup>. Nevertheless, it will be important to independently establish and confirm the structural stability of Watson-Crick base pairs using other experimental data, such as  $J_{\text{NN-COSY}}$  mentioned above which detect hydrogen bond alignments<sup>29,30</sup> (Figure 2.2a). Conversely, one should consider that a greater degree of local motions likely arises in Watson-Crick base pairs which flank non-canonical RNA motifs<sup>25</sup>.

## **2.4 Experimental determination and modeling of RNA helices**

### **2.4.1 Experimental validation of predicted Watson-Crick base pairs**

The RNA secondary structure is divided into constituent helical stems based on contiguous Watson-Crick base pairs flanked by other Watson-Crick pairs (Figure 2.2a). It assumes that  $^1\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$  resonances in A-form helices have been assigned using established methods<sup>48-50</sup>, and that experiments are carried out on uniformly (or residue specific)  $^{13}\text{C}/^{15}\text{N}$  labeled RNA samples (typically >0.2 mM). Imino proton line widths, NOE, and trans-hydrogen bond (e.g.  $J_{\text{NN-COSY}}$ )<sup>29,30</sup> connectivity are all appropriate NMR experimental parameters that may be used to validate hydrogen bond patterns in order to identify Watson-Crick base pairs (Figure 2.2a). However, while not expected, Watson-Crick base pairs exhibiting severely exchange broadened imino proton signals and/or unusually weak NOEs and trans-hydrogen bond connectivity should be excluded from this analysis even if they happen to satisfy above the criteria.

### **2.4.2 Measurement of RDCs**

The measurement of RDCs entails recording experiments that measure scalar couplings ( $J$ ) in unaligned media and a combination of scalar and dipolar couplings ( $J+D$ ) in alignment media (Figure 2.2b). Examples of such published NMR experiments and their reference pulse sequences can be found in Table 2.1, and expected values for scalar couplings can be found elsewhere<sup>51</sup>. Scalar couplings measured under isotropic conditions actually contain a minor contribution due to spontaneous alignment of RNA in the magnetic field ( $D_{\text{field}}$ ), which is typically <1 Hz for C-H bonds in RNAs <40 nt at 600 MHz. Spontaneous alignment occurs due to the

large magnetic susceptibility ( $\chi$ ) tensor that arises from constructive summation of individual nucleobase  $\chi$ -tensors<sup>52-54</sup>. For typical RNA samples (~40nt) at 600 MHz, the degree of alignment is in the range of 1-3 Hz. The magnitude of the dipolar contribution is field dependant and increases quadratically with increasing magnetic field strength (i.e. with  $B_0^2$ ).

The majority of RNA samples are aligned by dissolution into the appropriate ordering medium<sup>55</sup> (Table 2.2). Pf1 phage is the most popular commercially available medium for aligning nucleic acids<sup>56,57</sup>. In the case of small to moderately sized RNAs (25-50 nt), 18-25 mg/ml phage will yield an optimal level of alignment (~20 Hz maximum N-H)<sup>58</sup>. Typically, one adds a pre-concentrated RNA solution (~0.5-1.5 mM) in NMR buffer to a desired volume of Pf1 phage (50 mg/mL) in the same NMR buffer to an eppendorf tube. After mixing the phage/RNA components, the solution is then gently transferred into an NMR tube while cautiously avoiding the formation bubbles. A host of other acceptable media for aligning nucleic acids is given in Table 2.2.

Regardless of the alignment media chosen it is important to verify that it does not interfere with the RNA conformation. Typically, comparisons of chemical shifts obtained in the unaligned and aligned samples are enough to determine to what extent, if any, that the alignment media has distorted the RNA conformation. However, small variations in the chemical shifts of nucleobase carbon and nitrogen atoms are expected between unaligned and aligned samples due to incomplete averaging of the residual chemical shift anisotropy (RCSA)<sup>38,39,59</sup>. The RCSA

contributions scale linearly with the magnetic field and degree of order. Typical RCSA measurements are given in Table 2.3. Ideally, one will optimize the concentration of ordering medium required for each RNA sample. In some circumstances the necessary amount of alignment can be achieved using significantly less ordering media, as is the case for larger RNAs ( $>50$  nt). For systems in which a model or structure is available, one may utilize programs to predict alignment based on steric interactions with the alignment media<sup>60</sup>. Computational methods such as these can be exploited to estimate relative levels of order, and aid in determining optimal ordering medium concentrations.

The NMR experiments employed to measure splittings of the aligned sample should be the same as the unaligned. Conducting the experiments at the same magnetic field strength and on the same magnet will help ensure that differential contributions from magnetic field RDCs, relaxation interference effects, and even discrepancies in the timing of pulse sequences are minimized. Given these considerations, one can reliably measure RDCs from the subtraction of splittings measured in aligned (J+D) and unaligned (J) samples (Figure 2.2b). It is advisable to estimate the experimental RDC uncertainty from the standard deviation of duplicate measurements. Resonances exhibiting significant differences ( $> 3\sigma$ ) as a result of considerable broadening, overlap, presence/absence of unresolved multiplets should be discarded unless a weighted fit of those data points is implemented in the order tensor analysis. Depending upon the RNA sample, typically  $\sigma$  values will range between 0.5 and 3 Hz for single bond C-H RDCs.

### 2.4.3 Order tensor analysis

First, idealized A-form helices corresponding to the targeted Watson-Crick base pairs for each helix are constructed (Figure 2.2a). Each helix should conform to published parameters for A-form Watson-Crick base pairs<sup>25,28</sup>. If one builds helices using INSIGHT II 2000.1 (Molecular Simulations, Inc.), care should be taken to correct the propeller twist angles to their proper value of -14.50. Programs such as Curves 5.1<sup>61</sup>, FreeHelix98<sup>62</sup>, 3DNA<sup>63</sup>, SCHNAaP<sup>64</sup>, NUPARM and NUCGEN<sup>65</sup> can be used to compute the relevant helical parameters.

Next, the five order tensor elements of each A-form helix are computed. This is achieved by fitting experimentally measured RDCs for each respective helix to the proposed A-form structure. Several programs are available to carry out such calculations: ORDERTEN-SVD<sup>34</sup>, REDCAT<sup>66</sup>, PALES<sup>67,68</sup>, iDC<sup>69</sup>, CONFORMIST<sup>75</sup> and RAMAH<sup>39</sup>. Measured RDCs from non-ideal Watson-Crick base pairs as identified earlier in the analysis are excluded from this portion of the computation. Following calculation of each helix's order-tensor, the correlation between measured and calculated RDCs is examined to identify major outliers, which should be interrogated for possible measurement errors. However, in smaller data sets (< 11 RDCs), major outliers may not necessarily correspond to "bad" data. Rather, one must carefully interrogate all data, as points that appear to contribute toward a good fit may in fact be problematic. This concern highlights the need to independently identify poor RDC measurements as early in the analysis as possible. Of the methods one may choose, one way to identify poor RDC measurements is

through AFORM-RDC<sup>25</sup>, or any number of other approaches<sup>36</sup> capable of correctly estimating order tensor error due to structural noise and RDC uncertainty.

#### **2.4.4 Determining the average inter-helical alignment**

To determine the average inter-helical alignment, A-form RNA helices are superimposed onto their best-fit order tensor for each respective helix. This is done through a series of three rotations which orient the helix correctly within its principal axis system (PAS). The PAS frame is the orientation in which the principal axis directions  $S_{xx}$ ,  $S_{yy}$ , and  $S_{zz}$  are oriented along the x-, y- and z-axes, respectively, of the molecular frame (Figure 2.2c). Next the orientation of each helix relative to the other must be determined. Because the RDC reports on the relative angle of the bond to the magnetic field, it is insensitive to the positive/negative direction of axes. As a result there are four rotations that will correctly orient each helix within its PAS order tensor frame. The correct relative alignment of two helices involves properly determining which of three  $180^\circ$  rotations about the principal axis directions,  $S_{xx}$ ,  $S_{yy}$ , and  $S_{zz}$ , is appropriate. Due to this four-fold degeneracy,  $n$  helices can be assembled into  $4^{n-1}$  distinct structures that satisfy the measured RDCs<sup>70</sup> (Figure 2.2c).

In general, half these solutions are eliminated as they fail to satisfy constraints arising from the RNA secondary structure. The remaining two-fold degeneracy typically characterized by a  $180^\circ$  rotation about the long axis of the RNA, and is most often eliminated due to connectivity considerations of linking residues. Other experimental restraints, such as measuring RDCs in a second non-collinear



alignment medium<sup>70</sup> eliminate this degeneracy. Extended RNA conformations (e.g. those involving coaxially stacked helices) or RNA that contain an elongated helix will often have nearly axially symmetric ( $\eta \sim 0$ ) order/alignment tensors. For such cases, rotations of a helix around the effective long axis ( $S_{zz}$ ) of the RNA molecule will be ill defined. Thus, inter-helix bend angles tend to be much better defined than their corresponding inter-helical twist angles.

Having obtained the proper relative orientations, rotated helices can be assembled into an overall RNA structure. Any uncertainty in the structure can be derived from the known uncertainty in the PAS order tensor derived by using the program AFORM-RDC<sup>25</sup>, or as mentioned above by incorporating other approaches that are able to account for uncertainty in the relative orientation of RNA helices<sup>36</sup>. Regardless of the approach, it is useful to use a convention for specifying the relative orientation of helices. Starting with a coaxial alignment of helices  $i$  and  $j$  with helix axis oriented along the positive molecular  $z$ -direction, the orientation of helix  $i$  relative to a reference helix  $j$  can be specified using three Euler angles ( $\alpha_h, \beta_h, \gamma_h$ ) that transform (through the rotation matrix  $R_{ij}(\alpha_h, \beta_h, \gamma_h)$ ) helix  $i$  from a perfectly coaxial orientation with helix  $j$  to the RDC derived inter-helical structure. The angle  $\alpha_h$  defines a twist about helix  $j$ ,  $\beta_h$  defines the inter-helix bend angle, and  $\gamma_h$  defines a twist angle about helix  $i$ . The sum of  $\alpha_h + \gamma_h$  yields the inter-helix twist angle ( $\zeta_h$ ).

In this work, the 3' helix is used as the reference and positive angles refer to anti-clockwise rotation of the molecular frame (or clockwise rotation of the object). Thus, positive and negative inter-helix twist angles ( $\zeta_h$ ) correspond to over- and

under-twisting. Note  $R(\alpha_h, \beta_h, \gamma_h) = R(\alpha_h \pm 180^\circ, -\beta_h, \gamma_h \pm 180^\circ) = R(\alpha_h \mp 180^\circ, -\beta_h, \gamma_h \pm 180^\circ)$  and  $R_{ji} = R_{ij}^{-1} = R(-\alpha_h, -\beta_h, -\gamma_h)$ . The rotation matrix  $R_{ji}$  can be expressed in terms of the rotation matrices ( $R_i'$  and  $R_j'$ ) that diagonalize the helix  $i$  and  $j$  order tensors, respectively, obtained when employing helices whose global axis is coaxial with the molecular frame. The relation,  $R_{ij} = R_i'^{-1}R_j'$ , can be used to propagate the experimental error in  $R_i'$  and  $R_j'$  estimated using A-form-RDC into an error in the structural parameters defined by  $R_{ij}$ .

#### 2.4.5 Characterizing inter-helix motions

The motional amplitude between helices  $i$  and  $j$  can be obtained from the ratios of their respective GDOs ( $\vartheta$ ) as determined from the order tensor analysis,  $\vartheta_{int} = \vartheta_i/\vartheta_j$ , where  $\vartheta_i < \vartheta_j$ <sup>35</sup> (Figure 2.2c). The value of  $\vartheta_{int}$  ranges between 1 for inter-helix rigidity to 0 for maximum inter-helix motions. Note that while a  $\vartheta_{int} < 1$  implies inter-helix motions, the motional amplitudes are likely to be underestimated due to coupling of helical motions and overall alignment<sup>33,42</sup>. In contrast, a  $\vartheta_{int} \sim 1$  implies either the rigidity of helices or that inter-helix dynamics evade detection due to motional coupling. In either case, helix elongation can be used to resolve such ambiguities<sup>47</sup>.

#### 2.5 Structure and dynamics of the HIV-1 transactivation response element

As an example, the protocol is applied to determine the relative orientation and dynamics of two helices in the free state of the HIV-1 transactivation response element (TAR). The two TAR helices are linked by a trinucleotide pyrimidine bulge<sup>33</sup>. A total of 18 (12 base, 6 sugar) and 22 (13 base, 9 sugar) one-bond C–H

RDCs were measured in helices I and II, respectively, using  $\sim 22 \text{ mg ml}^{-1}$  of Pf-1 phage ordering medium<sup>33</sup>. The RDCs were used to determine order tensors for each helix using the program RAMAH. The order tensor frames and degree of order for each helix are shown in Figure 2.4. The large difference between the helix  $\vartheta$ s implies the presence of inter-helix motions. The error bars reflect a combination of RDC measurement uncertainty and A-form structural noise as implemented in the program A-form-RDC. The motional amplitudes are given by  $\vartheta_{\text{int}} = \vartheta_{\text{helix I}} / \vartheta_{\text{helix II}} = 0.56 \pm 5.2\%$ <sup>33</sup>. Following an initial superposition, three additional solutions are generated by rotation of one helix (helix II) relative to the other by  $180^\circ$  about the  $S_{xx}$ ,  $S_{yy}$  and  $S_{zz}$  directions, respectively. The helices are translated to satisfy a direct phosphodiester linkage between the two helices (distance between O3' atom of residue C39 in helix II and P atom of residue U40 in helix I is set to  $1.58 \text{ \AA}$ ). Two solutions ( $S_{xx}$  and  $S_{yy}$ ) are omitted as they lead to antiparallel helix orientations that are inconsistent with the TAR secondary structure. The third solution ( $S_{zz}$ ) is omitted because it leads to a distance between the O3' ribose oxygen of residue A22 and the P of residue G26 that cannot be satisfactorily linked by the trinucleotide bulge. The overall free TAR conformation is thus described by inter-helix bend ( $\beta_h$ ) and twist ( $\zeta_h$ ) angles of  $47^\circ \pm 4^\circ$  and  $61^\circ \pm 30^\circ$ , respectively, and a high degree of inter-helix flexibility ( $\vartheta_{\text{int}} = 0.56 \pm 5.2\%$ ).

## 2.6 Conclusions

In general, the relative orientation and dynamics of RNA helices will depend on sequence/structural context, temperature and pH, as well as presence/absence of metals and bound protein/ligand molecules. Application of the presented strategy

is beginning to illuminate salient trends and detailed relationships between inter-helical bend, twist and flexibility, which build on observations obtained previously by gel mobility measurements<sup>20,70,71</sup> and transient electric birefringence<sup>20,21</sup>. Here, we summarize some of the results obtained thus far, noting key trends and their possible interpretations.

Shown in Figure 2.5a is the relative orientation and dynamics of helices observed in three different RNA contexts in the absence and presence of  $Mg^{2+}$  ions. This includes HIV-1 TAR, the RNase P P4 helix containing a single pyrimidine bulge nucleotide<sup>72</sup>, and HIV-1 SL1 containing a purine-rich asymmetric four-nucleotide internal loop<sup>44</sup>. In Figure 2.5b, we show corresponding results for HIV-1 TAR bound to  $Mg^{2+}$  and four different small molecules containing a different number of cationic groups.

A general trend is observed between the degree of inter-helical bending ( $\beta_h$ ), helical over-twisting ( $\zeta_h$ ) and inter-helical flexibility ( $\vartheta_{int}$ ) (Figure 2.5). The greater the bend angle, the greater the degree of observed over-twisting and inter-helix flexibility. This behavior can be understood in terms of the stacking and electrostatic interactions at the bulge/internal loop that dictate the resulting orientation and dynamics of juxtaposed helices. Bulges and asymmetric internal loops induce inter-helical bending and over-twisting for two main reasons<sup>20,70,71</sup>. First, by extending the bulge/internal loop conformation, inter-helical bending alleviates electrostatic charge repulsion that would otherwise build up in coaxial structures owing to spatial confinement of bulge/internal loop phosphates. Second, inter-helical bending accommodates looped in conformations allowing favorable bulge/internal

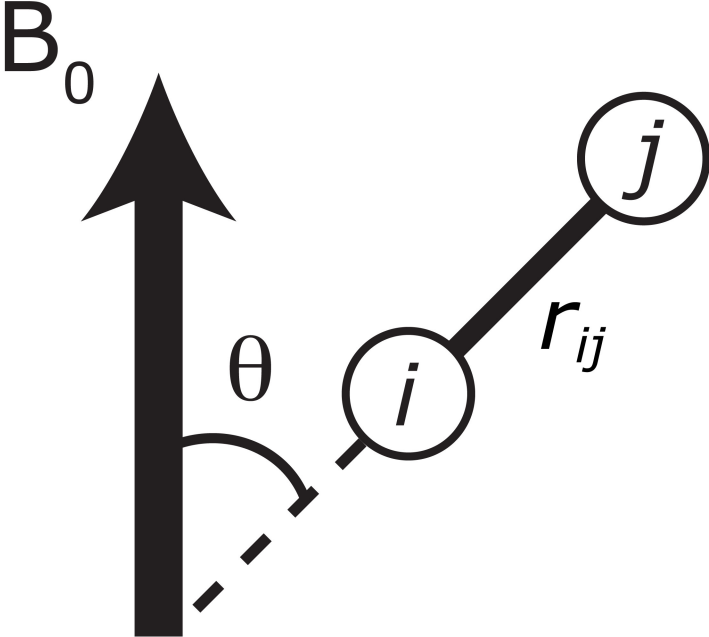
loop stacking interactions. It is the stacking interactions of bulge/internal loop residues that account for the observed inter-helical over-twisting. In the absence of tertiary contacts that stabilize the relative orientation of helices, the degree of inter-helical bending, over-twisting and inter-helix flexibility is expected to increase with the length of the bulge<sup>21,70,73-76</sup>, as is observed when comparing TAR and P4 both of which contain pyrimidine bulges (Figure 2.5a). Unopposed asymmetric bulges are expected to give rise to greater bending/flexibility compared to opposed symmetric internal loops that may have compensating bending effects<sup>21,70,73-76</sup>, as observed when comparing TAR and SL1 (Figure 2.5a).

By screening unfavorable electrostatic charge repulsion in and around the bulge/internal loops, divalent and monovalent ions (or small molecule containing positive groups) can help stabilize coaxial helical conformations<sup>20,21,77-80</sup>. However, this will often require the looping out of bulge/internal loop residues. Thus, the energetic gains owing to favorable coaxial helical stacking and metal binding have to offset the unfavorable loss of stacking interactions in bulge internal/loop residues. In the case of TAR, Mg<sup>2+</sup> binding induces a large structural transition toward a rigid coaxial inter-helical conformation (Figure 2.5a)<sup>81</sup>. This transition is accompanied by looping out of the otherwise stacked nucleobases of pyrimidine bulge residues U23 and C24<sup>81,82</sup>. In contrast to TAR, Mg<sup>2+</sup> binding has an insignificant effect on the P4 conformation for which favorable coaxial helical stacking and looping out of the uridine bulge is already observed in the absence of Mg<sup>2+</sup> (Figure 2.5a)<sup>72</sup>. Smaller conformational effects are also seen for SL1, which contains a purine-rich internal loop (Figure 2.5a)<sup>44</sup>.

Previous NMR studies have shown that TAR RNA undergoes conformational rearrangements upon binding to small molecule therapeutics bearing a different number and spatial arrangement of cationic groups<sup>83-87</sup>. Figure 2.5b shows these TAR conformational transitions as visualized through application of the presented RDC protocol<sup>33,81,88,89</sup>. Interestingly, one finds that molecules that contribute a larger number of cationic groups tend to stabilize more linear and rigid TAR conformations (Figure 2.5b) – in analogy to the trend observed when adding Mg<sup>2+</sup> metals (Figure 2.5a)<sup>89,90</sup>.

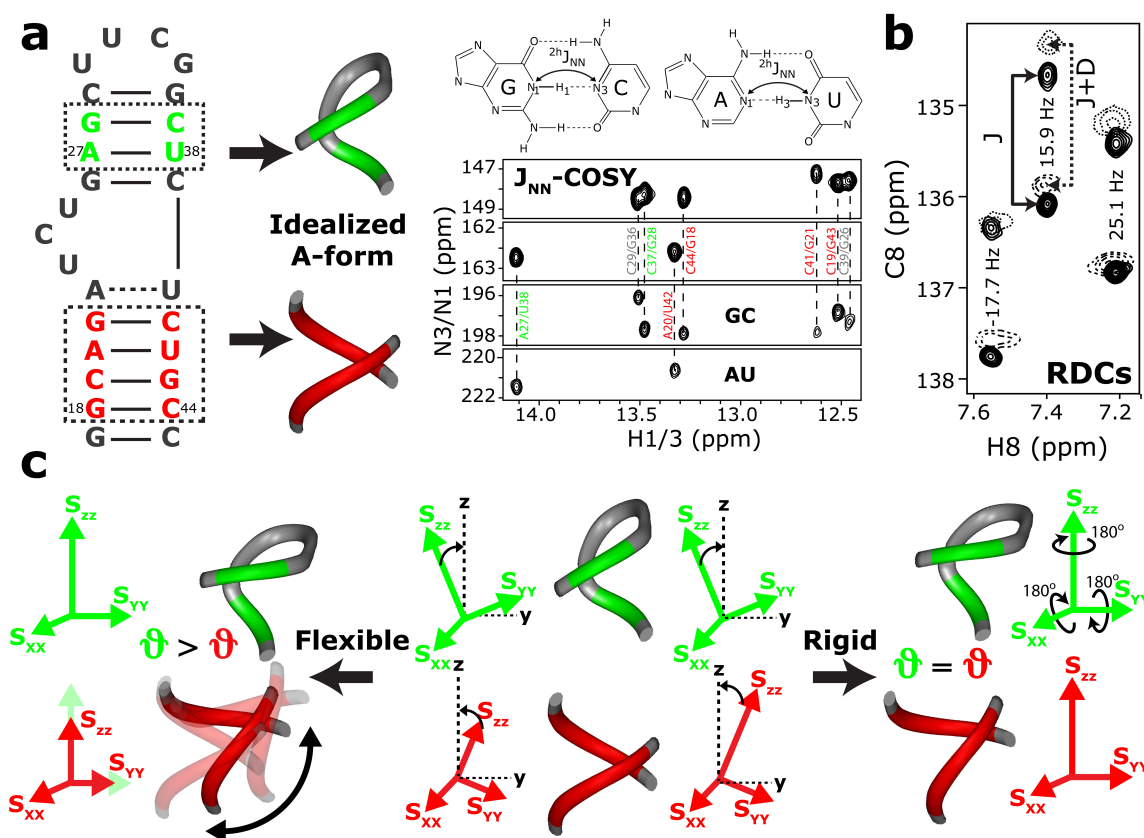
The above illustrates the diversity of conformations that can populate the RNA structural landscape and the possibility for systematic characterization using the described protocol. In the future, we expect applications to more complex RNA contexts, including those involving long-range tertiary contacts and inter-helical linkers composed of pseudo-knots and junctions under a wide range of physiologically relevant conditions.

This work was published in the journal *Nature Protocols*<sup>91</sup>. HM Al-Hashimi and MH Bailor conceived the idea. C Musselman developed the program A-FORM-RDC to propagate structural noise and RDC uncertainty into errors. AL Hansen developed the program RAMAH used to carry out SVD analysis of RDC measurements. MH Bailor calculated and collected Inter-helical Euler angles from RNA structures described within. HM Al-Hashimi and MH Bailor analyzed data.



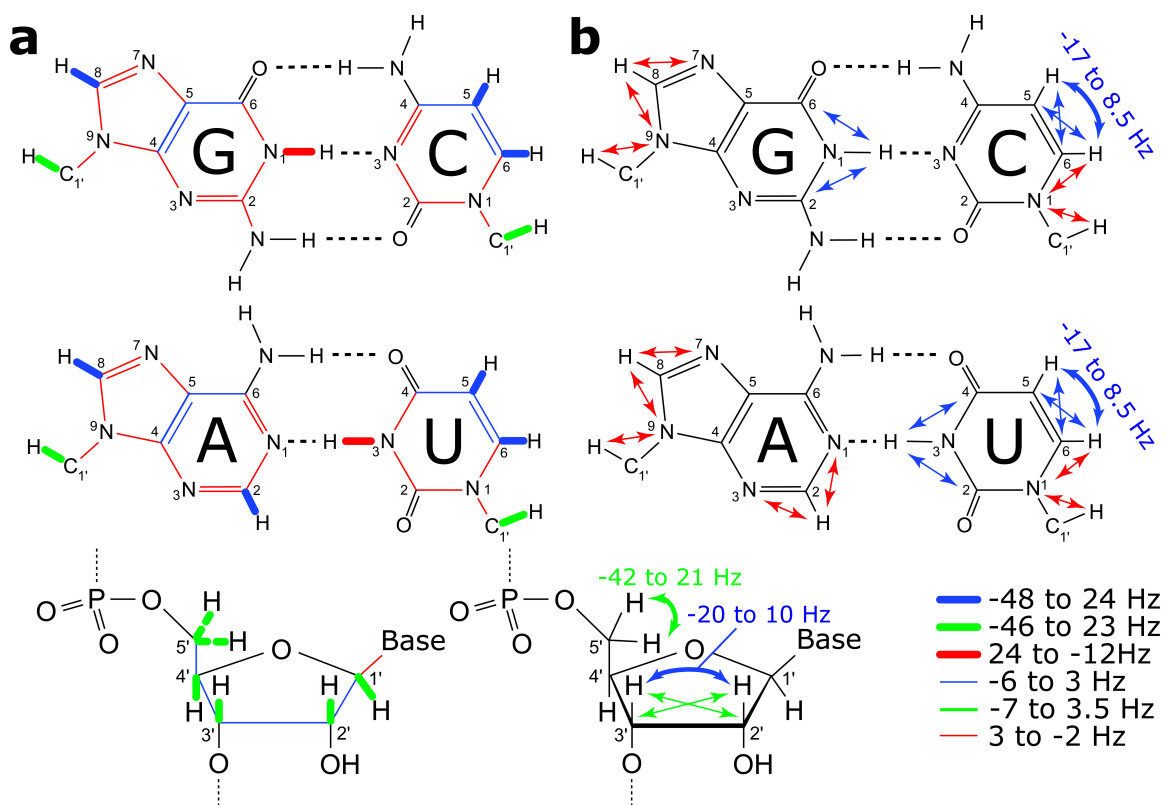
$$D_{ij} = -\frac{\gamma_i \gamma_j \mu_0 h}{(2\pi)^3 r_{ij}^3} \left\langle \frac{3\cos^2 \theta - 1}{2} \right\rangle$$

**Figure 2.1: The field dependant orientation and distance dependence of RDCs.** RDCs ( $D_{ij}$ ) between spins  $i$  and  $j$  provide long-range constraints on the average orientation ( $\theta$ ) of the internuclear bond vector relative to the applied magnetic field.

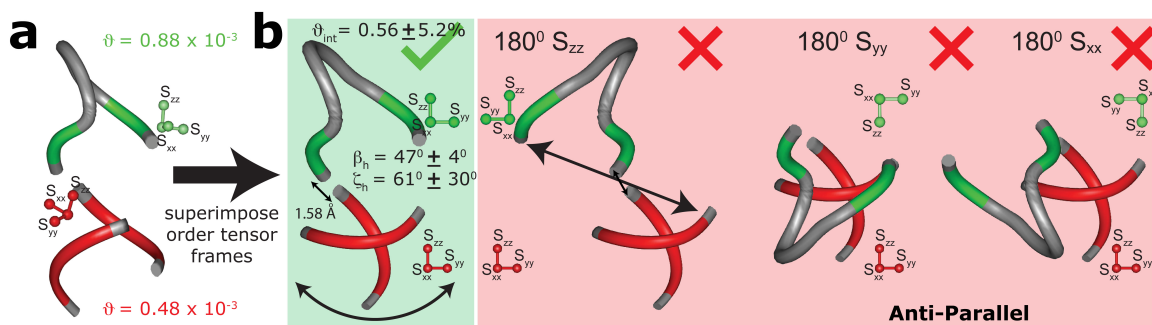


**Figure 2.2: Determining the relative orientation and dynamics of A-form helices using an order tensor analysis of RDCs.** (a) Watson-Crick base pairs that are flanked by other Watson-Crick base pairs are identified based on the predicted RNA secondary structural model. Resonance assignments in Watson-Crick base pairs are established using  $J_{NN}$ -COSY connectivity (shown in the Figure) or using through bond correlation experiments. The local structure of the experimentally verified Watson-Crick pairs is modeled using idealized A-form Helices. Next, NMR experiments (Table 2.1) are used to measure splittings between various nuclei under aligned (J + D) (Table 2.2) and unaligned (J) conditions. Note that differences in the chemical shifts (center of doublet) between aligned and unaligned conditions arise owing to a combination of RCSA contributions and different lock frequencies as a result of quadrupolar splitting of the  $D_2O$  signal in the aligned state. (b) RDCs are computed from the differences in these values and, together with the idealized A-form PDBs, are used to compute order tensors for each helix. (c) The helix order tensor frames ( $S_{xx}$ ,  $S_{yy}$ ,  $S_{zz}$ ) are superimposed to yield the relative orientation of helices subject to a  $4^{n-1}$ -fold degeneracy arising owing to allowed  $180^\circ$  inversions about the principal  $S_{xx}$ ,  $S_{yy}$ , and  $S_{zz}$  directions. Information about interhelix motions is obtained from the relative ratios of the generalized degree of order ( $\vartheta$ ) obtained for each helix ( $\vartheta_i/\vartheta_j = \vartheta_{int}$ ;  $\vartheta_i < \vartheta_j$ ). The  $\vartheta_{int}$  value ranges between 0 for maximum interhelix motions and 1 for interhelix rigidity.





**Figure 2.3: Typical RDCs measured in base and sugar moieties of RNA using the pulse sequences listed in Table 2.1.** (a) One-bond C-H and N-H RDCs are the most commonly targeted interactions owing to their favorable size but smaller one-bond C-C and C-N as well as (b) two- and three-bond RDCs can be measured. The motionally non-averaged C-H and N-H bond lengths used in the order tensor analysis are  $N_{1/3}-H_{1/3} \frac{1}{4} 1.01 \text{ \AA}$ ,  $C-H_{\text{base}} \frac{1}{4} 1.08 \text{ \AA}$ ,  $C-H_{\text{ribose}} \frac{1}{4} 1.09 \text{ \AA}$ <sup>92</sup>. All other bond lengths can be obtained from ref.<sup>93</sup>.



**Figure 2.4: Implementation of strategy to determine the relative orientation and dynamics of two helices for the free state of HIV-1 TAR.** (a) Idealized A-form helices are used to determine order tensors for each helix in TAR RNA, with 18 (12 base, 6 sugar) and 22 (13 base, 9 sugar) one-bond C-H RDCs used in the analysis of helices I and II, respectively. (b) Superposition of the experimentally determined order tensor frames yields one of four solutions for the relative orientation of helices. Three additional degenerate solutions ( $180^\circ S_{xx}$ ,  $180^\circ S_{yy}$ ,  $180^\circ S_{zz}$ ) are generated by subsequent rotation of a given helix (in this case helix II) by  $180^\circ$  about each of helix II three principal axes  $S_{xx}$ ,  $S_{yy}$ , and  $S_{zz}$ , respectively. In each case, the helices are translated/assembled by setting the distance between O3' of residue 39 and P of residue 40 equal to 1.58 Å. The  $180^\circ S_{zz}$  solution is discarded because it yields a distance between O3' of residue 22 and P of residue 40 that is long to be satisfactorily connected by a trinucleotide bulge, whereas solutions  $180^\circ S_{yy}$  and  $180^\circ S_{xx}$  are discarded because they lead to an antiparallel helix alignment that is inconsistent with the TAR secondary structure.



**Table 2.1: Pulse sequences implemented in the measurement of scalar and residual dipolar couplings in nucleic acids.**

Pulse Sequence	Reference	Type of RDCs	Comments
HC[C](C) hd-TROSY-E.COSY	51	$^1D_{C2H2}$ , $^1D_{C5H5}$ , $^1D_{C6H6}$ , $^1D_{C8H8}$ , $^1D_{C4C5}$ , $^1D_{C5C6}$ , $^2D_{C5H6}$ , $^2D_{C6H5}$ and $^2D_{C4H5}$	Pseudo-3D experiments for homonuclear decoupling employing TROSY and E.COSY elements. Demonstrated on a 24-nt RNA @ 25°C
CH <sub>2</sub> -S <sup>3</sup> E HSQC	96	$^1D_{(C5'H5'+C5'H5'')}$ and $^2D_{(H5'H5'')}$ (in DNA only $^1D_{(C2'H2'+C2'H2'')}$ and $^2D_{(H2'H2'')}$ )	2D experiments with spin-state selection for detection of up- or downfield carbon components of CH <sub>2</sub> spin states. Demonstrated on a 24-nt RNA @ 25°C
3D S <sup>3</sup> CT E.COSY	97	$^1D_{C4'H4'}$ , $^2D_{C5'H4'}$ , $^1D_{(C5'H5'+C5'H5'')}$ , ( $^1D_{C5'H5'[1]}$ - $^2D_{H5'H5'}$ ), $^2D_{C4'H5'+C4'H5'}$ , and $^3D_{H4'H5'[1]}$	3D experiments for measuring RDCs in methine-methylene C-H pairs. One experiment yields 8 splittings. Demonstrated on 24-nt RNA @ 25°C
H1C1C2 E.COSY	98	$^1D_{C1'H1'}$ , $^1D_{C2'H2'}$ , $^2D_{C1'H2'}$ , $^2D_{C2'H1'}$ , and $^3D_{H1'H2'}$	3D experiment utilizing E.COSY for measuring five splittings in one experiment. Demonstrated on a 24-nt RNA @ 25°C
IPAP HN-HSQC, IPAP H(N)C-HSQC	99	$^1D_{N1H1}$ , $^1D_{N3H3}$ , $^2D_{H1C2}$ , $^2D_{H1C6}$ , $^2D_{H3C2}$ , and $^2D_{H3C4}$	2D experiments yielding 1-2 couplings per experiment. Demonstrated on ubiquitin.
3D IPAP-HC <sub>2</sub> H-COSY 3D relay-HC <sub>2</sub> H-COSY	100	$^1D_{C2'H2'}$ and $^1D_{C3'H3'}$ ,	Uses C1'H1' to alleviate spectral overcrowding in the C2'H2' and C3'H3' region. Demonstrated on a 42-nt RNA @ 25°C
MQ-HCN	101	$^1D_{C1'H1'}$ , $^1D_{C1'N1/N9}$ , $^1D_{C1'C2'}$ , $^2D_{H1'N1/9}$ , $^2D_{H1'C2'}$ , $^2D_{H1'N1/9}$ , $^1D_{C6H6}$ , $^1D_{C6N1}$ , $^1D_{C6C5}$ , $^1D_{C8H8}$ , $^1D_{C8N9}$ , $^2D_{H8N9}$ , $^2D_{H6N1}$ , and $^2D_{H6C5}$	Suite of six MQ based 3D experiments. One - two splittings per experiment. Demonstrated on a 36-nt DNA in a 47 kDa complex
S <sup>3</sup> E IS[T]	102	$^1D$ and $^2D$	2D experiments for measuring most of the one and two bond splittings. Demonstrated on 24-nt DNA @ 15°C
<sup>13</sup> C- <sup>1</sup> H TROSY	103	$^1D_{C2H2}$ , $^1D_{C5H5}$ , $^1D_{C6H6}$ , and $^1D_{C8H8}$	Sensitivity enhanced using TROSY and native <sup>13</sup> C magnetization. Demonstrated on 15% randomly <sup>13</sup> C labeled 33-nt RNA @ 25°C.
3D MQ/TROSY-HCN-QJ	104	$^1D$ Pur. C1'N9, C8N9, C4N9, Pyr. C1'N1, C6N1, C2N1	3D quantitative J-modulated experiments for measuring one bond C-N splittings. Demonstrated on a 24-nt RNA @ 8°C

**Table 2.2: Established media used to align nucleic acids in structural studies.**

Medium	Reference	Temp. range (°C)	Features and limitations
DMPC:DHPC ("Bicelles")	105,106	27-45	Perpendicular alignment disc-like shape. Sensitive to ionic conditions.
Rod-shaped viruses (Pf1 phage and TMV)	56-58	5-60	Parallel alignment rod-like shape. Negatively charged. Most widely used.
Purple membrane	107,108	-269-69	Parallel alignment disc-like shape. Stable in pH range 2.5 to 10, and salt concentrations up to 5 M
Polyacrylamide Gels	109,110	5-45	Mechanical gel. Very stable and inert
n-Alkyl-poly(ethylene glycol)/n-alkyl alcohol or glucopone/n-hexanol (PEG)	111,112	0-40	Perpendicular alignment lamellar shape. In-sensitive to pH, and moderately sensitive to salt concentrations.

