USING RNA BACKBONE TORSIONS TO STUDY RNA STRUCTURE

by

DAVID IAN MICALLEF

A thesis submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Microbiology and Molecular Genetics and

The Graduate School of Biomedical Sciences

University of Medicine and Dentistry of New Jersey

Written under the direction of

Dr. Helen Berman

And approved by

_____

_____

_____

New Brunswick, New Jersey

[January, 2010]

ABSTRACT OF THE THESIS

USING RNA BACKBONE TORSIONS TO STUDY RNA STRUCTURE

By DAVID IAN MICALLEF

Thesis Director:

Dr. Helen Berman

Ribonucleic Acid (RNA) is an important cellular macromolecule vital to most if not all life on Earth. RNA has many different roles in the cell, most notably as the intermediary molecule that transfers genetic information from DNA to protein in translation. Recently, additional functions of RNA have been elucidated more clearly, such as catalyzing chemical reactions and regulating gene expression. These exciting new findings have shined a scientific spotlight on the field of RNA structure in order to better understand how the once mundane polynucleotide acts in such myriad ways.

An important factor in RNA's versatile nature is the inherent variation in its chemical structure. The hydroxyl group present on the ribose sugar of a ribonucleic acid makes the corresponding polynucleotide capable of chemical reaction, with itself or with other molecules in the cell. This hyper-reactivity allows RNA to form substantially unique structures, from the hammerhead ribozyme's helical shape from which it takes its name, to the L-shaped conformation common to all transfer RNAs. The problem at hand is thus to study RNA structure and determine if any new patterns can be discovered.

The work presented here centered on a collaborative effort to define a set of conformations common to two-nucleotide long sequences of RNA found in structures from the Protein Data Bank (PDB). This work contributed by clustering RNA di-nucleotides by their torsion angle space using a Fast Fourier averaging technique proven to be effective in clustering nucleotide structure. Each group in the collaboration used different methodologies to analyze the same RNA structural data, and yet found similar results. The collaboration ultimately produced a set of 46 consensus conformations defined by the seven dihedral angles of the sugar-to-sugar unit in a di-nucleotide RNA sequence.

To utilize this new set of RNA di-nucleotide conformations, a software tool was designed and developed to automatically assign the conformation nomenclature to input RNA structure. The program was successfully tested on the pilot study data. A test study was performed on a unique set of RNA structures. The results of this study demonstrated that the consensus conformation set can in fact be used to classify RNA structure.

Preface

The work presented here is an academic venture into the field of RNA structure. The background chapters are not meant to be of textbook quality, yet a concise synopsis of my understanding of the field over my years of graduate research at Rutgers University. The clustering of RNA di-nucleotides by torsion angles could never have been achieved without the mentoring of Drs. Helen Berman and Bohdan Schneider, for whom I will be forever grateful. As a computer science undergraduate student who branched into the new (at the time) field called computational biology, delving into RNA structure and biology has been a significant challenge. I hope to have come out a better person and a better scientist from this work and look forward to my next step in this road called life.

Acknowledgments

I would like to express my sincere appreciation to my advisor, Dr. Helen Berman. Helen has been the most supportive, understanding, and helpful advisor a graduate student can ever hope to have.

I also would like to expressly thank Dr. Bohdan Schneider, whose expertise on nucleotide torsion analysis provided the backbone (pun intended) of the work presented here.

Next, I wish to thank everyone at the PDB at Rutgers, starting with Dr. John Westbrook who lent his helpful hand on countless problems I faced over the years in my research. Thanks to Drs. Bill McLaughlin and Andrew Napoli for their guidance as graduate students. Thanks to Dr. Huanwang Yang for his help in x-ray crystallography and RNA. Thanks to Dr. Cathy Lawson for her x-ray crystallography help. Thanks to Dr. Andrei Kouranov, Dr. Wendy Tao, Raship Shah, and Raul Sala for being great CABM roommates and putting up with my thesis writing phase. Thanks to Bill Abbott for his reliable support and friendship. And thanks to Jim Croker, Lew Fernandez, James Chun, and Chris Suleski for their technical support.

A heartfelt thank you goes to my parents, George and Josephine Micallef, and my sister, Michelle. I could never have completed this work without your support.

And finally, thank you to my fiancée and my best friend, Kimberly Rivera. You have kept me going all these years. I look forward to what lies ahead of us as we continue our path together.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Ribonucleic acid (RNA) is an integral molecule involved in myriad cellular processes in every plant, animal, and microbe on Earth. Widely considered to have a leading role in the early evolution of life, the range of RNA's roles can be seen in genetics, health, disease, and the development of organisms. In addition, current research in RNA continues to discover new RNA molecules possessing novel biological functions, demonstrating that RNA plays far more roles than originally believed.

Statement of Problem

What gives RNA its versatility is its inherently flexible chemical structure. RNA is a much looser molecule compared to its stable, rigid cousin deoxyribose nucleic acid (DNA), and thus has proven difficult to understand at the structural level. This study aims to add to the RNA structure knowledge base in a small yet significant way.

Overview of the Study

The current study began with a partnership between the Berman lab at Rutgers and the Richardson lab at Duke, as both groups had separately worked on RNA backbone analysis and discovered similar results. The two research groups executed an across-the-database study of RNA structures, using structures from the public structural databases, namely the Nucleic Acid Data Bank (NDB) (Berman 1992) and the Protein Data Bank (PDB) (Berman 2000). The collaboration combined the two groups' independently developed approaches of RNA conformation analysis (Murray 2003; Schneider 2004) to cluster RNA conformations by their chemical torsion angles. From this collaboration, a set of consensus

RNA conformational families were determined and presented to the newly created RNA Ontology Consortium (ROC) for the benefit of the scientific community at large (Richardson 2008).

With this new RNA conformation consensus set in hand, a software tool was designed and developed for this study in order to automatically assign the novel classification nomenclature to input RNA structures. The program, named DiCAT for Di-nucleotide Conformation Assignment Tool, uses a decision tree logic based on the conformation torsion angle set to label each di-nucleotide pair in the RNA input with the appropriate consensus set name. The DiCAT tool was successfully tested on the original dataset used to produce the consensus conformational classes. Finally, an initial application of the DiCAT program to a set of RNA-protein structures demonstrated the utility of the conformation nomenclature to better understand RNA structure.

Extension of this study suggests further use of the DiCAT tool to find conformational motifs among RNA sequences and structures, such as those in RNA-protein binding sites. Such knowledge can develop a better understanding of RNA structure, with applications from aiding in drug design for RNA-binding proteins, to serving as a structural genomics prediction tool for proteins of unknown function that share a similar structural motif.