**Table 2.3: Chemical shift changes for protonated carbons and nitrogens due to RCSA contributions from alignment assuming  $1.0 \times 10^{-3}$  degree of order. Corresponding shifts for attached protons are  $\pm 8$  ppb.**

Atom	Base (ppb)		Sugar (ppb)			
	2/5/6/8	1'	2'	3'	4'	5'
<sup>13</sup> C	$\pm 170$	$\pm 23$	$\pm 20$	$\pm 67$	$\pm 67$	$\pm 40$
<sup>15</sup> N	$\pm 100$	NA	NA	NA	NA	NA

## 2.7 References

1. Williamson, J. R. Small subunit, big science. *Nature* **407**, 306-307 (2000).
2. Leulliot, N. & Varani, G. Current topics in RNA-protein recognition: control of specificity and biological function through induced fit and conformational capture. *Biochemistry* **40**, 7947-7956 (2001).
3. Al-Hashimi, H. M. Dynamics-based amplification of RNA function and its characterization by using NMR spectroscopy. *ChemBiochem* **6**, 1506-1519 (2005).
4. Serganov, A., Yuan, Y. R., Pikovskaya, O., Polonskaia, A., *et al.* Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem Biol* **11**, 1729-1741 (2004).
5. Williamson, J. R. Assembly of the 30S ribosomal subunit. *Q Rev Biophys* **38**, 397-403 (2005).
6. Schroeder, R., Barta, A. & Semrad, K. Strategies for RNA folding and assembly. *Nat Rev Mol Cell Biol* **5**, 908-919 (2004).
7. Mandal, M. & Breaker, R. R. Gene regulation by riboswitches. *Nat Rev Mol Cell Biol* **5**, 451-463 (2004).
8. Grundy, F. J. & Henkin, T. M. Regulation of gene expression by effectors that bind to RNA. *Curr Opin Microbiol* **7**, 126-131 (2004).
9. Hermann, T. Rational ligand design for RNA: the role of static structure and conformational flexibility in target recognition. *Biochimie* **84**, 869-875 (2002).
10. Tor, Y. Targeting RNA with small molecules. *ChemBiochem* **4**, 998-1007 (2003).
11. Vicens, Q. & Westhof, E. RNA as a drug target: the case of aminoglycosides. *ChemBiochem* **4**, 1018-1023 (2003).
12. D'Souza, V., Dey, A., Habib, D. & Summers, M. F. NMR structure of the 101-nucleotide core encapsidation signal of the Moloney murine leukemia virus. *J Mol Biol* **337**, 427-442 (2004).
13. Bax, A. & Grishaev, A. Weak alignment NMR: a hawk-eyed view of biomolecular structure. *Curr Opin Struct Biol* **15**, 563-570 (2005).
14. MacDonald, D. & Lu, P. Residual dipolar couplings in nucleic acid structure determination. *Curr Opin Struct Biol* **12**, 337-343 (2002).
15. Mollova, E. T., Hansen, M. R. & Pardi, A. Global structure of RNA determined with residual dipolar couplings. *J Am Chem Soc* **122**, 11561-11562 (2000).
16. Tjandra, N. & Bax, A. Measurement of dipolar contributions to  $1J_{CH}$  splittings from magnetic-field dependence of J modulation in two-dimensional NMR spectra. *J Magn Reson* **124**, 512-515 (1997).
17. Tolman, J. R., Flanagan, J. M., Kennedy, M. A. & Prestegard, J. H. Nuclear magnetic dipole interactions in field-oriented proteins: information for structure

- determination in solution. *Proc Natl Acad Sci U S A* **92**, 9279-9283 (1995).
18. Williamson, J. R. Induced fit in RNA-protein recognition. *Nat Struct Biol* **7**, 834-837 (2000).
  19. Lilley, D. M. The Varkud satellite ribozyme. *RNA* **10**, 151-158 (2004).
  20. Zacharias, M. & Hagerman, P. J. Bulge-induced bends in RNA: quantification by transient electric birefringence. *J Mol Biol* **247**, 486-500 (1995).
  21. Zacharias, M. & Hagerman, P. J. The influence of symmetric internal loops on the flexibility of RNA. *J Mol Biol* **257**, 276-289 (1996).
  22. Richards, R. J., Wu, H., Trantirek, L., O'Connor, C. M., *et al.* Structural study of elements of Tetrahymena telomerase RNA stem-loop IV domain important for function. *RNA* **12**, 1475-1485 (2006).
  23. Bothner-By, A. A. *In Encyclopedia of nuclear magnetic resonance* (Wiley, Chichester, 1995).
  24. Prestegard, J. H., al-Hashimi, H. M. & Tolman, J. R. NMR structures of biomolecules using field oriented media and residual dipolar couplings. *Q Rev Biophys* **33**, 371-424 (2000).
  25. Musselman, C., Pitt, S. W., Gulati, K., Foster, L. L., *et al.* Impact of static and dynamic A-form heterogeneity on the determination of RNA global structural dynamics using NMR residual dipolar couplings. *J Biomol NMR* **36**, 235-249 (2006).
  26. Tolman, J. R. & Ruan, K. NMR residual dipolar couplings as probes of biomolecular dynamics. *Chem Rev* **106**, 1720-1736 (2006).
  27. Al-Hashimi, H. M., Gorin, A., Majumdar, A., Gosser, Y. & Patel, D. J. Towards structural genomics of RNA: rapid NMR resonance assignment and simultaneous RNA tertiary structure determination using residual dipolar couplings. *J Mol Biol* **318**, 637-649 (2002).
  28. Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., *et al.* A standard reference frame for the description of nucleic acid base pair geometry. *J Mol Biol* **313**, 229-237 (2001).
  29. Klein, D. J., Schmeing, T. M., Moore, P. B. & Steitz, T. A. The kink-turn: a new RNA secondary structure motif. *EMBO J* **20**, 4214-4221 (2001).
  30. Dingley, A. J. & Grzesiek, S. Direct observation of hydrogen bonds in nucleic acid base pairs by internucleotide 2JNN couplings. *J Am Chem Soc* **120**, 8293-8297 (1998).
  31. Pervushin, K., Ono, A., Fernández, C., Szyperski, T., *et al.* NMR scalar couplings across Watson-Crick base pair hydrogen bonds in DNA observed by transverse relaxation-optimized spectroscopy. *Proc Natl Acad Sci U S A* **95**, 14147-14151 (1998).
  32. Saupe, A. Recent results in the field of liquid crystals. *Angew Chem Int E E* **7**, (1968).

33. Al-Hashimi, H. M., Gosser, Y., Gorin, A., Hu, W., *et al.* Concerted motions in HIV-1 TAR RNA may allow access to bound state conformations: RNA dynamics from NMR residual dipolar couplings. *J Mol Biol* **315**, 95-102 (2002).
34. Losonczi, J. A., Andrec, M., Fischer, M. W. F. & Prestegard, J. H. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J Magn Reson* **138**, 334-342 (1999).
35. Tolman, J. R., Al-Hashimi, H. M., Kay, L. E. & Prestegard, J. H. Structural and dynamic analysis of residual dipolar coupling data for proteins. *J Am Chem Soc* **123**, 1416-1424 (2001).
36. Zweckstetter, M. & Bax, A. Evaluation of uncertainty in alignment tensors obtained from dipolar couplings. *J Biomol NMR* **23**, 127-137 (2002).
37. Pervushin, K., Riek, R., Wider, G. & Wuthrich, K. Transverse relaxation-optimized spectroscopy (TROSY) for NMR studies of aromatic spin systems in <sup>13</sup>C-labeled proteins. *J Am Chem Soc* **120**, 6394-6400 (1998).
38. Ying, J., Grishaev, A., Bryce, D. L. & Bax, A. Chemical shift tensors of protonated base carbons in helical RNA and DNA from NMR relaxation and liquid crystal measurements. *J Am Chem Soc* **128**, 11443-11454 (2006).
39. Hansen, A. L. & Al-Hashimi, H. M. Insight into the CSA tensors of nucleobase carbons in RNA polynucleotides from solution measurements of residual CSA: towards new long-range orientational constraints. *J Magn Reson* **179**, 299-307 (2006).
40. Grishaev, A., Ying, J. & Bax, A. Pseudo-CSA restraints for NMR refinement of nucleic acid structure. *J Am Chem Soc* **128**, 10010-10011 (2006).
41. Briggman, K. B. & Tolman, J. R. De novo determination of bond orientations and order parameters from residual dipolar couplings with high accuracy. *J Am Chem Soc* **125**, 10164-10165 (2003).
42. Zhang, Q., Throolin, R., Pitt, S. W., Serganov, A. & Al-Hashimi, H. M. Probing motions between equivalent RNA domains using magnetic field induced residual dipolar couplings: accounting for correlations between motions and alignment. *J Am Chem Soc* **125**, 10530-10531 (2003).
43. Zhang, Q. & Al-Hashimi, H. M. Extending the NMR spatial resolution limit for RNA by motional couplings. *Nat Methods* **5**, 243-245 (2008).
44. Sun, X., Zhang, Q. & Al-Hashimi, H. M. Resolving fast and slow motions in the internal loop containing stem-loop 1 of HIV-1 that are modulated by Mg<sup>2+</sup> binding: role in the kissing-duplex structural transition. *Nucleic Acids Res* **35**, 1698-1713 (2007).
45. Leeper, T. C., Athanassiou, Z., Dias, R. L., Robinson, J. A. & Varani, G. TAR RNA recognition by a cyclic peptidomimetic of Tat protein. *Biochemistry* **44**, 12362-12372 (2005).
46. Chen, Y., Fender, J., Legassie, J. D., Jarstfer, M. B., *et al.* Structure of stem-loop IV



- of Tetrahymena telomerase RNA. *The EMBO Journal* **25**, 3156 (2006).
47. Zhang, Q., Sun, X., Watt, E. D. & Al-Hashimi, H. M. Resolving the motional modes that code for RNA adaptation. *Science* **311**, 653-656 (2006).
48. Varani, G., Aboul-ela, F. & Allain, F. H. T. NMR investigation of RNA structure. *Progress in Nuclear Magnetic Resonance Spectroscopy* **29**, 51-127 (1996).
49. Wijmenga, S. S. & van Buuren, B. N. M. The use of NMR methods for conformational studies of nucleic acids. *Progress in Nuclear Magnetic Resonance Spectroscopy* **32**, 287-387 (1998).
50. Fürtig, B., Richter, C., Wöhnert, J. & Schwalbe, H. NMR spectroscopy of RNA. *Chembiochem* **4**, 936-962 (2003).
51. Kontaxis, G., Clore, G. M. & Bax, A. Evaluation of cross-correlation effects and measurement of one-bond couplings in proteins with short transverse relaxation times. *J Magn Reson* **143**, 184-196 (2000).
52. Hansen, M. R., Hanson, P. & Pardi, A. Filamentous bacteriophage for aligning RNA, DNA, and proteins for measurement of nuclear magnetic resonance dipolar coupling interactions. *Methods Enzymol* **317**, 220-240 (2000).
53. Ottiger, M., Tjandra, N. & Bax, A. Magnetic Field Dependent Amide <sup>15</sup>N Chemical Shifts in a Protein-DNA Complex Resulting from Magnetic Ordering in Solution. *Chem Phys Lett* **87**, 192-196 (1982).
54. Kung, H. C., Wang, K. Y., Goljer, I. & Bolton, P. H. Magnetic alignment of duplex and quadruplex DNAs. *J Magn Reson, Series B* **109**, 323-325 (1995).
55. Redfield, A. G. On the theory of relaxation processes. *IBM J Res Dev* **1**, 19 (1957).
56. Alba, E. & Tjandra, N. On the accurate measurement of amide one-bond <sup>15</sup>N--<sup>1</sup>H couplings in proteins: Effects of cross-correlated relaxation, selective pulses and dynamic frequency shifts. *J Magn Reson* **183**, 160-165 (2006).
57. Boisbouvier, J., Bryce, D. L., O'neil-Cabello, E., Nikonowicz, E. P. & Bax, A. Resolution-optimized NMR measurement of (<sup>1</sup>)D(CH), (<sup>1</sup>)D(CC) and (<sup>2</sup>)D(CH) residual dipolar couplings in nucleic acid bases. *J Biomol NMR* **30**, 287-301 (2004).
58. Prestegard, J. H. & Kishore, A. I. Partial alignment of biomolecules: an aid to NMR characterization. *Curr Opin Chem Biol* **5**, 584-590 (2001).
59. Clore, G. M., Starich, M. R. & Gronenborn, A. M. Measurement of residual dipolar couplings of macromolecules aligned in the nematic phase of a colloidal suspension of rod-shaped viruses. *J Am Chem Soc* **120**, 10571-10572 (1998).
60. Tjandra, N., Omichinski, J. G., Gronenborn, A. M., Clore, G. M. & Bax, A. Use of dipolar <sup>1</sup>H-<sup>15</sup>N and <sup>1</sup>H-<sup>13</sup>C couplings in the structure determination of magnetically oriented macromolecules in solution. *Nat Struct Biol* **4**, 732-738 (1997).
61. Ravishanker, G., Swaminathan, S., Beveridge, D. L., Lavery, R. & Sklenar, H. Conformational and helicoidal analysis of 30 PS of molecular dynamics on the d

(CGCGAATTCGCG) double helix: "curves", dials and windows. *Journal of biomolecular structure & dynamics* **6**, 669 (1989).

62. Dickerson, R. E. DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res* **26**, 1906 (1998).

63. Lu, X. J. & Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res* **31**, 5108 (2003).

64. Lu, X. J., El Hassan, M. A. & Hunter, C. A. Structure and conformation of helical nucleic acids: analysis program (SCHNAaP). *J Mol Biol* **273**, 668-680 (1997).

65. Bansal, M., Bhattacharyya, D. & Ravi, B. NUPARM and NUCGEN: software for analysis and generation of sequence dependent nucleic acid structures. *Comput Appl Biosci* **11**, 281-287 (1995).

66. Valafar, H. & Prestegard, J. H. REDCAT: a residual dipolar coupling analysis tool. *J Magn Reson* **167**, 228-241 (2004).

67. Zweckstetter, M. & Bax, A. Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. *J Am Chem Soc* **122**, 3791-3792 (2000).

68. Zweckstetter, M., Hummer, G. & Bax, A. Prediction of charge-induced molecular alignment of biomolecules dissolved in dilute liquid-crystalline phases. *Biophys J* **86**, 3444-3460 (2004).

69. Wei, Y. & Werner, M. H. iDC: a comprehensive toolkit for the analysis of residual dipolar couplings for macromolecular structure determination. *J Biol NMR* **35**, 17-25 (2006).

70. Skrynnikov, N. R., Goto, N. K., Yang, D., Choy, W. Y., *et al.* Orienting domains in proteins using dipolar couplings measured by liquid-state NMR: differences in solution and crystal forms of maltodextrin binding protein loaded with  $\beta$ -cyclodextrin. *J Mol Biol* **295**, 1265-1273 (2000).

71. Al-Hashimi, H. M., Valafar, H., Terrell, M., Zartler, E. R., *et al.* Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings. *J Magn Reson* **143**, 402-406 (2000).

72. Getz, M. M., Andrews, A. J., Fierke, C. A. & Al-Hashimi, H. M. Structural plasticity and Mg<sup>2+</sup> binding properties of RNase P P4 from combined analysis of NMR residual dipolar couplings and motionally decoupled spin relaxation. *RNA* **13**, 251-266 (2007).

73. Tang, R. S. & Draper, D. E. Bulge loops used to measure the helical twist of RNA in solution. *Biochemistry* **29**, 5232-5237 (1990).

74. Bhattacharyya, A., Murchie, A. & Lilley, D. M. RNA bulges and the helical periodicity of double-stranded RNA. **343**, 484-487 (1990).

75. Bhattacharyya, A. & Lilley, D. M. Single base mismatches in DNA. Long- and short-range structure probed by analysis of axis trajectory and local chemical

reactivity. *J Mol Biol* **209**, 583-597 (1989).

76. Riordan, F. A., Bhattacharyya, A., McAteer, S. & Lilley, D. M. Kinking of RNA helices by bulged bases, and the structure of the human immunodeficiency virus transactivator response element. *J Mol Biol* **226**, 305-310 (1992).

77. Tang, R. S. & Draper, D. E. Bend and helical twist associated with a symmetric internal loop from 5S ribosomal RNA. *Biochemistry* **33**, 10089-10093 (1994).

78. Tang, R. S. & Draper, D. E. On the use of phasing experiments to measure helical repeat and bulge loop-associated twist in RNA. *Nucleic Acids Res* **22**, 835-841 (1994).

79. Kim, H. D., Nienhaus, G. U., Ha, T., Orr, J. W., *et al.* Mg<sup>2+</sup>-dependent conformational change of RNA studied by fluorescence correlation and FRET on immobilized single molecules. *Proc Natl Acad Sci U S A* **99**, 4284-4289 (2002).

80. Rueda, D., Wick, K., McDowell, S. E. & Walter, N. G. Diffusely Bound Mg<sup>2+</sup> Ions Slightly Reorient Stems I and II of the Hammerhead Ribozyme To Increase the Probability of Formation of the Catalytic Core†. *Biochemistry* **42**, 9924-9936 (2003).

81. Al-Hashimi, H. M., Pitt, S. W., Majumdar, A., Xu, W. & Patel, D. J. Mg<sup>2+</sup>-induced variations in the conformation and dynamics of HIV-1 TAR RNA probed using NMR residual dipolar couplings. *J Mol Biol* **329**, 867-873 (2003).

82. Ippolito, J. A. & Steitz, T. A. A 1.3-Å resolution crystal structure of the HIV-1 trans-activation response region RNA stem reveals a metal ion-dependent bulge conformation. *Proc Natl Acad Sci U S A* **95**, 9819-9824 (1998).

83. Du, Z., Lind, K. E. & James, T. L. Structure of TAR RNA complexed with a Tat-TAR interaction nanomolar inhibitor that was identified by computational screening. *Chem Biol* **9**, 707-712 (2002).

84. Faber, C., Sticht, H., Schweimer, K. & Rösch, P. Structural rearrangements of HIV-1 Tat-responsive RNA upon binding of neomycin B. *J Biol Chem* **275**, 20660-20666 (2000).

85. Aboul-ela, F., Karn, J. & Varani, G. Structure of HIV-1 TAR RNA in the absence of ligands reveals a novel conformation of the trinucleotide bulge. *Nucleic Acids Res* **24**, 3974-3981 (1996).

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

87. Murchie, A. I., Davis, B., Isel, C., Afshar, M., *et al.* Structure-based drug design targeting an inactive RNA conformation: exploiting the flexibility of HIV-1 TAR RNA. *J Mol Biol* **336**, 625-638 (2004).

88. Pitt, S. W., Majumdar, A., Serganov, A., Patel, D. J. & Al-Hashimi, H. M. Argininamide binding arrests global motions in HIV-1 TAR RNA: comparison with Mg<sup>2+</sup>-induced conformational stabilization. *J Mol Biol* **338**, 7-16 (2004).

89. Pitt, S. W., Zhang, Q., Patel, D. J. & Al-Hashimi, H. M. Evidence that electrostatic interactions dictate the ligand-induced arrest of RNA global flexibility. *Angew Chem*

*Int Ed Engl* **44**, 3412-3415 (2005).

90. Casiano-Negroni, A., Sun, X. & Al-Hashimi, H. M. Probing Na(+)-induced changes in the HIV-1 TAR conformational dynamics using NMR residual dipolar couplings: new insights into the role of counterions and electrostatic interactions in adaptive recognition. *Biochemistry* **46**, 6525-6535 (2007).

91. Bailor, M. H., Musselman, C., Hansen, A. L., Gulati, K., *et al.* Characterizing the relative orientation and dynamics of RNA A-form helices using NMR residual dipolar couplings. *Nat Protoc* **2**, 1536-1546 (2007).

92. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* **117**, 5179-5197 (1995).

93. Clowney, L., Jain, S. C., Srinivasan, A. R., Westbrook, J., *et al.* Geometric parameters in nucleic acids: nitrogenous bases. *J Am Chem Soc* **118**, 509-518 (1996).

94. Getz, M., Sun, X., Casiano-Negroni, A., Zhang, Q. & Al-Hashimi, H. M. NMR studies of RNA dynamics and structural plasticity using NMR residual dipolar couplings. *Biopolymers* **86**, 384-402 (2007).

95. Davis, B., Afshar, M., Varani, G., Murchie, A. I., *et al.* Rational design of inhibitors of HIV-1 TAR RNA through the stabilisation of electrostatic "hot spots". *J Mol Biol* **336**, 343-356 (2004).

96. Miclet, E., O'Neil-Cabello, E., Nikonowicz, E. P., Live, D. & Bax, A. 1H-1H dipolar couplings provide a unique probe of RNA backbone structure. *J Am Chem Soc* **125**, 15740-15741 (2003).

97. Miclet, E., Boisbouvier, J. & Bax, A. Measurement of eight scalar and dipolar couplings for methine-methylene pairs in proteins and nucleic acids. *J Biomol NMR* **31**, 201-216 (2005).

98. O'Neil-Cabello, E., Bryce, D. L., Nikonowicz, E. P. & Bax, A. Measurement of five dipolar couplings from a single 3D NMR multiplet applied to the study of RNA dynamics. *J Am Chem Soc* **126**, 66-67 (2004).

99. Ottiger, M., Delaglio, F. & Bax, A. Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra. *J Magn Reson* **131**, 373-378 (1998).

100. Vallurupalli, P. & Moore, P. B. Measurement of H2'-C2' and H3'-C3' dipolar couplings in RNA molecules. *J Biomol NMR* **24**, 63-66 (2002).

101. Yan, J., Corpora, T., Pradhan, P. & Bushweller, J. H. MQ-hCN-based pulse sequences for the measurement of 13C1'-1H1', 13C1'-15N, 1H1'-15N, 13C1'-13C2', 1H1'-13C2', 13C6/8-1H6/8, 13C6/8-15N, 1H6/8-15N, 13C6-13C5, 1H6-13C5 dipolar couplings in 13C, 15N-labeled DNA (and RNA). *J Biomol NMR* **22**, 9-20 (2002).

102. Zidek, L., Wu, H., Feigon, J. & Sklenar, V. Measurement of small scalar and dipolar couplings in purine and pyrimidine bases. *J Biomol NMR* **21**, 153 (2001).

103. Brutscher, B., Boisbouvier, J., Pardi, A., Marion, D. & Simorre, J. P. Improved Sensitivity and Resolution in 1H- 13C NMR Experiments of RNA. *J Am Chem Soc* **120**,

11845-11851 (1998).

104. Jaroniec, C. P., Boisbouvier, J., Tworowska, I., Nikonowicz, E. P. & Bax, A. Accurate measurement of  $^{15}\text{N}$ - $^{13}\text{C}$  residual dipolar couplings in nucleic acids. *J Biomol NMR* **31**, 231-241 (2005).

105. Tjandra, N. & Bax, A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* **278**, 1111-1114 (1997).

106. Ottiger, M. & Bax, A. Characterization of magnetically oriented phospholipid micelles for measurement of dipolar couplings in macromolecules. *J Biomol NMR* **12**, 361-372 (1998).

107. Sass, J., Cordier, F., Hoffmann, A., Rogowski, M., *et al.* Purple membrane induced alignment of biological macromolecules in the magnetic field. *J Am Chem Soc* **121**, 2047-2055 (1999).

108. Koenig, B. W., Hu, J. S., Ottiger, M., Bose, S., *et al.* NMR measurement of dipolar couplings in proteins aligned by transient binding to purple membrane fragments. *J Am Chem Soc* **121**, 1385-1386 (1999).

109. Tycko, R., Blanco, F. J. & Ishii, Y. Alignment of Biopolymers in Strained Gels: A New Way To Create Detectable Dipole- Dipole Couplings in High-Resolution Biomolecular NMR. *J Am Chem Soc* **122**, 9340-9341 (2000).

110. Sass, H. J., Musco, G., Stahl, S. J., Wingfield, P. T. & Grzesiek, S. Solution NMR of proteins within polyacrylamide gels: diffusional properties and residual alignment by mechanical stress or embedding of oriented purple membranes. *Journal of Biomol NMR* **18**, 303-309 (2000).

111. Ruckert, M. & Otting, G. Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR experiments. *J Am Chem Soc* **122**, 7793-7797 (2000).

112. Alvarez-Salgado, F., Desvaux, H. & Boulard, Y. NMR assessment of the global shape of a non-labelled DNA dodecamer containing a tandem of GT mismatches. *Magnetic Resonance in Chemistry* **44**, 1081 (2006).

## Chapter 3

### Linking RNA Topology to Structure and Dynamics

#### 3.1 Introduction

Structure prediction, as well as the rational manipulation of atomic structures of biomolecules are two aims that are anticipated to significantly advance applications within drug discovery and bioengineering. The problem is exceptionally challenging for exceedingly flexible RNAs, which are able to adopt wide-ranging conformations in response to cellular cues or changes in physicochemical conditions<sup>1-4</sup>. The thermodynamic principles that relate RNA primary sequence to secondary structure are well established and routinely used in secondary structure prediction<sup>5-7</sup>. However, current approaches for predicting 3D structures of RNA ignore flexibility and conformational adaptation, and strongly rely on homology modeling for identifying long-range contacts rather than on principles that are encoded at the secondary structural level<sup>8,9</sup>.

The global conformation of RNA is largely defined by the relative arrangement of A-form helices linked by flexible pivot points. Of these flexible pivots, approximately 70% consist of two-way junctions such as bulges and internal loops<sup>1-3,10,11</sup> (Figure 3.1a). Recently, it was shown that a combination of domain-elongation NMR spectroscopy<sup>12</sup> and molecular dynamics (MD) simulations<sup>13</sup> allowed for the

construction of ensembles of atomic resolution structures for the transactivation response element (TAR) RNA<sup>14</sup> from the human immunodeficiency virus type-1 (HIV-1) and type-2 (HIV-2) with timescale sensitivity encompassing pico- to milliseconds.

The structural ensembles, generated from this procedure, revealed collective three-dimensional rigid-body motions of A-form helices about a trinucleotide (HIV-1) or dinucleotide (HIV-2) bulge as shown in three-dimensional cube maps (Figure 3.1). The twist angles about each helix ( $\alpha_h$  and  $\gamma_h$ ) and the inter-helical bend angle ( $\beta_h$ ) are specified for each conformer in the ensemble (Figure 3.1). For both HIV-1 and HIV-2 TAR, the RNA conformers sample <5% of possible inter-helical orientations and trace out a spatially non-random trajectory in which the two helices twist in a correlated manner while bending (Figure 3.1b, in blue). Increasing the bulge length from two to three nucleotides resulted in an increase in the sampled inter-helical orientations and a decrease in the correlations between the twist angles<sup>13</sup> (Figure 3.1b, in blue). Seven distinct ligand-bound HIV-1 TAR conformations were represented in this narrow spatially anisotropic distribution indicating that ligands induce structural adaptation by capturing pre-existing conformations<sup>13</sup>. Despite its persistent occurrence in TAR, the physical basis for this spatially anisotropic inter-helical confinement remains unknown.

### **3.2 Defining two-way junctions**

In order to explore the generality of the anisotropic inter-helical confinement, we devised an approach to measure and compare inter-helical angles across any type of

two-way junction. The first step entailed developing a generalized definition to identify any and all two-way junction types. Building on a previous convention<sup>15</sup>, we designate two-way junctions  $H_iS_XH_jS_Y$  in which  $i$  and  $j$  specify the length of 5' and 3' helices, respectively. The single stranded nucleotides in 5' and the 3' strands, respectively are specified by  $X$  and  $Y$  ( $X \geq Y$ ), respectively (Figure 3.1a). Using the above definition, we measured inter-helical angles for all  $H_{\geq 3}S_XH_{\geq 3}S_Y$  junctions (Table 3.1) in the Protein Data Bank (PDB)<sup>16</sup> as described below. Similar results were obtained for  $H_{\geq 4}S_XH_{\geq 4}S_Y$  junctions (data not shown). For simplicity we will refer to  $H_{\geq 3}S_XH_{\geq 3}S_Y$  junctions here after as  $S_XS_Y$ . The attraction of using such a convention for junction nomenclature is that it easily adapts to higher-order junction types (e.g. 3-way or 4-way junctions).

### 3.3 PDB search of RNA two-way junctions

In order to discern the orientation occupied by the helices of two-way junctions, a comprehensive search of RNA two-way junctions in the RCSB Protein Data Bank was preformed on January 11, 2009 using the online database query system RNA FRABASE (RNA FRAGments search engine and dataBASE; <http://rnafrabase.ibch.poznan.pl>)<sup>17</sup>. The search for junction types within the database was done through the use of dot-bracket notation to designate junction types. In dot-bracket notation<sup>18</sup>, '(' and ')' are used to represent 5' and 3' Watson-Crick base pair nucleotides, respectively, and '.' to represent a lone or non- Watson-Crick base paired nucleotide. Thus, an  $H_4S_3H_4S_0$  bulge junction type is represented as (((...((( ))))). For asymmetric junctions, such as  $H_4S_2H_4S_1$ , the search was conducted twice using ((((.((( )))..))) and (((.((( )))..))). The reason is that RNA



fragments as listed in RNA FRABASE reflect the orientation of the junction type within the context of the entire RNA. Thus, in order to find all occurrences of a particular junction type two searches had to be conducted as illustrated above.

In the case of higher order structure types, such as kissing complexes or pseudo-knots, brackets '[', ']', '{' and '}' are used to signify nucleotides participating in Watson-Crick base pairs that create those structure types. In the current discussion of two-way junctions, we focused on simple two-way junction types and neglected all higher order junction types mentioned above. All searches were conducted without a sequence constraint and done iteratively for all  $H_{\geq 3}S_XH_{\geq 3}S_Y$  and  $H_{\geq 4}S_XH_{\geq 4}S_Y$  junctions with  $X$  and  $Y$  ranging between 0 and 10. Queried results were output into CSV file format and processed using in-house Perl scripts to extract coordinates from the PDB<sup>16</sup> for subsequent calculations of inter-helical angles ( $\alpha_h$   $\beta_h$   $\gamma_h$ ) as described below. The PDB accession numbers used in our survey are listed in Table 3.1.

### **3.3.1 Computing inter-helical angles for arbitrary RNA two-way junctions**

For the purpose of this work, two-way RNA junctions are defined as lone nucleotide(s) and/or non-Watson-Crick base pair(s) that adjoin two A-form helices each consisting of at least three Watson-Crick base pairs. A common reference frame was developed in order to compute and compare inter-helical angles across arbitrary two-way junctions. Each two-way junction was aligned in such a way that the longest single strand corresponding to the  $X$ -residues runs along the +ve  $z$ -direction of the molecular frame from the 5' to the 3' end. This allows for the unique

assignment of a “lower” 5' helix, referred to hereafter as helix I (H1), and an “upper” 3' helix, referred to hereafter as helix II (HII), which are shown in Figure 3.1. In the case of symmetric junctions where the length of each single strand is equal ( $X = Y$ ), the helices were arbitrarily oriented such that the strand with the lowest chain letter/ residue number in the PDB file runs along the +ve z-direction from the 5' to the 3' end. Also, note that +ve Euler angles correspond to a clockwise rotation. The inter-helical twist angle ( $\zeta_h$ ) is computed using  $\zeta_h = \alpha_h + \gamma_h$  such that minimum over- and under-twisting of the two helices correspond to negative and positive  $\zeta_h$  values, respectively. The computed Euler angles ( $\alpha_h \beta_h \gamma_h$ ) are degenerate with respect to the following possible angular sets,  $(\alpha_h \pm 180^\circ, -\beta_h, \gamma_h \pm 180^\circ)$ , and  $(\alpha_h \pm 180^\circ, -\beta_h, \gamma_h \mp 180^\circ)$ . For the special case of RNA structures with perfectly oriented parallel or anti-parallel inter-helical conformations, there is a continuous degeneracy defined by  $(\alpha_h \pm D, \beta_h = 0^\circ \text{ or } 180^\circ, \gamma_h \mp D)$  where D is a constant. A unique set of inter-helical Euler angles was selected by (i) limiting the values of  $\alpha_h \beta_h \gamma_h$  to within  $\pm 180^\circ$  and (ii) selecting the one degenerate solution that minimized the angular magnitude,  $\delta_{EA}$ , calculated as follows,  $\delta_{EA} = \sqrt{\alpha_h^2 + \beta_h^2 + \gamma_h^2}$ .

While a single solution can also be selected by limiting the value of  $\beta$  between  $0^\circ$  and  $180^\circ$ , this leads to artificial correlations between the angles  $\alpha_h$  and  $\gamma_h$  due to selection of degenerate solutions that differ by  $\pm 180^\circ$ . This result yields distributions that are extended in  $\alpha_h \beta_h \gamma_h$  space. Nevertheless, the two different representations of data yield equivalent conclusions. A similar procedure was used to compute the  $H_{\geq 4}S_X H_{\geq 4}S_Y$  inter-helical angles.

The inter-helical angles were computed by adapting the procedure previously described in chapter 2<sup>19</sup>. This approach built on a statistical survey of RNA structure to show that Watson-Crick base pairs flanked by other Watson-Crick base pairs adopt local conformations that strongly conform with idealized A-form geometry<sup>20</sup>. A reference two-way junction consisting of two coaxially stacked and continuous idealized A-form helices (iHI and iHII) was constructed using Insight II (Molecular Simulations, Inc.) as described previously<sup>19</sup>. Each helix contained four Watson-Crick base pairs and the helix axis was oriented along the z direction of the molecular frame such that helices run toward the +ve Z direction from iHI to iHII. The rotation around the helix axis was fixed to an arbitrary reference orientation in which the angle between the closing base pair y-axis, as defined by Westhof and coworkers<sup>21</sup>, and the y-axis of the molecular frame is  $-54^\circ$ . For every  $H_{\geq 3}S_X H_{\geq 3}S_Y$  two-way junction, the sugar and phosphate backbone heavy atoms of two central Watson-Crick base pairs in HI were superimposed onto the idealized iHI reference helix. The iHII helix was then independently superimposed onto HII.

An Euler rotation  $R(\alpha_h \beta_h \gamma_h)$  was computed using the in-house program EULER-RNA<sup>19</sup> that transforms iHII from the orientation observed in the junction to the reference co-axial alignment. The inter-helical angles,  $\alpha_h$ ,  $\beta_h$  and  $\gamma_h$ , were deduced directly from the resulting rotation matrix as shown in equation 3.1.

$$R(\alpha_h \beta_h \gamma_h) = \begin{bmatrix} -\sin \alpha_h \sin \gamma_h + \cos \alpha_h \cos \beta_h \cos \gamma_h & -\sin \alpha_h \cos \gamma_h - \cos \alpha_h \cos \beta_h \sin \gamma_h & \cos \alpha_h \sin \beta_h \\ \cos \alpha_h \sin \gamma_h + \sin \alpha_h \cos \beta_h \cos \gamma_h & \cos \alpha_h \cos \gamma_h - \sin \alpha_h \cos \beta_h \sin \gamma_h & \sin \alpha_h \sin \beta_h \\ -\cos \gamma_h \sin \beta_h & \sin \gamma_h \sin \beta_h & \cos \beta_h \end{bmatrix} \quad (3.1)$$

In the vast majority of cases, the superpositions yielded heavy atom root-mean-square deviations (rmsd) that are less than 1Å. A total of 66 junctions with an rmsd superposition >2Å rmsd were omitted from analysis.

### 3.4 Computing topologically allowed inter-helical orientations

The topologically allowed inter-helical orientations were computed for bulge junctions using in-house software. Trial inter-helical orientations were generated starting with a reference coaxially stacked helix consisting of two idealized A-form helices (iHI and iHII) each containing 11 Watson-Crick base pairs. The P atom in iHI in the  $Y = 0$  strand was set as the pivot point by translation to the origin. Trial inter-helical orientations that exhaustively sampled all of the  $\alpha_h \beta_h \gamma_h$  space in increments of 5° were generated by rotation of iHII. Bulge residues were not included in this analysis. Inter-helical orientations were rejected if they resulted in inter-helix steric collisions from an assumed Van der Waals radii of 1.40 Å (excluding hydrogen atoms) and/or if the distance between the O3'(i) and P(i+1) in the 5' (iHI) and 3' (iHII) helices, respectively, exceeded a given cut-off value specified by the specific bulge length under investigation. A distance cut-off of 4.9 Å per bulge nucleotide was used, which corresponds to the average distance per nucleotide observed over all 379 bulge  $H_{\geq 4}S_X H_{\geq 4}S_Y$  junctions in our PDB database. Thus, cut-off distances of 4.9 Å, 9.8 Å and 14.7 Å were used for one, two and three bulge nucleotides, respectively.

The  $\alpha_h \beta_h \gamma_h$  angles for all accepted orientations were processed to yield a single unique solution among degenerate sets as described above. Due to minor differences in steric collisions, results varied slightly depending on the length of the

idealized helices used (Figure 3.2). The topologically allowed inter-helical orientations for non-bulge asymmetric junctions ( $Y \neq 0$ ;  $X > 0$ ;  $X > Y$ ) were computed by applying a correction to the bulge-derived distributions to take into account inter-helical over-twisting induced by the maximum number of non-canonical junction base pairs. Each asymmetric junction was converted into an X-0 bulge by maximizing the number of non-canonical base pairs. For example, a 3-1 junction was reduced into a 2'-0 bulge, and 4-1 into a 3'-0 bulge. This was motivated by the observation of  $\sim 34^\circ$  systematic shift in the inter-helical twist angle with increasing  $Y$  in the PDB database (Figure 3.3) which suggests that residues within junctions induce similar inter-helical twisting as might be expected from a typical Watson-Crick base pair most likely by stacking inside helices. The X'-bulge nucleotides then define the effective bulge length and the bulge distribution to which the correction in twist angles is applied. In particular, the values of  $\alpha_h$  and  $\gamma_h$  in the X'-0 bulge distribution were each shifted by an amount  $X+Y \times 8.5^\circ$  to account for an average increase in the inter-helical twist angle ( $\alpha_h + \gamma_h$ ) of  $\sim 34^\circ$  per base pair.

The fraction of space that is topologically sampled by a given junction was computed by dividing the number of topologically allowed inter-helical orientations by the total number of orientations sampled in the search. The fraction of the space sampled by junctions in the PDB was computed by counting the number of orientations in the grid search that are  $10^\circ$  from orientations in the PDB. The  $10^\circ$  cut-off takes into account both the  $5^\circ$  increments of the grid ( $\sim 8.7^\circ$ ) as well as  $\sim 5^\circ$  errors in computing the inter-helical angles. A similar approach was used to

compute the overlap between the PDB orientations and the topologically computed distributions.

### **3.5 Distribution of bulge-type two-way junctions**

Strikingly, the inter-helical orientations observed for all 148 tri-nucleotide and 275 dinucleotide bulges, free and protein/ligand bound, are also confined to narrow anisotropic distributions (Figure 3.1b, in red) that sample <5% of possible orientations. Interestingly, the range in distribution of inter-helical angles feature 15 and 10 unique sequences within the set of trinucleotide and dinucleotide bulges, respectively, seemingly hinting that while sequence may play a role in biasing the distributions of inter-helical angles, it does not limit that range. The distributions from the PDB are akin to the distributions observed for HIV-1 and HIV-2 TAR and feature similar variations with bulge length (Figure 3.1b).

We examined if the anisotropic inter-helical confinement observed across bulges arises from simple topological forces that restrict the allowed range of inter-helical orientations. To this end, we computed the allowed inter-helical orientations across bulges of varying lengths subject to two trivial constraints: (i) helices cannot sterically clash and (ii) the distance between O3'(i) and P(i+1) in the 3' and 5' helices, respectively, cannot exceed the average bulge linker length (4.9 Å per nucleotide). Remarkably, implementation of these two simple topological constraints was enough to generate inter-helical orientations that quantitatively reproduce the PDB-derived and TAR dynamic distributions, as well as their variation with bulge length (Figure 3.1b, shown in grey).

Of the allowed possible inter-helical orientations available, the PDB-derived junctions sample only 4%-20% of the space, yet account for > 85% of the observed orientations (Figure 3.1b). For bulges  $\leq 4$  nucleotides long, the confinement is dominated by connectivity (excludes 73%-90% of orientations) rather than steric (excludes  $\sim 49\%$  of orientations) constraints. However, the connectivity constraints decrease in a gradual fashion with increasing bulge length, and are insignificant at seven nucleotides. In contrast, steric constraints are independent of helix length (Figure 3.2a). However, due to the strong anisotropy of the inter-helical distributions and  $34^\circ$  helical pitch, changing the length of a helix can lead to large ( $34^\circ$  per base pair) changes in the relative arranged twist for neighboring helices (Figure 3.2b).

### **3.6 Prediction of internal loop conformations**

Amazingly, the spatially anisotropic inter-helical confinement is observed throughout symmetric and asymmetric internal loops (Figure 3.3). The PDB-derived internal loop distributions are strikingly similar to those observed for bulges with one simple exception: they feature a systematic shift in the  $\alpha_h$  and  $\gamma_h$  twist angles of  $\sim 17^\circ$  and thus an inter-helical over-twisting ( $\alpha_h + \gamma_h$ ) of  $\sim 34^\circ$  with each  $Y$  increment. The  $Y$  value specifies the number of nucleotides in the shorter strand (Figure 3.1a) and therefore the maximum number of non-canonical base pairs that can insert between helices at the junction (Figure 3.3a). Thus,  $Y$ -dependent variations in  $\alpha_h$  and  $\gamma_h$  are as expected if one assumes that residues within the junction

preferentially adopt looped-in, stacked conformations that maintain the canonical RNA A-form helical pitch of  $\sim 33^\circ$  (Figure 3.3a).

By introducing a correction to the inter-helical twist angles ( $\alpha_h$  and  $\gamma_h$ ) in the bulge computed topological distributions, which assume that base pairing is maximized within the junction through the formation of non-canonical base pairing, as described above, we were able to quantitatively reproduce the inter-helical distributions observed for all symmetric and asymmetric internal loops (Figure 3.3b, in gray). It is remarkable that even though the topologically computed distributions for the 14 different junction types (Figures 3.1b and 3.3) sample only 4%-20% and on average 7% of possible 3D orientations, they accommodate >85% of the inter-helical orientations observed in the PDB.

Major outliers are junctions that exceeded our assumed average inter-helical linker constraint that we implemented in the distributions calculations. All outliers are readily accommodated by employing a linker length constraint as a distribution of lengths that accounts for scenarios other than maximizing base pairing formation within the junction. Furthermore, the topological constraints are also expected to increase with an increasing RNA size due to growing self-avoidance constraints, leading to further definition of RNA global architecture.

### **3.7 Conclusions**

The study and analysis of RNA inter-helical angles presents a new take on a characteristic basic and ubiquitous characteristic of RNA regarded to be universal: the modular nature of its structural motifs. Here, we have shown that based



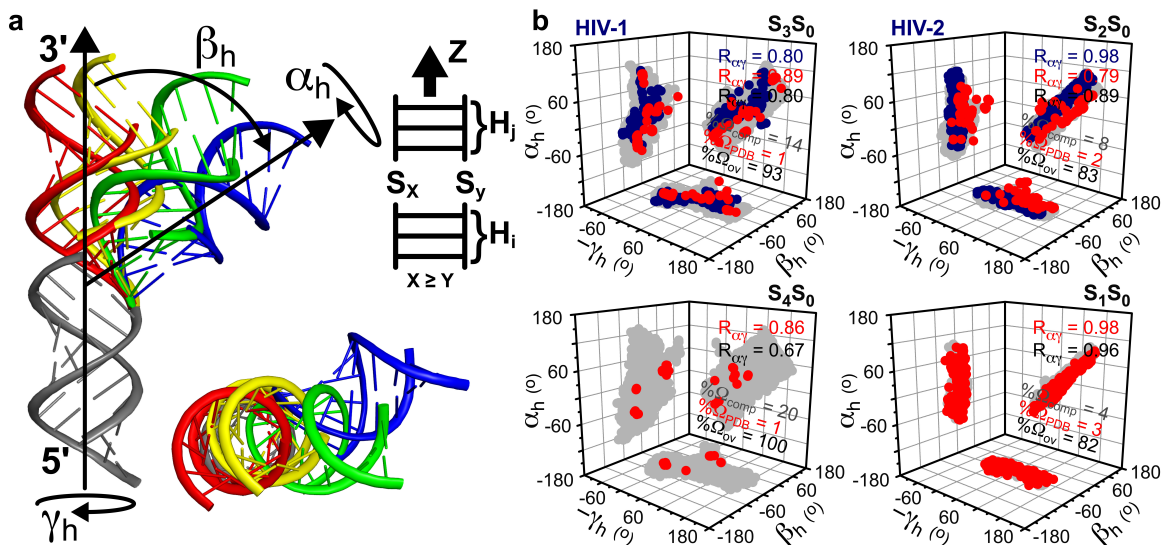
exclusively on knowledge obtained from a secondary sequence, one can definitively define the available conformational space of a two-way junction. Unlike other methods that have been implemented to describe and predict RNA structures, our method scales with the number of junctions and not atoms. Therefore the task of de novo structure prediction for large RNAs becomes, invariably, a more feasible endeavor. Additionally, the modular nature of such a structure-prediction algorithm greatly reduces the number of structural variables to consider, especially in the case of dynamic and structurally adaptive systems.

This work, short of providing a complete description for all RNA structural motifs, suggests that some topological considerations are likely at play in larger, more complicated architectural organization. Moreover, the results presented within suggest that tertiary contacts and other intermolecular interactions act to selectively stabilize specific conformations within the narrow topologically allowed ensemble. Tertiary interactions and other types of higher-order structural organization are likely characteristic of the many processes that involve RNA, such as folding and tertiary capture mechanisms involving the recognition of small molecules and proteins. Computation studies have already demonstrated the importance of topological features to RNA folding<sup>22-24</sup>. In particular, these works<sup>22-24</sup> illustrate how sequence, secondary and tertiary structure are intrinsically connected and point to an underlying and universal code of structure, dynamics and adaptation.

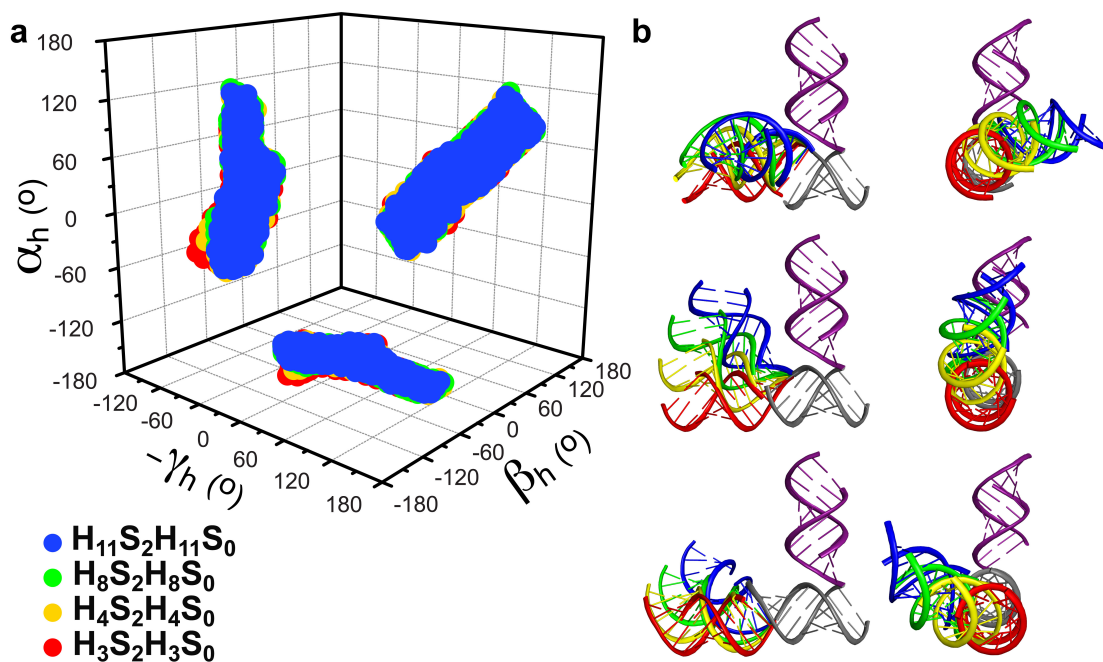
As presented, the current work does not feature a rationalization for the importance of sequence in determining structure or dynamics of a particular RNA. In part, this is due to the limited structural and experimental interrogations of such parameters. To date, the majority of structures inhabiting the PDB are heavily weighted toward a few RNA elements and limited number of classes. This homogeneity limits the statistical structure mining and data collection that can be done. Nevertheless, the ever-expanding nature of the database means that future investigations will benefit greatly from the continued progress and pursuit of experimentally derived RNA structures. Correspondingly, this pursuit requires a concomitant and corresponding effort to expand upon the types and classes of RNA structures currently known.

Moreover, the vast majority of structures deposited in the PDB are derived from crystal structures, which are often obtained by submitting the RNA to conditions of high salt and low temperatures. Thus, there will need to be an additional effort to incorporate experimentally obtained data of RNA under more physiologically relevant conditions. Studies that seek to integrate computational methods with established and budding structure determination techniques, like those incorporating molecular dynamics simulations with residual dipolar coupling analysis<sup>13,25</sup>, will be integral to expanding our understanding of RNA structure and dynamics, and critical to exploring the range of solution and cellular conditions that play a part in the functionality of RNA.

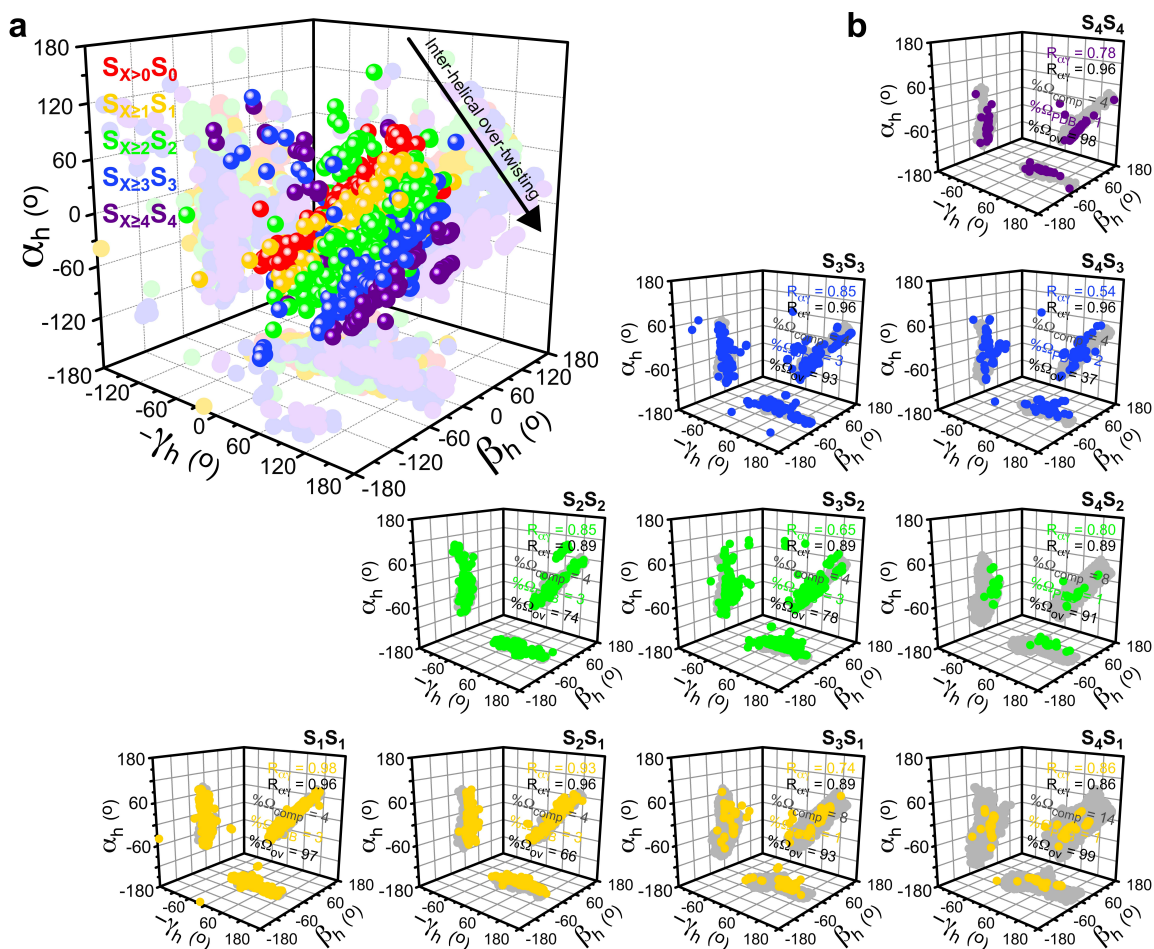
This work was published in the journal *Science*<sup>26</sup>. HM Al-Hashimi and MH Bailor conceived the idea. MH Bailor measured, calculated and collected all inter-helical Euler data, and calculated small molecule parameters. HM Al-Hashim and MH Bailor analyzed data.



**Figure 3.1: Topological confinement of RNA inter-helical orientations across  $H_{X-3}S_X H_{X-3}S_0$  bulges.** (a) Nomenclature used to designate 5' and 3' helices and inter-helical angles ( $\alpha_h \beta_h \gamma_h$ ) across two-way junctions containing  $X$  and  $Y$  ( $X \geq Y$ ) single stranded (S) residues. The helices depict the topologically allowed range of orientations for a dinucleotide bulge. (b) 3D inter-helical orientation maps showing individual 2D projections along each plane together with the associated correlation coefficients ( $R$ ) between the inter-helical twist angles for bulges that are four ( $S_4S_0$ ), three ( $S_3S_0$ , HIV-1 TAR), two ( $S_2S_0$ , HIV-2 TAR), and one ( $S_1S_0$ ) nucleotide long. The NMR-MD, PDB-derived, and topologically computed inter-helical distributions are shown in blue, red, and grey, respectively. There are a total of 751, 275, 148 and 21 PDB-derived entries for  $S_1S_0$ ,  $S_2S_0$ ,  $S_3S_0$  and  $S_4S_0$  type junctions, respectively. The % of 3D inter-helical orientations sampled by the PDB-derived and topologically computed distributions is indicated ( $\Omega_{PDB}$  and  $\Omega_{comp}$ , respectively) along with the fraction of the PDB-derived orientations that falls within  $10^\circ$  of the topologically allowed distribution ( $\Omega_{ov}$ ).



**Figure 3.2: Effects of topological constraints on helix length and relative orientation of sequentially arranged two-way junctions.** (a) The topologically allowed inter-helical orientations for a dinucleotide bulge as a function of helix length. (b) Dependence of the inter-helical orientations on helix length. Shown are the changes in the inter-helical spatial distribution for a dinucleotide bulge (shown in color) relative to a reference helix (in purple) as a function of increasing the length of an intervening helix (in grey).



**Figure 3.3: Topological confinement of RNA inter-helical orientations across  $H_3S_xH_3S_y$  internal loops.** 3D inter-helical orientation maps showing individual 2D projections along each plane together with the associated correlation coefficients ( $R$ ) between the inter-helical twist angles for different types of internal loops. (a) The PDB-derived inter-helical distributions for various families of Y-junctions. (b) The PDB-derived and topologically computed inter-helical distributions for different types of internal loops are shown in color and grey, respectively. There are a total of [374, 471, 133, 130], [230, 309 32], [455, 155] and [104] PDB-derived entries for  $[S_1S_1, S_2S_1, S_3S_1, S_4S_1]$ ,  $[S_2S_2, S_3S_2, S_4S_2]$ ,  $[S_3S_3, S_4S_3]$  and  $[S_4S_4]$  type junctions, respectively. The % of inter-helical orientations sampled by the PDB-derived and topologically computed distributions is indicated ( $\Omega_{\text{PDB}}$  and  $\Omega_{\text{comp}}$ , respectively) along with the fraction of the PDB-derived orientations that falls within  $10^\circ$  of the topologically allowed distribution ( $\Omega_{\text{ov}}$ ).

**Table 3.1: PDB accession numbers used to compute inter-helical angles for  $H_{\geq 3}S_X H_{\geq 3}S_Y$  junctions. <sup>a</sup>PDBs that were excluded in the  $H_{\geq 4}S_X H_{\geq 4}S_Y$  series.**

1A1T <sup>a</sup>	1G1X	1L8V	1O9M	1TFW	1XNR	205D	2F4S	2IXY	2PJP	397D	3D5C
1A3M	1GID	1L9A	1OLN <sup>a</sup>	1TUT	1XSG	280D	2F4T	2J00	2PN3	3BBN	3D5D
1A4D	1GRZ	1LC4	1OND	1TXS	1XSH	28SP <sup>a</sup>	2F4U	2J01	2PN4	3BBO	3DEG <sup>a</sup>
1AJU	1HC8 <sup>a</sup>	1LC6 <sup>a</sup>	100A	1U3K	1XST	28SR <sup>a</sup>	2F4V	2J02	2PWT	3BBX	3DF1
1AKX	1HNW	1LMV	1OSW <sup>a</sup>	1U63 <sup>a</sup>	1XSU	2A04	2F88	2J03	2PXB <sup>a</sup>	3BNL	3DF2
1ARJ	1HNX	1LNG	1OW9 <sup>a</sup>	1U6B <sup>a</sup>	1Y39 <sup>a</sup>	2A2E	2F8S	2J28	2PXD <sup>a</sup>	3BNN	3DF3
1B36	1HNZ	1LPW	1P5M	1U6P	1Y69	2A64 <sup>a</sup>	2FDT	2J37	2PXE <sup>a</sup>	3BNO	3DF4
1BVJ <sup>a</sup>	1HQ1 <sup>a</sup>	1LVJ	1P5N	1ULL <sup>a</sup>	1Y6S	2AAR	2FEY <sup>a</sup>	2JL5	2PXF <sup>a</sup>	3BNP	3DIG <sup>a</sup>
1BYJ	1HR0	1M1K	1P5O	1UN6	1Y6T	2ADT	2FQN	2JL6	2PXK <sup>a</sup>	3BNQ	3DIL <sup>a</sup>
1C04 <sup>a</sup>	1HR2	1M5L	1P5P	1UTS	1Y73	2AHT	2FRL <sup>a</sup>	2JL7	2PXL <sup>a</sup>	3BNR	3DIM <sup>a</sup>
1C2W	1HWQ	1M82 <sup>a</sup>	1P9X	1UUD	1Y90	2AU4	2G5K	2JL8	2PXP <sup>a</sup>	3BNS	3DIO <sup>a</sup>
1C2X	1I94	1M90	1PNS	1UUI	1Y95	2AVY	2G5Q	2JSE <sup>a</sup>	2PXQ <sup>a</sup>	3BO2 <sup>a</sup>	3DIQ <sup>a</sup>
1C4L <sup>a</sup>	1I95	1MFK <sup>a</sup>	1PNU	1VOQ	1Y99	2AW4	2GDI <sup>a</sup>	2JUK	2PXT <sup>a</sup>	3BO3 <sup>a</sup>	3DIR <sup>a</sup>
1CQ5 <sup>a</sup>	1I96	1MFQ	1PNX	1VOR	1YHQ	2AW7	2GIO	2JWV	2PXU <sup>a</sup>	3BO4 <sup>a</sup>	3DIS <sup>a</sup>
1CQL <sup>a</sup>	1I97	1MFY <sup>a</sup>	1PNY	1VOS	1YI2	2AWB	2GIP	2JXS	2PXV <sup>a</sup>	3CC2	3DIX <sup>a</sup>
1D6K	1I9X	1MJI	1Q7Y	1VOU	1YIJ	2B64	2GIS <sup>a</sup>	2JXV	2QBZ	3CC4	3DIY
1DK1 <sup>a</sup>	1IBK	1MMS <sup>a</sup>	1Q81	1VOV	1YIT	2B66	2GJW	2JYF	2QEK	3CC7	3DIZ <sup>a</sup>
1DUL <sup>a</sup>	1IBL	1MNX	1Q82	1VOW	1YJ9	2B9M	2GM0	2JYH	2QEX	3CCE	3DJ0 <sup>a</sup>
1DZ5	1IBM	1MWL	1Q86	1VOX	1YJN	2B9N	2G05	2K3Z	2QUX	3CCJ	3DJ2 <sup>a</sup>
1E4P <sup>a</sup>	1J2B <sup>a</sup>	1N32	1QA6 <sup>a</sup>	1VOY	1YJW	2B90	2GRW	2K41	2R1S	3CCL	3DLL
1ELH <sup>a</sup>	1J5A	1N33	1QBP	1VOZ	1YL3	2B9P	2GV4 <sup>a</sup>	2K7E	2R20	3CCM	3DVV
1EOR <sup>a</sup>	1J5E	1N34	1QD3	1VPO	1YL4	2BE0	2GY9	2NOK	2R21	3CCQ	3E5C
1EXY <sup>a</sup>	1J7T	1N36	1QVF	1VQ6	1YLG	2BEE	2GYA	2NOQ	2R22	3CCR	3E5E
1F6X	1JID	1N66 <sup>a</sup>	1QVG	1VQ8	1YNC	2CD1	2GYB	2O3V	2R8S	3CCS	3E5F
1F6Z <sup>a</sup>	1JJ2	1N8R	1R2P	1VQ9	1YNE	2CD3	2GYC	2O3W	2RKJ <sup>a</sup>	3CCU	3F1E
1F78	1J07 <sup>a</sup>	1N8X	1R3E	1VQK	1YNG	2CD5	2HEM <sup>a</sup>	2O3X	2UU9	3CCV	3F1F
1F79	1JU1	1NBK	1R7W	1VQN	1YRJ	2CD6	2HGI	2O3Y	2UUA	3CD6	3F1G

1F7F	1JZX	1NBR	1R7Z <sup>a</sup>	1VQO	1YRM	2CKY	2HGJ	2O43	2UUB	3CF5	3F1H
1F7G <sup>a</sup>	1JZY	1NBS	1RFR	1VS5	1YSH	2CSX <sup>a</sup>	2HGP	2O45	2UUC	3CGP	406D <sup>a</sup>
1F7I	1JZZ	1NJI	1RY1	1VS6	1YY0	2D17	2HGQ	2OE5	2UWM	3CGQ	420D
1F7Y <sup>a</sup>	1K01	1NJM	1S1H	1VS7	1YZ9	2D18	2HGR	2OE6	2V3C	3CGR	422D <sup>a</sup>
1F9L <sup>a</sup>	1K73	1NJN	1S1I	1VS8	1ZZJ	2D1A	2HGU	2OE8	2VHM	3CGS	429D
1FCW <sup>a</sup>	1K8A	1NJO	1S2F <sup>a</sup>	1VSA	1Z31 <sup>a</sup>	2D1B	2HHH	2OGM	2VHN	3CJZ	462D
1FFK	1K8S	1NJP	1S34 <sup>a</sup>	1W2B	1Z43	2D30	2HW8 <sup>a</sup>	2OGN	2VHO <sup>a</sup>	3CPW	
1FFZ	1K9M	1NKW	1S72	1WSU	1Z58	2DR5 <sup>a</sup>	2I2P	2OGO	2VHP <sup>a</sup>	3CUN	
1FG0	1KD1	1NLC	1S9S	1WVD	1Z79	2E5L	2I2T	2OIJ	2VPL	3D0U <sup>a</sup>	
1FJG	1KOG	1NUJ	1SM1	1X8W	1Z7F	2ESI	2I2U	2OYI	2VQE	3D0X <sup>a</sup>	
1FMN <sup>a</sup>	1KQS	1NUV	1SY4 <sup>a</sup>	1XBP	1ZC5	2ESJ	2I2V	2OJ0	2VQF	3D2G	
1FOQ <sup>a</sup>	1KUQ <sup>a</sup>	1NWX	1SYZ <sup>a</sup>	1XHP	1ZHO <sup>a</sup>	2ET3	2I7E	2OM7	2ZJP	3D2V	
1FQZ <sup>a</sup>	1KXK	1NWY	1T0D	1XMO	1ZX7	2ET4	2I7Z	2OW8	2ZJQ	3D2X	
1FYO	1L1C	1NZ1 <sup>a</sup>	1T0E	1XMQ	1ZZ5	2ET5	2IHX	2OZB <sup>a</sup>	2ZJR	3D5A	
1FYP	1L1W	1O3Z	1T0K	1XNQ	1ZZN <sup>a</sup>	2ET8	2IRN <sup>a</sup>	2PCV <sup>a</sup>	2ZKO	3D5B	

---

### 3.8 References

1. Serganov, A. & Patel, D. J. Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nat Rev Genet* **8**, 776-790 (2007).
2. Al-Hashimi, H. M. & Walter, N. G. RNA dynamics: it is about time. *Curr Opin Struct Biol* **18**, 321-329 (2008).
3. Cruz, J. A. & Westhof, E. The dynamic landscapes of RNA architecture. *Cell* **136**, 604-609 (2009).
4. Schwalbe, H., Buck, J., Fürtig, B., Noeske, J. & Wöhnert, J. Structures of RNA switches: insight into molecular recognition and tertiary structure. *Angew Chem Int Ed Engl* **46**, 1212-1219 (2007).
5. Mathews, D. H. & Turner, D. H. Prediction of RNA secondary structure by free energy minimization. *Curr Opin Struct Biol* **16**, 270-278 (2006).
6. Shapiro, B. A., Yingling, Y. G., Kasprzak, W. & Bindewald, E. Bridging the gap in RNA structure prediction. *Curr Opin Struct Biol* **17**, 157-165 (2007).
7. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research* **31**, 3406 (2003).
8. Parisien, M. & Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**, 51-55 (2008).
9. Das, R. & Baker, D. Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci U S A* **104**, 14664 (2007).
10. Lilley, D. M. J. Structures of helical junctions in nucleic acids. *Quarterly Reviews of Biophysics* **33**, 109-159 (2000).
11. Bindewald, E., Hayes, R., Yingling, Y. G., Kasprzak, W. & Shapiro, B. A. RNAJunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign. *Nucleic Acids Res* **36**, D392-D397 (2008).
12. Zhang, Q., Stelzer, A. C., Fisher, C. K. & Al-Hashimi, H. M. Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature* **450**, 1263-1267 (2007).
13. Frank, A. T., Stelzer, A. C., Al-Hashimi, H. M. & Andricioaei, I. Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition. *Nucleic Acids Res* **37**, 3670-3679 (2009).
14. Puglisi, J. D., Tan, R., Calnan, B. J., Frankel, A. D. & Williamson, J. R. Conformation of the TAR RNA-arginine complex by NMR spectroscopy. *Science* **257**, 76-80 (1992).
15. Lilley, D. M., Clegg, R. M., Diekmann, S., Seeman, N. C., *et al.* A nomenclature of junctions and branchpoints in nucleic acids. *Nucleic Acids Res* **23**, 3363-3364 (1995).
16. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).



17. Popena, M., Blazewicz, M., Szachniuk, M. & Adamiak, R. W. RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures. *Nucleic Acids Res* **36**, D386-D391 (2008).
18. Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J. & Giegerich, R. RNASHAPES: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* **22**, 500 (2006).
19. Bailor, M. H., Musselman, C., Hansen, A. L., Gulati, K., *et al.* Characterizing the relative orientation and dynamics of RNA A-form helices using NMR residual dipolar couplings. *Nat Protoc* **2**, 1536-1546 (2007).
20. Musselman, C., Pitt, S. W., Gulati, K., Foster, L. L., *et al.* Impact of static and dynamic A-form heterogeneity on the determination of RNA global structural dynamics using NMR residual dipolar couplings. *J Biomol NMR* **36**, 235-249 (2006).
21. Yang, H., Jossinet, F., Leontis, N., Chen, L., *et al.* Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res* **31**, 3450-3460 (2003).
22. Sorin, E. J., Nakatani, B. J., Rhee, Y. M., Jayachandran, G., *et al.* Does native state topology determine the RNA folding mechanism? *J Mol Biol* **337**, 789-797 (2004).
23. Lin, J. C. & Thirumalai, D. Relative stability of helices determines the folding landscape of adenine riboswitch aptamers. *J Am Chem Soc* **130**, 14080-14081 (2008).
24. Cho, S. S., Pincus, D. L. & Thirumalai, D. Assembly mechanisms of RNA pseudoknots are determined by the stabilities of constituent secondary structures. *Proc Natl Acad Sci U S A* **106**, 17349-17354 (2009).
25. Stelzer, A. C., Frank, A. T., Bailor, M. H., Andricioaei, I. & Al-Hashimi, H. M. Constructing atomic-resolution RNA structural ensembles using MD and motionally decoupled NMR RDCs. *Methods* **49**, 167-173 (2009).
26. Bailor, M. H., Sun, X. & Al-Hashimi, H. M. Topology links RNA secondary structure with global conformation, dynamics, and adaptation. *Science* **327**, 202-206 (2010).

## Chapter 4

### RNA Topology and Adaptation: The Rules Behind RNA-Ligand

#### Selection

##### 4.1 Introduction

RNA-ligand interactions are a fundamental aspect of many regulated biological processes. In a number of these processes small molecule binding triggers conformational transition, or traps a specific conformer from within a greater continuum<sup>1,2</sup>. Thus, mechanistic regulation is achieved in one of two ways: induced fit or tertiary capture. However, RNA-ligand recognition is still currently poorly understood. Riboswitches are RNA elements commonly found in the untranslated regions of non-coding RNAs (ncRNAs) and messenger RNAs (mRNAs), and structural studies of such RNA elements have proved invaluable toward illuminating the important features influencing recognition and binding of cognate metabolites. For example, the thiamine pyrophosphate (TPP) binding riboswitch, which is able to gene dependently regulate both translation and transcription<sup>3,4</sup>, functions by recognizing TPP and related derivatives through a structurally preformed pyrimidine-sensor helix and initially unstructured pyrophosphate-sensor helix<sup>3,5-7</sup>. The TPP riboswitch highlights a number of important structural features relevant to RNA-ligand recognition. Yet, like the majority of known RNA-bound structures it

fails to disclose the determinants of conformational and structural negotiation that must occur to achieve the observed bound conformer.

Much of what is currently known about RNA binding is centered around establishing favorable electrostatic<sup>8,9</sup> and hydrogen bond interactions<sup>2,10,11</sup> to specifically interact with a target. However, electrostatic interactions are notoriously non-specific, and a number of ligands are known to bind promiscuously for such reasons<sup>12,13</sup>. Additionally, of the known RNA-ligand bound structures very few display any chemical structural continuity among ligand structures, as well as sequence and secondary structure among RNA elements. For example, the structure of the transactivation response element (TAR) of HIV-1 has been solved in eight contexts<sup>14-19</sup>(Figure 4.1). Thus, it is difficult to derive a common picture of the driving force behind RNA recognition. This is all further complicated by the presence of structural adaptation which occurs concurrently with binding of the small molecule. While it is clear each bound structure of a TAR-ligand complex exhibits different global conformations with respect to the orientation of helices, it is difficult to rationalize how the specific interactions of a particular ligand are able to select for the RNA conformer observed.

Other structural factors known to contribute to protein-ligand recognition, such as shape-complimentarily and the like<sup>20-23</sup>, have been observed in RNA aptamers<sup>24</sup>. However, for RNA elements in cellular contexts, where ligand and other bimolecular interactions are tied to a specific function, the compromises between binding affinity/specificity and function are not as well defined. Recent work from the Al-

Hashimi lab has shown TAR undergoes large amplitude motions ( $\sim 160^\circ$ ), in which each of the three inter-helical Euler angles changes in a correlated manner<sup>25</sup>. Moreover, it was shown that the shortest pathway linking all three conformers through single axis rotations overlapped with each known TAR ligand-bound conformation<sup>25</sup> (Figure 3.1). Thus, given that ligand binding to TAR, and RNA in general, is known to localize preferentially at the junction, it provided an ideal framework to explore the rules of global structural adaptation<sup>24,26</sup> (Figure 4.1).

In order to gain insights into the conformational selection rules that guide interactions of RNA and small-molecules, we used NMR spectroscopy to structurally characterize HIV-1 TAR (Figure 4.2a) bound to five aminoglycoside small molecules (Figure 4.2b). The strategy implemented was nearly identical to that described in chapter 2. However, in addition to determining the inter-helical angles of TAR, we also sought to compare those parameters with measurable properties of the ligands. To this end, five aminoglycosides were chosen that differed in charge, shape, and size, and which bound TAR with dissociation constants ( $K_d$ ) in the micromolar range<sup>27</sup>.

#### **4.2 Preparation and purification of $^{13}\text{C}/^{15}\text{N}$ labeled HIV-1 TAR RNA**

HIV-1 TAR samples for NMR studies were prepared by *in vitro* transcription using T7 RNA polymerase (Takara Mirus Bio, Inc.), uniformly  $^{13}\text{C}/^{15}\text{N}$  labeled nucleotide triphosphates (ISOTEC, Inc.), unlabeled (Sigma) nucleotide triphosphates, and synthetic DNA templates (Integrated DNA Technologies, Inc.) containing the T7 promoter and sequence of interest. The RNAs were purified by 15% (w/v)

denaturing polyacrylamide gel electrophoresis, using 8M urea and TBE. The RNA was eluted from the gel in 20 mM Tris pH 8 buffer followed by ethanol precipitation. The RNA pellet was dissolved in water and exchanged into NMR buffer (15 mM sodium phosphate, 0.1 mM EDTA, and 25 mM NaCl at pH ~6.4) multiple times using a Centricon Ultracel YM-3 concentrator (Millipore Corp.). All aminoglycosides were purchased from Sigma-Aldrich and were used as is. Stock solutions of aminoglycosides (100 mM – 300 mM) dissolved in NMR buffer were used in the titration experiments.

### 4.3 Chemical shift perturbation mapping and titrations

The first step in the analysis of RNA-ligand complexes by NMR is the measurement of chemical shift titrations, which will ultimately yield a set of chemical shift perturbations and dissociation constants ( $K_d$ s). Chemical shift perturbation mapping is an informative method that quickly allows one to qualitatively identify the sites of interest. For this work, total chemical shift perturbations were quantified using the relationship in equation 4.1.

$$\Delta\delta_{total} = \sqrt{\Delta\delta_H^2 + \left(\frac{\gamma_H}{\gamma_{C/N}}\right)^2 \Delta\delta_{C/N}^2}, \quad (4.1)$$

where  $\Delta\delta_i$  is the change in chemical shift for the  $i$  nuclei type, and  $\gamma_i$  its corresponding gyromagnetic ratio. Using this procedure, chemical shift mapping was carried out for each aminoglycoside complex.

The difficulty associated with chemical shift perturbations is that they arise from a number of different sources. For instance, localization of the ligand, as well as

structural perturbations induced by the ligand will each yield changes in peak position. Nonetheless, chemical shift perturbation analysis can greatly help to focus subsequent experimental efforts in order to resolve ambiguities. Additionally, one can calculate site-specific  $K_d$  values. These  $K_d$ 's contain information about the perturbation energy localized at specific locations. Hence, there exists a great deal of information that one can gather from this simple data.

With the aid of two-dimensional (2D) HSQC spectra recorded following incremental increases in the aminoglycoside concentration, we were able to measure chemical shifts occurring for a variety of resonances for the concentration ranges below: [NeoB] (0.02, 0.04, 0.08, 0.16, 0.32, 0.64 and 0.80 mM against 0.2 mM TAR); [Par] (0.04, 0.08, 0.16, 0.32, 0.64, 1.28 and 2.56 mM against 0.4 mM TAR); [Rib] (0.10, 0.20, 0.40, 0.80, 1.60, 3.20 and 6.40 mM against 0.1 mM TAR); [KanB] (0.10, 0.20, 0.40, 0.80, 1.20, 1.60 and 3.20 mM against 0.1 mM TAR); [Tob] (0.06, 0.12, 0.18, 0.24, 0.74, 1.76, 2.50 and 3.78 mM against 0.3 mM TAR) (Figure 4.3).

For each aminoglycoside, the concentration at which RDC measurements were recorded are underlined. Under the conditions described above, 98%, 97%, 98%, 97%, and 95% of TAR is estimated to be in the aminoglycoside bound state for NeoB, Par, KanB, Tob, and Rib, respectively. Apparent dissociation constants ( $K_d$ ) were obtained by fitting the observed changes from chemical shift experiments to the following equation<sup>28</sup>,

$$\delta_{obs} = \delta_{free} + \frac{\Delta\delta_T \{ ([L]_T + [RNA]_T + K_d) - \sqrt{([L]_T + [RNA]_T + K_d)^2 + (4 \cdot [L]_T \cdot [RNA]_T)} \}}{2 \cdot [RNA]_T} \quad (4.2)$$

where  $[L]_T$  is the total concentration of aminoglycoside,  $[RNA]_T$  is the TAR concentration based on UV absorbance at 260 nm,  $\Delta\delta_T$  is the difference in chemical shifts between the “free” and bound states (in ppm),  $\delta_{obs}$  is the observed chemical shift (in ppm), and  $\delta_{Free}$  is the chemical shift in the “free” state (in ppm). The data was fitted using Origin software (OriginLab Corporation), with  $\Delta\delta_T$  and  $K_d$  allowed to vary during fitting (Figure 4.3b).

As shown in Figures 4.2a and 4.3b, the majority of chemical shift perturbations localize in and around the bulge for each aminoglycoside, in good agreement with previous work<sup>16</sup>. The chemically similar neomycin B (NeoB) and paromomycin (Par) induce similar chemical shift perturbations, which are particularly pronounced in and around the bulge (Figures 4.2a-b and 4.3). Likewise, we observe similar perturbations among the chemically similar tobramycin (Tob) and kanamycin B (KanB) (Figures 4.2a-b and 4.3). However, the NeoB/Par chemical shift perturbations observed differed from those of Tob/KanB, and an entirely unique set of chemical shifts were observed for ribostamycin (Rib) (Figures 4.2a-b and 4.3). Thus, it appears that the aminoglycosides either bind to distinct TAR conformations or interacted uniquely with a common TAR structure.