Chapter 2: Literature Review

Background and Significance

RNA Biological Overview

Research into ribonucleic acids, long considered simply the middle man in the central dogma of molecular biology (Figure 1), has exploded over the past decade, as the scientific community discovers the role of this molecular building block in genetics, health, disease, and the development of organisms (DeJong 2002; Colegrove-Otero 2005). The inherent variation of the molecule, modeled as the daily currency of the cell versus the vaulted gold standard that is DNA, allows for RNA to wear many hats. It can provide structural stability to large RNA-protein complexes such as the ribosome, take on very specific functions with highly conserved tertiary structures such as in tRNA and catalytic ribozymes, and finally, serve its most famous role as the interpreter that allows genes in DNA to be converted into viable proteins. Additionally, the current hot topic in RNA research involves the relatively recent discovery of small interfering RNA (siRNA) and micro RNA (miRNA), which both play exciting enzymatic roles in post-transcriptional gene regulation. RNA is clearly involved in the vitality of all life on earth, with its hand in many vital aspects of cellular life, and so learning as much as we can about its structure and how it interacts with protein in the cell is of utmost scientific importance.

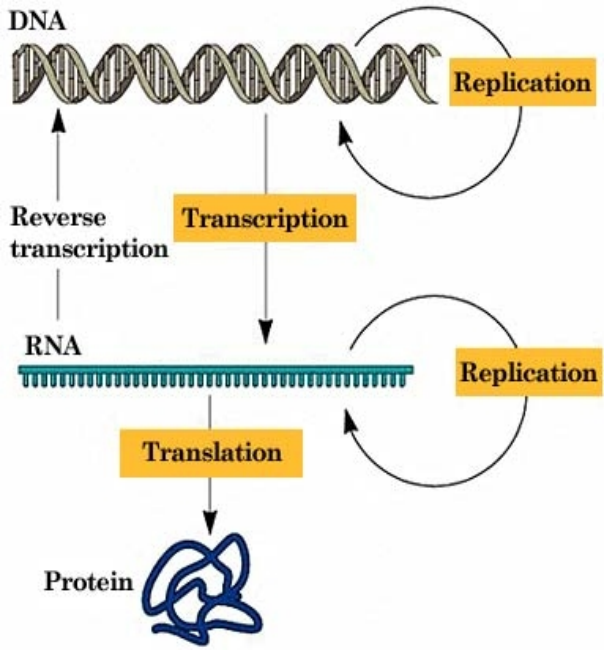


Figure 1 – The Central Dogma of Molecular Biology

(http://www.bioinfocreator.com/images/central_dogma.gif)

RNA Structure

RNA structure differs from DNA in two small, but deeply profound ways: Uracil replaces thymine bases in RNA sequences (Figure 2). Uracil lacks the methyl group of thymine at the C5 position, but maintains the same hydrogen bonding with adenine. Methylation is a cellular means of protection of DNA, DNA bases are methylated to ward off nucleases, a function not needed by RNA. Also, the loss of the hydrophobic moiety allows uracil to hydrogen bond in non-Watson-Crick fashion, giving RNA increased variation.

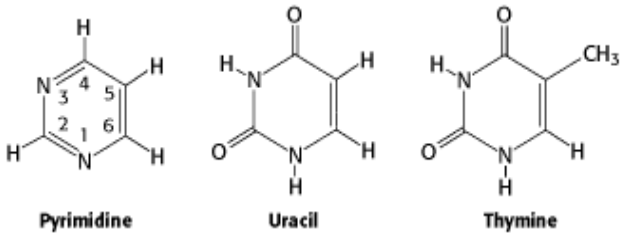


Figure 2 – Uracil and Thymine

The sugar group in RNA is ribose, which has a hydroxyl group on the 2' position, as opposed to the deoxyribose of DNA that has a lone hydrogen atom at the 2’ locus (Figure 3). Sugar atoms are numbered with primes to distinguish them from atoms in bases.

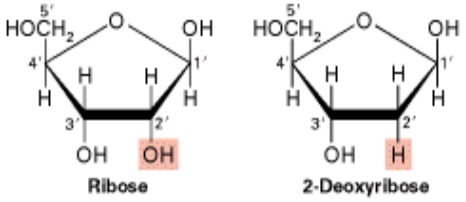


Figure 3 – Ribose and Deoxyribose

The presence of that one hydroxyl group in its sugar gives RNA vastly different chemical properties from DNA, explaining how it can be such a versatile molecule with relatively high turnover. The hydroxyl increases the chemical reactivity of RNA, and also makes the macromolecule more susceptible to degradation and hydrolysis. RNA often reacts with itself, as it can form various secondary structures such as loops, bulges, and pseudoknots (Figure 4).

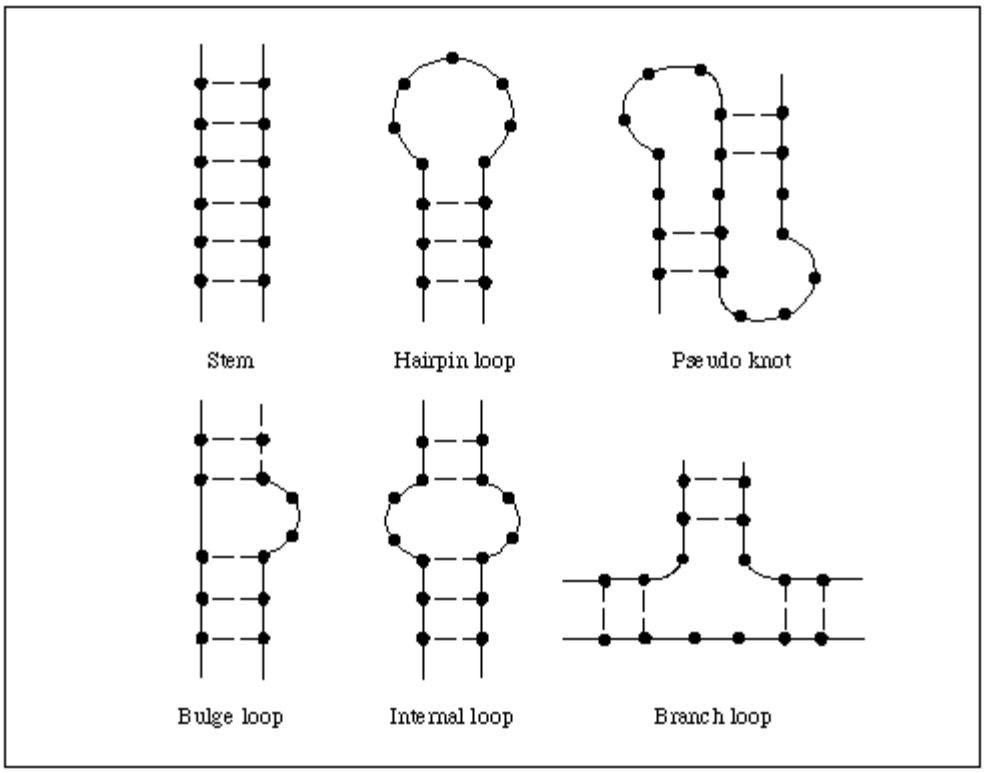


Figure 4 – RNA Secondary Structure Motifs

A hairpin loop (or terminal loop) consists of a double-stranded stem and a single-stranded loop that bridges one end of the stem. Hairpins are building blocks of complex RNA structures. Hairpins also serve as distinct sites for protein recognition, by presenting their loop regions for base pairing with other RNAs, or, as in mRNA, by creating an energetic barrier resistant to rapid read-through by ribosomes (Shen 1995).