#### **4.4 Chemical shift intensities**

Another NMR observation used in this analysis was chemical shift intensity measurements, which report on the relative structural/dynamical timescales of specific sites. For this work, chemical shift intensities provide qualitative information on the local structural dynamics occurring in each of the RNA-ligand

complexes. As discussed in earlier work<sup>29</sup>, chemical shift intensities can be an informative first measure of structure and dynamics which allows one to ascertain the presence of potential domain motions. By normalizing intensities measured for each of the five aminoglycosides we were able to compare and contrast measurements for each of the complexes. In agreement with our observations described in section 4.3, residues in the bulge, as well as the loop, were found to exhibit higher dynamics than intensities measured from either helix.

Chemical shift intensities were determined by measuring intensities from peaks picked by NMRview. Resonances were sorted according to their respective nucleobase and atom type. Within a set of like-resonances, a single peak intensity from a known Watson-Crick base pair was chosen from the helix that dominates tumbling, and its intensity was arbitrarily set to 0.1. All other resonances within that same set were normalized against this reference intensity (Figure 4.4b). Using this procedure, intensities greater than 0.1 are likely resonances from nucleotides exhibiting motions greater than those experienced by the helical domain. For example, atoms of the sugar and nucleobases for bulge and loop regions reveal the presence of faster motions given their higher intensities, which is consistent with previous studies of TAR's local dynamics<sup>29</sup> (Figure 4.4b). On the other hand, intensities below 0.1 are likely involved in exchange broadening processes<sup>29</sup>. Examples of such processes are exclusively located within helix 1, which corresponds to the larger chemical shift perturbations previously described (Figure 4a). This likely arises as a result of dynamic localization processes associated with the presence of an aminoglycoside molecule. However, just as with the chemical



shift perturbation analysis, we are unable to completely distinguish between scenarios where the aminoglycoside binds a distinct or common TAR conformer.

#### **4.5 Measurement of residual dipolar couplings and order-tensor analysis**

The alignment of samples was prepared by the addition of Pf1 phage<sup>30</sup> into NMR buffer with the TAR-aminoglycoside complexes to yield a final Pf1 phage concentration of ~19 mg/mL and final TAR concentrations ranging between 0.15 and 0.17 mM. The addition of phage did not affect the TAR-aminoglycoside conformation as judged from careful comparison of chemical shifts in the absence and presence of phage. As mentioned above, RDCs were measured under conditions in which 98%, 97%, 98%, 97%, and 95% of TAR is estimated to be bound to NeoB, Par, KanB, Tob and Rib, respectively.

The RDC experiments were measured at 298 K on an Avance Bruker 600 MHz spectrometer equipped with a triple-resonance cryogenic probe for NeoB, Par and Tob, and on an 800 MHz Varian Inova spectrometer equipped with a triple-resonance Z-gradient probe for KanB and Rib. 2D  $^{13}\text{C}$ - $^1\text{H}$  (or  $^{15}\text{N}$ - $^1\text{H}$ ) S3E HSQC experiments were used to measure one bond  $^1D_{\text{C}_6\text{H}_6}$ ,  $^1D_{\text{C}_8\text{H}_8}$ ,  $^1D_{\text{C}_5\text{H}_5}$ ,  $^1D_{\text{C}_2\text{H}_2}$ ,  $^1D_{\text{C}_1'\text{H}_1'}$ , and  $^1D_{\text{N}_1/3\text{H}_1/3}$  RDCs by computing the difference in splittings along the  $^{13}\text{C}$  (or  $^{15}\text{N}$ ) dimension as observed in the presence and absence of Pf1 phage<sup>30</sup>. RDC measurement error was estimated from duplicate experiments that yielded splittings along the  $^1\text{H}$  or  $^{13}\text{C}/^{15}\text{N}$  dimension except in the case of Par and Tob where only the  $^{13}\text{C}$  RDCs were measured (Figure 4.4c). Measured RDCs for each aminoglycoside complex are listed in Table 4.1.

The RDCs measured in non-terminal Watson-Crick base pairs of the 5' and 3' helices of TAR, also known as helices I and II, respectively, were subjected to an order tensor analysis, as described in chapter 2, using co-axial idealized A-form helices as input coordinates<sup>31</sup>. In all cases, RDCs from flexible base pairs G17-C45 and A22-U40 were excluded from the order tensor analysis<sup>31</sup>. Key statistics pertaining to the order tensor analysis are shown in Table 4.2. The helices used in the order tensor analysis were constructed using the Biopolymers module in Insight II (Molecular Simulations, Inc.) as previously described in chapter 2. Following construction of each helix a correction to the propeller twist angles was applied to change the parameter from +15° to an idealized A-form RNA value of -15°<sup>31</sup>. A crystal structure of the UUCG loop<sup>32</sup> was appended to helix II by superimposing the closing base pair of the x-ray structure as previously described<sup>33</sup>.

The RDCs measured in structurally stable residues U31 and G34<sup>34,35</sup> of the loop and non-terminal base pairs in the idealized A-form helices were included in the determination of best-fit order tensors for each domain using singular value decomposition as implemented by the in-house program RAMAH<sup>36,37</sup>. Order tensor errors due to “A-form structural noise” and RDC uncertainty were estimated using the program AFORM-RDC<sup>31</sup>. Order tensor analysis of the <sup>1</sup>H/<sup>13</sup>C average and <sup>13</sup>C measured RDCs for NeoB, Rib and KanB using AFORM-RDC yielded nearly identical results within experimental error.

#### 4.6 Determination of inter-helical Euler angles

Using an order tensor analysis<sup>31</sup> of NMR residual dipolar couplings (RDCs)<sup>38</sup> (Tables 4.1-4.2 and Figure 4.4c), we determined the relative orientation of TAR helices in each aminoglycoside complex. As described in chapter 2, the overall structure of each TAR-aminoglycoside complex was assembled by rotating each domain into its principal axis system (PAS) of the best-fit order tensor and by bringing the O3' and P atoms from the two helices in the non-bulge strand with a distance of 1.59 Å. In each case, three of the four degenerate inter-helical orientations were omitted because they resulted in steric collisions and/or a distance between A22(O3')-G26(P) that exceeds 28 Å and that could not be satisfactorily linked using only three bulge nucleotides, as previously described<sup>39</sup>. The high uncertainty in the twist angles ( $\alpha_h$  and  $\gamma_h$ ) arises in part due to near axial symmetry of the order tensor ( $\eta$  ranged between 0.15 and 0.50).

Results confirmed that NeoB/Par, KanB/Tob, and Rib bind and stabilize different TAR conformations within the trinucleotide bulge-encoded distribution (Figure 4.2d). Best-fit order tensor frames specified by rotation matrices ( $R_i'$  and  $R_j'$ ) that diagonalize the order tensors of the 5' and 3' helices I and II respectively were used to compute the rotation  $R_{ij}(\alpha_h \beta_h \gamma_h) = R_i'^{-1}R_j'$  from which the inter-helical angles  $\alpha_h$ ,  $\beta_h$  and  $\gamma_h$  were deduced. The error in  $R_i'$  and  $R_j'$  obtained from AFORM-RDC was propagated into the angles  $-\gamma_h -\beta_h -\alpha_h$  using the  $R_{ij} = R_i'^{-1}R_j'$  relation.

#### 4.7 TAR-Ligand relationships

Interestingly, the aminoglycoside-bound TAR conformations trace a linear conformational pathway along the bulge-encoded distribution (Figure 4.2d). In order to gain a more complete understanding of this RNA-ligand relationship, we compared the inter-helical Euler angle measurements above with the physical properties of the small molecules. All small molecule properties (e.g. charge, volume, solvent accessible surface-area) for each ligand were computed using the chemical calculator plugins program – cxcalc – from ChemAxon (<http://www.chemaxon.com>). Using these values we were able to fit the chemical properties output by the chemical calculator against the  $\alpha_h$   $\beta_h$   $\gamma_h$  values from the RNA in an attempt to derive simple yet revealing relationships.

Remarkably, the specific position of a bound TAR conformation along the conformational pathway is quantitatively encoded by the aminoglycoside size. We observe a strong correlation between all three TAR inter-helical angles and the aminoglycoside solvent accessible surface area (SAS) (or volume, data not shown), with the larger aminoglycosides favoring more bent and twisted conformations (Figure 4.2e). A correlation is also observed between the aminoglycoside SAS and the NMR derived binding affinities (Figure 4.3), with tighter binding resulting in more bent conformations (Figure 4.2e). Thus, changing the size of aminoglycosides allows selective and tunable capture of distinct TAR inter-helical orientations spanning a total of  $\sim 160^\circ$  in  $\alpha_h$   $\beta_h$   $\gamma_h$  space.

#### 4.8 Universality of size-encoded conformational selection

Given the trend described in section 4.7 for TAR-aminoglycoside complexes, one has to wonder if this trend holds true for other types of RNA-ligand interactions. Another search of the PDB<sup>40</sup> for RNA structures bound to three or more small molecules brought to light examples of five other RNAs. Analysis of these additional RNA structures revealed that the trends observed for TAR are a general feature of RNA small molecule structural adaptation. For a variety of small molecule (Figure 4.5a-b) bound RNA junctions with known x-ray crystal structures<sup>41-45</sup>, small molecules capture inter-helical orientations that trace linear pathways along the junction-encoded topological distribution (Figures 4.6a and 4.7). In all cases, increasing the small molecule size results in increased inter-helical bending and correlated clockwise or anti-clockwise helical twisting.

The size-encoded relations observed in A-site constructs containing an  $S_2S_1$  type junction (Figure 4.6) are preserved in the larger ribosome context (Figure 4.8). Interestingly, reversing the directionality of the A-site sequence, as occurs in the dimerization initiation site (DIS) leads to a flip in the clockwise versus anti-clockwise sense of the twist angles (Figure 4.6). Large variations in inter-helical angles are also observed for  $S_9S_2$  and  $S_3S_3$ -type junctions in the Thi-Box riboswitch when bound to thiamine phosphate analogues that differ in size but otherwise contain identical charge<sup>46</sup> (Figures 4.5 and 4.6). Conversely, the variations are insignificant for TPP  $S_9S_2$  and  $S_3S_3$ -type junctions when bound to thiamine phosphate analogues that have similar size<sup>47</sup> (Figure 4.7). In general, weaker

correlations are observed with other small molecules properties such as charge, as described in section 4.7 (Figure 4.9).

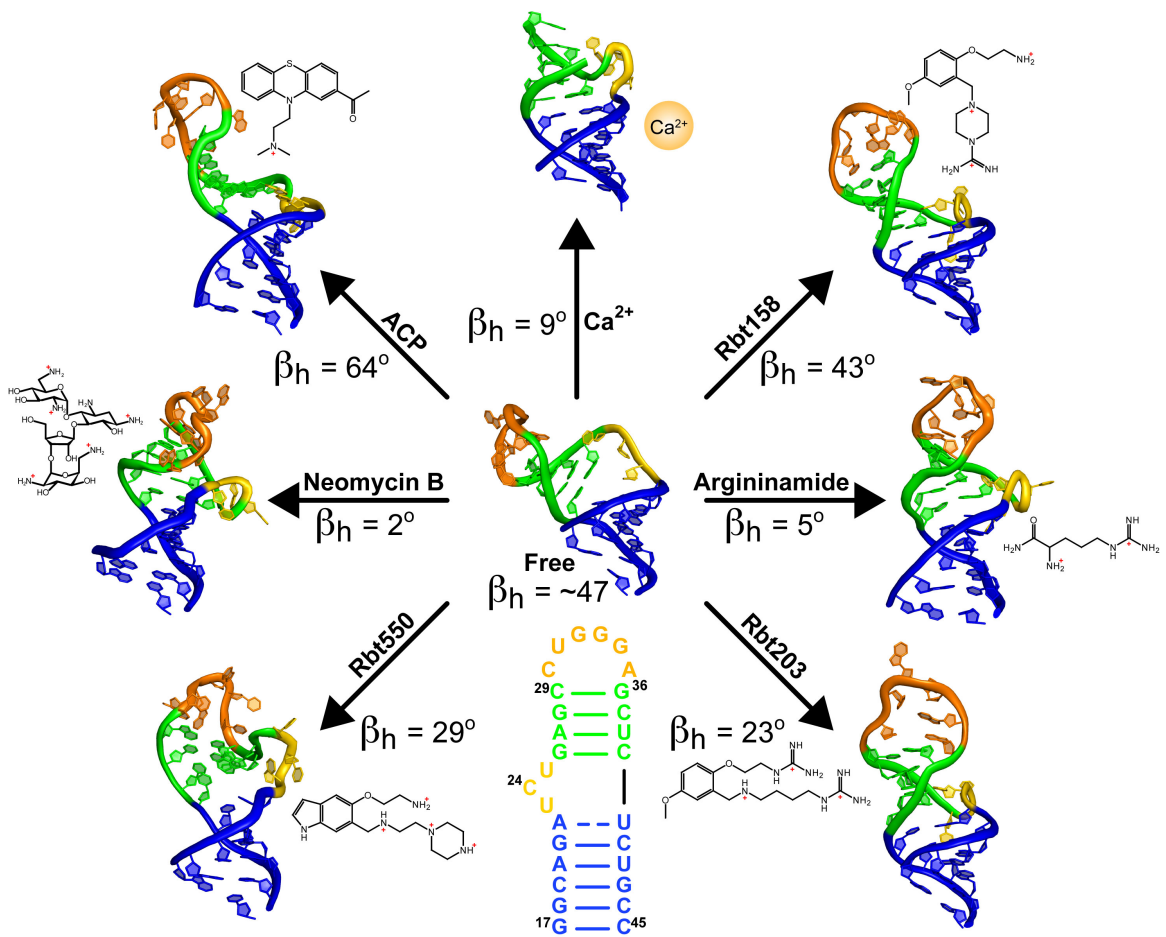
#### **4.9 Conclusions**

What is the molecular basis for size-encoded RNA conformational selection? The X-ray structures of RNA complexes reveal that junctions are enlarged by variable degrees of inter-helical bending and twisting so as to accommodate small molecules of different sizes with optimal packing (Figures 4.6). The strong correlations observed between inter-helical bending and twisting are similar to that observed for highly topologically confined junctions (e.g.  $S_1S_0$ , Figure 3.1a), and likely reflect added topological constraints arising from insertion of the small molecules within the junctions. Thus, topological constraints are also seen to contribute toward defining the global conformation and dynamic adaptation of RNA.

Herein we have shown that the size of a small-molecule is a major determinant in determining what the overall global conformation of a RNA will be. However, there are still a number of other factors which are important and that do make contributions toward binding and recognition of the RNA target. Ultimately we do not fully understand how each of these different characteristics of RNA ligands function together. Going forward, we will need to determine more characteristics that contribute to the binding and recognition of RNA elements. Moreover, it will be necessary to determine how these different features of the ligand relate to the global and local topology of the RNA.

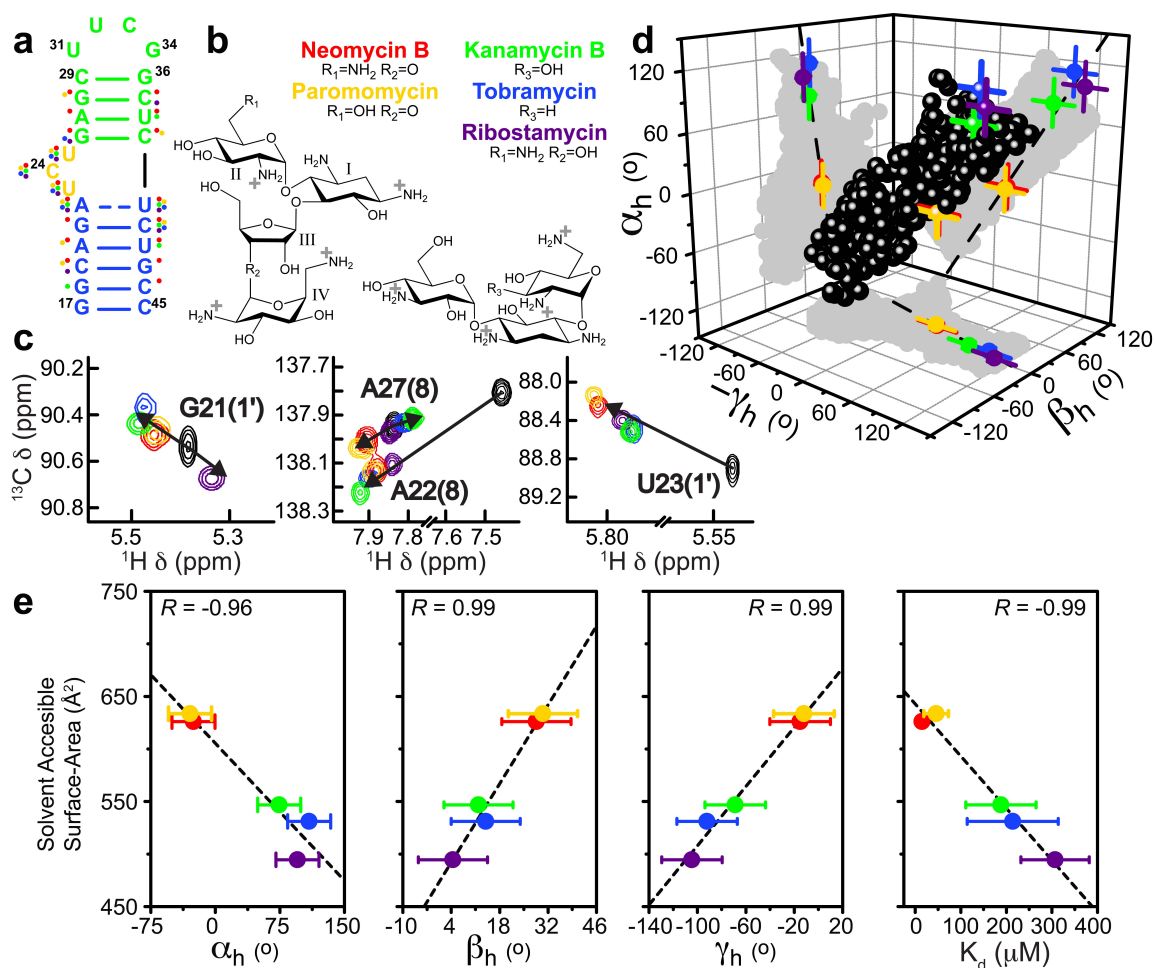
The ability to successfully find appropriate RNA targets and predict their free and ligand-bound conformations, or binding pose, will be the next step in this process. Related work from this lab has already shown that determination of specific RNA-ligand bound conformations is a challenge best met with new approaches to structure determination, and will likely require integration of both computational and experimental methods<sup>48,49</sup>. However, if ultimately successful, the manipulation of RNA using small-molecules will present us with an unparalleled ability to develop novel therapeutics and further expand our biochemical tool-kit and our knowledge of RNA function. Thus, it would seem that the rational manipulation of RNA structure, and possibly activity, with small molecules appears to be within our reach.

This work was published in the journal *Science*<sup>50</sup>. HM Al-Hashimi and MH Bailor conceived the idea. MH Bailor and X Sun prepared RNA Samples. MH Bailor collected and measured residual dipolar couplings, dissociation constants, chemical shift perturbations and chemical shift intensities. HM Al-Hashim and MH Bailor analyzed data.

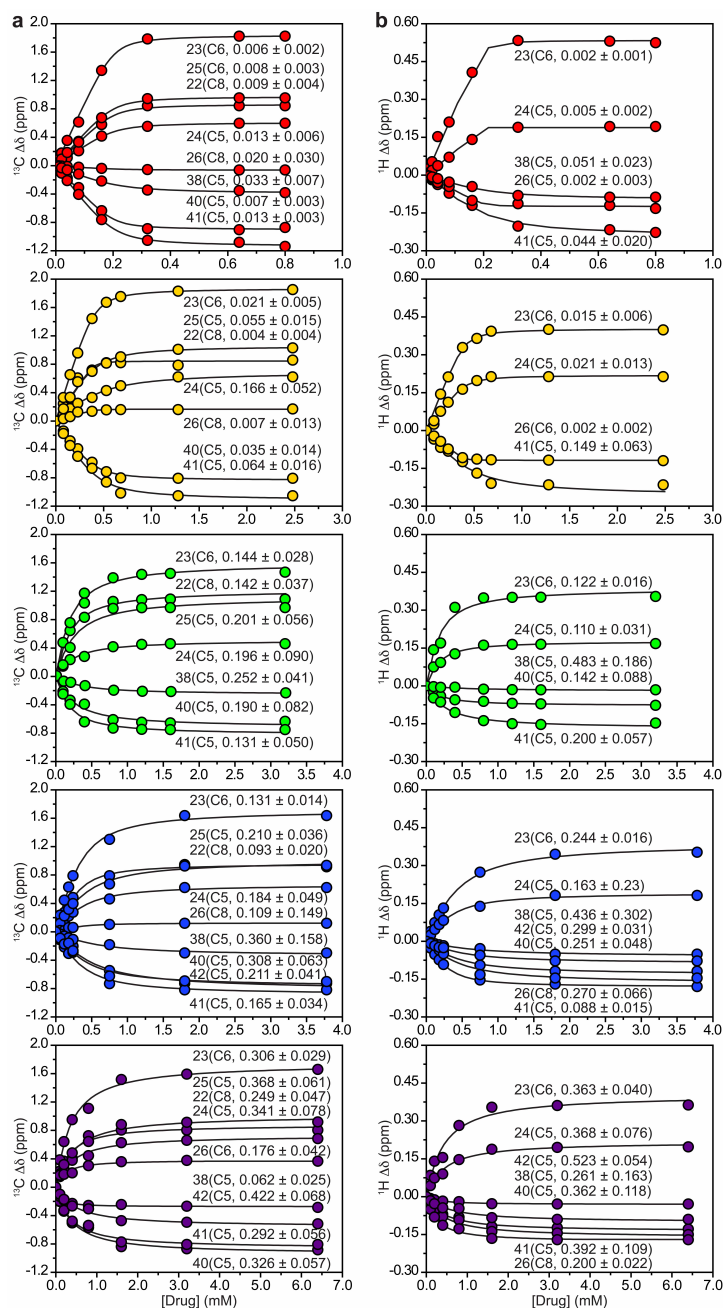


**Figure 4.1: HIV-1 transactivation element (TAR) structure adaptation pinwheel.** High-resolution structures of the TAR free and bound to a variety of small molecules. Each TAR-bound conformer displays a different inter-helical bend angle ( $\beta_h = 2^\circ - 64^\circ$ ) with regard to each ligand (PDB accession numbers 1ANR, 1ARJ, 1LVJ, 1QD3, 1UUD, 1UUI, 1UTS, 397D).

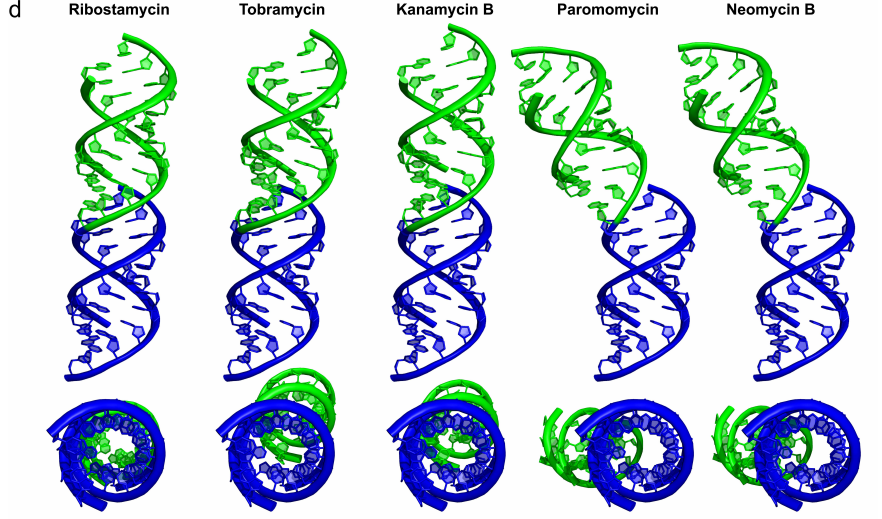
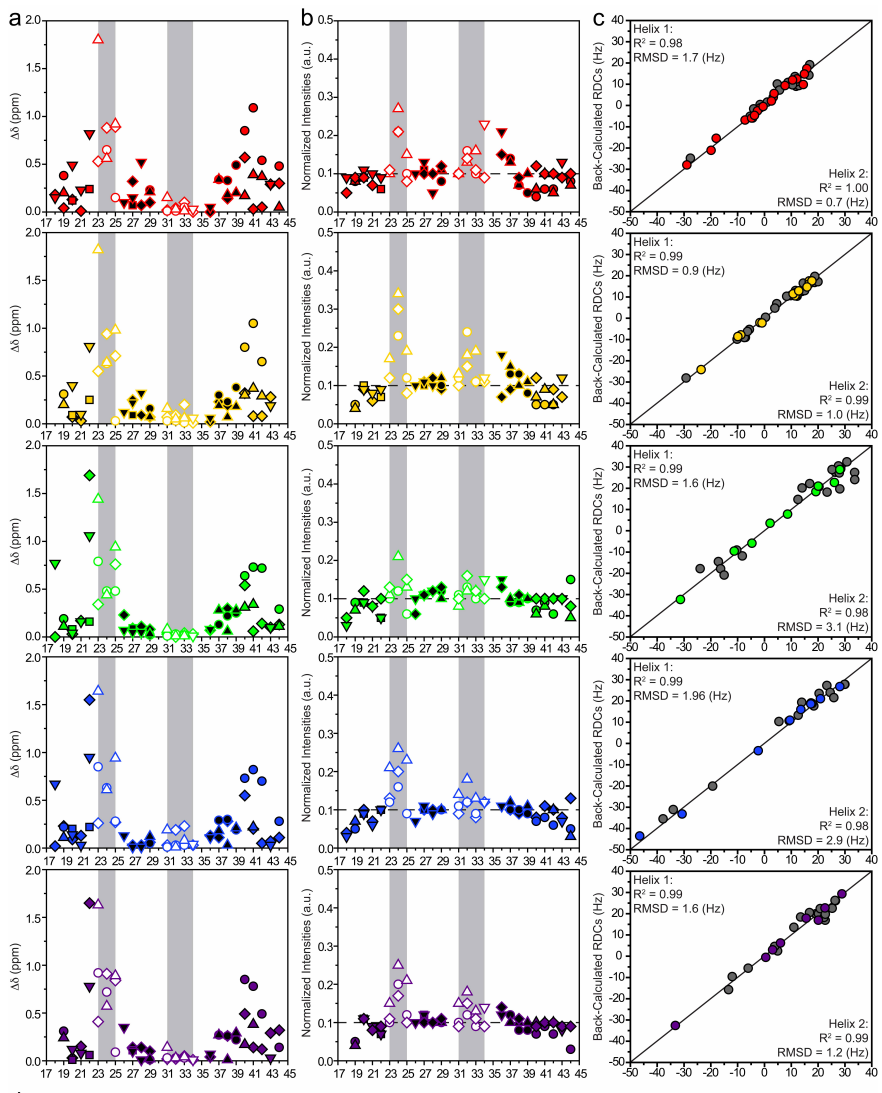




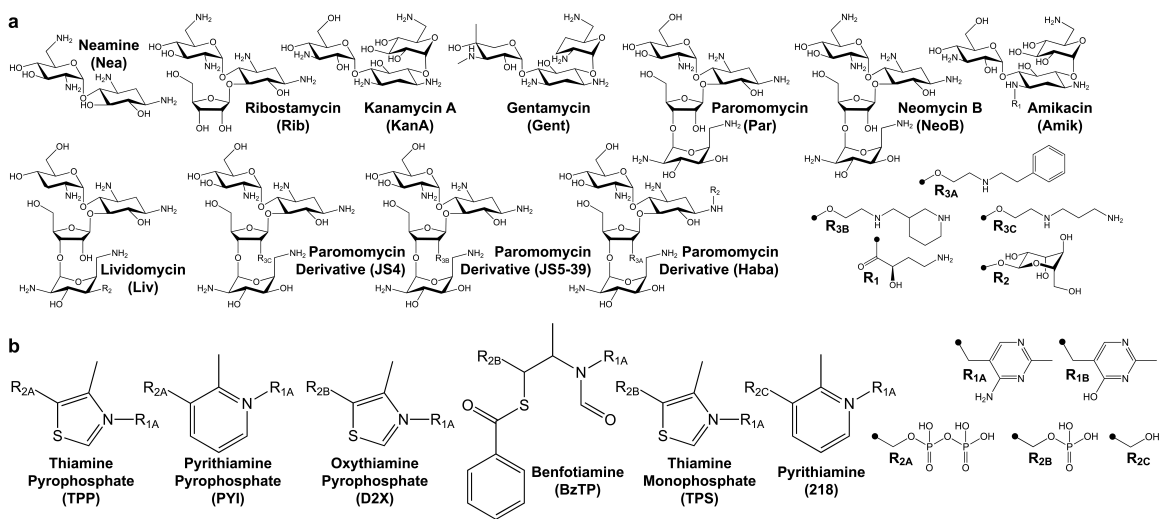
**Figure 4.2: Size-encoded selection of TAR inter-helical orientations using aminoglycosides.** (a) Secondary structure of the TAR construct used in NMR studies in which the wild-type apical loop has been replaced with a UUCG loop. Residues that undergo significant chemical shift perturbations ( $\Delta\delta_{total} > 0.30$  ppm) are highlighted using circles that are color coded according to aminoglycoside. (b) Chemical structure of the five aminoglycosides. (c) Examples of TAR NMR chemical shift perturbations highlighting differences in the aminoglycoside binding modes. (d) Inter-helical orientational maps showing the aminoglycoside-bound TAR conformations (color-coded according to aminoglycoside) and the computed topologically allowed inter-helical orientations for trinucleotide bulges (black). Shown are 2D projections of the 3D best-fit straight line through the aminoglycoside-bound TAR conformations. (e) Correlation plots between the solvent accessible surface area (SASA) of the aminoglycosides versus the bound TAR inter-helical angles (left three panels) and the dissociation constant ( $K_d$ , right panel). Points are color-coded accordingly to aminoglycoside. The best-fit line is shown in each case along with the correlation coefficient ( $R$ ).



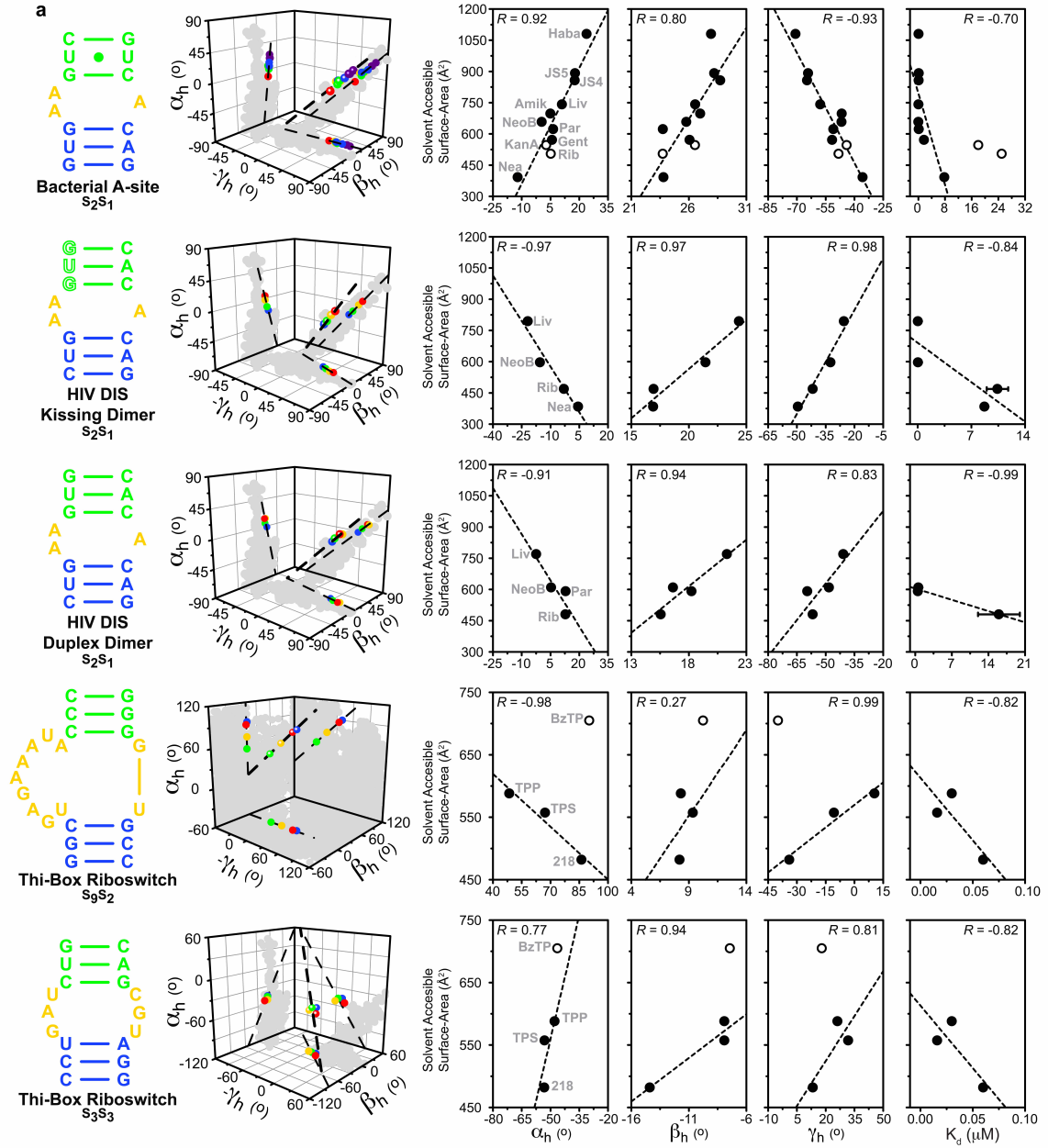
**Figure 4.3: Titration curves as a function of total aminoglycoside concentration.** TAR was titrated against each of the five aminoglycosides, which are color-coded according to color scheme used in Figure 4.2b with the best-fitted  $K_d$  indicated next to each curve. (a-b) Chemical shifts of aminoglycosides in the  $^{13}\text{C}$ - and  $^1\text{H}$ -dimensions, respectively, with the calculated apparent dissociation constant and associated error.



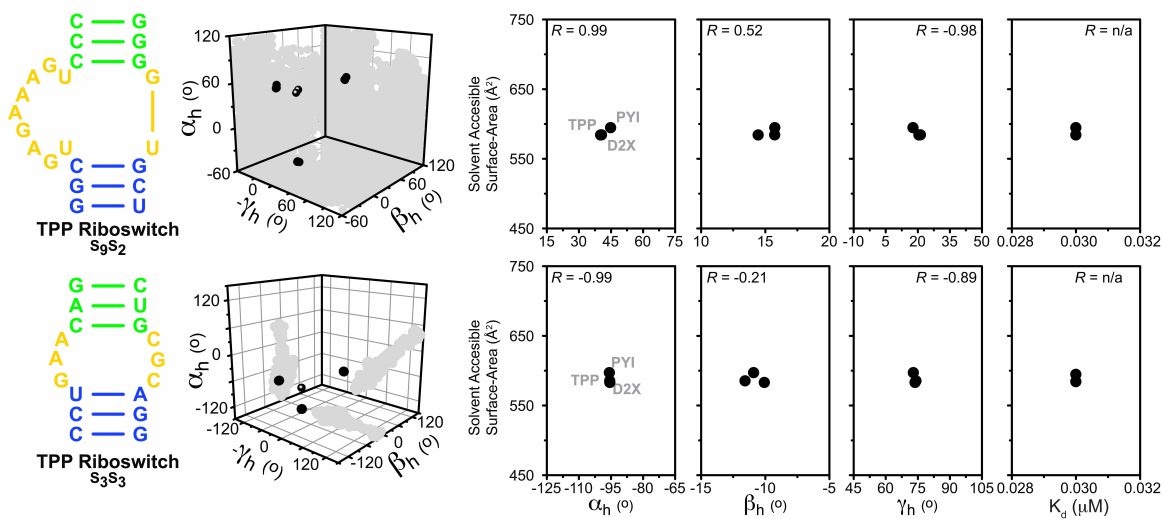
**Figure 4.4: Chemical shift perturbations, intensities and residual dipolar couplings and derived models for aminoglycoside-bound TAR complexes.** (a) Total chemical shift perturbations for TAR-aminoglycoside complexes are shown for each residue for a variety of nucleobase and sugar resonances, which were used to deduce the binding of ligands and for the calculations of dissociation constants ( $K_d$ ). (b) Chemical shift intensities for TAR-aminoglycoside complexes are shown for each residue for a variety of nucleobase and sugar resonances, which qualitatively report on the presence of local dynamics or exchanging processes. (c) Correlation plots between measured RDCs and values back-calculated when independently fitting helix order tensors to an idealized A-form geometry for TAR bound to neomycin (red), paromomycin (yellow), kanamycin B (green), tobramycin (blue) and ribostamycin (purple). Helix I and II RDCs are shown in color and grey respectively. The root-mean-square-deviation (RMSD) and correlation coefficient ( $R$ ) are shown on each plot. (d) RDC-derived inter-helical orientation of TAR bound to each aminoglycoside. Diamonds, squares, circles, up-triangle and down-triangle correspond to C1'H1', C2H2, C5H5, C6H6 and C8H8, respectively.



**Figure 4.5: Chemical structures of ligands from complexes of bound two-way junction RNA elements. (a-b) Chemical structures of the small molecules in Figure 4.6-4.8.**

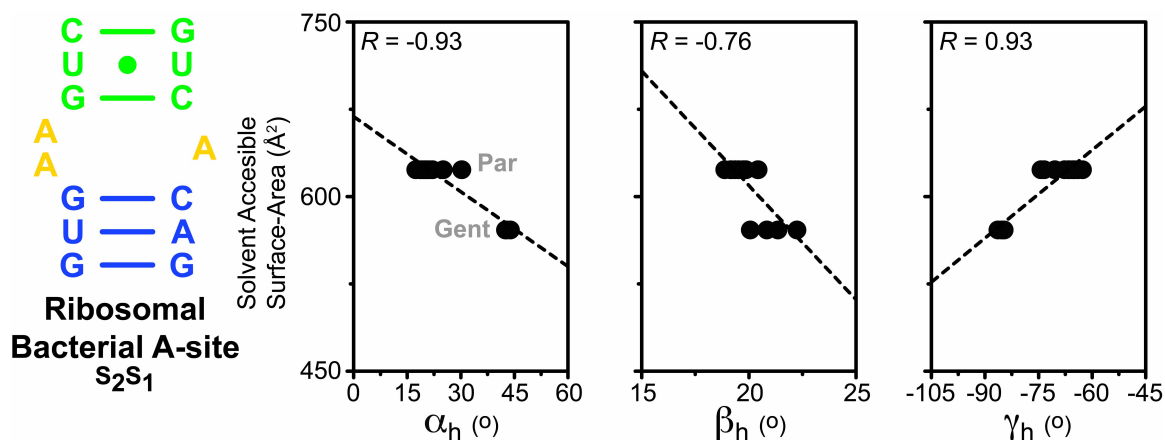


**Figure 4.6: Generalized size-encoded RNA conformational selection using small molecules.** (a) Inter-helical orientation maps depicting small molecule bound RNA inter-helical orientations color-coded according to small molecule solvent accessible surface area (SAS, ranging from red = 400 Å<sup>2</sup> to purple = 1100 Å<sup>2</sup>) and the junction-allowed topological distribution (in grey) for diverse RNA two-way junctions (shown and labeled to the left) including aminoglycosides and derivatives bound to A-site (1J7T, 2BE0, 2BEE, 2ESI, 2ESJ, 2ET3, 2ET4, 2ET5, 2ET8, 2G5Q, 2PWT), HIV kissing dimers (2FCX, 2FCY, 2FCZ, 2FD0); thiamine phosphate analogues bound to *E. coli* Thi-Box (2HOJ, 2HOK, 2HOL, 2HOM, 2HOO, 2HOP); and for aminoglycosides bound to HIV duplex dimers (3C3Z, 3C44, 3C5D, 3C7R). Shown to the right are correlation plots between the small molecule SAS and bound RNA inter-helical angles ( $\alpha_h\beta_h\gamma_h$ ) and  $K_d$ . Outliers shown in open symbols represent cases in which more than one small-molecule is bound to the RNA or in which added functional groups protrude out in solution and do not directly contact the RNA. (b) A representative series of X-ray structures of two-way RNA junction bound to small molecules of increasing size. The specific example is of the DIS kissing dimer bound to aminoglycosides (PDB accession 2FCX, 2FCY, 2FCZ, 2FD0). The aminoglycosides are shown in red and residues in the junction in yellow.



**Figure 4.7: Control examples of generalized size-encoded RNA conformational selection using small molecules.** Correlation plots between the small molecule SAS and bound RNA inter-helical angles ( $\alpha_h\beta_h\gamma_h$ ) and  $K_d$  for thiamine phosphate analogues bound to *E. coli* Thi-Box (2HOJ, 2HOK, 2HOL, 2HOM, 2HOO, 2HOP) and eukaryotic TPP riboswitch (3D2G, 3D2V, 3D2X).





**Figure 4.8: Inter-helical Euler angles of A-site in the context of the 16S ribosome bound to a variety of aminoglycosides.** Correlation plots are shown for each individual inter-helical Euler angle and the corresponding small molecule solvent accessible surface area (SAS) (PDB accession numbers 1FJG, 1IBK, 1N32, 1N33, 1XMO, 2J00, 2J02, 2QB9, 2QBB, 2QBH, 2QBJ, 2UU9, 2UUA, 2UUB, 2UUC). Correlation coefficients ( $R$ ) are shown in each panel.



<b>Residue</b>	<b>NeoB</b>	<b>Par</b>	<b>Rib</b>	<b>Tob</b>	<b>KanB</b>
18 (C8H8)	-8.9	n/a	n/a	n/a	1.9
18 (N1H1)	2.4	-3.1	n/a	n/a	n/a
18 (C1'H1')	-18.2	n/a	n/a	n/a	n/a
19 (C6H6)	-1.7	-5.2	n/a	-2.5	n/a
19 (C1'H1')	-5.4	n/a	n/a	-8.2	n/a
19 (C5H5)	-7.4	n/a	n/a	n/a	-3.4
20 (C8H8)	3.1	11.2	11.1	9.3	26.0
20 (C2H2)	-2.8	-1.9	0.3	3.3	8.5
20 (C1'H1')	-29.1	-32.2	-44.6	-46.7	-44.2
21 (C8H8)	14.3	16.5	19.9	n/a	32.5
21 (C1'H1')	-20.1	-23.8	-33.4	-30.9	-31.5
21 (N1H1)	-5	-9	n/a	n/a	n/a
22 (C8H8)	15.6	26.9	25.9	16.3	22
22 (C2H2)	12.6	16.3	22.8	12.2	25.4
22 (C1'H1')	n/a	n/a	2.4	1.9	0.3
23 (C6H6)	1.7	2.1	5.4	5.3	0.7
23 (C1'H1')	3.3	4.9	1.2	-2.6	-2.2
23 (C5H5)	n/a	n/a	3.6	1.8	1.2
24 (C6H6)	-2.1	-1.1	-0.9	-4.7	0.1
24 (C1'H1')	-1.8	n/a	n/a	n/a	n/a
24 (C5H5)	0.1	-0.1	-2.6	n/a	-3.1
25 (C6H6)	-1.8	-0.7	-2.1	4.1	4.6
25 (C1'H1')	0.1	-1.3	-2.4	-9.7	-13.5
25 (C5H5)	5.6	11	5.9	3.9	8.5
26 (C8H8)	14.3	18.5	26.3	29.8	24.8
26 (C1'H1')	n/a	-3.2	-6.3	-9.7	-8.5
26 (N1H1)	-2.1	-5.8	n/a	n/a	n/a
27 (C8H8)	11.2	12	21.5	20.2	22.8
27 (C2H2)	16.7	15.6	26.2	23.1	23.2
27 (C1'H1')	15.4	11.4	4.7	n/a	-17.4
28 (C8H8)	9.4	11.4	19.0	26.4	26.4
28 (C1'H1')	11	4.4	3.6	-15.6	-16.6
28 (N1H1)	-5	-7.3	n/a	n/a	n/a
29 (C6H6)	8.6	14.3	20.0	18.2	28.3
29 (C1'H1')	n/a	0.2	n/a	n/a	-24.2
29 (C5H5)	n/a	12	26.9	25.7	20.2
31 (C6H6)	4.6	11.4	13.3	12.4	25.0
31 (C1'H1')	-4.2	-10.4	-12.2	-16.0	-15.2
31 (C5H5)	10.3	14.8	22.0	13.7	27.9
32 (C6H6)	10.7	13.2	15.9	17.7	15.1
32 (C1'H1')	6.5	9.5	15.3	17.9	22.7
32 (C5H5)	n/a	3.7	12.9	14.6	20.3
33 (C6H6)	0.0	1.8	6.8	6.2	10.4
33 (C1'H1')	-22.4	-26.1	-27.9	-27.2	-16.75
33 (C5H5)	-10.7	-12.9	-11.8	-13.3	-3
34 (C8H8)	5.4	8.1	15.7	17.6	25.9
34 (C1'H1')	0.9	3.7	10.8	9.0	16.6
34 (N1H1)	n/a	n/a	n/a	n/a	n/a
36 (C8H8)	11.6	15.7	22.5	26.3	33.5
36 (C1'H1')	-27.8	-35.6	n/a	n/a	n/a
36 (N1H1)	-4.6	-7.7	n/a	n/a	n/a

37 (C6H6)	10.4	19.8	16.6	14.3	26.3
37 (C1'H1')	n/a	-29.4	n/a	-34.2	n/a
37 (C5H5)	12.7	n/a	16.8	n/a	27.7
38 (C6H6)	12.7	13.8	n/a	n/a	13.9
38 (C1'H1')	n/a	-15.6	-13.6	-19.5	-10.5
38 (C5H5)	n/a	n/a	20.6	17.2	30.6
38 (N3H3)	-5.5	-6.5	n/a	n/a	n/a
39 (C6H6)	16.4	9.6	29.1	24.3	12.3
39 (C1'H1')	n/a	n/a	n/a	n/a	n/a
39 (C5H5)	11.8	18.6	25.1	n/a	27.4
40 (C6H6)	16.2	27.2	n/a	n/a	20.7
40 (C1'H1')	n/a	9.3	-21.1	n/a	-20.5
40 (C5H5)	13.4	16.2	17.5	19.7	18.2
41 (C6H6)	11.9	16.4	n/a	n/a	n/a
41 (C1'H1')	n/a	n/a	n/a	-19.2	n/a
41 (C5H5)	14.7	12.6	22.4	n/a	19.9
42 (C6H6)	7.5	17.5	n/a	n/a	n/a
42 (C1'H1')	15.6	15.7	n/a	n/a	-4.9
42 (C5H5)	n/a	n/a	28.7	20.7	25.9
42 (N3H3)	-3.9	-10.1	n/a	n/a	n/a
43 (C8H8)	3.4	1.6	5.8	13.4	19
43 (C1'H1')	10.3	10.5	2.9	-15.4	-11.5
43 (N1H1)	-0.7	-1.1	n/a	n/a	n/a
44 (C6H6)	-9.1	n/a	n/a	n/a	n/a
44 (C1'H1')	n/a	n/a	n/a	n/a	n/a
44 (C5H5)	n/a	n/a	15.4	n/a	27.9
45 (C6H6)	n/a	n/a	n/a	n/a	n/a
45 (C1'H1')	-15.6	n/a	-26.4	-11.5	-30.1
45 (C5H5)	4.8	n/a	n/a	n/a	n/a

**Table 4.2: Statistics for order tensor analysis of RDCs measured in HIV-1 TAR bound to aminoglycosides.** Shown are the number of RDCs used in the analysis (N), the condition number (CN) describing the orientational distribution of the RDC dataset<sup>51</sup>, the rootmean-square-deviation (rmsd) and correlation coefficient ( $R^2$ ) between measured and bestfitted RDC values, the asymmetry parameter ( $\eta$ ), generalized degree of order ( $\vartheta$ ) describing the degree of helix alignment<sup>51</sup>, the internal generalized degree of order ( $\vartheta_{\text{int}}$ ) describing the amplitude of inter-helical motions<sup>51</sup>. The corresponding inter-helical angles and errors are shown in Figure 4.2 of the main manuscript.

Complex	Helix	N	CN	rmsd (Hz)	$R^2$	$\eta$	$\vartheta(10^{-3})$	$\vartheta_{\text{int}}$
NeoB	I	18	2.8	1.7	0.98	$0.33 \pm 0.04$	$0.67 \pm 0.04$	$0.99 \pm 0.10$
	II	19	2.9	1.8	0.99	$0.39 \pm 0.07$	$0.68 \pm 0.06$	
Par	I	12	3.2	0.9	0.99	$0.15 \pm 0.07$	$0.80 \pm 0.02$	$0.91 \pm 0.05$
	II	23	2.9	1.5	0.99	$0.49 \pm 0.04$	$0.88 \pm 0.04$	
Rib	I	8	3.4	1.6	0.99	$0.25 \pm 0.09$	$1.03 \pm 0.05$	$0.97 \pm 0.09$
	II	16	2.5	2.2	0.99	$0.17 \pm 0.05$	$1.06 \pm 0.09$	
Tob	I	8	4.5	2.0	0.99	$0.24 \pm 0.18$	$1.03 \pm 0.05$	$0.97 \pm 0.07$
	II	16	3.9	3.9	0.98	$0.16 \pm 0.09$	$1.06 \pm 0.06$	
KanB	I	9	2.6	1.6	0.99	$0.30 \pm 0.10$	$1.25 \pm 0.05$	$0.99 \pm 0.08$
	II	20	2.6	4.6	0.98	$0.22 \pm 0.08$	$1.24 \pm 0.08$	

#### 4.10 References

1. Nagel, J. H. & Pleij, C. W. Self-induced structural switches in RNA. *Biochimie* **84**, 913-923 (2002).
2. Serganov, A., Yuan, Y. R., Pikovskaya, O., Polonskaia, A., *et al.* Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem Biol* **11**, 1729-1741 (2004).
3. Edwards, T. E., Klein, D. J. & Ferré-D'Amaré, A. R. Riboswitches: small-molecule recognition by gene regulatory RNAs. *Curr Opin Struct Biol* **17**, 273-279 (2007).
4. Serganov, A. The long and the short of riboswitches. *Curr Opin Struct Biol* **19**, 251-259 (2009).
5. Serganov, A., Polonskaia, A., Phan, A. T., Breaker, R. R. & Patel, D. J. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature* **441**, 1167-1171 (2006).
6. Miranda-Ríos, J. The THI-box riboswitch, or how RNA binds thiamin pyrophosphate. *Structure* **15**, 259-265 (2007).
7. Lang, K., Rieder, R. & Micura, R. Ligand-induced folding of the thiM TPP riboswitch investigated by a structure-based fluorescence spectroscopic approach. *Nucleic Acids Res* **35**, 5370-5378 (2007).
8. Wilson, C., Nix, J. & Szostak, J. Functional requirements for specific ligand recognition by a biotin-binding RNA pseudoknot. *Biochemistry* **37**, 14410-14419 (1998).
9. Pitt, S. W., Zhang, Q., Patel, D. J. & Al-Hashimi, H. M. Evidence that electrostatic interactions dictate the ligand-induced arrest of RNA global flexibility. *Angew Chem Int Ed Engl* **44**, 3412-3415 (2005).
10. Gilbert, S. D., Reyes, F. E., Edwards, A. L. & Batey, R. T. Adaptive ligand binding by the purine riboswitch in the recognition of guanine and adenine analogs. *Structure* **17**, 857-868 (2009).
11. Fritsch, V. & Westhof, E. Molecular adaptation in RNA complexes. *Structure* **17**, 784-786 (2009).
12. Hermann, T. Rational ligand design for RNA: the role of static structure and conformational flexibility in target recognition. *Biochimie* **84**, 869-875 (2002).
13. Blount, K. F., Zhao, F., Hermann, T. & Tor, Y. Conformational constraint as a means for understanding RNA-aminoglycoside specificity. *J Am Chem Soc* **127**, 9818-9829 (2005).
14. Aboul-ela, F., Karn, J. & Varani, G. The structure of the human immunodeficiency virus type-1 TAR RNA reveals principles of RNA recognition by Tat protein. *J Mol Biol* **253**, 313-332 (1995).
15. Aboul-ela, F., Karn, J. & Varani, G. Structure of HIV-1 TAR RNA in the absence of

- ligands reveals a novel conformation of the trinucleotide bulge. *Nucleic Acids Res* **24**, 3974-3981 (1996).
16. Faber, C., Sticht, H., Schweimer, K. & Rösch, P. Structural rearrangements of HIV-1 Tat-responsive RNA upon binding of neomycin B. *J Biol Chem* **275**, 20660-20666 (2000).
  17. Davis, B., Afshar, M., Varani, G., Murchie, A. I., *et al.* Rational design of inhibitors of HIV-1 TAR RNA through the stabilisation of electrostatic "hot spots". *J Mol Biol* **336**, 343-356 (2004).
  18. Murchie, A. I., Davis, B., Isel, C., Afshar, M., *et al.* Structure-based drug design targeting an inactive RNA conformation: exploiting the flexibility of HIV-1 TAR RNA. *J Mol Biol* **336**, 625-638 (2004).
  19. Ippolito, J. A. & Steitz, T. A. A 1.3-Å resolution crystal structure of the HIV-1 trans-activation response region RNA stem reveals a metal ion-dependent bulge conformation. *Proc Natl Acad Sci U S A* **95**, 9819-9824 (1998).
  20. Chen, K. & Kurgan, L. Investigation of Atomic Level Patterns in Protein—Small Ligand Interactions. *PLoS One* **4**, (2009).
  21. Kortagere, S., Krasowski, M. D. & Ekins, S. The importance of discerning shape in molecular pharmacology. *Trends Pharmacol Sci* **30**, 138-147 (2009).
  22. Luo, W., Pei, J. & Zhu, Y. A fast protein-ligand docking algorithm based on hydrogen bond matching and surface shape complementarity. *J Mol Model* **16**, 903-913 (2010).
  23. Ebalunode, J. O. & Zheng, W. Molecular Shape Technologies in Drug Discovery: Methods and Applications. *Curr Top Med Chem* (2010).
  24. Hermann, T. & Patel, D. J. Adaptive recognition by nucleic acid aptamers. *Science* **287**, 820-825 (2000).
  25. Zhang, Q., Stelzer, A. C., Fisher, C. K. & Al-Hashimi, H. M. Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature* **450**, 1263-1267 (2007).
  26. Thomas, J. R. & Hergenrother, P. J. Targeting RNA with small molecules. *Chem Rev* **108**, 1171-1224 (2008).
  27. Blount, K. F. & Tor, Y. Using pyrene-labeled HIV-1 TAR to measure RNA-small molecule binding. *Nucleic Acids Res* **31**, 5490 (2003).
  28. Zuideweg, E. R. Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry* **41**, 1-7 (2002).
  29. Zhang, Q., Sun, X., Watt, E. D. & Al-Hashimi, H. M. Resolving the motional modes that code for RNA adaptation. *Science* **311**, 653-656 (2006).
  30. Hansen, M. R., Mueller, L. & Pardi, A. Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nat Struct Biol* **5**, 1065-1074 (1998).

31. Musselman, C., Pitt, S. W., Gulati, K., Foster, L. L., *et al.* Impact of static and dynamic A-form heterogeneity on the determination of RNA global structural dynamics using NMR residual dipolar couplings. *J Biomol NMR* **36**, 235-249 (2006).
32. Ennifar, E., Nikulin, A., Tishchenko, S., Serganov, A., *et al.* The crystal structure of UUCG tetraloop. *J Mol Biol* **304**, 35-42 (2000).
33. Al-Hashimi, H. M., Gorin, A., Majumdar, A., Gosser, Y. & Patel, D. J. Towards structural genomics of RNA: rapid NMR resonance assignment and simultaneous RNA tertiary structure determination using residual dipolar couplings. *J Mol Biol* **318**, 637-649 (2002).
34. Duchardt, E. & Schwalbe, H. Residue specific ribose and nucleobase dynamics of the cUUCGg RNA tetraloop motif by MNMR <sup>13</sup>C relaxation. *J Biomol NMR* **32**, 295-308 (2005).
35. Hansen, A. L., Nikolova, E. N., Casiano-Negroni, A. & Al-Hashimi, H. M. Extending the range of microsecond-to-millisecond chemical exchange detected in labeled and unlabeled nucleic acids by selective carbon R(1rho) NMR spectroscopy. *J Am Chem Soc* **131**, 3818-3819 (2009).
36. Losonczi, J. A., Andrec, M., Fischer, M. W. F. & Prestegard, J. H. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J Mag Reson* **138**, 334-342 (1999).
37. Hansen, A. L. & Al-Hashimi, H. M. Insight into the CSA tensors of nucleobase carbons in RNA polynucleotides from solution measurements of residual CSA: towards new long-range orientational constraints. *J Magn Reson* **179**, 299-307 (2006).
38. Tjandra, N. & Bax, A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* **278**, 1111-1114 (1997).
39. Al-Hashimi, H. M., Gosser, Y., Gorin, A., Hu, W., *et al.* Concerted motions in HIV-1 TAR RNA may allow access to bound state conformations: RNA dynamics from NMR residual dipolar couplings. *J Mol Biol* **315**, 95-102 (2002).
40. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).
41. François, B., Russell, R. J., Murray, J. B., Aboul-ela, F., *et al.* Crystal structures of complexes between aminoglycosides and decoding A site oligonucleotides: role of the number of rings and positive charges in the specific binding leading to miscoding. *Nucleic Acids Res* **33**, 5677-5690 (2005).
42. Vicens, Q. & Westhof, E. Crystal structure of paromomycin docked into the eubacterial ribosomal decoding A site. *Structure* **9**, 647-658 (2001).
43. Kondo, J. *et al.* Crystal structure of the bacterial ribosomal decoding site complexed with a synthetic doubly functionalized paromomycin derivative: a new specific binding mode to an a-minor motif enhances in vitro antibacterial activity.



*ChemMedChem* **2**, 1631-1638 (2007).

44. Kondo, J., François, B., Russell, R. J., Murray, J. B. & Westhof, E. Crystal structure of the bacterial ribosomal decoding site complexed with amikacin containing the gamma-amino-alpha-hydroxybutyryl (haba) group. *Biochimie* **88**, 1027-1031 (2006).
45. François, B., Szychowski, J., Adhikari, S. S., Pachamuthu, K., *et al.* Antibacterial aminoglycosides with a modified mode of binding to the ribosomal-RNA decoding site. *Angew Chem Int Ed Engl* **43**, 6735-6738 (2004).
46. Edwards, T. E. & Ferré-D'Amaré, A. R. Crystal structures of the thi-box riboswitch bound to thiamine pyrophosphate analogs reveal adaptive RNA-small molecule recognition. *Structure* **14**, 1459-1468 (2006).
47. Thore, S., Leibundgut, M. & Ban, N. Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. *Science* **312**, 1208-1211 (2006).
48. Frank, A. T., Stelzer, A. C., Al-Hashimi, H. M. & Andricioaei, I. Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition. *Nucleic Acids Res* **37**, 3670-3679 (2009).
49. Stelzer, A. C., Frank, A. T., Bailor, M. H., Andricioaei, I. & Al-Hashimi, H. M. Constructing atomic-resolution RNA structural ensembles using MD and motionally decoupled NMR RDCs. *Methods* **49**, 167-173 (2009).
50. Bailor, M. H., Sun, X. & Al-Hashimi, H. M. Topology links RNA secondary structure with global conformation, dynamics, and adaptation. *Science* **327**, 202-206 (2010).
51. Tolman, J. R., Al-Hashimi, H. M., Kay, L. E. & Prestegard, J. H. Structural and dynamic analysis of residual dipolar coupling data for proteins. *J Am Chem Soc* **123**, 1416-1424 (2001).

## Chapter 5

### Conclusions and Future Directions

#### 5.1 Conclusions and Future

RNA's function, and that of all other biological macromolecule, is derived from a dynamic heterogeneous milieu. Temperature, metals, small molecules, proteins and other RNAs live and interact in a complex and temporal medium that elicits the necessary and critical functions that drive cellular processes. This is life; science is the test-tube. Our comprehension of RNA is still at a beginning, and yet there is much we have learned. Since Tom Cech and Sidney Altman first demonstrated the catalytic abilities of RNA in the 1980's our view of biology has been altered dramatically<sup>1-3</sup>. The research described herein reveals a simply truth: knowledge of RNA secondary structure for two-way junctions is sufficient to infer its structural arrangement, dynamics and adaptation. While no direct evidence has been presented, it is likely that these same constraints exist for pseudoknots, kissing complexes, three-way, four-way and other higher-order junctions.

As was discussed in chapter 2, the determination of RNA conformations at atomic resolution faces many challenges. Current structure determination techniques, while ultimately insightful, face significant limitations to the size and type of the RNAs that can be targeted. For instance, X-ray crystallography is challenged by less

ordered RNA conformers, or which there are many. And NMR is currently limited both by size of  $\sim 100$  nt<sup>4</sup> and the quality of force fields that guide the structure elucidation process. Moreover, other spectroscopic techniques which might be utilized in structure determination/inference strategies are also hindered by many of the physical characteristics of RNA. For instance, small-angle scattering (SAS) is a very promising spectroscopic technique, and amenable to a variety of physiological conditions, yet conformational dynamics hinders its utility<sup>5</sup>. As a result the majority of studies that produce physical models are limited to conditions involving ligand-binding or high-salt, which is known to dampen domain motions<sup>6</sup>.

For structure determination/inference strategies to expand beyond their current limitations, new computational techniques will need to be developed. For example, work from this lab has shown that the integration of molecular dynamics (MD) trajectories with residual dipolar coupling (RDC) measurements is capable of producing physical descriptions of the conformational dynamics occurring in solution<sup>7,8</sup>. Moreover, others have demonstrated the ability and utility to integrate rigid body<sup>9</sup> and short MD simulations<sup>10</sup> of proteins with SAS measurements. Given our results illustrating the coupling of topology with structure and dynamics, it seems a natural next step that future strategies would seek to integrate the information gained from knowledge of the secondary structure with other computational methods and experimental data. The coupling of theoretical and experimental should alleviate current limitations faced by employing singular methods to elucidate biological questions. For instance, the three-dimensionally

resolution of the ~9,000 nucleotide genome of HIV-1<sup>11</sup> will require more ingenuity than current structure determination techniques permit.

Hopefully, in pushing the envelop of structure determination and inference we will also push the boundaries of biotechnological applications. Future applications, such as RNA therapeutic development and gene therapy present exciting new avenues to applications of basic research whose potential appears to be almost limitless. Early work has already demonstrated the utility of mobile genetic RNA elements in targeting and treating diseases like cancer<sup>12-17</sup> and HIV<sup>18,19</sup>. In addition, other research has uncovered links to therapeutic resistance in bacteria<sup>20</sup> and the identification of prostate cancer markers<sup>21,22</sup> that will certainly aid methods for early detection. However our ability to make these applications feasible will largely depend our capacity to target specific genetic sequences from an enormous range of genes and cellular pathways. Thus, our ability to unify and rationalize the underlying relationship of sequence, secondary and tertiary structure, and RNA dynamics with its cellular function will be pivotal as we move forward.

## 5.2 References

1. Gold, L. Catalytic RNA: a Nobel Prize for small village science. *New Biol* **2**, 1-4 (1990).
2. North, G. Nobel prizes: chemistry. RNA's catalytic role. *Nature* **341**, 556 (1989).
3. Waldrop, M. M. Catalytic RNA wins chemistry Nobel. *Science* **246**, 325 (1989).
4. D'Souza, V., Dey, A., Habib, D. & Summers, M. F. NMR structure of the 101-nucleotide core encapsidation signal of the Moloney murine leukemia virus. *J Mol Biol* **337**, 427-442 (2004).
5. Lipfert, J., Ouellet, J., Norman, D. G., Doniach, S. & Lilley, D. M. The complete VS ribozyme in solution studied by small-angle X-ray scattering. *Structure* **16**, 1357-1367 (2008).
6. Casiano-Negroni, A., Sun, X. & Al-Hashimi, H. M. Probing Na(+)-induced changes in the HIV-1 TAR conformational dynamics using NMR residual dipolar couplings: new insights into the role of counterions and electrostatic interactions in adaptive recognition. *Biochemistry* **46**, 6525-6535 (2007).
7. Frank, A. T., Stelzer, A. C., Al-Hashimi, H. M. & Andricioaei, I. Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition. *Nucleic Acids Res* **37**, 3670-3679 (2009).
8. Stelzer, A. C., Frank, A. T., Bailor, M. H., Andricioaei, I. & Al-Hashimi, H. M. Constructing atomic-resolution RNA structural ensembles using MD and motionally decoupled NMR RDCs. *Methods* **49**, 167-173 (2009).
9. Petoukhov, M. V. & Svergun, D. I. Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophys J* **89**, 1237-1250 (2005).
10. Ahn, S., Kim, K. H., Kim, Y., Kim, J. & Ihee, H. Protein tertiary structural changes visualized by time-resolved X-ray solution scattering. *J Phys Chem B* **113**, 13131-13133 (2009).
11. Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., *et al.* Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**, 711-716 (2009).
12. Song, M. S. & Lee, S. W. Cancer-selective induction of cytotoxicity by tissue-specific expression of targeted trans-splicing ribozyme. *FEBS Lett* **580**, 5033-5043 (2006).
13. Ban, G., Song, M. S. & Lee, S. W. Cancer Cell Targeting with Mouse TERT-Specific Group I Intron of *Tetrahymena thermophila*. *J Microbiol Biotech* **19**, 1070 (2009).
14. Moreau, D., Jacquot, C., Tsita, P., Chinou, I., *et al.* Original triazine inductor of new specific molecular targets, with antitumor activity against nonsmall cell lung cancer. *Int J Cancer* **123**, 2676-2683 (2008).

15. Wei, M. Q., Mengesha, A., Good, D. & Anné, J. Bacterial targeted tumour therapy-dawn of a new era. *Cancer Lett* **259**, 16-27 (2008).
16. Wei, M. Q., Metharom, P., Ellem, K. A. & Barth, S. Search for "weapons of mass destruction" for cancer -- immuno/ gene therapy comes of age. *Cell Mol Immunol* **2**, 351-357 (2005).
17. Metharom, P., Ellem, K. A. & Wei, M. Q. Gene transfer to dendritic cells induced a protective immunity against melanoma. *Cell Mol Immunol* **2**, 281-288 (2005).
18. Nazari, R. & Joshi, S. Exploring the potential of group II introns to inactivate human immunodeficiency virus type 1. *J Gen Virol* **89**, 2605-2610 (2008).
19. Guo, H., Karberg, M., Long, M., Jones, J. P., *et al.* Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science* **289**, 452-457 (2000).
20. Centron, D. & Roy, P. H. Presence of a group II intron in a multiresistant *Serratia marcescens* strain that harbors three integrons and a novel gene fusion. *Antimicrobial Agents Chemo* **46**, 1402 (2002).
21. Romanuik, T. L., Ueda, T., Le, N., Haile, S., *et al.* Novel biomarkers for prostate cancer including noncoding transcripts. *Am J Pathol* **175**, 2264-2276 (2009).
22. Ding, Y., Larson, G., Rivas, G., Lundberg, C., *et al.* Strong signature of natural selection within an FHIT intron implicated in prostate cancer risk. *PLoS One* **3**, e3533 (2008).

## Appendix A

### EULER-RNA-AFORM.pl

```
#!/usr/bin/perl -w

use strict;
use Math::Trig;
use Math::Complex;

use constant N      => 6.022045e23;
use constant Hplank => 6.626208e-34;
use constant gH     => 26.7522128e07;
use constant gC     => 6.7282840e07;
use constant gN     => -2.71261804e07;
use constant kbol   => 1.38066e-23;
use constant mu0    => 4*pi*10**-07;

#####
#This program uses two pdbs and a residue number, input by the user to
#create a matrix that represents the structure for each PDB and then
#finds the corresponding rotation matrix. Finally, euler angles are
#calculated to determine the phase and bend of the two structures with
#respect to the molecular frame of the first PDB.
#Program Name: EulerRNA
#Created 11-22-06 by Maximillian H. Bailor
#####

#Global variables
my ($pdb_param_ref,@pdb_parameters,@pdb_data);
my ($write_file,$degx);
my ($pdbfile_1,$pdbfile_2,@bond_1,@bond_2,$residue);
my (@struc_vec_comp1,@struc_vec_comp2,@struc_vec_comp3,@struc_vec_comp4,@dcos);

if (!$ARGV[0]) {
    print "The first pdb file, reference structure:";
    $pdbfile_1 = <STDIN>;
    chomp $pdbfile_1;

    print "The second pdb file:";
    $pdbfile_2 = <STDIN>;
    chomp $pdbfile_2;

    print "Please list a specific residue (example, 17):";
    $residue = <STDIN>;
    chomp $residue;
}
```

```

} elsif (($#ARGV < 2) || ($#ARGV > 2)) {
    print "Your entry is incorrect, please enter in the following:\n";
    print "Example: test.pdb test_2.pdb res#\n";
    exit;
} else {
    $pdbfile_1 = $ARGV[0];
    $pdbfile_2 = $ARGV[1];
    $residue = $ARGV[2];
}

open(FH, "$pdbfile_1") || die "could not find file: $!";
my @pdbcontent_1 = <FH>;
close(FH);

open(GH, "$pdbfile_2") || die "could not find file: $!";
my @pdbcontent_2 = <GH>;
close(GH);

@bond_1 = ($residue, "C1", "H1");
@bond_2 = ($residue, "C2", "H2");

@struc_vec_comp1 = pdb_find($bond_1[0], $bond_1[1], $bond_1[2], @pdbcontent_1);
@struc_vec_comp2 = pdb_find($bond_2[0], $bond_2[1], $bond_2[2], @pdbcontent_1);
die "could not find the first atom in the first bond\n" unless $struc_vec_comp1[5];
die "could not find the second atom in the first bond\n" unless $struc_vec_comp2[5];

my @TrMatEV = pdb_find(46, "A", "A", @pdbcontent_1);

@struc_vec_comp3 = pdb_find($bond_1[0], $bond_1[1], $bond_1[2], @pdbcontent_2);
@struc_vec_comp4 = pdb_find($bond_2[0], $bond_2[1], $bond_2[2], @pdbcontent_2);
die "could not find the first atom in the second bond\n" unless $struc_vec_comp3[5];
die "could not find the second atom in the second bond\n" unless $struc_vec_comp4[5];

my @abg2 = Struc_MAT(@struc_vec_comp1, @struc_vec_comp2);
my @abg3 = Struc_MAT(@struc_vec_comp3, @struc_vec_comp4);

my @abg5 = inverse_mat(3, @abg3);

my @abg9 = mat_mult(3, @abg3, @abg5);
my $apple2 = mat_norm(@abg9);

my @abg13 = mat_scalar((1/$apple2), 9, @abg5);

my @abg14 = mat_mult(3, @abg13, @abg2);

my @spot = abg(@abg14);
printf "%5.2f\t%5.2f\t%5.2f\t%5.2f\t", $spot[0], $spot[1],
    $spot[2], $spot[0] + $spot[2];

@dcos = dircos_mat(@spot);
$dex = rotation_mat_x(@abg14, @dcos);

if ($dex == 1.00) { print "Degeneracy Found and Uncorrected!\n";
} elsif ($dex == 2.00) { ; }

```



```

print "\n";

sub abg {
  my (@a) = @_;
  my (@b,$a,$b,$g,@d,$d,$e,@newd,@newa);

  if ((($a[2] > 0.00) || ($a[6] < 0.00)) && !(abs($a[8]) > 1.00)) {
    $b = -1*acos($a[8])*180/pi;
  } elsif (abs($a[8]) > 1.00) {
    $b= acos(1.00)*180/pi;
  } else {
    $b = acos($a[8])*180/pi;
  }

  $a = alphgamcalc($a[6],$a[7],0.00,1.000,1.0e-6,1.0e
18,$b,2.50,$a[1],$a[3],"alpha",$a[0],$a[4]);
  $g = alphgamcalc($a[2],$a[5],0.00,1.000,1.0e-6,1.0e-
18,$b,2.50,$a[1],$a[3],"gamma",$a[0],$a[4]);
  $g *= -1;

  @b = ($a,$b,$g);

  @d = dircos_mat($a,$b,$g);
  foreach $d (@d) {
    $e = ones($d);
    push @newd, $e;
  }
  foreach $d (@a) {
    $e = ones($d);
    push @newa, $e;
  }

  if ((($newa[0] != $newd[0]) && ($newa[4] != $newd[4])) && (($newa[1] == $newd[1]) &&
($newa[3] == $newd[3]))) {
    $a = alphgamcalc($a[6],$a[7],0.00,1.000,1.0e-6,1.0e-
18,$b,1.00,$a[1],$a[3],"alternate",$a[0],$a[4]);
    $g = alphgamcalc($a[2],$a[5],0.00,1.000,1.0e-6,1.0e-
18,$b,1.00,$a[1],$a[3],"gamma",$a[0],$a[4]);
    $g *= -1;
    @b = ($a,$b,$g);
    return @b;
  } elsif (($newa[0] != $newd[0]) && ($newa[1] != $newd[1]) && ($newa[3] != $newd[3]) &&
($newa[4] != $newd[4])) {
    #corrects for degeneracies in alpha
    if (($newa[6] != $newd[6]) && ($newa[7] != $newd[7])) {
      $a -= 180;
      @b = ($a,$b,$g);
      return @b;
    } elsif (($newa[2] != $newd[2]) && ($newa[5] != $newd[5])) {
      $g -= 180;
      @b = ($a,$b,$g);
      return @b;
    } else {
      print "warning: one of your angles may contain a degeneracy!\n";
      return @b;
    }
  }
}

```

```

    }
  } else { return @b; }
}

sub dircos_mat {
  my ($a, $b, $g, @trash) = @_;
  $a = ($a*pi)/180.00;
  $b = ($b*pi)/180.00;
  $g = ($g*pi)/180.00;

  my $m11 = -sin($g)*sin($a) + cos($b)*cos($a)*cos($g);
  my $m12 = sin($g)*cos($a) + cos($b)*sin($a)*cos($g);
  my $m13 = -cos($g)*sin($b);
  my $m21 = -cos($g)*sin($a) - cos($b)*cos($a)*sin($g);
  my $m22 = cos($g)*cos($a) - cos($b)*sin($a)*sin($g);
  my $m23 = sin($g)*sin($b);
  my $m31 = sin($b)*cos($a);
  my $m32 = sin($b)*sin($a);
  my $m33 = cos($b);

  my @m = ($m11, $m12, $m13, $m21, $m22, $m23, $m31, $m32, $m33);
  return @m;
}

sub rotation_mat_x {
  my @a = @_;
  my (@b,@c,@d,$i,$j);

  $j = 0;
  @b = @a[0..8];
  @c = @a[9..17];

  for ($i = 0; $i < 9; $i++) {
    if ((abs($b[$i]) < 1.0e-8) && (abs($c[$i]) < 1.0e-8)) { $d[$i] = 1.00;
    } elsif (abs($b[$i]) <= abs($c[$i])) { $d[$i] = $b[$i]/$c[$i];
    } elsif (abs($b[$i]) > abs($c[$i])) { $d[$i] = $c[$i]/$b[$i];
    }
  }

  for ($i = 0; $i < 9; $i++) {
    if (abs($d[$i]) < 0.95 || !($d[$i])) {
      $j = 1;
      return ($j);
    } elsif (abs($d[$i]) > 0.95) { ; }
  }
  $j = 2;
  return ($j);
}

sub ones {
  my @a = @_;
  my $a = 0;
  if ($a[0] < -1.0e-12) {
    $a = -1.000;
  } else { $a = 1.000; }
}