An internal loop is an interruption in a double strand caused by nucleotides on both strands that cannot form Watson-Crick or wobble G●U base pairs. Internal loops provide recognition sites for RNA-RNA and RNA-protein binding, and are also sites of ribozyme cleavage. The unpaired bases in internal loops are not unpaired and free, as the canonical

2D depiction suggests. These loop regions are often highly structured, as bases within the loop form non-Watson-Crick base pairs to stabilize the loop.

Bulges occur in double-stranded RNAs, where one strand of the duplex has a sequence of unpaired nucleotides. They too serve as protein recognition sites, such as in the binding of HIV-1 Tat protein to the bulge in TAR RNA, where a single arginine residue in Tat recognizes the TAR bulge (Puglisi 1992).

RNA pseudoknots are tertiary structural elements that result when a secondary structural loop base pairs with a complementary sequence outside the loop. Pseudoknots are involved in such processes as RNA self-splicing, translational autoregulation, and ribosomal frameshifting (ten Dam 1992). The best understood pseudoknot motif is the simple hairpin-type, where a hairpin loop pairs with a complementary sequence to form a stem directly adjacent to a hairpin stem (Figure 5).

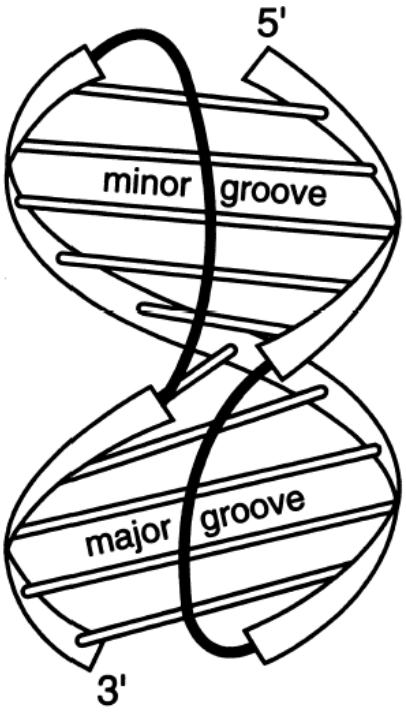


Figure 5 – Hairpin-Type Pseudoknot with Directly Adjacent Helical Stems

RNA regularly forms double helices, with base pairs inclined from the center of the double stranded RNA, akin to A-type DNA (Figure 6). Hence, double stranded RNA is called A-RNA. The RNA duplex is dominant in RNA stems, accounting for as much as 50% of the residues in the average non-mRNA (Moore 1999).

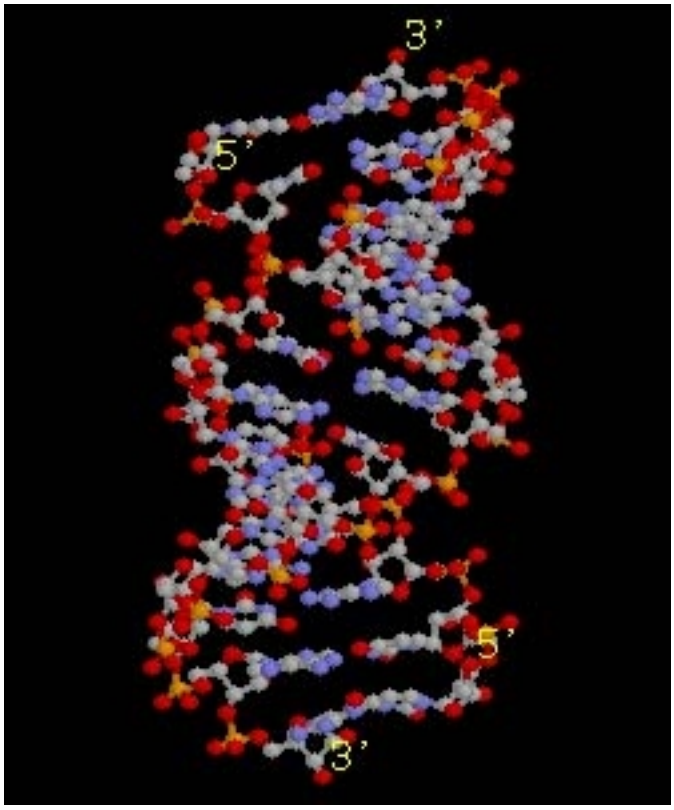


Figure 6 – A-RNA Double Helix

Duplex RNA is conformationally rigid, with a narrow and deep major groove and a wide, shallow minor groove. A-RNA base pairs are only slightly twisted, and when observed down the helical axis the base pairs are inclined to and displaced from the helix axis (Figure 7).

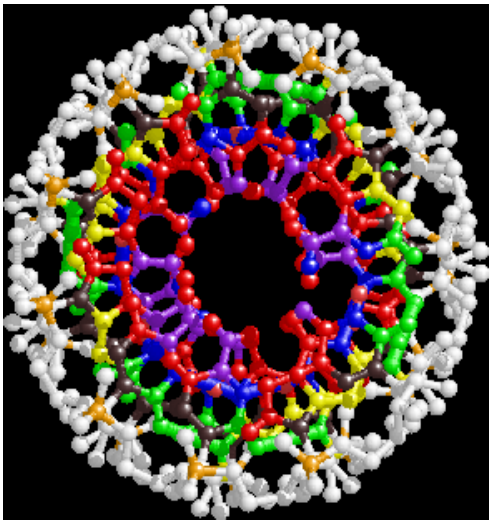


Figure 7 – View Down A-RNA Double Helix

The energy involved in the formation of RNA secondary structure is substantially larger than that of RNA tertiary structure (Tinoco 1999). Thus, RNA secondary structure is usually assembled prior to tertiary structure. By contrast, the formation of protein secondary and tertiary structures is generally intimately linked, making protein folding the result of a complex balance of the energies associated with secondary and tertiary structure formation. The strongly hierarchical nature of RNA folding implies that RNA secondary structure (helices, bulges, loops, and junctions) is generally preformed, and does not require the stabilizing presence of RNA tertiary structure or protein. This suggests that RNA-binding proteins recognize RNA nucleotides in the context of stable secondary structures (Leulliot 2001). As will be discussed below, RNA and protein interact via a mutual induced fit partnership. On the RNA end, this induced fit does not imply the disruption of RNA secondary structure, but the reorganization of local elements of secondary structure, the formation of a defined structure for disordered single-stranded elements, and the stabilization of a defined three-dimensional RNA conformation.

The ribose sugar of RNA is more rigid than the deoxyribose of DNA, and is most commonly found in the C3'-endo pucker when found in helical form. Ribose's 2'OH group would otherwise sterically clash with the base attached to C1' of the sugar. This rigidity translates into the conformational rigidity of RNA double helices, which can only take on the A-form, compared to the relatively flexible polymorphism of the DNA double helix (e.g. A-, B-, and Z-form helices). RNA ribose can also take on the C2'-endo pucker, but the C2' sugar pucker is much more rare and found in RNA nucleotides that partake in bulges, loops, or other non-helical structures. Figure 8 below depicts the C3' and C2' RNA ribose sugar puckers.