```

```

        return ($a);
    }

    sub betacalc {
        my (@a) = @_;
        my ($i);

        if (($a[1] > $a[3]) || ($a[2] < $a[3])) { $i = -1*acos($a[0])*180/pi;
        } elsif ($a[8] > $a[4]) { $i = acos($a[4])*180/pi;
        } else { $i = acos($a[0])*180/pi;
        } return ($i);
    }

    sub alphgamcalc {
        my (@a) = @_;
        my ($i);

        if ((abs($a[6]) < $a[7]) && ($a[10] eq "alpha")) {
            $i = asin(($a[8]-$a[9])/2)*180/pi; #gives sum of alpha and gamma
        } elsif ((abs($a[6]) < $a[7]) && ($a[10] eq "alternate")) {
            $i = acos(($a[11]+$a[12])/2)*180/pi; #gives sum of alpha and gamma
        } elsif ((abs($a[6]) < $a[7]) && ($a[10] eq "gamma")) {
            $i = 0.000; #gamma when beta is zero
        } elsif ((abs($a[1]/$a[0]) < $a[3]) && (abs($a[0]) > $a[4])) {
            $i = atan($a[1]/$a[0])*180/pi;
        } elsif ((abs($a[0]/$a[1]) < $a[3]) && (abs($a[1]) > $a[4])) {
            $i = acot($a[0]/$a[1])*180/pi;
        } elsif ((abs($a[1]/$a[0]) < $a[3]) && (abs($a[0]) < $a[4])) {
            $i = atan($a[5])*180/pi;
        } elsif ((abs($a[0]/$a[1]) < $a[3]) && (abs($a[1]) < $a[4])) {
            $i = acot($a[5])*180/pi;
        } else {
            $i = atan($a[5])*180/pi;
        } return ($i);
    }

    #Calculates the cross product of two vectors
    sub cross_prod {
        my (@a) = @_;
        my (@b);
        $b[0] = ($a[1]*$a[5] - $a[2]*$a[4]);
        $b[1] = ($a[2]*$a[3] - $a[0]*$a[5]);
        $b[2] = ($a[0]*$a[4] - $a[1]*$a[3]);
        return (@b);
    }

    #normalizes a vector
    sub normalize {
        my (@a) = @_;
        my (@b,$b);
        $b = sqrt($a[0]**2+$a[1]**2+$a[2]**2);
        @b = ($a[0]/$b,$a[1]/$b,$a[2]/$b);
        return (@b);
    }

```

```

#multiplies a matrix by some scalar
sub mat_scalar {
    my ($a,$j,@a) = @_;
    my (@b,$i);
    for($i = 0; $i < $j; $i++) {
        $b[$i] = $a*$a[$i];
    } return (@b);
}

#extracts the scalar from an identity matrix
sub mat_norm {
    my (@a) = @_;
    my $i = ($a[0] + $a[4] + $a[8] )/3;
    return ($i);
}

#creates a 3x3 matrix from two bonds to use as a
#structural representation of the molecule.
sub Struc_MAT {
    my (@vectors) = @_;
    my $v1 = $vectors[0] - $vectors[3];
    my $v2 = $vectors[1] - $vectors[4];
    my $v3 = $vectors[2] - $vectors[5];
    my $w1 = $vectors[6] - $vectors[9];
    my $w2 = $vectors[7] - $vectors[10];
    my $w3 = $vectors[8] - $vectors[11];
    my @v_mat = normalize($v1,$v2,$v3);
    my @w_mat = normalize($w1,$w2,$w3);
    my @x = cross_prod(@v_mat,@w_mat);
    my @y = cross_prod(@x,@v_mat);
    my @x_mat = normalize(@x);
    return (@x,@y,@v_mat);
}

#transposes a matrix
sub transpose_mat {
    my ($a,@a) = @_;
    my (@b,$i,$j);
    for(my $i = 0; $i < $a; $i++) {
        for(my $j = 0; $j < $a; $j++) {
            $b[3*$i+$j] = $a[3*$j+$i];
        }
    }
    return (@b);
}

#multiplies two 3x3 matrices and outputs the product
sub mat_mult {
    my ($x,@a) = @_;
    my (@b,@c,@d,$d,@e,$i,$j,$k);
    @b = @a[0..8];
    @c = @a[9..17];
    for($i = 0; $i < $x; $i++) {
        for($j = 0; $j < $x; $j++) {
            for($k = 0; $k < $x; $k++) {

```

```

        $d[3*$i+$j] += $b[3*$i+$k]*$c[3*$k+$j];
    }
}
} return (@d);
}

#calculates the cofactor matrix of a matrix A
sub cofact_mat {
    my ($k,$l,$n,@a) = @_;
    my (@b,@c,$i,$j,$m,$p);
    for($i = 0; $i < $k; $i++) {
        for($j = 0; $j < $k; $j++) {
            $m = 3*$i+$j-$l*3;
            $p = 3*$j+$i-$n*3;
            if(!($i == $p) && !($j == $m)) {
                push @c, $a[3*$i+$j];
            }else {
                ;
            }
        }
    }
    return (@c);
}

#calculates the determinant of a 2x2 matrix
sub det2x2 {
    my (@a) = @_;
    my $b = $a[0]*$a[3] - $a[1]*$a[2];
    return $b;
}

#calculates an inverse matrix
sub inverse_mat {
    my ($a,@a) = @_;
    my (@b,$b,@c,$c,@d,$d,$e,$i,$j,$k);
    for($i = 0; $i < $a; $i++) {
        for($j = 0; $j < $a; $j++) {
            @b = cofact_mat($a,$i,$j,@a);
            $b = det2x2(@b);
            $b *= (-1)**($i+$j);
            push @c, $b;
        }
    }

    $c = $a[0]*$c[0] - $a[1]*$c[1] + $a[2]*$c[2];
    $e = 1/$c;

    @d = transpose_mat(3,@c);
    @d = mat_scalar($e,9,@d);
    return (@d);
}

#searches an input pdb and outputs the atomic coordinates associated with
#the atoms of interest.
sub pdb_find {
    my ($res_num, $C_atom, $H_atom, @input) = @_;

```

```

my (@line,@output);
ONE:  foreach my $a (@input) {
    my $i = 0;

    @line = split /\s+/, $a;
    my $count = scalar(@line);

    if ($count == 11) {
        $i = 0;
    } elsif ($count == 12) {
        $i = 1;
    } else {
        ;
    }

    if ($line[0] eq "END") {
        last ONE;
    }
    elsif ($line[0] eq ("ATOM" || "HETATM")) {
        TRMAT: {
            if (($res_num  == $line[4+$i]) && ($line[2] =~
m/^( $C_atom|$H_atom)\s?$/)) {
                my @list;
                push @list, ($line[5+$i],$line[6+$i],$line[7+$i]);
                push @output, @list;
                last TRMAT;
            } else {
                ;
            }
        }
    } else {
        next ONE;
    }
} return @output;
}

```

## Appendix B

### newmax\_basics.pm

```
#####  
# newmax_basics package  
# Basic functions for manipulation of PDB files.  
# Created 01/2009 by Maximillian H. Bailor  
# Al-Hashimi Lab, University of Michigan  
#####  
  
package newmax_basics;  
use FileHandle;  
use Storable qw(dclone);  
#use Math::Matrix;  
#use Carp;  
use Math::Trig;  
#use Math::Complex;  
use warnings;  
use strict;  
  
use constant pi      => 4*atan2(1,1);  
use constant kb      => 1.3806504e-23;  
use constant N       => 6.02214179e23;  
use constant Hplank  => 6.626208e-34;  
use constant deg_2_rad => pi/180.0;  
use constant rad_2_deg => 180.0/pi;  
  
require Exporter;  
our @ISA = qw(Exporter);  
our %EXPORT_TAGS = ( 'all' => [ qw(  
  ] ),  
];  
  
our @EXPORT_OK = ( @{ $EXPORT_TAGS{'all'} } );  
our @EXPORT = qw(  
);  
  
our $VERSION = '0.20';  
$VERSION = eval $VERSION;  
use constant pi      => 4*atan2(1,1);  
  
sub new {  
    my $invocant = shift;  
    my $class = ref($invocant) || $invocant;  
    my $self = {
```

```

        #      PDB Descriptors
        atms  => [],
        atmpck => [],
        atmtyp => {},
        chnnmb=> {},
        mdlmb=> {},
        resnmb => {},
        restyp => {},
        #      ??? Descriptors
        mmdl  => 0,
        pdb => "",
        eul => [0,0,0],
        rot => [1,0,0,0,1,0,0,1],
        trans => [0,0,0],
        @_
    };
    bless $self, $class;
    $self->PDBread($self->{pdb},$self->{mmdl}) if $self->{pdb};
    return $self;
}

sub copy {
    my $oldmodel = shift;
    my $class = ref $oldmodel;
    my $newmodel = bless {%{$oldmodel}}, $class;
    return dclone($newmodel);
}

sub copypdb {
    my ($self,$reslist,$atmtyp,$seltyp) = @_;
    $reslist = _allres($self) if $reslist eq "all";

    $self->selectatms($atmtyp,$reslist) if $seltyp;
    $self->selectatms2($atmtyp,$reslist) if !$seltyp;

    my $ref = $self->copy;
    $self->atmclnup('atmpck');
    $ref->atmclnupall;
    return dclone($ref);
}

sub copypdbmdl {
    my ($self,$mdl) = @_;
    my $new = $self->selectmdl($mdl);
    return dclone($new);
}

sub atmclnup {
    my ($self,$set) = @_;
    delete $self->{$set};
}

sub atmclnupall {
    my $self = shift;
    my $tmp = $self->{atmpck};

```



```

my $all = ['atms','atmtyp','chnnmb','mdlmb','resnmb','restyp','atmpck'];
foreach my $i (@$all) { delete $self->{$i}; }
$self->{atms} = $tmp;
}

##      Object Accessor Methods
sub addatms {
    my ($self,$data) = @_;
    push @{$self->{atms}}, $data;
}

sub addatmtyp {
    my ($self,$atm,$data) = @_;
    $atm=~s/\s+//g;
    if(!exists $self->{atmtyp}->{$atm}) { $self->{atmtyp}->{$atm} = []; }
    push @{$self->{atmtyp}->{$atm}}, $data;
}

sub addchain {
    my ($self,$chn,$data) = @_;
    $chn=~s/\s+//g;
    if(!exists $self->{chnnmb}->{$chn}) { $self->{chnnmb}->{$chn} = []; }
    push @{$self->{chnnmb}->{$chn}}, $data;
}

sub addhlx {
    my ($self,$hlx,$data) = @_;
    if(!exists $self->{hlxnmb}->{$hlx}) { $self->{hlxnmb}->{$hlx} = []; }
    push @{$self->{hlxnmb}->{$hlx}}, $data;
}

sub addmdl {
    my ($self,$mdl,$data) = @_;
    if(!exists $self->{mdlmb}->{$mdl}) { $self->{mdlmb}->{$mdl} = []; }
    push @{$self->{mdlmb}->{$mdl}}, $data;
}

sub addresnmb {
    my ($self,$res,$data) = @_;
    $res=~s/\s{1,}//g;
    if(!exists $self->{resnmb}->{$res}) { $self->{resnmb}->{$res} = []; }
    push @{$self->{resnmb}->{$res}}, $data;
}

sub addrestyp {
    my ($self,$res,$data) = @_;
    $res=~s/\s{1,}//g;
    if(!exists $self->{restyp}->{$res}) { $self->{restyp}->{$res} = []; }
    push @{$self->{restyp}->{$res}}, $data;
}

sub chngatmtyp {
    my ($self,$atmo,$atmn) = @_;
    my $oldatms = $self->gettype('atmtyp',$atmo);
    foreach my $i (@$oldatms) { $i->{atmtyp} = $atmn; }
}

```

```

}

sub chngeul {
    my ($self,$eul) = @_;
    $self->{eul} = $eul;
    $self->_eul2rot($eul);
}

sub chngrot {
    my ($self,$rot) = @_;
    $self->{rot} = $rot;
    $self->{eul} = _rot2eul($rot);
}

sub chngtrans {
    my ($self,$trans) = @_;
    $self->{trans} = $trans;
}

sub getcoord {
    my ($self,$chn,$res,$atm) = @_;
    my ($coord,$out) = ($self->gettype('atms'),[]);
    foreach my $i (@$coord) {
        next unless $i->{atmtyp} =~ /$atm/i;
        next unless $i->{resnmb} =~ /$res/i;
        next unless $i->{chnltr} =~ /$chn/i;
        if($i->{atmtyp} =~ /^N\d$/i && $atm eq 'N') {
            next unless ($i->{atmtyp} =~ /9/ && $i->{restyp} =~ /[ag]/i) || ($i->{atmtyp} =~ /1/ && $i->{restyp} =~ /[cu]/i);
        }
        elsif($i->{atmtyp} =~ /^C\d$/i && $atm eq 'C') {
            next unless ($i->{atmtyp} =~ /^C4$/i && $i->{restyp} =~ /[ag]/i) || ($i->{atmtyp} =~ /^C2$/i && $i->{restyp} =~ /[cu]/i);
        }
        elsif($i->{atmtyp} =~ /^[cn]\d\D$/i && $atm =~ /^[cn]$/i) { next; }
        push @$out, ($i->{x},$i->{y},$i->{z});
        return $out;
    } return [0,0,0];
}

sub getcoordatms {
    my ($self,$chn,$res,$atm) = @_;
    my ($coord,$out) = ($self->gettype('atmtyp',$atm),[]);
    foreach my $i (@$coord) {
        next unless $i->{resnmb} =~ /$res/i;
        next unless $i->{chnltr} =~ /$chn/i;
        push @$out, ($i->{x},$i->{y},$i->{z});
        return $out;
    } return [0,0,0];
}

sub getcoordmatrix {
    my ($self,$chn,$res) = @_;
    my ($coord,$out) = ($self->gettype('atms'),[]);
    foreach my $i (@$coord) {
        next unless $i->{resnmb} =~ /$res/i;

```

```

        next unless $i->{chnltr}=~/\$chn/i;

        push @$out, ($i->{x},$i->{y},$i->{z});
        return $out;
    } return [0,0,0];
}

sub geteul {
    my $self = shift;
    return $self->{eul};
}

sub getrot {
    my $self = shift;
    return $self->{rot};
}

sub gettrans {
    my $self = shift;
    return $self->{trans};
}

sub gettype {
    my ($self,$type,$ind) = @_ ;
    if(!$type) { return $self->{atms}; }
    elsif($type eq 'atmtyp') { return $self->{atmtyp}->{$ind} if $self->{atmtyp}->{$ind}; }
    elsif($type eq 'atmpck') { return $self->{atmpck} if $self->{atmpck}; }
    elsif($type eq 'chnnmb') { return $self->{chnnmb}->{$ind} if $self->{chnnmb}->{$ind}; }
    elsif($type eq 'hlxnmb') { return $self->{hlxnmb}->{$ind} if $self->{hlxnmb}->{$ind}; }
    elsif($type eq 'mdl nmb') { return $self->{mdl nmb}->{$ind} if $self->{mdl nmb}->{$ind}; }
    elsif($type eq 'resnmb') { return $self->{resnmb}->{$ind} if $self->{resnmb}->{$ind}; }
    elsif($type eq 'restyp') { return $self->{restyp}->{$ind} if $self->{restyp}->{$ind}; }
    elsif($type eq 'sngatm') { return shift @{$self->{atms}} if $self->{atms}; }
    elsif($type eq 'atms') { return $self->{atms} if $self->{atms}; }
    else { return 0; }
}

sub openpdbout {
    my ($self,$data,$append) = @_ ;
    my $fh;
    if($data && !$append) { $fh = new FileHandle "> $data"; }
    elsif($data && $append) { $fh = new FileHandle ">> $data"; }
    else { $fh = new FileHandle ">&\*STDOUT" or die "could not open STDOUT: $!\n"; }
    return $fh;
}

sub openpdb {
    my ($self,$data) = @_ ;
    die "No file to open has been given: $!" unless $data;
    $self->{pdbin} = $data;
    my $fh = new FileHandle $self->{pdbin};
    die "Could not find file: $!" unless $fh;
    return $fh;
}

```

```

sub selectatms {
  my ($self,$type,@reslist) = @_;
  my ($mr,$ma) = ([],[]);
  my (@res,%ma);
  $mr = _reslist(@reslist);
  $ma = _getatmset($type);

  # organize and remove duplicate atoms
  @$ma = sort { $a cmp $b } @$ma;
  %ma = map { $_ => undef } @$ma;
  @res = sort { $a <=> $b } keys %$mr;

  foreach my $r (@res) {
    my ($tmp,$chn);
    push @$tmp, @{$self->gettype('resnmb',$r)} unless !$self->gettype('resnmb',$r);
    $chn = $mr->{$r};
    foreach my $ln (@$tmp) {
      next unless $ln->{chnltr} eq $chn;
      next unless exists $ma{$ln->{atmtyp}};
      push @{$self->{atmpck}}, $ln;
    }
  }
}

sub selectatms2 {
  my ($self,$type,@reslist) = @_;
  my ($mr,$ma,$reslist) = ([],[],'');

  $reslist = join ', ', @reslist;
  @reslist = split /,/, $reslist;
  foreach my $res (@reslist) { $self->selectatms($type,$res); }
}

sub selectmdl {
  my ($self,$mdl) = @_;
  my $new = newmax_basics->new();
  my $mdlslct = [ @{$self->gettype('mdlmb',$mdl)}];

  return 0 if $$mdlslct==-1;
  foreach my $i (@$mdlslct) {
    my $ln = {};

    $ln->{head}=$i->{head};
    $ln->{atmnmb}=$i->{atmnmb};
    $ln->{atmtyp}=$i->{atmtyp};
    $ln->{restyp}=$i->{restyp};
    $ln->{chnltr}=$i->{chnltr};
    $ln->{resnmb}=$i->{resnmb};
    $ln->{x}=$i->{x};
    $ln->{y}=$i->{y};
    $ln->{z}=$i->{z};
    $ln->{aux1}=$i->{aux1};
    $ln->{aux2}=$i->{aux2};
    $ln->{seg}=$i->{seg};
  }
}

```

```

        $new->addatms($ln);
        $new->addatmtyp($ln->{atmtyp},$ln);
        $new->addchain($ln->{chnltr},$ln);
        $new->addresnmb($ln->{resnmb},$ln);
        $new->addrestyp($ln->{restyp},$ln);
    } return $new;
}

##      Utility Accessor Methods
sub PDBavecoord {
    my $self = shift;
    my ($ave,$cnt) = ([0.000,0.000,0.000],0);

    my $atms = $self->gettype('atms');
    foreach my $i (@$atms) {
        #      printf "%s %s %s\n",$i->{x},$i->{y},$i->{z};
        $ave->[0] += $i->{x};
        $ave->[1] += $i->{y};
        $ave->[2] += $i->{z};
        $cnt++;
    }
    die "No atoms!: $" unless $cnt;
    foreach my $a (@$ave) { $a /= $cnt; }
    return $ave;
}

sub PDBconcatenate {
    my $self = shift;
    my $othr = shift;
    return 0;
}

sub PDBdist {
    my ($ref,$cmp) = @_;

    my $tmp1 = shift @{$ref->gettype};
    my $tmp2 = shift @{$cmp->gettype};
    my $out = 0.0;

    foreach my $i ('x','y','z'){ $out += ($tmp1->{$i}-$tmp2->{$i})**2; }
    return sqrt($out);
}

sub PDBgetatm {
    my ($self,$res,$atm) = @_;
    my $tmpr = $self->gettype('resnmb',$res);
    print $tmpr,"\n";
    foreach my $i (@$tmpr) { if($i->{atmtyp} =~ /$atm/gi) { return $i; } }
    return 0;
}

sub PDBgetheader {
    my $self = shift;
    my $filenm = shift;
    my ($Hd,$fh);

```

```

my $header = {};

$fh = $self->openpdb($filenm);
PDB: while(<$fh>) {
    my ($ln,$k) = ({});

    ($Hd=substr($_,0,6))=~s/ //g;
    last PDB if ($Hd=~ /END/);
    next if ($Hd eq "ATOM" || $Hd eq "HETATM" || $Hd eq "TER");
    chomp $_;
    if(exists $header->{$Hd}) {
        push @{$header->{$Hd}}, $_;
    } else {
        $header->{$Hd} = [];
        push @{$header->{$Hd}}, $_;
    }
} close $fh;
}

sub PDBread {
my ($self,$filenm,$multmdl) = @_ ;
my ($Hd,$MDL,$Atmnm,$Rstp,$Sg);
my $fh = $self->openpdb($filenm);
PDB: while(<$fh>) {
    my $ln = {};

    ($Hd=substr($_,0,6))=~s/ //g;
    last PDB if ($Hd=~ /END/) && !$multmdl;
    last PDB if ($Hd=~ /END$/) && $multmdl;
    if($Hd=~ /MODEL/) { $MDL=substr($_,6)+0; next; }
    $MDL=1 unless $MDL;
    next unless ($Hd eq "ATOM" || $Hd eq "HETATM");
    ($Atmnm=substr($_,12,4))=~s/\s+//g;
    ($Rstp=substr($_,16,4))=~s/\s+//g;
    ($Sg=substr($_,72))=~s/[ \n]//g;

    $ln->{head}=$Hd;
    $ln->{atmnm} = substr($_,6,5)+0;
    $ln->{atmtyp}=$Atmnm=~s/\s+//g;
    $ln->{restyp}=$Rstp;
    $ln->{chnltr} = substr($_,21,1) =~s/\s+/_//g;
    $ln->{resnmb} = substr($_,22,5);
    $ln->{x} = substr($_,28,10)+0.0;
    $ln->{y} = substr($_,38,8)+0.0;
    $ln->{z} = substr($_,46,8)+0.0;
    $ln->{aux1} = substr($_,54,6)+0.0;
    $ln->{aux2} = substr($_,60,6)+0.0;
    $ln->{seg}=$Sg;

    $self->addatms($ln);
    $self->addatmtyp($ln->{atmtyp},$ln);
    $self->addchain($ln->{chnltr},$ln);
    $self->addmdl($MDL,$ln);
    $self->addresnmb($ln->{resnmb},$ln);
    $self->addrestyp($ln->{restyp},$ln);
}
}

```

```

    }
    ## corrections to atom types
    chngatmtyp($self,"OP1","O1P");
    chngatmtyp($self,"OP2","O2P");
    close $fh;
}

sub PDBrmsd {
    my ($self,$other) = @_;
    my ($k,$l,$rmsd,$cnt,$tmpr,$tmpcA,$tmpcB) = (0,0,0.0,0,[],[],[]);
    foreach (@{$self->{atms}}) { push @$tmpcA, ${$_}{chnltr}; }
    foreach (@{$other->{atms}}) { push @$tmpcB, ${$_}{chnltr}; }
    foreach my $a (@{$self->{atms}}) { push @$tmpr,($a->{resnmb})-${other->{atms}}[$l-
->{resnmb}]; $l++; }

    for my $i (@{$self->gettype('atms')}) {
        for my $j (@{$other->gettype('atms')}) {
            if(($i->{atmtyp} eq $j->{atmtyp}) && ($tmpr->[$k] == ($i->{resnmb})-$j-
->{resnmb}))) {
                next unless (($i->{chnltr} eq $tmpcA->[$k]) && ( $j->{chnltr} eq
$tmpcB->[$k]));
                $rmsd += ($i->{x}-$j->{x})**2+($i->{y}-$j->{y})**2+($i->{z}-$j-
->{z})**2;
                $cnt++; $k++;
            }
        }
    } return sqrt($rmsd/($cnt-1));
}

sub PDBrotate {
    my $self = shift;
    my $rot = $self->getrot;
    foreach my $i (@{$self->gettype('atms')}) {
        my $tmp = [0.00,0.00,0.00];
        $tmp->[0] = $i->{x}$rot->[0]+$i->{y}$rot->[1]+$i->{z}$rot->[2];
        $tmp->[1] = $i->{x}$rot->[3]+$i->{y}$rot->[4]+$i->{z}$rot->[5];
        $tmp->[2] = $i->{x}$rot->[6]+$i->{y}$rot->[7]+$i->{z}$rot->[8];
        $i->{x} = $tmp->[0];
        $i->{y} = $tmp->[1];
        $i->{z} = $tmp->[2];
    }
}

sub PDBtorsionang {
    my ($self,$type) = @_;
    my ($a,$b,$c,$d);
    my ($aa,$atms) = ([],[ 'P','O5','C5','C4','C3','O3']);
    if($type=~/^alpha/) {
        foreach my $atm (@$atms[-1..2]) { push @$aa, $self->{atmtyp}->{$atm}; }
        $a = $self->{atmtyp}->{O3}; $b = $self->{atmtyp}->{P}; $c = $self->{atmtyp}->{O5};
        $d = $self->{atmtyp}->{C5};
        shift @{$a}; shift @{$b};
    } elsif($type=~/^beta/) {
        foreach my $atm (@$atms[0..3]) { push @$aa, $self->{atmtyp}->{$atm}; }
    }
}

```

```

    $a = $self->{atmtyp}->{P}; $b = $self->{atmtyp}->{O5}; $c = $self->{atmtyp}->{C5};
$d = $self->{atmtyp}->{C4};
    shift @{$b};
    } elsif($type=~/^gamma/) {
        foreach my $atm (@$atms[1..4]) { push @$aa, $self->{atmtyp}->{$atm}; }
        $a = $self->{atmtyp}->{O5}; $b = $self->{atmtyp}->{C5}; $c = $self->{atmtyp}->{C4};
$d = $self->{atmtyp}->{C3};
    } elsif($type=~/^delta/) {
        foreach my $atm (@$atms[2..5]) { push @$aa, $self->{atmtyp}->{$atm}; }
        $a = $self->{atmtyp}->{C5}; $b = $self->{atmtyp}->{C4}; $c = $self->{atmtyp}->{C3};
$d = $self->{atmtyp}->{O3};
    } elsif($type=~/^epsilon/) {
        foreach my $atm (@$atms[3..5,0]) { push @$aa, $self->{atmtyp}->{$atm}; }
        $a = $self->{atmtyp}->{C4}; $b = $self->{atmtyp}->{C3}; $c = $self->{atmtyp}->{O3};
$d = $self->{atmtyp}->{P};
        shift @{$a}; shift @{$b}; shift @{$c};
    } elsif($type=~/^zeta/) {
        foreach my $atm (@$atms[4..5,0..1]) { push @$aa, $self->{atmtyp}->{$atm}; }
        $a = $self->{atmtyp}->{C3}; $b = $self->{atmtyp}->{O3}; $c = $self->{atmtyp}->{P};
$d = $self->{atmtyp}->{O5};
        shift @{$a}; shift @{$b};
    } elsif($type=~/^chi/) {
        $a = $self->{atmtyp}->{O4}; $b = $self->{atmtyp}->{C1}; $c = $self->{atmtyp}-
>{N19}; $d = $self->{atmtyp}->{C24};
        } return _torang2($a,$b,$c,$d);
}

```

```

sub PDBtorsionang2 {
    my ($self,$res,$type) = @_;
    my ($a,$b,$c,$d);
    my ($resa,$resb,$resc,$resd,$chn);
    my ($atm,$bkbn) = ([,['P','O5\','C5\','C4\','C3\','O3\']]);

    $chn = $res;
    $res =~s/\w\.\d+/$1/;
    $chn =~s/(\w)\.\d+/$1/;
    $resa = $resb = $resc = $resd = $res;

    if($type=~alpha/i) {
        $resa--;
        foreach my $z (@$bkbn[-1..2]) { push @$atm, $z; }
    } elsif($type=~beta/i) {
        foreach my $z (@$bkbn[0..3]) { push @$atm, $z; }
    } elsif($type=~epsilon/i) {
        $resd++;
        foreach my $z (@$bkbn[3..5,0]) { push @$atm, $z; }
    } elsif($type=~zeta/i) {
        $resc++; $resd++;
        foreach my $z (@$bkbn[4..5,0..1]) { push @$atm, $z; }
    } elsif($type=~chi/i) {
        $a = getcoord($self,$chn,$res,'O4\');
        $b = getcoord($self,$chn,$res,'C1\');
        $c = getcoord($self,$chn,$res,'N');
        $d = getcoord($self,$chn,$res,'C');
    }
}

```



```

        return _torang($a,$b,$c,$d);
    } elseif($type=~~/gamma/i) {
        foreach my $z (@$bkbn[1..4]) { push @$atm, $z; }
    } elseif($type=~~/delta/i) {
        foreach my $z (@$bkbn[2..5]) { push @$atm, $z; }
    } elseif($type=~~/^eta$/i) {
        $resa--; $resd++;
        push @$atm, ('C4\'','P','C4\'','P');
    } elseif($type=~~/theta/i) {
        $resc++; $resd++;
        push @$atm, ('P','C4\'','P','C4\'');
    } else {
        die "I don't know that torsion angle\n";
    }
}

$a = getcoord($self,$chn,$resa,$atm->[0]);
$b = getcoord($self,$chn,$resb,$atm->[1]);
$c = getcoord($self,$chn,$resc,$atm->[2]);
$d = getcoord($self,$chn,$resd,$atm->[3]);

return _torang($a,$b,$c,$d);
}

sub PDBtorsionang3 {
    my ($self,$res,$taset) = @_ ;
    my ($atm,$bkbn,$outset) = ([],[ 'P','O5\'','C5\'','C4\'','C3\'','O3\''],[]);

    my $chn = $res;
    $res =~s/\w\.\d+/$1/;
    $chn =~s/(\w)\.\d+/$1/;
TA:  foreach my $type (@$taset) {
        my ($a,$b,$c,$d);
        my ($resa,$resb,$resc,$resd);
        $resa = $resb = $resc = $resd = $res;

        if($type=~~/alpha/i) {
            $resa--;
            foreach my $z (@$bkbn[-1..2]) { push @$atm, $z; }
        } elseif($type=~~/beta/i) {
            foreach my $z (@$bkbn[0..3]) { push @$atm, $z; }
        } elseif($type=~~/epsilon/i) {
            $resd++;
            foreach my $z (@$bkbn[3..5,0]) { push @$atm, $z; }
        } elseif($type=~~/zeta/i) {
            $resc++; $resd++;
            foreach my $z (@$bkbn[4..5,0..1]) { push @$atm, $z; }
        } elseif ($type=~~/chi/i) {
            $a = getcoord($self,$chn,$res,'O4\'');
            $b = getcoord($self,$chn,$res,'C1\'');
            $c = getcoord($self,$chn,$res,'N');
            $d = getcoord($self,$chn,$res,'C');
            push @$outset, _torang($a,$b,$c,$d);
            next TA;
        } elseif($type=~~/gamma/i) {
            foreach my $z (@$bkbn[1..4]) { push @$atm, $z; }
        }
    }
}

```

```

    } elsif($type=~~/delta/i) {
        foreach my $z (@$bkbn[2..5]) { push @$atm, $z; }
    } elsif($type=~~/^eta$/i) {
        $resa--; $resd++;
        push @$atm, ('C4\'','P','C4\'','P');
    } elsif($type=~~/theta/i) {
        $resc++; $resd++;
        push @$atm, ('P','C4\'','P','C4\'');
    } else {
        die "I don't know that torsion angle\n";
    }

    $a = getcoord($self,$chn,$resa,$atm->[0]);
    $b = getcoord($self,$chn,$resb,$atm->[1]);
    $c = getcoord($self,$chn,$resc,$atm->[2]);
    $d = getcoord($self,$chn,$resd,$atm->[3]);
    push @$outset, _torang($a,$b,$c,$d);
}
return $outset;
}

sub PDBsugtorang {
    my ($self,$res,$type) = @_;
    my ($a,$b,$c,$d);
    my ($P,$v) = (0,[]);
    my ($chn,$aset) = ($res,['C4\'','O4\'','C1\'','C2\'','C3\'']);
    $res =~ s/\w\.\d+/$1/;
    $chn =~ s/(\w)\.\d+/$1/;

    foreach my $nu (0..4) {
        my $aslct = [];
        if($type=~~/^nu0/ || ($type=~~/^P|numax/ && $nu == 0)){ foreach my $i
(@$aset[0..3]) { push @$aslct, $i; } }
        elsif($type=~~/^nu1/ || ($type=~~/^P|numax/ && $nu == 1)){ foreach my $i
(@$aset[1..4]) { push @$aslct, $i; } }
        elsif($type=~~/^nu2/ || ($type=~~/^P|numax/ && $nu == 2)){ foreach my $i
(@$aset[2..4,0]) { push @$aslct, $i; } }
        elsif($type=~~/^nu3/ || ($type=~~/^P|numax/ && $nu == 3)){ foreach my $i
(@$aset[3..4,0..1]) { push @$aslct, $i; } }
        elsif($type=~~/^nu4/ || ($type=~~/^P|numax/ && $nu == 4)){ foreach my $i
(@$aset[4,0..2]) { push @$aslct, $i; } }
        $a = getcoord($self,$chn,$res,$aslct->[0]);
        $b = getcoord($self,$chn,$res,$aslct->[1]);
        $c = getcoord($self,$chn,$res,$aslct->[2]);
        $d = getcoord($self,$chn,$res,$aslct->[3]);
        my $ta = _torang($a,$b,$c,$d);
        return $ta unless $type=~~/^P|numax/;
        push @$v,$ta;
    }
    $P = atan2((($v->[4]+$v->[1]-$v->[3]-$v->[0]),(2.0*$v->[2]*
(sin(0.628319)+sin(1.256637))));
    return ($P*rad_2_deg) if $type=~~/P/;
    return ($v->[2]/cos($P)) if $type=~~/numax/;
}

```

```

sub PDBtranslate {
    my $self = shift;
    my $trans = shift;
    $self->chgtrans($trans) if $trans;
    my $coord = $self->gettrans;
    foreach my $i (@{$self->gettype('atms')}) {
        $i->{x} -= $coord->[0];
        $i->{y} -= $coord->[1];
        $i->{z} -= $coord->[2];
    }
}

sub PDBwrite {
    my ($self,%optns) = @_;
    my $type = $optns{seltyp} || "";
    my $ind = $optns{pck} || "";
    my $filem = $optns{out} || "";

    my ($i,$atms,$fh) = (0,$self->gettype($type,$ind),$self->openpdbout($filem,$optns{app}));
    die "NO atoms to write\n" if !$atms;
    if(exists $optns{mdl} && exists $optns{ter} && $optns{ter}==1) { print $fh "MODEL
$optns{mdl}\n"; }
    elsif(exists $optns{mdl} && !exists $optns{ter}) { print $fh "MODEL $optns{mdl}\n"; }

    foreach my $j (@{$atms}) { _pdpline($j,$i,$fh); $i++; }
    if($i && exists $optns{ter} && $optns{ter} == 1 && !exists $optns{end}) { print $fh "TER\n"; }

    elsif($i && exists $optns{mdl} && !(exists $optns{end})) { print $fh "ENDMDL\n"; }
    elsif($i && exists $optns{mdl} && exists $optns{end}) { print $fh "ENDMDL\nEND\n"; }
    elsif(exists $optns{end} && !exists $optns{mdl}) { print $fh "END\n"; }
    elsif($i && exists $optns{end} && $optns{end}==1) { print $fh "END\n"; }
    close $fh;
}

#####
sub _allres {
    my $self = shift;
    my @res;
    foreach my $i (sort {$a <=> $b} keys %{$self->{resnmb}}) { push @res, $i; }
    my $out = sprintf "%s.%s-%s",${$self->{resnmb}}->{$res[0]}}[0]-
>{chnltr},$res[0],$res[$#res];
    return $out;
}

sub _atmset {
    my ($atmtyp,$atmnmb,$mod) = @_;
    my ($c,$d) = ([],[]);

    if(!$atmnmb) { return $atmtyp; }
    foreach my $a (@$atmtyp) {
        push @$c, map { $a.$_ } @$atmnmb;
        if($mod) { push @$d, map { $_.$mod } @$c; }
        else { push @$d, @$c; }
    } return $d;
}

```

```

sub _dotprod {
    my ($a,$b,$vecl) = @_ ;

    my ($tmp,$maga,$magb,$i) = (0,0,0,0);
    $maga = _vecnorm($a);
    $magb = _vecnorm($b);
    while($i < $vecl) { $tmp += $maga->[$i]*$magb->[$i]; $i++; }
    return ($tmp);
}

sub _eul2rot {
    my $self = shift;
    my $eul = shift;
    foreach (@{$eul}) { $_ *= (pi/180.0); }
    my $r = [];

    $r->[0] = -sin($eul->[2])*sin($eul->[0]) + cos($eul->[1])*cos($eul->[0])*cos($eul->[2]);
    $r->[1] = sin($eul->[2])*cos($eul->[0]) + cos($eul->[1])*sin($eul->[0])*cos($eul->[2]);
    $r->[2] = -cos($eul->[2])*sin($eul->[1]);
    $r->[3] = -cos($eul->[2])*sin($eul->[0]) - cos($eul->[1])*cos($eul->[0])*sin($eul->[2]);
    $r->[4] = cos($eul->[2])*cos($eul->[0]) - cos($eul->[1])*sin($eul->[0])*sin($eul->[2]);
    $r->[5] = sin($eul->[2])*sin($eul->[1]);
    $r->[6] = sin($eul->[1])*cos($eul->[0]);
    $r->[7] = sin($eul->[1])*sin($eul->[0]);
    $r->[8] = cos($eul->[1]);

    foreach (@{$eul}) { $_ *= (180.0/pi); }
    $self->chngrrot($r);
}

sub _getatmset {
    my $type = shift;
    my $ma = [];
    if(!$type || $type eq "heavy") {
        $ma = _atmset(['C','N','O'],[1..9]);
        push @$ma, @_{_atmset(['P','O1P','O2P','OP1','OP2'])};
        push @$ma, @_{_atmset(['C','N','O'],[1..9],'\')};
    }
    elsif($type eq "all") {
        $ma = _atmset(['C','H','N','O'],[1..9]);
        push
@_{_atmset(['P','O1P','O2P','OP1','OP2','1H5\'','2H5\'','HO2\'','1H2','2H2','1H6','2H6'])}; @ $ma,
        push @$ma, @_{_atmset(['C','H','N','O'],[1..9],'\')};
        push @$ma, @_{_atmset(['1H','2H'],[2,4,5])};
    }
    elsif($type eq "base") { $ma = _atmset(['C','N','O'],[1..9]); }
    elsif($type eq "bkbn") {
        $ma = _atmset(['C','O'],[1..5],'\');
        push @$ma, @_{_atmset(['P'])};
    }
    elsif($type eq "phos") { $ma = _atmset(['P']); }
    elsif($type eq "sug") { $ma = _atmset(['C','O'],[1..5],'\'); }
    elsif($type eq "sugC") { $ma = _atmset(['C'],[1..5],'\'); }
    elsif($type eq "sugO") { $ma = _atmset(['O'],[1..5],'\'); }
}