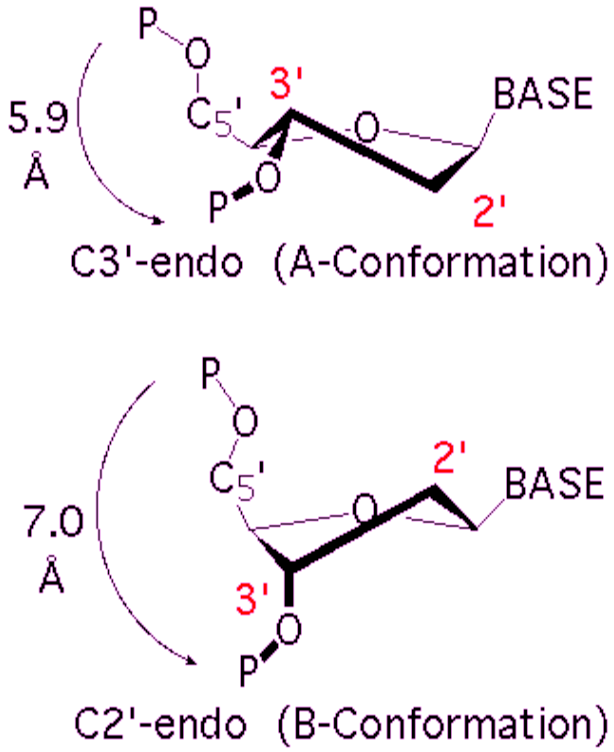


Figure 8 – Ribose C3'-Endo and C2'-Endo Pucker Conformations

A-RNA helices are often interrupted by bulges and internal loops, which play roles in the specific function of RNA double helices. The role of RNA helices alone in protein binding is unclear, as A-RNA has a deep, inaccessible major groove and shallow minor groove. However, the recent structure of a slight variant of the A-RNA helix, named, A'-RNA, showed a conformational difference that created a wider major groove in the double helix, suggesting a role in RNA-protein interactions (Tanaka 1999).

The 2'-OH group of RNA also can play a role in protein interactions, as it can form stabilizing hydrogen bonds, either with protein residues or RNA phosphates. Additionally, the extra hydroxyl group can aid the protein in discriminating between DNA and RNA when searching for its appropriate substrate (Antson 2000).

RNA Backbone Structure

There are several chemical properties of RNA that can be used to classify the molecule's conformation. One can analyze RNA by its primary sequence; its secondary structure, such as base pairing, loops, bulges, etc.; base and base pair variation (twist, buckle, roll, etc.); and finally, the backbone and glycosidic bond torsion angles and their correlated variation. For this study, the focus will be on the latter for several reasons. With the continual growth of RNA-containing structures in the NDB (Murthy 2003), there is a vast array of three-dimensional structures to compare with each other. For each nucleotide in every RNA structure in the NDB, we can represent its local geometry as a set of flexible parameters, the torsion angles, something that could not be done prior without such a large and reliable database. Thus we can compare all RNA on a basic structural level, classify common

modes of such structure, and then discover what specific biological functions are prevalent among such structural motifs.

A nucleotide has seven degrees of torsional freedom. The phosphodiester backbone has six variable angles, designated alpha, beta, gamma, delta, epsilon, and zeta; the glycosidic bond angle is named chi (Figure 9).

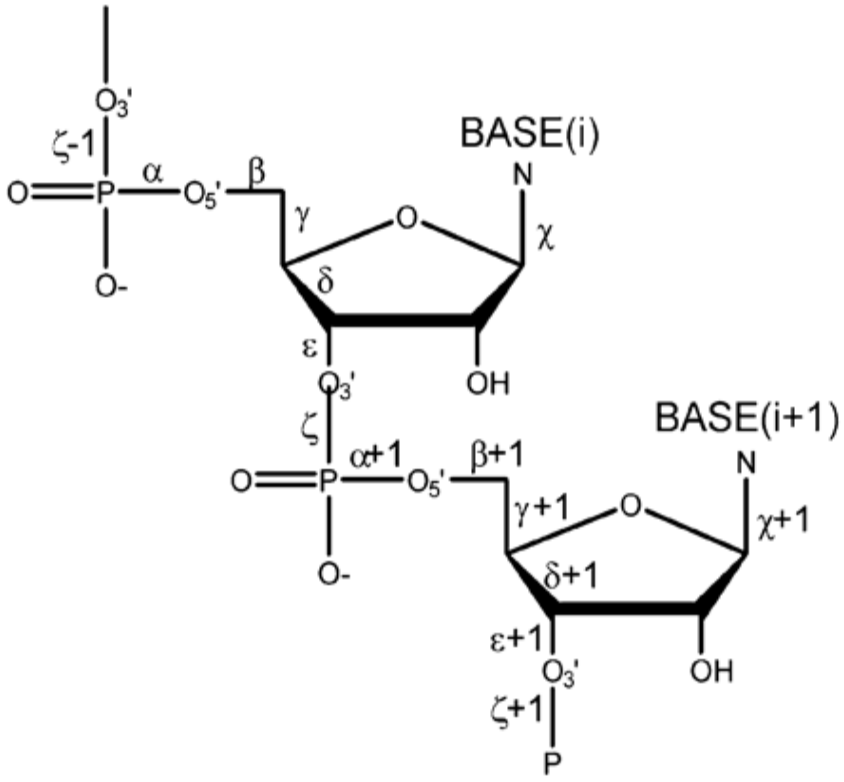


Figure 9 – Dinucleotide Fragment with Backbone and Glycosidic Bond Angle Names

These seven angles can be used to classify a section of an RNA backbone on the nucleotide level. Steric considerations alone dictate that the backbone angles are restricted to discrete ranges (Olson 1982; Sundaralingam 1969), and are accordingly not free to adopt any value

between 0 and 360°. In fact, they have highly correlated values, as can be shown graphically using the RNA in the 50S ribosomal subunit (Figure 10) (Ban 2000; Schneider 2004).

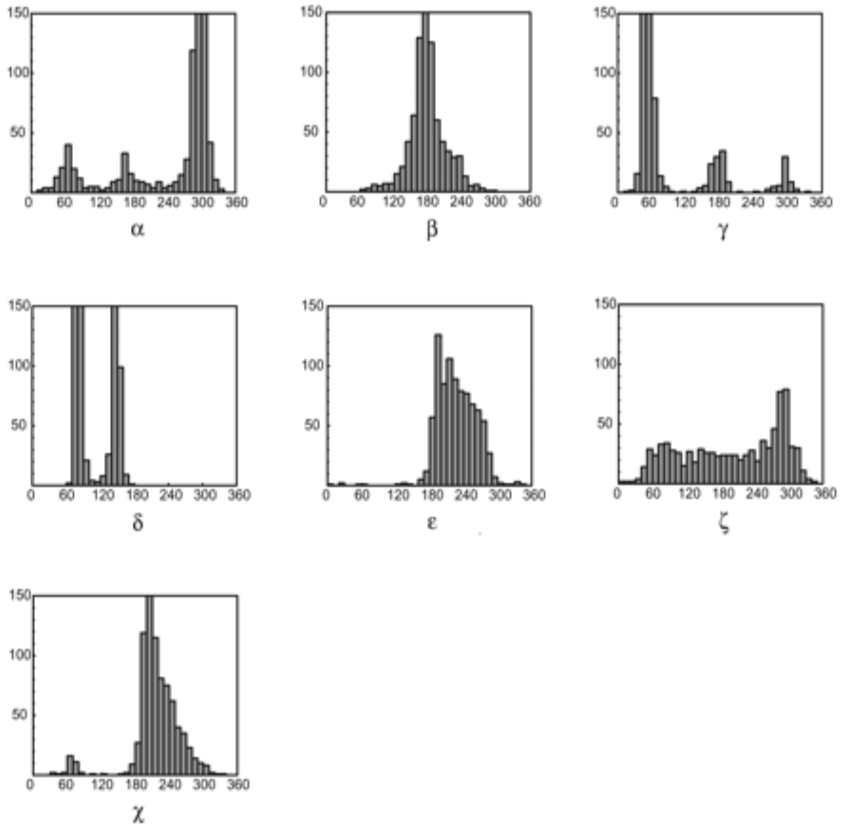


Figure 10 – Histograms of the Seven Torsion Angles of the 50S Ribosomal Subunit (NDB Code RR0033) (Schneider 2004)

There are a number of well-established correlations involving pairs of these backbone torsion angles, as well as sugar pucker and glycosidic angle. These have been observed in mononucleosides and nucleotides (inherently more flexible in solution as well as subject to packing forces in a crystal), as well as in oligonucleotides (Schneider 1997; Packer 1998).

Such correlations show that atomic variation in oligo- and polynucleotides can be classified. In general, such correlations are due to the reduction of non-bonded contacts that occur with particular conformations.

However, one of the leading problems involved with studying the torsion conformational space of nucleotides is the inherent multidimensionality of the data. Studying one nucleotide presents seven angles at a time, while looking at two nucleotides doubles that to 14 parameters. Early work indicated that a key part of the conformational behavior of the RNA backbone lies in the two torsion angles involved at the phosphodiester link, namely zeta of nucleotide *i* and alpha of nucleotide *i*+1. Empirical analysis of a very limited set of crystal data revealed that the behavior of the phosphodiester link leads to seven major conformational classes (Kim 1973). This insight was used to focus on the correlations between these two angles and the other backbone torsion angles (Schneider 2004), and our current collaborative study has expanded to include several other combinations of three torsion angles to supplement the phosphodiester-centric study.

RNA-Protein Interactions

RNA-binding proteins have a central role in many aspects of genetic activity within an organism, such as regulation, transcription, and cell development. Thus, it is extremely important to examine the nature of complexes that are formed between proteins and nucleic acids, as they form the basis of our understanding of how these processes take place. Over the past decade, the world has witnessed a great expansion in the determination of high-quality structures of nucleic acid-binding proteins. As a result, the number of such structures has seen a constant increase in the PDB and the NDB. These structures have provided valuable insight into the stereochemical principles of binding, including how particular base sequences are recognized and how the RNA structure is quite often stabilized upon protein binding.

When RNA and protein bind each other, recognition occurs almost invariably by "induced fit" rather than by rigid "lock-and-key" docking (Varani 1997; Williamson 2000). The protein, the RNA, and sometimes both undergo large conformational changes, leading to large changes in the local as well as global properties of the interacting components. RNA-binding by protein requires not only a cluster of specific nucleotides for chemical recognition, but also cofolding of the RNA and peptide between each other (Frankel 1998). The RNA refolds itself around an unstructured peptide, inducing conformational changes in the RNA and protein and locking both ligands into more rigid conformations. Removing the protein component of an RNA-protein structure will be instructive, as RNA adapts very unusual conformations when bound to proteins. These conformations, often characterized by bases splayed out into the solvent, would not be stable unless the bases were deeply buried against the protein surface (Figure 11) (Leulliot 2001). Thus, mining the structural

databases for RNA when in complex with protein will provide a glimpse into these novel RNA conformations.

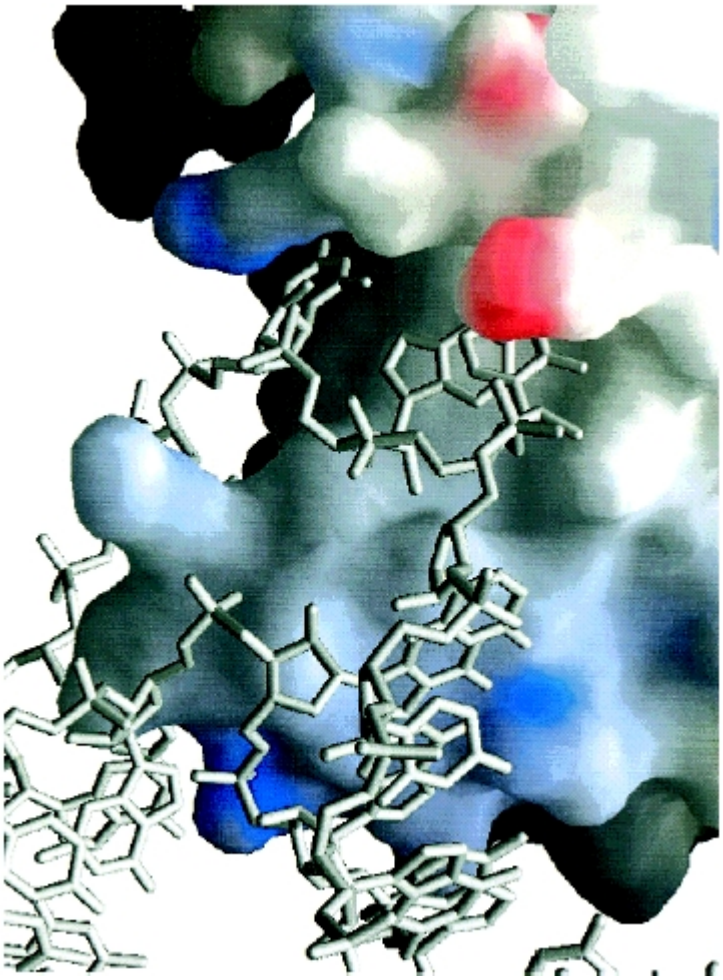


Figure 11 – RNA Bound to Protein Adopts Novel Conformations

(Leulliot 2001)

RNA tertiary structure is known to be malleable, such as structural modification in response to protein binding. For example, several ribozymes are inactive until protein cofactors bind and stabilize the active tertiary conformation of the enzyme (Weeks 1996;

Caprara 1996). Similarly, poorly ordered loops between secondary structural elements of RNA-binding proteins are often remodeled upon ligand binding. Large-scale conformational changes are observed only when either molecule has several independent subunits, such as multiple protein domains or RNA helices anchored at multiple junctions. Since the relative orientation of these subunits is often weakly defined in the free molecule (Crowder 1999), both the RNA and protein molecules must be easily rearranged upon their union. There are several examples of RNA-protein recognition by induced fit, some of which will be described below.