```

```

elseif($type eq "vr8") { $ma = _atmset(['C','H'],[1,2],''); }
elseif($type eq "wcbp") { $ma = _atmset(['N'],[1,3]); }
elseif($type eq "coll") {
    $ma = _atmset(['C','O'],[1..9]);
    push @$ma, @[_atmset(['O1P','O2P','OP1','OP2'])];
    push @$ma, @[_atmset(['C','O'],[1..9],'')];
}
elseif($type eq "xx") { $ma = _atmset(['C4\','P','C3\']); }
else {
    $type=~s/\ *//g;
    $ma = _atmset([$type]);
} return $ma;
}

sub _pdpline {
    my ($pk,$atms,$fh) = @_;

    my $ln = sprintf("%s", $pk->{head});
    if(length($pk->{head}) == 4) {
        $pk->{atmnmb} = sprintf "%4s",$atms+1;
        $ln = sprintf("%s %6s ", $ln, $pk->{atmnmb});
    } else {
        $pk->{atmnmb} = sprintf "%5s",$atms+1;
        $ln = sprintf("%s%5s ", $ln, $pk->{atmnmb});
    }

    my $atm = $pk->{atmtyp};
    if(length($atm) == 1) {
        $ln = sprintf("%11s %-3s",$ln,$atm);
    } elseif(length($atm) == 2) {
        $ln = sprintf("%11s %-3s",$ln,$atm);
    } elseif(length($atm) == 3) {
        $ln = sprintf("%11s %-3s",$ln,$atm);
    } elseif(length($atm) == 4) {
        $ln = sprintf("%11s%-3s",$ln,$atm);
    }
    $ln = sprintf("%16s %-4s%1s%4s  %8.3f%8.3f%8.3f %5.2f %5.2f      %s\n",$ln,$pk->{restyp},
        $pk->{chnltr},$pk->{resnmb},$pk->{x},$pk->{y},$pk->{z},$pk->{aux1},$pk->{aux2},$pk->{seg});
    $ln=~s/_/ /;
    print $fh $ln;
}

sub _reslist {
    my @reslist = @_;
    my $res = {};
    foreach my $resln (@reslist) {
        my ($num,$oth);
        while($resln=~/*(\w{1})\.(\d+)(^\d+)**/g) { $res->{$2} = $1; }
        while($resln=~/(\w{1})\.(\d+)-(\d+)/g) { foreach my $r ($2..$3) { $res->{$r} = $1; } }
    }
    while($resln=~/(\w{1})\.(\d+)-\w{1}\.(\d+)/g) { foreach my $r ($2..$3) { $res->{$r} = $1; } }
    } return ($res);
}

```

```

}

sub _rot2eul {
    my $rot = shift;

    my $eul = [];
    $eul->[0] = (180.0/pi)*atan2($rot->[7],$rot->[6]);
    $eul->[2] = (180.0/pi)*atan2($rot->[5],$rot->[2]*-1.0);

    my $tmp = sin($eul->[2]*(pi/180.0))?($rot->[5]/sin($eul->[2]*(pi/180.0))):0;
    $eul->[1] = (180.0/pi)*atan2($tmp,$rot->[8]);
    return $eul;
}

sub _torang {
    my ($p1,$p2,$p3,$p4) = @_;
    if(!$p1 || !$p2 || !$p3 || !$p4) { return 0.00; }

    my ($a,$b,$c) = ([],[],[]);
    my $i = 0; while($i < 3) { $a->[$i] = $p2->[$i] - $p1->[$i]; $i++; }
    $i = 0; while($i < 3) { $b->[$i] = $p3->[$i] - $p2->[$i]; $i++; }
    $i = 0; while($i < 3) { $c->[$i] = $p4->[$i] - $p3->[$i]; $i++; }

    my $nab = _xprod($a,$b);
    my $nbc = _xprod($b,$c);
    my $cos = _dotprod($nab,$nbc,3);
    my $sin = sqrt(1.0-($cos**2));
    my $sgn = _dotprod($c,$nab,3);
    $sgn *= 1/sqrt($sgn**2);
    return atan2($sin,$cos)*rad_2_deg*$sgn;
}

sub _torang2 {
    my ($a,$b,$c,$d) = @_;
    my ($out,$cnt) = ([],0);
    while($d->[$cnt]) {
        push @{$out}, _torang($a->[$cnt],$b->[$cnt],$c->[$cnt],$d->[$cnt]);
        $cnt++;
    } return $out;
}

sub _vecnorm {
    my $a = shift;
    my ($out,$mag) = ([],0);
    foreach (@{$a}) {$mag += ($_**2); }
    $mag = sqrt($mag);
    foreach (@{$a}) { push @{$out}, $_/$mag; }
    return $out;
}

sub _vector1 {
    my ($p1,$p2) = @_;
    my ($a,$i) = ([],0);

    while($i < 3) { $a->[$i] = $p2->[$i] - $p1->[$i]; $i++; }
}

```

```

        return vecnorm($a);
    }

    sub _vector2 {
        my ($p1,$p2,$p3) = @_ ;
        my ($a,$b,$i) = ([],[],0);

        while($i < 3) { $a->[$i] = ($p2->[$i] + $p1->[$i])/2.0; $i++; }
        $i = 0; while($i < 3) { $b->[$i] = $p3->[$i] - $a->[$i]; $i++; }
        return vecnorm($b);
    }

    sub _xprod {
        my $a = shift;
        my $b = shift;
        my ($c,$mag,$i) = ([],0,0);

        $c->[0] = ($a->[1]*$b->[2] - $a->[2]*$b->[1]);
        $c->[1] = ($a->[2]*$b->[0] - $a->[0]*$b->[2]);
        $c->[2] = ($a->[0]*$b->[1] - $a->[1]*$b->[0]);
        return _vecnorm($c);
    }

    2;

```

## Appendix C

### newmax\_basics2.pm

```
#####  
# newmax_basics2 package  
# Basic Computational functions for manipulation of PDB files.  
# Created 01/2009 by Maximillian H. Bailor  
# Al-Hashimi Lab, University of Michigan  
#####  
  
package newmax_basics2;  
use newmax_basics;  
  
#use Math::Matrix;  
#use Math::Quaternion;  
use FileHandle;  
use Storable qw(dclone);  
#use Carp;  
use PDL;  
use PDL::Basic;  
use PDL::Ops;  
use PDL::Core;  
use PDL::MatrixOps;  
#use PDL::Slices;  
use PDL::Slatec;  
#use PDL::Ufunc;  
#use PDL::Math;  
use warnings;  
use strict;  
  
use constant pi => 4*atan2(1,1);  
use constant kb => 1.3806504e-23;  
use constant N => 6.02214179e23;  
use constant Hplank => 6.626208e-34;  
use constant deg_2_rad => pi/180.0;  
use constant rad_2_deg => 180.0/pi;  
  
require Exporter;  
our @ISA = qw(Exporter);  
  
our %EXPORT_TAGS = ( 'all' => [ qw(  
)],  
);
```



```

our @EXPORT_OK = ( @{ $EXPORT_TAGS{'all'} } );
our @EXPORT = qw(
);

our $VERSION = '0.20';
$VERSION = eval $VERSION;

use constant N          => 6.02214179e23;
use constant Hplank     => 6.626208e-34;
use constant gH         => 26.7522128e07;
use constant gC         => 6.7282840e07;
use constant gN         => -2.71261804e07;
use constant kbol       => 1.38066e-23;
use constant pi         => 4*atan2(1,1);
use constant mu0        => 4*pi*10**-07;

sub getatmmtrx {
    my ($self,$data,$type) = @_;
    my ($out,$ln) = ([],$self->newmax_basics::gettype($data,$type));
    foreach my $i (@{$ln}) { push @{$out}, [$i->{x},$i->{y},$i->{z}]; }
    return pdl $out;
}

sub PDBfit {
    my ($obj,$ref,$cmp) = @_;

    my $averef = newmax_basics::PDBavecoord($ref);
    my $avecmp = newmax_basics::PDBavecoord($cmp);
    my $aveobj = newmax_basics::PDBavecoord($obj);

    my ($dif,$i) = ([],0);
    while($i < 3) { $dif->[$i] = $avecmp->[$i]-$aveobj->[$i]; $i++; }

    $ref->newmax_basics::chngrans($averef);
    $cmp->newmax_basics::chngrans($avecmp);
    $obj->newmax_basics::chngrans($aveobj);
    $ref->newmax_basics::PDBtranslate;
    $cmp->newmax_basics::PDBtranslate;
    $obj->newmax_basics::PDBtranslate;
    my $rot = _getR($ref,$cmp);

    $cmp->newmax_basics::chngrrot($rot);
    $obj->newmax_basics::chngrrot($rot);
    $cmp->newmax_basics::PDBrotate; # this is here to properly calc rmsd
    $obj->newmax_basics::PDBrotate;

    my $rotcmp = _matxvec($rot,$dif,3,3);
    my $newxyz = _matadd($rotcmp,$averef,sgn=>-1);

    $obj->newmax_basics::chngrans($newxyz);
    $obj->newmax_basics::PDBtranslate;
    return $rot;
}

sub PDBfit2 {

```

```

my ($obj,$ref,$cmp) = @_;

my $averef = newmax_basics::PDBavecoord($ref);
my $avecmp = newmax_basics::PDBavecoord($cmp);
my $aveobj = newmax_basics::PDBavecoord($obj);

my ($dif,$i) = ([],0);
while($i < 3) { $dif->[$i] = $aveobj->[$i]-$aveobj->[$i]; $i++; }

$ref->newmax_basics::chngrans($averef);
$cmp->newmax_basics::chngrans($avecmp);
$obj->newmax_basics::chngrans($avecmp);
$ref->newmax_basics::PDBtranslate;
$cmp->newmax_basics::PDBtranslate;
$obj->newmax_basics::PDBtranslate;
my $rot = _getR($ref,$cmp);

$cmp->newmax_basics::chngrrot($rot);
$obj->newmax_basics::chngrrot($rot);
$cmp->newmax_basics::PDBrotate; # this is here to properly calc rmsd
$obj->newmax_basics::PDBrotate;

my $rotcmp = _matxvec($rot,$dif,3,3);
my $newxyz = _matadd($rotcmp,$averef,sgn=>-1);

$obj->newmax_basics::chngrans($newxyz);
$obj->newmax_basics::PDBtranslate;
return $rot;
}

sub PDBfit3 {
my ($ref,$cmp) = @_;

my $averef = newmax_basics::PDBavecoord($ref);
my $avecmp = newmax_basics::PDBavecoord($cmp);

$ref->newmax_basics::chngrans($averef);
$cmp->newmax_basics::chngrans($avecmp);
$ref->newmax_basics::PDBtranslate;
$cmp->newmax_basics::PDBtranslate;
my $rot = _getR($ref,$cmp);

$cmp->newmax_basics::chngrrot($rot);
$cmp->newmax_basics::PDBrotate; # this is here to properly calc rmsd
return 0;
}

sub PDBfindR {
my $ref = shift;
my $cmp = shift;

my $averef = newmax_basics::PDBavecoord($ref);
my $avecmp = newmax_basics::PDBavecoord($cmp);

$ref->newmax_basics::chngrans($averef);

```

```

$cmp->newmax_basics::chngrans($avecmp);
$ref->newmax_basics::PDBtranslate;
$cmp->newmax_basics::PDBtranslate;
my $rot = _getR($ref,$cmp);
$cmp->newmax_basics::chngrrot($rot);
$cmp->newmax_basics::PDBrotate; # this is here to properly calc rmsd
return $rot;
}

sub _getR {
my $ref = shift;
my $cmp = shift;
my ($M,$m,$ev,$e,$pk,$cnt,$eig,@tmpr,@tmpcA,@tmpcB);
my $cov = [];
## need to take these variables out!
my ($keyr,$keyc);
my $ln1 = $ref->gettype('atms');
my $ln2 = $cmp->gettype('atms');

die "not the same number of atoms!\n" unless ($#{ $ln1 } == $#{ $ln2 });
my ($j,$jj,$newcmp) = (0,0,[]);
foreach (@$ln1) { push @tmpcA, ${$_}{chnltr}; }
foreach (@$ln2) { push @tmpcB, ${$_}{chnltr}; }
while($#{ $ln2 }[$j]) { push @tmpr, ($ln1->[$j]->{resnmb} - $ln2->[$j]->{resnmb}); $j++; }
$j = 0;
for my $i (@$ln1) {
    for my $k (@$ln2) {
        if(($i->{atmty} eq $k->{atmty}) && ($i->{resnmb} == ($k->{resnmb}+$tmpr[$j]))) {
            next unless ($i->{chnltr} eq $tmpcA[$j] && $k->{chnltr} eq
$tmpcB[$j]);

            $newcmp->[$j]->{x} = $k->{x};
            $newcmp->[$j]->{y} = $k->{y};
            $newcmp->[$j]->{z} = $k->{z};
            $j++;
        }
    }
}

for (my $i1 = 0; $i1 < 3; $i1++) {
    $cov->[$i1] = [];
    for (my $i2 = 0; $i2 < 3; $i2++) { $cov->[$i1]->[$i2] = 0.0; }
} $j = 0;

for my $i (@$ln1) {
    $cov->[0]->[0] += $newcmp->[$j]->{x}*$i->{x};
    $cov->[1]->[0] += $newcmp->[$j]->{x}*$i->{y};
    $cov->[2]->[0] += $newcmp->[$j]->{x}*$i->{z};

    $cov->[0]->[1] += $newcmp->[$j]->{y}*$i->{x};
    $cov->[1]->[1] += $newcmp->[$j]->{y}*$i->{y};
    $cov->[2]->[1] += $newcmp->[$j]->{y}*$i->{z};

    $cov->[0]->[2] += $newcmp->[$j]->{z}*$i->{x};
    $cov->[1]->[2] += $newcmp->[$j]->{z}*$i->{y};

```

```

        $cov->[2]->[2] += $newcmp->[$j]->{z}*i->{z};
        $j++;
    }

    $M->[0]->[0] = $cov->[0]->[0]+$cov->[1]->[1]+$cov->[2]->[2];
    $M->[0]->[1] = $cov->[1]->[2]-$cov->[2]->[1];
    $M->[0]->[2] = $cov->[2]->[0]-$cov->[0]->[2];
    $M->[0]->[3] = $cov->[0]->[1]-$cov->[1]->[0];

    $M->[1]->[0] = $M->[0]->[1];
    $M->[1]->[1] = $cov->[0]->[0]-$cov->[1]->[1]-$cov->[2]->[2];
    $M->[1]->[2] = $cov->[0]->[1]+$cov->[1]->[0];
    $M->[1]->[3] = $cov->[2]->[0]+$cov->[0]->[2];

    $M->[2]->[0] = $M->[0]->[2];
    $M->[2]->[1] = $M->[1]->[2];
    $M->[2]->[2] = -1*$cov->[0]->[0]+$cov->[1]->[1]-$cov->[2]->[2];
    $M->[2]->[3] = $cov->[1]->[2]+$cov->[2]->[1];

    $M->[3]->[0] = $M->[0]->[3];
    $M->[3]->[1] = $M->[1]->[3];
    $M->[3]->[2] = $M->[2]->[3];
    $M->[3]->[3] = -1*$cov->[0]->[0]-$cov->[1]->[1]+$cov->[2]->[2];

    $m = pdl $M;
    ($ev,$e) = eigens_sym($m);

    $pk = $cnt = $eig = 0;
    while($pk < 4) {
        if(index($e,$pk) > $eig) {
            $cnt = $pk;
            $eig = index($e,$cnt);
        } $pk++;
    }
    my $vec = $ev->dice([$cnt])->clump(-1);

    my $q = [];
    $q->[0] = index($vec,0);
    $q->[1] = index($vec,1);
    $q->[2] = index($vec,2);
    $q->[3] = index($vec,3);

    my $rot = [];
    $rot->[0] = $q->[0]*$q->[0]+$q->[1]*$q->[1]-$q->[2]*$q->[2]-$q->[3]*$q->[3];
    $rot->[1] = 2.0*($q->[1]*$q->[2]-$q->[0]*$q->[3]);
    $rot->[2] = 2.0*($q->[1]*$q->[3]+$q->[0]*$q->[2]);
    $rot->[3] = 2.0*($q->[2]*$q->[1]+$q->[0]*$q->[3]);
    $rot->[4] = $q->[0]*$q->[0]-$q->[1]*$q->[1]+$q->[2]*$q->[2]-$q->[3]*$q->[3];
    $rot->[5] = 2.0*($q->[2]*$q->[3]-$q->[0]*$q->[1]);
    $rot->[6] = 2.0*($q->[3]*$q->[1]-$q->[0]*$q->[2]);
    $rot->[7] = 2.0*($q->[3]*$q->[2]+$q->[0]*$q->[1]);
    $rot->[8] = $q->[0]*$q->[0]-$q->[1]*$q->[1]-$q->[2]*$q->[2]+$q->[3]*$q->[3];

    my $u = _mattrtranspose($rot);
    return $u;

```

```

}

sub DistCutOff {
    my $ref = shift;
    my $cmp = shift;
    my ($i,$j,$k);
    my ($atmA,$atmB) = ([],[]);

    ## need to take these variables out!
    my ($key1,$key2,$chain);
    my $Hlx01 = getatmmtrx($ref->{$key1}->[$chain]->{atms},'atms');
    my $Hlx02 = getatmmtrx($cmp->{$key2}->[$chain]->{atms},'atms');
    my $cnt = $Hlx01->dim(1);

    while($cnt) {
        my $tmp01 = $Hlx01->dice_axis(1,[$cnt-1])->clump(-1)->copy();
        my $tmp02 = $Hlx02->copy()-$tmp01;
        $tmp02 = sqrt($tmp02**2);
        print min($tmp02->sumover)."\n";
        $cnt--;
    }
    return 0;
}

sub DistCutOff2 {
    my %optns = @_;
    my ($i,$j,$k,$fh,$pts,$HLX);
    my ($atmA,$atmB,$a,$b,$g) = ([],[],0,0,0);

    $fh = new FileHandle "> $optns{fileout}";
    my $hlx = (defined $optns{hlx})?$optns{hlx}:die "Need to input helices for simulation!\n";
    my $dist = (defined $optns{distcutoff})?$optns{distcutoff}:0;
    my $linkdist = (defined $optns{linkcutoff})?$optns{linkcutoff}:999999;
    my $stp = (defined $optns{stepsize})?$optns{stepsize}:0;
    if(defined $optns{link}) { foreach my $a (@{$optns{link}}) { push @$pts, pdl($a); }
    } else { foreach (0..#{ $hlx}){ push @$pts, zeroes(3); }

    foreach my $a (@$hlx) { push @$HLX, getatmmtrx($a,'atms'); }
    while($a*$stp < 360) {
        while($b*$stp <= 180) {
            ## RtMt multiplied by the position vectors
            LOOP: while($g*$stp < 360) {
                my ($Hlx01,$Hlx02,$Hlx03,$pt1,$pt2,$ptn,$link,$cnt);
                $Hlx03 = $HLX->[1] x
                alpbetgam(alpha($a*$stp),beta($b*$stp),gamma($g*$stp));
                $pt1 = $pts->[0];
                $ptn = $pts->[1] x
                alpbetgam(alpha($a*$stp),beta($b*$stp),gamma($g*$stp));
                ## RtMt multiplied by the position vectors
                $cnt = $HLX->[0]->dim(1);
                $link = ($ptn - $pt1);
                $link = sqrt(sum($link**2));
                if($link > $linkdist) { $g++; next LOOP; }
                while($cnt) {

```

```

1)->copy;
my $tmp01 = $HLX->[0]->dice_axis(1,[$cnt-1])->clump(-
my $tmp02 = ($Hlx03 - $tmp01);
$tmp02 = $tmp02**2;
if(min(sqrt($tmp02->sumover)) < $dist) { $g++; next
LOOP; }
$cnt--;
}
printf $fh
"%d\t%d\t%d\t%d\t%.2f\n",$g,($a*$stp),($b*$stp),($g*$stp),$link;
$g++;
} $b++; $g=0;
} $a++; $b=0; $g=0;
} close $fh;
return 0;
}

sub DistCutOff3 {
my %optns = @_;
my ($i,$j,$k,$fh,$pts,$HLX,@linkdist);
my ($atmA,$atmB,$a,$b,$g,$a2,$b2,$g2) = ([],[],0,0,0,0,0);

$fh = new FileHandle "> $optns{fileout}";
my $hlx = (defined $optns{hlx})?$optns{hlx}:die "Need to input helices for simulation!\n";
my $dist = (defined $optns{distcutoff})?$optns{distcutoff}:0;
push @linkdist, (defined
$optns{linkcutoff})?@{$optns{linkcutoff}}:(999999,999999,999999);
my $stp = (defined $optns{stepsize})?$optns{stepsize}:0;
if(defined $optns{link}) { foreach my $a (@{$optns{link}}) { push @$pts, pdl($a); }
} else { foreach (0..$#{ $hlx}){ push @$pts, zeroes(3); }

foreach my $a (@$hlx) { push @$HLX, getatmmtrx($a,'atms'); }
my ($pt1,$pt2);
$pt1 = $pts->[0];
$pt2 = $pts->[5];
while($a*$stp < 360) {
while($b*$stp <= 180) {
## RtMt multiplied by the position vectors
LOOPA: while($g*$stp < 360) {
my ($Hlx01,$Hlx02,$Hlx03,$ptn,$pto,$linka,$cnt1,$cnt2,$cnt3);
$Hlx02 = $HLX->[1] - $pt1;
$Hlx02 = $Hlx02 x
alpbetgam(alpha($a*$stp),beta($b*$stp),gamma($g*$stp));
$Hlx02 = $HLX->[1] + $pt1;
$ptn = $pts->[1] x
alpbetgam(alpha($a*$stp),beta($b*$stp),gamma($g*$stp));
$pto = $pts->[2] x
alpbetgam(alpha($a*$stp),beta($b*$stp),gamma($g*$stp));
$cnt1 = $HLX->[0]->dim(1);
$cnt2 = $HLX->[0]->dim(1);
$cnt3 = $HLX->[1]->dim(1);
$linka = ($ptn - $pt2);
$linka = sqrt(sum($linka**2));
if($linka > $linkdist[0]) { $g++; $a2=$b2=$g2=0; next LOOPA; }
while($a2*$stp < 360) {

```

```

while($b2*$stp <= 180) {
  ##      RtMt multiplied by the position vectors
  LOOPB: while($g2*$stp < 360) {
    ##      RtMt multiplied by the position vectors
    my ($ptp,$ptq,$linkb,$linkc);
    $Hlx03      =      $HLX->[2]      x
  alpbetgam(alpha($a2*$stp),beta($b2*$stp),gamma($g2*$stp));
    $ptp      =      $pts->[3]      x
  alpbetgam(alpha($a2*$stp),beta($b2*$stp),gamma($g2*$stp));
    $ptq      =      $pts->[4]      x
  alpbetgam(alpha($a2*$stp),beta($b2*$stp),gamma($g2*$stp));

    $linkb = ($pto - $ptp);
    $linkb = sqrt(sum($linkb**2));
    if($linkb > $linkdist[1]) { $g2++; next
LOOPB; }

    $linkc = ($ptp - $pt1);
    $linkc = sqrt(sum($linkc**2));
    if($linkc > $linkdist[2]) { $g2++; next
LOOPB; }

interactions of helix 2 and 3
>dice_axis(1,[$cnt3-1])->clump(-1)->copy;

< $dist) { $g2++; next LOOPB; }

    $cnt3--;
  }
  while($cnt1) { ## checks for vdw
    my $tmp01 = $HLX->[1]-
    my $tmp02 = ($Hlx03 - $tmp01);
    $tmp02 = $tmp02**2;
    if(min(sqrt($tmp02->sumover))
    $cnt1--;
  }
  while($cnt1) { ## checks for vdw
    my $tmp01 = $HLX->[0]-
    my $tmp02 = ($Hlx02 - $tmp01);
    $tmp02 = $tmp02**2;
    if(min(sqrt($tmp02->sumover))
    $cnt1--;
  }
  while($cnt2) { ## checks for vdw
    my $tmp01 = $HLX->[0]-
    my $tmp02 = ($Hlx03 - $tmp01);
    $tmp02 = $tmp02**2;
    if(min(sqrt($tmp02->sumover))
    $cnt2--;
  }
  }
  printf                                $fh
"%d\t%d\t%d\t%d\t%d\t%d\t%.2f\t%.2f\t%.2f\n",($a*$stp),($b*$stp),($g*$stp),($a2*$stp),($b2*
$stp),($g2*$stp),$linka,$linkb,$linkc;

  $g2++;
} $b2++; $g2=0;

```

```

        } $a2++; $b2=$g2=0;
      } $g++; $a2=$b2=$g2=0;
    } $b++; $g=$a2=$b2=$g2=0;
  } $a++; $b=$g=$a2=$b2=$g2=0;
} close $fh;
return 0;
}

sub alpbetgam {
  my ($a,$b,$g) = @_;
  my $tmp = matmult($b,$a);
  my $rot = matmult($g,$tmp);
  return $rot;
}

sub alpha {
  my $a = shift;
  $a *= (pi/180.0);
  return pdl [[cos($a),sin($a),0],[-sin($a),cos($a),0],[0,0,1]];
}

sub beta {
  my $b = shift;
  $b *= (pi/180.0);
  return pdl [[cos($b),0,-sin($b)],[0,1,0],[sin($b),0,cos($b)]];
}

sub gamma {
  my $g = shift;
  $g *= (pi/180.0);
  return pdl [[cos($g),sin($g),0],[-sin($g),cos($g),0],[0,0,1]];
}

sub IdentMat {
  return pdl [[1,0,0],[0,1,0],[0,0,1]];
}

sub _mattranspose {
  my $mat = shift;
  my $out = [];
  $out->[0] = $mat->[0];
  $out->[1] = $mat->[3];
  $out->[2] = $mat->[6];
  $out->[3] = $mat->[1];
  $out->[4] = $mat->[4];
  $out->[5] = $mat->[7];
  $out->[6] = $mat->[2];
  $out->[7] = $mat->[5];
  $out->[8] = $mat->[8];
  return $out;
}

sub _matxvec {
  my $mat1 = shift;
  my $mat2 = shift;

```



```

my $i = shift;
my $j = shift;
my ($out);

for (my $i1 = 0; $i1 < $i; $i1++) {
    $out->[$i1] = 0.0;
    for (my $j1 = 0; $j1 < $j; $j1++) {
        $out->[$i1] += $mat1->[$j*$i1+$j1] * $mat2->[$j1];
    }
}
return $out;
}

sub _matadd {
my $mat1 = shift;
my $mat2 = shift;
my %optns = @_ ;
my $tmp = [];
die "matrix elements are not equal!\n" unless ($#{ $mat1 } == $#{ $mat2 });
$optns{sgn} = 1 unless $optns{sgn};

my $j = 0;
for my $i (@{ $mat1 }) { ${ $tmp }[$j] = $mat2->[$j]*$optns{sgn}+$i; $j++; }
my $out = $tmp;
return $out;
}

sub _vectorrot {
## reference CG&A (FEB 1984 pg 31)
my ($a,$b,$t) = @_ ;

$t *= pi/180;
my $v = pdl $a;
my $l = pdl $b;
my $I = identity(3);
my $L = pdl [[0,$l->at(2),-1*$l->at(1)],[-1*$l->at(2),0,$l->at(0)],[$l->at(1),-1*$l->at(0),0]];
my $d = sqrt(sum($l**2));
my $nv = $v x ($I + sin($t)/$d*$L + ((1-cos($t))/($d**2)*($L x $L));
return list($nv);
}

sub _snnerf {
## Parsons J et al, J. Comput. Chem. 26: 1063-8, 2005.
my ($Rbnd,$ang,$tor,$vecAB,$vecBC,$atmC) = @_ ;
my $nm = pdl [@{newmax_basics::_xprod($vecAB,$vecBC)}];
my $Mx = $nm;
my $C2 = pdl $atmC;
return (list($d2 x $Mx + $C2));
}
2;

```

## Appendix D

### newlsf4.pl

```
#!/usr/bin/perl

#####
#####
#Program Name: newlsf4.pl
#Created 06/2008 by Maximillian H. Bailor
#Description: This program superimposes two pdbs based on residue number(s).
#Updated 02/2009 by Maximillian H. Bailor
#    new format for more efficient command line implementation and
#    expanded features(e.g. superimposition and pdb output of new orientation).
#####
#####

#use lib '/local/home/bailor/BinProg';
use lib '/Users/maximillianbailor/RED/BinProg';
use Getopt::Long;
use newmax_basics;
use newmax_basics2;
use Math::Trig;
use Math::Complex;
use warnings;
use strict;

sub usage {
    printf STDERR "usage:  newlsf4.pl [options]\n";
    printf STDERR "options:\n";
    printf STDERR " \t--atoms all|sug|base|bkbn\n";
    printf STDERR " \t--dist\n";
    printf STDERR " \t--euler\n";
    printf STDERR " \t--input pdb1 pdb2 etc...\n";
    printf STDERR " \t--res min-max[,...]\n";
    printf STDERR " \t--output file\n";
    printf STDERR " \t--move translate rna pdb\n";
    printf STDERR " \t--numbering 1|100\n";
    printf STDERR " \t--rotate\n";
    printf STDERR " \t--superimpose\n";
    printf STDERR " \t--torsionang\n";
    printf STDERR " \t--write\n";
    printf STDERR " \t--help\n";
    exit 0;
}
```

```

## Global Variables
my ($i,@chnid,@outreslist,@structure);
## Set Defaults and implement GetOptions()
my (@atmtyp,$dist,$euler,@fileinput,$fileout,@rotate,$overlay);
my (@reslist,$renmb,$superimpose,$ta,$sugta,@trans,$write,$xx);

GetOptions(
    'atoms=s{1,}'    => \@atmtyp,
    'dist+'          => \$dist,
    'euler+'        => \$euler,
    'res=s{1,}'     => \@reslist,
    'help'          => \&usage,
    'input=s{1,}'   => \@fileinput,
    'move=f{3}'     => \@trans,
    'numbering=i'   => \$renmb,
    'output=s'      => \$fileout,
    'qrotate=f{3}'  => \@rotate,
    'superimpose+' => \$superimpose,
    'torsionang=s'  => \$ta,
    'usugtorang=s'  => \$sugta,
    'write+'        => \$write,
    'xxxx=s'        => \$xx,
    'vverlay+'      => \$overlay,
    "               => ,
) or die &usage;

$atmtyp[0] = "sug" unless @atmtyp;
$atmtyp[1] = "all" unless $atmtyp[1];
&usage unless (@reslist || @trans);

if(@fileinput) {
    $i = 0;
    foreach my $pdb (@fileinput) { $structure[$i] = newmax_basics->new(pdb=>$pdb); $i++; }
} else { &usage; }

if($dist) {
    ## I think the first one should work, don't know if the second is necessary
    die "Need to include residue selection\n" unless @reslist;
    die "Need to select atom types for distance calculation\n" unless $atmtyp[1];
    my ($cmp,$ref);
    if($#fileinput == 0) {
        $ref = newmax_basics::copypdb($structure[0],$reslist[0],$atmtyp[0]);
        $cmp = newmax_basics::copypdb($structure[0],$reslist[1],$atmtyp[1]);
    } elsif ($#fileinput == 1) {
        $ref = newmax_basics::copypdb($structure[0],$reslist[0],$atmtyp[0]);
        $cmp = newmax_basics::copypdb($structure[1],$reslist[1],$atmtyp[1]);
    }
    my $link = newmax_basics::PDBdist($ref,$cmp);
    printf "PDB: %s Linker distance = %.3f\n",$ref->{pdb},$link;
    exit 0;
}

if($euler) {
    ## I think this should work, but need to find suiTable example
    if($#reslist == 1) {
        my $ref = newmax_basics::copypdb($structure[0],$reslist[0],$atmtyp[0]);
        my $cmp = newmax_basics::copypdb($structure[1],$reslist[1],$atmtyp[0]);
    }
}

```

```

my $R = newmax_basics2::PDBfindR($ref,$cmp);
my $rmsd = newmax_basics::PDBrmsd($ref,$cmp);

printf "RMSD Fit = %7.3f ", $rmsd;
my $eul = newmax_basics::_rot2eul($R);
printf "Euler Angles: %8.3f %8.3f %8.3f %8.3f", $eul->[0], $eul->[1], $eul->[2], $eul-
>[0]+$eul->[2];
printf "\n";
} elsif($#reslist == 3) {
my $ref1 = newmax_basics::copypdb($structure[0], $reslist[0], $atmtyp[0]);
my $cmp1 = newmax_basics::copypdb($structure[1], $reslist[2], $atmtyp[0]);
my $ref2 = newmax_basics::copypdb($structure[0], $reslist[1], $atmtyp[0]);
my $cmp2 = newmax_basics::copypdb($structure[1], $reslist[3], $atmtyp[0]);

$cmp2->newmax_basics2::PDBfit($ref1, $cmp1);
my $rmsd = newmax_basics::PDBrmsd($ref1, $cmp1);
printf "RMSD Fit = %7.3f ", $rmsd;

my $R = newmax_basics2::PDBfindR($cmp2, $ref2);
$rmsd = newmax_basics::PDBrmsd($cmp2, $ref2);
printf "RMSD Fit = %7.3f ", $rmsd;
my $eul = newmax_basics::_rot2eul($R);
printf "Euler Angles: %8.3f %8.3f %8.3f %8.3f", $eul->[0], $eul->[1], $eul->[2], $eul-
>[0]+$eul->[2];
printf "\n";
}
exit 0;
}

if($#rotate==2) {      ## I think the first one should work, don't know if the second is necessary
my ($ref);
$ref = newmax_basics::copypdb($structure[0], $reslist[0], 'all');
$ref->newmax_basics::chngeul(\@rotate);
$ref->newmax_basics::PDBrotate(\@rotate);
$ref->PDBwrite(out=>$fileout);
exit 0;
}

if($superimpose) {   ## I think the first one should work
my ($ref, $cmp, $rmsd);
if($#reslist == 1) {
$ref = newmax_basics::copypdb($structure[0], $reslist[0], $atmtyp[0]);
$cmp = newmax_basics::copypdb($structure[1], $reslist[1], $atmtyp[0]);
newmax_basics2::PDBfindR($ref, $cmp);
$rmsd = newmax_basics::PDBrmsd($ref, $cmp);
printf "RMSD Fit = %.3f\n", $rmsd;
}
exit 0;
}

if($overlay) {      ## I think the first one should work
my ($ref, $cmp, $obj, $rmsd);
$ref = newmax_basics::copypdb($structure[0], $reslist[0], $atmtyp[0]);

```

```

$cmp = newmax_basics::copypdb($structure[1],$reslist[0],$atmtyp[0]);
$obj = newmax_basics::copypdb($structure[1],$reslist[1],$atmtyp[1]);
newmax_basics2::PDBfit2($obj,$ref,$cmp);
$rmsd = newmax_basics::PDBrmsd($ref,$cmp);
printf "RMSD Fit = %.3f\n", $rmsd;
$obj->PDBwrite(out=>$fileout);
exit 0;
}

if($ta) {
  my ($ref);
  $atmtyp[0] = "all";

  my $ressel = reslistaug($reslist[0],1);
  my $tares = [split /,/, reslistaug3($reslist[0],0)];

  foreach my $i (@$tares) {
    my ($torang,$taset);
    $ref = newmax_basics::copypdb($structure[0],$ressel,$atmtyp[0],1);
    printf "%s\t",$i;
    if($ta=~/,/) {
      $taset = [split /,/, $ta];
      foreach my $ang (@$taset) {
        $torang = newmax_basics::PDBtorsionang2($ref,$i,$ang);
        $torang+=360 if $torang < 0;
        printf "%s:\t%.1f\t",$ang,$torang;
      }
    } else {
      $torang = newmax_basics::PDBtorsionang2($ref,$i,$ta);
      $torang+=360 if $torang < 0;
      printf "%s:\t%.1f",$ta,$torang;
    } print "\n";
  }
  exit 0;
}

if($sugta) {
  my ($ref);
  $atmtyp[0] = "all";

  my $ressel = reslistaug($reslist[0],1);
  my $sugtares = [split /,/, reslistaug3($reslist[0],0)];

  foreach my $i (@$sugtares) {
    my ($torang,$sugtaset);
    $ref = newmax_basics::copypdb($structure[0],$ressel,$atmtyp[0],1);
    printf "%s\t",$i;
    if($sugta=~/,/) {
      $sugtaset = [split /,/, $sugta];
      foreach my $ang (@$sugtaset) {
        $torang = newmax_basics::PDBsugtorang($ref,$i,$ang);
        $torang+=360 if $torang < 0;
        printf "%s:\t%.1f\t",$ang,$torang;
      }
    } else {

```

```

        $torang = newmax_basics::PDBsugtorang($ref,$i,$sugta);
        $torang+=360 if $torang < 0;
        printf "%s:\t%.1f", $sugta, $torang;
    } print "\n";
}
# $ref->PDBwrite if $write;
exit 0;
}

if($xx) {
    my ($ref);
    $atmtyp[0] = "all";

    my $ressel = reslistaug($reslist[0],1);
    my $tares = [split /,/, reslistaug3($reslist[0],0)];

    foreach my $i (@$tares) {
        my ($torang,$taset);
        $ref = newmax_basics::copypdb($structure[0],$ressel,$atmtyp[0],1);
        if($xx=~/,/) {
            $taset = [split /,/, $xx];
            foreach my $ang (@$taset) {
                $torang = newmax_basics::PDBtorsionang3($ref,$i,$ang);
                $torang+=360 if $torang < 0;
                printf "%s:\t%.1f\t", $ang, $torang;
            }
        } else {
            $torang = newmax_basics::PDBtorsionang3($ref,$i,$xx);
            $torang+=360 if $torang < 0;
            printf "%s:\t%.1f", $xx, $torang;
        } print "\n";
    }
    exit 0;
}

if($#trans!=-1) {
    my ($ref);

    $ref = newmax_basics->new(pdb=>$fileinput[0]);

    my $trans = \@trans;
    $ref->chngrans(\@trans);
    $ref->PDBtranslate($trans);
    $ref->PDBwrite(out=>$fileout);
    exit 0;
}

if($write) {
    my ($ref);
    $ref = newmax_basics::copypdb($structure[0],$reslist[0],$atmtyp[0],1);
    $ref->PDBwrite(out=>$fileout);
    exit 0;
}

exit 0;