RNA-Binding Domain (RBD)

The RNA-binding domain (RBD), or RNA-recognition motif (RRM), is the most common structural class of protein motifs that bind single-stranded RNA (Hall 2002), and one of the most common protein folds in eukaryotic genomes (Varani 1998). The canonical α/β fold of the RBD is compact and globular, with a four-strand antiparallel β-sheet packed against two α-helices. One noteworthy feature of RBD-containing proteins is the frequent presence of multiple non-identical RBD subunits. When these proteins bind single-stranded RNAs, the binding of any one RBD is weaker and not as specific as that of the total set of RBDs in the functional protein. Thus the domains function to work greater together than the sum of their parts individually. Most RBDs use the solvent-exposed surface of the β-sheet as an RNA-binding platform. RNA binds to this side of β-sheet, while the other side is buried inside the protein domain by the two α-helices connecting the β strands. However, there is a surprising diversity of binding mechanisms among RBD-containing proteins. The preference for RNA-binding by the more stable β-sheet over the usually flexible α-helix

could be the nature of its ssRNA target, which is extended and flexible itself (Antson 2000).

Proteins containing the RBD are involved in processing of pre-rRNA and -mRNA in the nucleus. More RBDs, as well as novel RNA-binding proteins, that recognize single-stranded RNA sequences are certain to be found at the exon-exon junction, where many proteins are deposited or recruited after splicing the pre-mRNA (Le Hir 2001). Other RNA-binding proteins will likely be associated with cytoplasmic mRNA and small RNAs (Argaman 2001). Prediction of proteins with the RBD fold poses a problem because the RBD α/β fold is a protein superfamily (Orengo 1993), and thus there are many proteins with the fold that do not bind RNA. This suggests that there is more involved to the mechanism of RNA-binding by proteins than simply containing the RBD motif.

Double Stranded RNA Binding Motif (dsRBM)

The double stranded RNA binding motif (dsRBM) is the most common protein domain that binds RNA duplexes (Hall 2002). This small globular domain folds to form a three-stranded antiparallel β-sheet with two α-helices positioned on one side. It is found in virtually all organisms (Fierro-Monti 2000), and is best observed in the enzymes adenosine deaminase (ADAR1 and ADAR2), in the protein kinase PKR, and in the dsRNA-specific endoribonuclease RNase III. In the crystal structure of a single dsRBM bound to a 16-base pair RNA duplex, the RNA interacted with the α-helical side of the protein, with the β-sheet surface free and exposed (Ryter 1998). This is opposite to the arrangement between RNA and the secondary structures of the RBD motif.

A source of novel double stranded RNA-binding proteins is likely to evolve from the world of microRNAs (miRNAs), which form imperfect duplexes at specific mRNA sites to repress translation (Lagos-Quintana 2001; Ruvkun 2001). miRNA duplexes are distinct from the perfect duplex structures of RNA interference (RNAi), which trigger mRNA degradation (Bass 2000). However, both miRNA and RNAi duplexes are processed by Dicer nuclease, which contains a single dsRBM (Grishok 2001). The small duplex regions formed by miRNAs and their mRNA targets are certainly bound by proteins (Hall 2002). Because these duplexes are imperfect, their stability will be low. Thus, a bound protein will increase their stability, with the deformed RNA double helix allowing proteins to make sequence-specific interactions. This supports the induced fit partnership model between RNA and protein.

RNA-Protein Structure Examples

As of September 16, 2009, there are 4,373 total structures in the NDB, 1,589 of which contain RNA. 1,148 were determined by X-ray crystallography; 441 by NMR. Among these, 739 are specifically RNA-protein complexes. 687 were determined by X-ray crystallography; 52 by NMR. In most of these structures, the RNA strands are folded into secondary structures. Hairpin loops are the simplest such motif (Valegard 1994; Oubridge 1994; Price 1998). More complex RNA structures include tRNA molecules (Rould 1991; Cusack 1996; Goldgur 1997; Cusack 1998; Sankaranarayanan 1999) and the hepatitis delta virus ribozyme (Ferre-D'Amare 1998). The largest structures, and perhaps most complex, are those of the ribosomal subunits, large (Ban 2000) and small (Tocilj 1999; Wimberly 2000), and the complete ribosome itself (Yusupova 2001). The tertiary structures of the RNA in these examples are primarily formed prior to protein binding. However, as

described above, the RNA's secondary structure and point of impact often change upon binding. The protein acts as an influential stabilizing force in the RNA-protein interactions (Antson 2000).

Examples of Induced Fit RNA-Protein Interactions

HIV-1 Virus Tat-TAR Interaction

Transcriptional elongation of the HIV-1 promoter is regulated through a mechanism dependent on the recognition of an RNA regulatory element, the transactivator response element (TAR) RNA, by the virally encoded transactivator protein Tat (Jones 1994; Gait 1993). The bulge region of the TAR RNA binding site changes its conformation upon protein binding, as observed in studies using peptide derivatives of Tat (Long 1995; Aboul-ela 1995), and a single arginine amino acid (Puglisi 1992). The free and bound structures of HIV-1 TAR RNA represent two different ways of accommodating three bulged residues inside an RNA double helix (Figure 12).

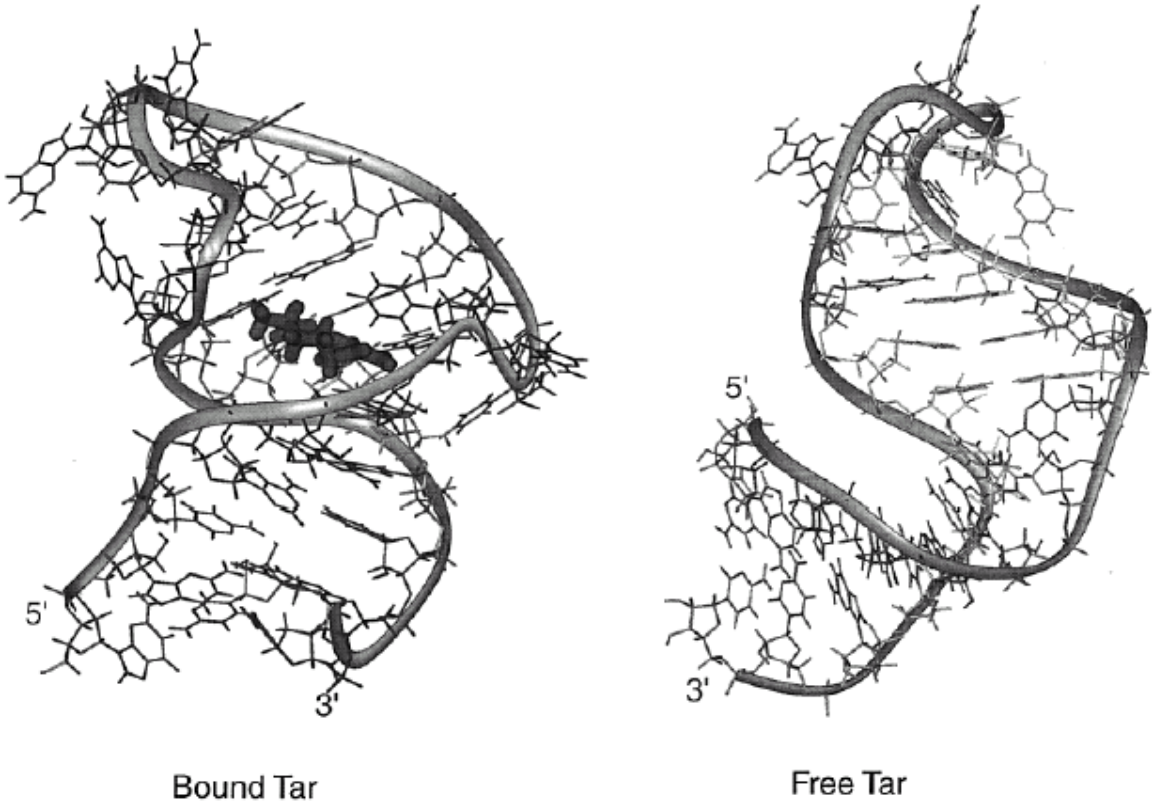


Figure 12 – HIV-1 TAR RNA Conformational Change upon Binding

(Leulliot 2001)

In the free RNA, U23 is looped out of the helix, and the two other bulged nucleobases assume continuous stacking interactions (Aboul-ela 1996), inducing a kink in the RNA helix (Riordan 1992). In the bound RNA, a uracil interacts instead with an arginine through hydrogen bonding and electrostatic interactions (Aboul-ela 1995). This conformational change relieves the helical twist and rise induced by the continuous stacking of the bulged residues and straightens the double helix. Studies of global properties of TAR have demonstrated that the bend of the RNA also changes upon ligand binding (Zacharias 1995).

HIV-1 Virus Rev-RRE Interaction

The Rev-RRE complex controls the export of HIV-1 viral RNA from the nuclei of infected cells (Fischer 1995; Stutz 1995). Peptides derived from Rev protein, corresponding to the arginine-rich RNA binding domain of Rev, are only partially helical when free of RNA, but stabilize their α-helical conformation upon binding to the Rev-response element (RRE) RNA (Tan 1994; Battiste 1996). Furthermore, the peptides took on different folds when bound to different aptamers similar to RRE (Figure 13) (Ye 1996).

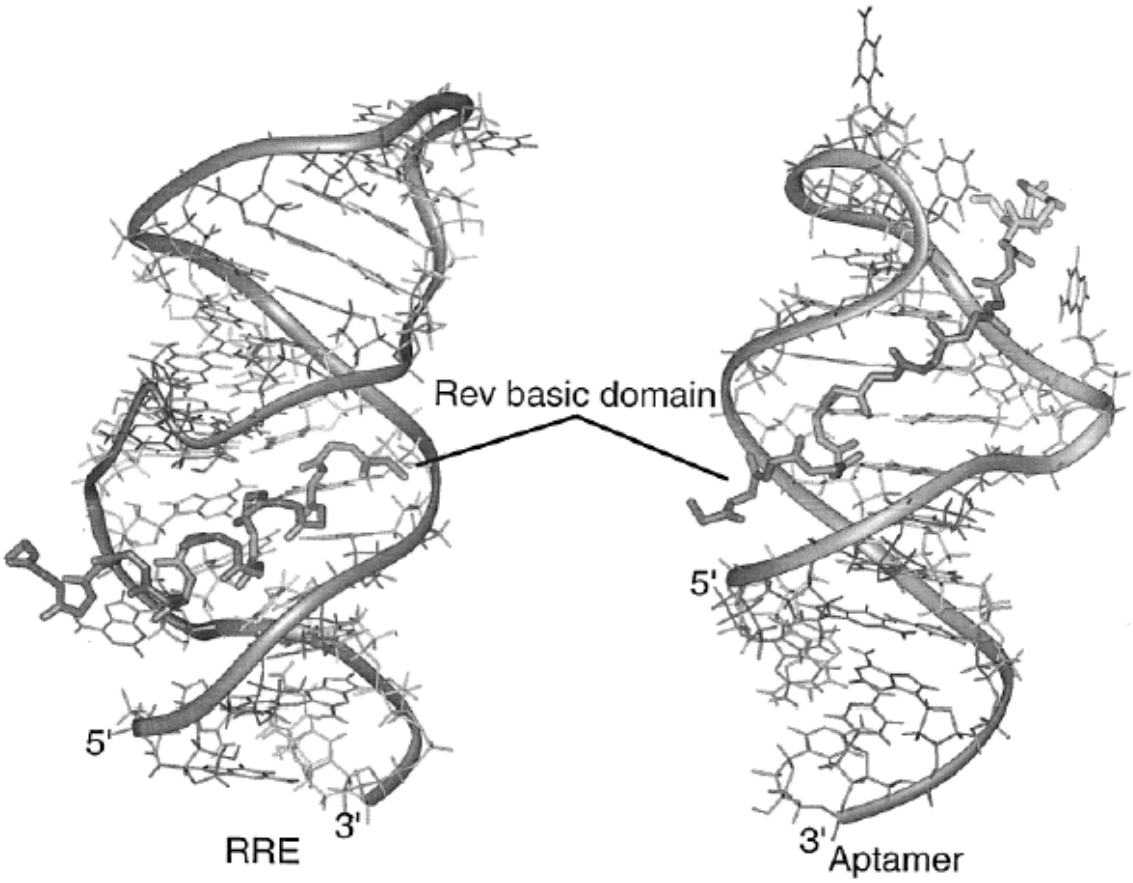


Figure 13 – Protein has Different Conformations when Binding Different RNA Aptamers (Leulliot 2004)

The conformation of the RRE RNA also exhibits different conformations when free and bound to protein. First, upon protein binding, the major groove becomes open compared to that of regular A-form RNA. Second, in free RNA, the non-canonical G●G base pair adopts a symmetrical G(*anti*)●G(*syn*) conformation (Peterson 1994). After Rev has bound, the *syn* guanine flips to an *anti* conformation, resulting in a new base pair and the local reversal of the backbone chain (Peterson 1996). The variation of a uracil nucleotide looped out of the helix allows for this conformational change. This base can be replaced by a

propyl linker, suggesting that the uracil specifically has no effect on Rev binding, but is vital to RRE's mobility in accepting its protein ligand. The G●G base pair also does not interact with Rev, as its importance is in widening the major groove to ease the protein into the RNA.

Examples of Proteins with RBD

Sex-Lethal Protein

The sex-lethal protein from *D. melanogaster* is involved in the sex-determination process by binding tightly to a U-rich segment of an intron of the *transformer* pre-mRNA, thereby regulating alternative splicing of the gene (Sosnowski 1989; Inoue 1990). The protein contains two RBDs, arranged in tandem and separated by a 10-residue linker. The two domains work together to bind RNA stronger than as individuals (Kanaar 1995). Free of RNA, the linker is disordered and the two RBDs do not interact with each other (Crowder 1999). When bound to RNA, the two domains form a V-shaped cavity lined by their β-sheet surfaces (Figure 14) (Handa 1999).