```

```

##      subroutines      ##
sub reslistaug {
    my ($reslist,$shift) = @_ ;
    my $newlist = "";

    while($reslist=~/(\\w{1})\\.(\d+)[^-]*/g) { $newlist .= $1.".".$2-$shift)."-"($2+$shift).";" ; }
    while($reslist=~/(\\w{1})\\.(\d+)-(\d+)/g) { $newlist .= $1.".".$2-$shift)."-"($3+$shift).";" ; }
    while($reslist=~/(\\w{1})\\.(\d+)-\\w{1}\\.(\d+)/g) { $newlist .= $1.".".$2-$shift)."-"($3+$shift).";" ; }
    $newlist =~s/,,$//;
    return $newlist;
}

sub reslistaug2 {
    my $reslist = shift;
    my $newlist = "";

    while($reslist=~/(\\w{1})\\.(\d+)[^-]*/g) { $newlist .= $1.".".$2.";" ; }
    while($reslist=~/(\\w{1})\\.(\d+)-(\d+)/g) { foreach my $i ($2-1..$3+1) { $newlist .= $1.".".$i.";" ; } }
    while($reslist=~/(\\w{1})\\.(\d+)-\\w{1}\\.(\d+)/g) { foreach my $i ($2..$3) { $newlist .= $1.".".$i.";" ; } }
    $newlist =~s/,,$//;

    my @res = split /,/ , $newlist;
    my %res = map { $_ => undef } @res;
    $newlist = join ', ' , sort keys %res;
    print $newlist,"\n";

    return $newlist
}

sub reslistaug3 {
    my $reslist = shift;
    my $newlist = "";

    while($reslist=~/(\\w{1})\\.(\d+)[^-]*/g) { $newlist .= $1.".".$2.";" ; }
    while($reslist=~/(\\w{1})\\.(\d+)-(\d+)/g) { foreach my $i ($2..$3) { $newlist .= $1.".".$i.";" ; } }
    while($reslist=~/(\\w{1})\\.(\d+)-\\w{1}\\.(\d+)/g) { foreach my $i ($2..$3) { $newlist .= $1.".".$i.";" ; } }
    $newlist =~s/,,$//;
    my @res = split /,/ , $newlist;
    my %res = map { $_ => undef } @res;
    $newlist = join ', ' , sort keys %res;
    return $newlist;
}

```

## Appendix E

### abgconNEWFRAME.pl

```
#!/usr/bin/perl

#####
#Program Name: abgconNEWFRAME.pl
#Description: Program modifies ranges of alpha, beta and gamma.
#Created 11/2008 by Maximillian H. Bailor
#####

use lib '/Users/maximillianbailor/RED/BinProg';
use max_sets;
use FileHandle;
use Getopt::Long;
use Math::Trig;
use Math::Complex;
use warnings;
use strict;

use constant pi => 4*atan2(1,1);

sub usage {
    printf STDERR "usage:  abgconNEWFRAME.pl [options]\n";
    printf STDERR "options: --input input pdb file name\n";
    printf STDERR "      --output base output file name\n";
    exit 0;
}

## Global Variables
my (@filein,$fileout,$fh,@rval,@colstat,@acolstat);
my ($shift,$min,$pred,$fswtch,$add) = (0,0,0,0,0);
srand;

GetOptions(
    'help'          => \&usage,
    'add+'          => \&$add,
    'colstat=i{1,}' => \&@colstat,
    'input=s{1,}'   => \&@filein,
    'fswitch+'      => \&$fswtch,
    'min=f'         => \&$min,
    'output=s'      => \&$fileout,
    'pred=f'        => \&$pred,
    'rvalue=i{2}'   => \&@rval,
    'shift=f'       => \&$shift,
```



```

        'vcstat=i{1,}'    => \@acolstat,
        "                =>,
) or die &usage;

if($fileout) { $fh = new FileHandle "> $fileout"; }
else { $fh = \*STDOUT; }
open FH, $filein[0] || die "could not find file!\n";
if($filein[1]) { open GH, $filein[1] || die "could not find file!\n"; }

my @range = grep { $_%$min == 0 } (0..359) if $min;

if($min) {
    my @abg_set = <FH>;
    my ($a,$b,$c,@trsh);
    @trsh = @abg_set;
    @$a = grep { s/\w+\s+([+]?[d+\.]*\d*)\s+[+]?[d+\.]*\d*\s+[+]?[d+\.]*\d*/$1/ } @trsh;
    @trsh = @abg_set;
    @$b = grep { s/\w+\s+[+]?[d+\.]*\d*\s+[+]?[d+\.]*\d*\s+[+]?[d+\.]*\d*/$1/ } @trsh;
    @trsh = @abg_set;
    @$c = grep { s/\w+\s+[+]?[d+\.]*\d*\s+[+]?[d+\.]*\d*\s+[+]?[d+\.]*\d*/$1/ } @trsh;

    my ($ap,$bp,$cp) = ($a,$b,$c);
    my ($Czmin,$Czmax,$Crmin,$Crmax,$Zmin,$Zmax,$rmin,$rmax) = (-1,-1,-1,-1,1e10,-
1e10,1e10,-1e10);
    foreach my $cnt (@range) {
        my ($i,$r,$z,$abgR) = (0,0,0,[]);
        while($i<${$a}+1) {
            my $tmp = abgminimize($a->[$i],$b->[$i],$c->[$i],$cnt);

            push @{$abgR->[0]},$tmp->[0];
            push @{$abgR->[1]},$tmp->[1];
            push @{$abgR->[2]},$tmp->[2];

            $r += $tmp->[0]**2+$tmp->[1]**2+$tmp->[2]**2;
            $i++;
        }
        $z = Zfact_deprel($abgR->[0],$abgR->[1],$abgR->[2]);
        $r = sqrt($r);
        if($z < $Zmin) { $Zmin = $z; $Czmin = $cnt; }
        if($z > $Zmax) { $Zmax = $z; $Czmax = $cnt; }
        if($r < $rmin) { $rmin = $r; $Crmin = $cnt; }
        if($r > $rmax) { $rmax = $r; $Crmax = $cnt; }
        printf "%d %.1f %.2f\n", $cnt,$r,$z;
        $cnt+=5;
    }
    printf "Zmin\toffset\tZmax\toffset\n";
    printf "%.1f\t%d\t%.1f\t%d\n",$Zmin,$Czmin,$Zmax,$Czmax;
    printf "Rmin\toffset\tRmax\toffset\n";
    printf "%.1f\t%d\t%.1f\t%d\n",$rmin,$Crmin,$rmax,$Crmax;
    close FH;
    exit 0;
}

if($#rval==1) {
    my ($i,$j) = @rval;

```

```

my (@i,@j);
while(<FH>) {
    next if $_ =~ /^#/;
    my @ln = split('\s+',$_);

    push @i, $ln[$i];
    push @j, $ln[$j];
}
printf "R^2: %5.2f\tR: %5.2f\n",Rxysq(\@i,\@j),Rxy(\@i,\@j);
exit 0;
}

if(@colstat) {
my (@i,@j,@col);
my ($col,$ave,$sdv) = (0,[],[]);

while(<FH>) {
    next if $_ =~ /Illegal division by zero at/;
    next if $_ =~ /Could not find file/;
    last if $_ =~ /^END$/;
    my @ln = split('\s+',$_);
    $col = 0;
    if(!$add) { foreach my $i (@colstat) { push @{$col[$col]}, $ln[$i]; $col++; } }
    else { push @{$col[$col]}, $ln[$colstat[0]]+$ln[$colstat[1]]; $col++; }
#
    foreach my $i (@colstat) { push @{$col[$col]}, $ln[$i]; $col++; }
}
foreach my $i (@col) { push @$ave, ave(@$i); }
foreach my $i (@$ave) { printf "average: %5.2f\tst. deviation %5.2f\n", $i,stddev(shift @col,$i);
}
exit 0;
}

if(@acolstat) {
my (@i,@j,@col);
my ($col,$ave,$sdv) = (0,[],[]);

while(<FH>) {
    next if $_ =~ /Illegal division by zero at/;
    next if $_ =~ /Could not find file/;
    last if $_ =~ /^END$/;
    my @ln = split('\s+',$_);
    $col = 0;
    if(!$add) { foreach my $i (@acolstat) { push @{$col[$col]}, abs($ln[$i]); $col++; } }
    else { push @{$col[$col]}, abs($ln[$acolstat[0]]+$ln[$acolstat[1]]); $col++; }
}
foreach my $i (@col) { push @$ave, ave(@$i); }
foreach my $i (@$ave) { printf "average: %5.2f\tst. deviation %5.2f\n", $i,stddev(shift @col,$i);
}
exit 0;
}

if($pred) {
while(<FH>) {
    my @ln = split('\s+',$_);
    $ln[1]+=$pred/2;
}
}

```

```

        $ln[3]+=$pred/2;
        my $abg = abgminimize($ln[1],$ln[2],$ln[3],$shift);

        printf $fh "%s\t",$ln[0];
        for(1..3) { printf $fh "%.2f\t", shift @$abg; }
        printf $fh "%s", $ln[4] if $ln[4];
        print $fh "\n";
    }
}

if($fswtch) {
    while(<FH>) {
        my @ln = split('\s+',$_);
        my $tmp = $ln[1];
        $ln[1] = $ln[3]*-1.0;
        $ln[2] *= -1.0;
        $ln[3] = $tmp*-1.0;
        my $abg = abgminimize($ln[1],$ln[2],$ln[3],$shift);

        printf $fh "%s\t",$ln[0];
        for(1..3) { printf $fh "%.2f\t", shift @$abg; }
        printf $fh "%s", $ln[4] if $ln[4];
        print $fh "\n";
    }
}

while(<FH>) {
    my @ln = split('\s+',$_);

    my $abg = abgminimize($ln[1],$ln[2],$ln[3],$shift);

    if($ln[4] && $ln[4]=~/\d+/) {
        $ln[4]+=360.0 if $ln[4]<-180;
        $ln[4]-=360.0 if $ln[4]>180;
        $ln[4]+=360.0 if $ln[4]<-180;
        $ln[4]-=360.0 if $ln[4]>180;
        $ln[4] = 0.01 if !$ln[4];
    }

    printf $fh "%s\t",$ln[0];
    for(1..3) { printf $fh "%.2f\t", shift @$abg; }
    foreach my $i (4..$#ln+1) { printf $fh "%s\t", $ln[$i] if($ln[$i]); }
    print $fh "\n";
}
close FH; close GH; close $fh;
exit 0;

sub abgminimize {
    my ($a,$b,$g,$shift) = @_;
    $a+=$shift;
    $g-=$shift;

    if($a < -180) { $a+=360; } elsif($a > 180) { $a-=360; }
    if($b > 180) { $b-=360; } elsif($b < -180) { $b+=360; }
    if($g < -180) { $g+=360; } elsif($g > 180) { $g-=360; }
}

```

```

my ($abg,$d) = ([],[]);

$abg->[0] = [$a,$b,$g];
$abg->[1] = [$a-180,$b*-1,$g+180];
$abg->[2] = [$a+180,$b*-1,$g-180];
$abg->[3] = [$a+180,$b*-1,$g+180];
$abg->[4] = [$a-180,$b*-1,$g-180];

$d->[0] = sqrt($a**2+$b**2+$g**2);
$d->[1] = sqrt(($a-180)**2+($b*-1)**2+($g+180)**2);
$d->[2] = sqrt(($a+180)**2+($b*-1)**2+($g-180)**2);
$d->[3] = sqrt(($a+180)**2+($b*-1)**2+($g+180)**2);
$d->[4] = sqrt(($a-180)**2+($b*-1)**2+($g-180)**2);

my ($pck,$i,$j) = (1e+10,0,0);
foreach my $k (@$d) { if($pck > $k) { $pck = $k; $i = $j; }; $j++; }
return ($abg->[$i]);
}

sub ave {
my @val = @_;
my ($a,$cnt) = (0,0);
foreach my $i (@val) { $a += $i; $cnt++; }
return ($a/$cnt);
}

sub cov_dep {
my ($r1,$r2,$r3) = @_;
my $cov_r1r2 = $r3*(1.0-$r1**2-$r2**2);
$cov_r1r2 -= 0.5*($r1*$r2)*(1-$r1**2-$r2**2-$r3**2);
$cov_r1r2 /= ((1-$r1**2)*(1-$r2**2));
return $cov_r1r2;
}

sub fisher {
my $x = shift @_;
return (0.5*log((1+$x)/(1-$x)));
}

sub Rxy {
my ($x,$y) = @_;
my ($ssx,$ssy,$ssxy) = (sumsq($x,$x),sumsq($y,$y),sumsq($x,$y));
# printf "%.2f %.2f \n",$ssy**2,$ssy;
return (($ssxy)/sqrt(abs($ssx*$ssy)));
}

sub Rxysq {
my ($x,$y) = @_;
my ($ssx,$ssy,$ssxy) = (sumsq($x,$x),sumsq($y,$y),sumsq($x,$y));
return (($ssxy**2)/($ssx*$ssy));
}

sub sdiff_deprel {
my ($cov,$n) = @_;

```

```

        return sqrt((2.0-2.0*$cov)/($n-3));
    }

    sub stddev {
        my ($data,$ave) = @_;
        my ($a,$cnt) = (0,0);
        foreach my $i (@$data) { $a += ($i-$ave)**2; $cnt++; }
        return (sqrt($a/($cnt-1)))
    }

    sub sumsq {
        my ($x,$y) = @_;
        my ($xave,$yave,$n) = (ave(@$x),ave(@$y),${#$x}+1);
        my ($ss,$i) = (0.0,0);
        foreach (@$x) { $ss += (($x->[$i]-$xave)*($y->[$i]-$yave)); $i++; }
        return $ss;
    }

    sub Zfact_deprel {
        my ($x,$y,$z) = @_;
        my ($ab,$bc,$ag,$n) = (Rxy($x,$y),Rxy($y,$z),Rxy($x,$z),${#$x}+1);
        my $cov = cov_dep($ab,$bc,$ag);
        return ((fisher($ab)-fisher($bc))/sdiff_deprel($cov,$n));
    }

```

## Appendix F

### abgconverter.pl

```
#!/usr/bin/perl

#####
#Program Name: abgconverter.pl
#Description: Program modifies ranges of alpha, beta and gamma.
#Created 11/2008 by Maximillian H. Bailor
#####

use lib '/Users/maximillianbailor/RED/BinProg';
use max_sets;
use FileHandle;
use Getopt::Long;
use warnings;
use strict;
use constant pi => 4*atan2(1,1);

sub usage {
    printf STDERR "usage:  abgconverter.pl [options]\n";
    printf STDERR "options: --alpha range for alpha values\n";
    printf STDERR "      --beta range for beta values\n";
    printf STDERR "      --gamma range for gamma values\n";
    printf STDERR "      --difference ?????????\n";
    printf STDERR "      --input input pdb file name\n";
    printf STDERR "      --output base output file name\n";
    printf STDERR "      --randsel randomly which abg are flipped\n";
    printf STDERR "      --trim randomly chooses which set of abg w/i a cap\n";
    printf STDERR "      --union ?????????\n";
    printf STDERR "      --help\n";
    exit 0;
}

## Global Variables
my (@filein,$fileout,$fh,@alpha,@beta,@gamma,$trim,$randsel);
my (@lines,@plus);
my ($diff,$union,$zone,$excise) = (0,0,0,0);
srand;

GetOptions(
    'help'          => \&usage,
    'alpha=i{2}'    => \@alpha,
    'beta=i{2}'     => \@beta,
    'gamma=i{2}'    => \@gamma,
```

```

'difference+' => \$diff,
'excise+'     => \$excise,
'input=s{1,}' => \@filein,
'output=s'    => \$fileout,
'plus=i{3}'   => \@plus,
'randsel+'    => \$randsel,
'trim=i'      => \$trim,
'union+'      => \$union,
'zone+'       => \$zone,
" => ,

```

) or die &usage;

```

if($fileout) { $fh = new FileHandle "> $fileout"; }
else { $fh = \*STDOUT; }
open FH, $filein[0] || die "could not find file!\n";
if($filein[1]) { open GH, $filein[1] || die "could not find file!\n"; }

```

```

if($trim) {
    while(<FH>) {
        if(@lines < $trim) { push @lines, $_; }
        elsif(rand($.) < $trim) {
            splice @lines,rand(@lines),1;
            push @lines, $_;
        }
    }
    foreach (@lines) { print $fh "$_"; }
    exit 0;
}

```

```

if($randsel) {
    $alpha[0] = $beta[0] = $gamma[0] = -360;
    $alpha[1] = $gamma[1] = 360;
    $beta[1] = 0;
    while(<FH>) {
        my @ln = split('\s{1,}', $_);
        if(int(rand 2) == 1) {
            if($ln[1] < 0) {
                $ln[1] += 180; $ln[2] *= -1.0;
                if($ln[3] > 0) { $ln[3] -= 180; } else { $ln[3] += 180; }
            } elsif($ln[1] > 0) {
                $ln[1] -= 180; $ln[2] *= -1.0;
                if($ln[3] > 0) { $ln[3] -= 180; } else { $ln[3] += 180; }
            }
        }

        if($ln[1] < -180) { $ln[1] += 360; }
        if($ln[1] > 180) { $ln[1] -= 360; }
        if($ln[3] < -180) { $ln[3] += 360; }
        if($ln[3] > 180) { $ln[3] -= 360; }

        if(($ln[1] > 45) && ($ln[3] > 45)) {
            if($ln[1] > $ln[3]) { $ln[1] -= 360; $ln[1] += 180; $ln[2] *= -1.0; $ln[3] -= 180; }
            else { $ln[3] -= 360; $ln[1] -= 180; $ln[2] *= -1.0; $ln[3] += 180; }
        } elsif(($ln[1] < -45) && ($ln[3] < -45)) {
            if($ln[1] < $ln[3]) { $ln[1] += 360; $ln[1] -= 180; $ln[2] *= -1.0; $ln[3] += 180; }

```

```

        } else { $ln[3] += 360; $ln[1] += 180; $ln[2] *=-1.0; $ln[3] -= 180; }
    }
    for(0..$#ln) { printf $fh "%s\t", shift @ln; }
    print $fh "\n";
}
exit 0;
}

if($zone) {
    die "Need abg values to use this function\n" unless $alpha == 1 && $beta == 1 &&
    $gamma == 1;
    while(<FH>) {
        my @ln = split("\s{1},$_");

        if(($ln[1] > $alpha[0]) && ($ln[1] < $alpha[1])) {
            if(($ln[2] > $beta[0]) && ($ln[2] < $beta[1])) {
                if(($ln[3] > $gamma[0]) && ($ln[3] < $gamma[1])) {
                    if($ln[1] > 0) { $ln[1] -= 180; } else { $ln[1] += 180; }
                    $ln[2] *=-1.0;
                    if($ln[3] > 0) { $ln[3] -= 180; } else { $ln[3] += 180; }
                }
            }
        }

        for(0..$#ln) { printf $fh "%s\t", shift @ln; }
        print $fh "\n";
    }
    exit 0;
}

if($excise) {
    die "Need abg values to use this function\n" unless $alpha == 1 && $beta == 1 &&
    $gamma == 1;
    while(<FH>) {
        my @ln = split("\s{1},$_");

        if(($ln[1] > $alpha[0]) && ($ln[1] < $alpha[1])) {
            if(($ln[2] > $beta[0]) && ($ln[2] < $beta[1])) {
                if(($ln[3] > $gamma[0]) && ($ln[3] < $gamma[1])) {
                    for(0..$#ln) { printf $fh "%s\t", shift @ln; }
                    print $fh "\n";
                }
            }
        }
    }
    exit 0;
}

if($#plus==2) {
    die "Need abg values to use this function\n" unless $alpha == 1 && $beta == 1 &&
    $gamma == 1;
    while(<FH>) {
        my @ln = split("\s{1},$_");
        if(($ln[1] > $alpha[0]) && ($ln[1] < $alpha[1])) {
            if(($ln[2] > $beta[0]) && ($ln[2] < $beta[1])) {

```



```

        if(($ln[3] > $gamma[0]) && ($ln[3] < $gamma[1])) {
            $ln[1]+= $plus[0]; $ln[2]+= $plus[1]; $ln[3]+= $plus[2];
        }
    }
}

for(0..$#ln) { printf $fh "%s\t", shift @ln; }
print $fh "\n";
}
exit 0;
}

if((@alpha || @beta || @gamma) && !$trim && !$diff && !$union) {
    if(!@alpha) { $alpha[0] = -360; $alpha[1] = 360; }
    if(!@beta) { $beta[0] = -360; $beta[1] = 360; }
    if(!@gamma) { $gamma[0] = -360; $gamma[1] = 360; }

    while(<FH>) {
        my @ln = split('\s{1,}', $_);

        if($ln[1] < $alpha[0]) {
            $ln[1]+=180; $ln[2]*=-1.0;
            if($ln[3] > 0) { $ln[3]-=180; } else { $ln[3]+=180; }
        } elsif($ln[1] > $alpha[1]) {
            $ln[1]-=180; $ln[2]*=-1.0;
            if($ln[3] > 0) { $ln[3]-=180; } else { $ln[3]+=180; }
        }
        if($ln[3] < $gamma[0]) {
            $ln[3]+=180; $ln[2]*=-1.0;
            if($ln[1] > 0) { $ln[3]-=180; } else { $ln[3]+=180; }
        } elsif($ln[3] > $gamma[1]) {
            $ln[3]-=180; $ln[2]*=-1.0;
            if($ln[1] > 0) { $ln[3]-=180; } else { $ln[3]+=180; }
        }
        if($ln[2] < $beta[0]) {
            if($ln[1] > 0) { $ln[3]-=180; } else { $ln[3]+=180; }
            $ln[2]*=-1.0;
            if($ln[3] > 0) { $ln[3]-=180; } else { $ln[3]+=180; }
        } elsif($ln[2] > $beta[1]) {
            if($ln[1] > 0) { $ln[3]-=180; } else { $ln[3]+=180; }
            $ln[2]*=-1.0;
            if($ln[3] > 0) { $ln[3]-=180; } else { $ln[3]+=180; }
        }
    }

    if($ln[1] < -180) { $ln[1]+=360; }
    if($ln[1] > 180) { $ln[1]-=360; }
    if($ln[3] < -180) { $ln[3]+=360; }
    if($ln[3] > 180) { $ln[3]-=360; }

    if(($ln[1] > 45) && ($ln[3] > 45)) {
        if($ln[1] > $ln[3]) { $ln[1] -= 360; $ln[1]+=180; $ln[2]*=-1.0; $ln[3]-=180; }
        else { $ln[3] -= 360; $ln[1]-=180; $ln[2]*=-1.0; $ln[3]+=180; }
    } elsif(($ln[1] < -45) && ($ln[3] < -45)) {
        if($ln[1] < $ln[3]) { $ln[1] += 360; $ln[1]-=180; $ln[2]*=-1.0; $ln[3]+=180; }
        else { $ln[3] += 360; $ln[1]+=180; $ln[2]*=-1.0; $ln[3]-=180; }
    }
}

```

```

    }

    for(0..$#ln) { printf $fh "%s\t", shift @ln; }
    print $fh "\n";
}
} elseif(!$trim && !$diff && !$union) {
    while(<FH>) {
        my @ln = split("\s{1},",$_);

        if($ln[1] < -180) { $ln[1]+=360; }
        if($ln[1] > 180) { $ln[1]-=360; }
        if($ln[3] < -180) { $ln[3]+=360; }
        if($ln[3] > 180) { $ln[3]-=360; }

        if(($ln[1] > 45) && ($ln[3] > 45)) {
            if($ln[1] > $ln[3]) { $ln[1] -= 360; $ln[1]+=180; $ln[2]*=-1.0; $ln[3]-=180;
            } else { $ln[3] -= 360; $ln[1]-=180; $ln[2]*=-1.0; $ln[3]+=180; }
        } elseif(($ln[1] < -45) && ($ln[3] < -45)) {
            if($ln[1] < $ln[3]) { $ln[1] += 360; $ln[1]-=180; $ln[2]*=-1.0; $ln[3]+=180;
            } else { $ln[3] += 360; $ln[1]+=180; $ln[2]*=-1.0; $ln[3]-=180; }
        }

        for(0..$#ln) { printf $fh "%s\t", shift @ln; }
        print $fh "\n";
    }
}

if($diff) {
    my (%abg1,%abg2);
    while(<FH>) {
        my @ln = split /\s{1},/, $_;
        my $key = join(' ',@ln[1..3]);
        $abg1{$key} = undef;
    }
    while(<GH>) {
        my @ln = split /\s{1},/, $_;
        my $key = join(' ',@ln[1..3]);
        $abg2{$key} = undef;
    }
    my $intersect = max_sets::intersection_hash(\%abg1,\%abg2);
    print $fh "intersection\t",join(" \nintersection\t", keys %{$intersect}),"\n";
}

if($union) {
    my (%abg1,%abg2);
    while(<FH>) {
        my @ln = split /\s{1},/, $_;
        my $key = join(' ',@ln[1..3]);
        $abg1{$key} = undef;
    }
    while(<GH>) {
        my @ln = split /\s{1},/, $_;
        my $key = join(' ',@ln[1..3]);
        $abg2{$key} = undef;
    }
}

```

```
    my $union = max_sets::intersection_hash(\%abg1,\%abg2);
    print $fh "union\t",join(" \nunion\t", keys %{$union}),"\n";
}

close FH; close GH; close $fh;
exit 0;
```

## Appendix G

### maxpdbfitgenerator2.pl

```
#!/usr/bin/perl

use Getopt::Long;
use Math::Trig;
use warnings;
use strict;

sub usage {
    printf STDERR "usage: perl maxpdbfitgenerator.pl [options]\n";
    printf STDERR "options: -input => input list file\n";
    printf STDERR "        -output => output euler angle file\n";
    printf STDERR "        -templateref => reference pdb file\n";
    printf STDERR "        -seekresidues => ref helix1\n";
    printf STDERR "        -refresidues => ref helix2\n";
    printf STDERR "        -help\n";
    exit 0;
}

my ($file,$new,$ref,$refstm2,$refstm1,$atm,$mod,@refbp);
my ($st5p,$st3p,$refhlx,$buf,$shft,$cplx) = ([11,12],[33,34],1,1,1,0);

GetOptions(
    'help'          => \&usage,
    'atoms+'        => \&$atm,
    'basepairref=i{1,3}' => \@refbp,
    'complex+'      => \&$cplx,
    'filetype=i'    => \&$shft,
    'input=s'       => \&$file,
    'modpdb+'       => \&$mod,
    'output=s'      => \&$new,
    'templateref=s' => \&$ref,
    'seekresidues=s' => \&$refstm1,
    'refresidues=s' => \&$refstm2,
) or die &usage;

die "I don't have a file to print to" unless $new;

open FH, $file || die "could not find file!:\n";
open GH, "> $new" || die "could not open file!:\n";

if(@refbp) {
    my (@Tst3p,@Tst5p,@HL,$S1,$S2);
```

```

@HL = (length($refbp[0])>1)?split //, $refbp[0]:($refbp[0],$refbp[0]);
$refhlx = (!$refbp[1])?1:$refbp[1];
$buf = $refbp[2]?$refbp[2]:1;
@Tst5p = ($st5p->[0]-$HL[0],$st5p->[0]-$buf,$st5p->[1]+$buf,$st5p->[1]+$HL[1]);
@Tst3p = ($st3p->[0]-$HL[1],$st3p->[0]-$buf,$st3p->[1]+$buf,$st3p->[1]+$HL[0]);
foreach my $i (@Tst5p) { $i = 'Z'.".".$i; }
foreach my $i (@Tst3p) { $i = 'Z'.".".$i; }
$S1 = join ',',(join '-',@Tst5p[0..1]),(join '-',@Tst3p[2..3]);
$S2 = join ',',(join '-',@Tst5p[2..3]),(join '-',@Tst3p[0..1]);

$refstm2 = ($refhlx==1)?$S1:$S2;
$refstm1 = ($refhlx==1)?$S2:$S1;
}

if($shft != 2) {
    $shft = 1 if $shft == 1;
}

while (<FH>) {
    next if m/^\n$/;
    next if m/^\#/;
    my @tmp = split /\s{1,}/, $_;
    $tmp[0]=~s/\.pdb//;
    $tmp[0] = uc $tmp[0] unless $mod;
    $tmp[0] .= "\.pdb";

    my $com1;
    if(!$cmplx) {
        $com1 = sprintf "~ /RED/BinProg/newsf4.pl -r %s %s %s %s -i %s %s -
e",$refstm2,$refstm1,$tmp[$shft],$tmp[$shft+1],$ref,$tmp[0];
        $com1 = sprintf "~ /RED/BinProg/newsf4.pl -r %s %s %s %s -i %s %s -e -a
bkbn",$refstm2,$refstm1,$tmp[$shft],$tmp[$shft+1],$ref,$tmp[0] if $atm;
    } else {
        $com1 = sprintf "~ /RED/BinProg/newsf4.pl -r %s %s %s %s -i %s %s -
e",$refstm2,$refstm1,$tmp[$#tmp-1],$tmp[$#tmp],$ref,$tmp[0];
        $com1 = sprintf "~ /RED/BinProg/newsf4.pl -r %s %s %s %s -i %s %s -e -a
bkbn",$refstm2,$refstm1,$tmp[$#tmp-1],$tmp[$#tmp],$ref,$tmp[0] if $atm;
    }

    my $output = qx($com1 2>&1);
    if(!$tmp[3]) { $tmp[3] = 'na'; }
    elsif($tmp[1]=~/^\w\.\d/) { $tmp[1] = 'na'; }
    printf GH "%s %s %s",$tmp[1],$tmp[0],$output if !$cmplx;
    printf GH "%s %s",chomp $_,$output if $cmplx;
}

close GH;
close FH;

exit 0;

```

## Appendix H

### maxdistgenerator.pl

```
#!/usr/bin/perl
use Getopt::Long;
use Math::Trig;
use warnings;
use strict;

sub usage {
    printf STDERR "usage: perl maxpdbfitgenerator.pl [options]\n";
    printf STDERR "options: -input => input list file\n";
    printf STDERR "        -output => output euler angle file\n";
    printf STDERR "        -help\n";
    exit 0;
}

my ($file,$new,$ref,$refstm2,$refstm1);
GetOptions(
    'help'           => \&usage,
    'input=s'        => \$file,
    'output=s'       => \$new,
) or die &usage;
die "I don't have a file to print to" unless $new;
open FH, $file || die "could not find file!:$!\n";
open GH, "> $new" || die "could not open file!:$!\n";

while (<FH>) {
    next if m/^\n$/;
    next if m/^\#/;
    my @tmp = split /\s{1,}/, $_;
    $tmp[0]=~/s/\.pdb//;
    $tmp[0] = uc $tmp[0];
    $tmp[0] .= "\.pdb";

    $tmp[2]=~/s/\w\.\d+-(\w\.\d+),\w\.\d+-\w\.\d+/$1/;
    $tmp[3]=~/s/(\w\.\d+)-\w\.\d+,\w\.\d+-\w\.\d+/$1/;

    my @res1 = $tmp[2]=~/\w\.(d+)/;
    my @res2 = $tmp[3]=~/\w\.(d+)/;
    my @chn1 = $tmp[2]=~/(\w)\.\d+/;
    my @chn2 = $tmp[3]=~/(\w)\.\d+/;

    $tmp[2] = $chn1[0]."."($res1[0]+1);
    $tmp[3] = $chn2[0]."."($res2[0]-1);
}
```

```
        my $com1 = sprintf "~/RED/BinProg/newsf4.pl -r %s %s -a O3* P -i %s -
d",$tmp[2],$tmp[3],$tmp[0];
        my $output = qx{$com1 2>&1};
        printf GH "%s %s",$tmp[1],$output;
    }
    close GH; close FH;
    exit 0;
```

## Appendix I

### maxtoranggenerator2.pl

```
#!/usr/bin/perl
use Getopt::Long;
use Math::Trig;
use warnings;
use strict;

sub usage {
    printf STDERR "usage: perl maxtoranggenerator.pl [options]\n";
    printf STDERR "options: -input => input list file\n";
    printf STDERR "        -output => output euler angle file\n";
    printf STDERR "        -torang => list of torsion angles\n";
    printf STDERR "        -help\n";
    exit 0;
}

my ($file,$new,$sugta,$torang,$mod);
GetOptions(
    'help'           => \&usage,
    'input=s'        => \$file,
    'output=s'       => \$new,
    'sugta=s'        => \$sugta,
    'torang=s'       => \$torang,
    'mod+'           => \$mod,
) or die &usage;
die "I don't have a file to print to" unless $new;
open FH, $file || die "could not find file!:\n";
open GH, "> $new" || die "could not open file!:\n";

while (<FH>) {
    next if m/^\n$/;
    next if m/^\#/;
    my @tmp = split /\s{1,}/, $_;
    $tmp[0]=~/\.\pdb//;
    $tmp[0] = uc $tmp[0] unless $mod;
    $tmp[0] .= "\.pdb";
    my $com1;
    $com1 = sprintf "~/RED/BinProg/newlsf4.pl -r %s -i %s -t %s",$tmp[1],$tmp[0],$torang if
$torang;
    $com1 = sprintf "~/RED/BinProg/newlsf4.pl -r %s -i %s -u %s",$tmp[1],$tmp[0],$sugta if
$sugta;
    my $output = qx($com1 2>&1);
    if($output=~/\n/) { $output = join "$tmp[0]\n",(split /\n/, $output, -1),; }
```



```
        else { $output = $tmp[1]."\t".$tmp[0]."\t".$output; }
        if(!$tmp[2]) { $tmp[2] = 'na'; }
        printf GH "%s", $output;
    }
    close GH; close FH;
    exit 0;
```

## Appendix J

### rnaconformations2.pl

```
#!/usr/bin/perl

#####
#Program Name: rnaconformations.pl
#Description: ??????????????????
#Created 11/2008 by Maximillian H. Bailor
#####

use Getopt::Long;
#use lib '/local/home/bailor/BinProg'; ## for when I'm at work
use lib '/Users/maximillianbailor/RED/BinProg'; ## for when I'm at home
use newmax_basics;
use newmax_basics2;
use warnings;
use strict;

use constant pi => 4.0*atan2 1,1;
use constant tpsi => 1.0/sqrt(2.0*pi);

sub usage {
    print STDERR "usage: rnaconformations.pl [options] \n";
    print STDERR "options: --help \n";
    print STDERR "options: --bulgecutoff => bulges cut-off distance\n";
    print STDERR "options: --distcutoff => vdw's cut-off distance\n";
    print STDERR "options: --input => pdb file of reference helix\n";
    print STDERR "options: --junction => junction type to model\n";
    print STDERR "options: --linker => two residues linked by a bulge\n";
    print STDERR "options: --output => output file base name\n";
    print STDERR "options: --residues => used to speed up computation\n";
    print STDERR "options: --stepsize => stepsize to use in grid search\n";
    print STDERR "options: --template => pdb file of observation helix\n";
    exit 0;
}

## Global variables
my ($i,$j,$k) = (0,0,0);

## Set Defaults and implement GetOptions()
my ($fileout,@filein,$tmplt,$fileprint,$jnc,$refhlx,$ref,@strct,@pdsim,@res);
my (@chnid,@reslist,@linker,@linklist,@linkchnid,@link,$pt1,$pt2,$pts,@bulg);
my ($cnt,$eulerang,$neweulang,$step,$dist) = (0,[],[],5,2.0);
my ($add,$cnta,$cntb,$cntg) = (0,0,0,0);
```

```

GetOptions(
  'help' => \&usage,
  'addhelix+' => \&add,
  'bulgcutoff=f{1,3}' => \@bulg,
  'distcutoff=f' => \&dist,
  'input=s{2,}' => \@filein,
  'junction=i' => \&jnc,
  'linker=s{1,}' => \@linker,
  'output=s' => \&fileout,
  'residues=s{1,}' => \@res,
  'stepsize=i' => \&step,
  'template=s' => \&tmplt,
) or die &usage;

if(( $#filein == -1 ) || $step) { &usage; }
$bulg[0] = 999999 if ! $bulg[0];
my $atmtyp = "all";
## creating reference and observation helices
if(( $#res+1 > 0 ) {
  $k = 0; foreach my $f (@filein) { $strct[$k] = newmax_basics->new(pdb=>$f); $k++; }
  $k = 0; foreach my $f (@strct) { push @pdbsim,
newmax_basics::copypdb($f,$res[$k],$atmtyp); $k++; }
} else { &usage; }
if(@linker) {
  my @latmlist = substr($linker[0],2)<substr($linker[1],2)?('O3\','P'):( 'P','O3\');
  push @latmlist, substr($linker[2],2)<substr($linker[3],2)?('O3\','P'):( 'P','O3\') if $#linker
== 5;
  push @latmlist, substr($linker[4],2)<substr($linker[5],2)?('O3\','P'):( 'P','O3\') if $#linker
== 5;
  my @cnt = ($#linker == 5)?(0,1,1,2,2,0):(0,1);
  foreach my $i (@cnt) {
    push @latmlist, @pts,
newmax_basics::getcoord($strct[$i],substr($linker[$i],0,1),substr($linker[$i],2),$latmlist[$i]);
  }
}
#print $fileout,"\n";
my $st = time();
newmax_basics2::DistCutOff2(hlx=>\@pdbsim,distcutoff=>$dist,stepsize=>$step,fileout=>$fileout,li
nkcutoff=>$bulg[0],link=>$pts) unless $add;
newmax_basics2::DistCutOff3(hlx=>\@pdbsim,distcutoff=>$dist,stepsize=>$step,fileout=>$fileout,li
nkcutoff=>\@bulg,link=>$pts) if $add;
my $et = time();
printf "simulation time: %.2f (sec)\t%.2f (min)\t%.2f (hrs)\n",($et-$st),($et-$st)/60,($et-
$st)/60/60;
exit 0;

#####
## local subroutines
#####

sub residuelist {
  my $reslst = shift;
  my ($stp,$lst,$i,@list,@chn);
  my ($outreslst,$outchnlst) = ([],[]);

```

```

@list = ($reslst =~ /\w\.\d+\D\w\.\d+/gm);
if($1) { @chn = ($reslst =~ /(\w)\.\d+\D(\w)\.\d+/gm); }
@list = ($reslst =~ /(\d+)\D(\d+)/gm) unless $1;

$i = 0;
while($list[$i]) {
    $stp = $list[$i];
    $lst = $list[$i+1];
    while($stp <= $lst) {
        push @{$outreslst}, $stp;
        if(defined $chn[$i]) { push @{$outchnlst}, $chn[$i]; }
        else { push @{$outchnlst}, "all"; }
        $stp++;
    } $i += 2;
} return ($outchnlst,$outreslst);
}

```