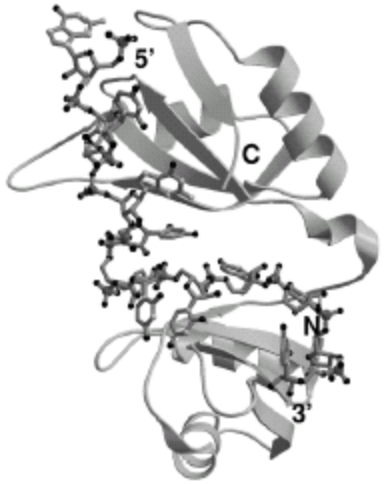


Figure 14 – Sex-Lethal Protein-ssRNA Complex. 5' and 3' of RNA

C and N termini of protein are labeled

(Antson 2000)

In this structure, nine nucleotides bind to the protein cavity, with a sharp turn of the RNA coinciding with the bottom of the cavity. This turn is stabilized by protein-RNA interactions as well as three hydrogen bonds formed between phosphates and 2' sugar hydroxyls in the RNA backbone within the turn. Three other nucleotides stabilize the binding with 2'-OH hydrogen bonding to protein residues. The importance of the ribose sugar thus plays an important role in discrimination between DNA and RNA for this protein. Other protein-RNA interactions include hydrogen bonds and salt bridges between backbone phosphates and the protein, and further hydrogen bonding and stacking interactions between bases and residues. The conformation of the RNA backbone is largely dictated by these interactions, with the backbone torsion angles significantly different from A-RNA (Dock-Bregeon 1988).

U1A Protein

The first structures published of RBD-RNA complexes were that of human U1A protein bound to a stem loop structure derived from U1 snRNA (Oubridge 1994) and an internal loop regulatory element (Allain 1996). Again, both the protein and RNA exhibit conformational change upon binding. The protein repositions the C-terminal helix away from the RNA-binding surface (Avis 1996), while the single-stranded RNA's bases flip inside out to interact with the protein instead of other RNA bases (Figure 15) (Gubser 1996).

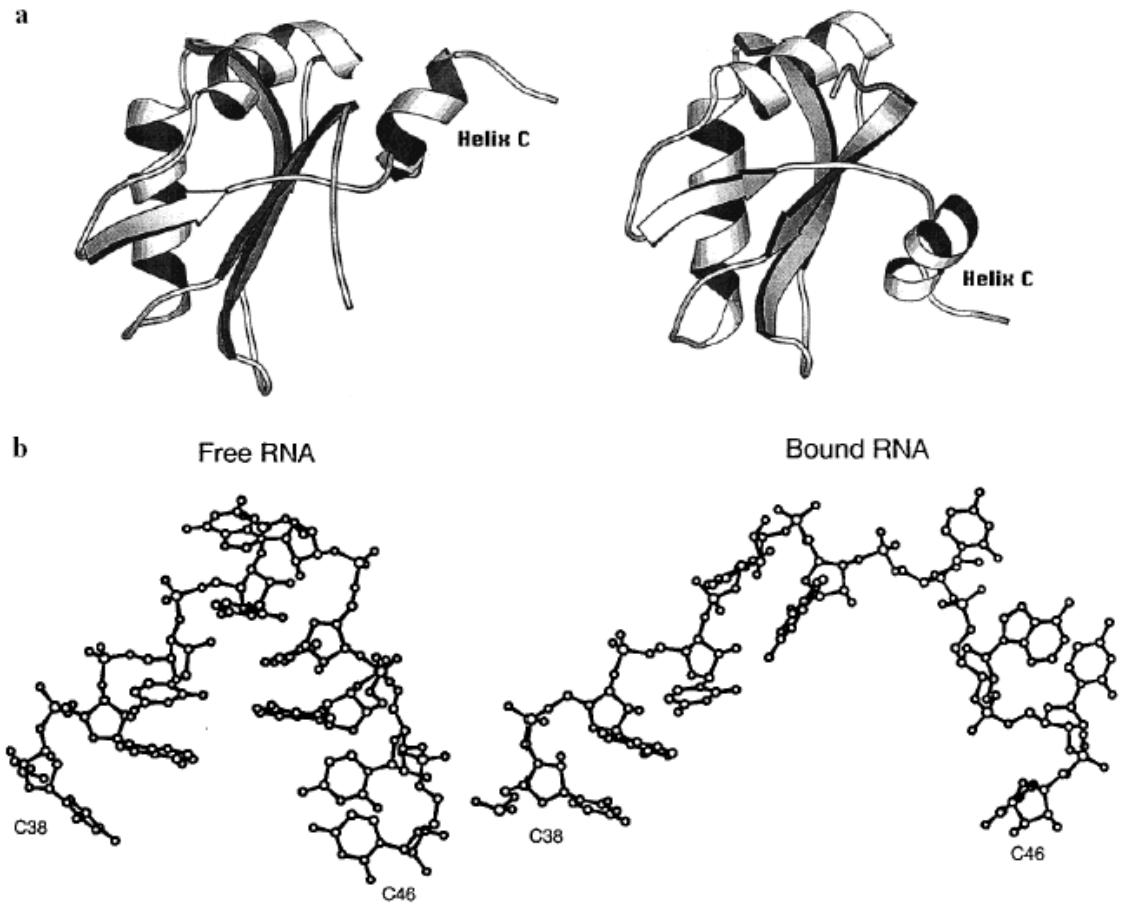


Figure 15 – Free and Bound Conformations of (a) U1A Protein and its (b) RNA Substrate (Leulliot 2001)

The major differences between the free and bound structures of U1A are found at the C-terminal end of helix C. When bound, residues immediately preceding helix C are involved in extensive interactions with three single-stranded nucleotides, but no direct contacts are made between the helix and the RNA. However, deletion of helix C abolishes specific RNA binding (Scherly 1991; Zeng 1997). This conformational shift of helix C exposes the RBD β sheet surface for binding, suggesting that intramolecular interactions involving helix C and the protein's RBD act as a placeholder for the RNA. Thus, the helix acts as a

competitive ligand for the protein, implying how its loss can affect the specificity of the protein for its natural target.

The region of the protein most critical for specificity, loop 3, also changes significantly due to RNA binding (Kranz 1999). In the free protein, the loop is flexible, as it acts as a probe to find its target RNA. When bound, the loop becomes rigid, with protein-RNA interactions solidifying the complex (Mittermaier 1999). The RNA also exhibits a change from freedom to rigidity when bound to U1A. This shift is observed in its pattern of base-stacking interactions (Gubser 1996). In the free RNA, most single-stranded bases are oriented toward the inside of the loop, filling the cavity created by the sugar-phosphate backbone of the double-helical stem. Protein residues in loop 3 then overtake this space in the cavity, while the RNA bases move outward to face the protein surface. Thus, the protein and RNA are able to bind each other in a well-choreographed conformational dance.

Nucleolin

The protein nucleolin is thought to correct folding and packaging of pre-rRNA in the nucleolus. These functions are mediated through the direct binding of nucleolin to two unrelated RNA sequences (Ginisty 2001). The central domain of nucleolin contains four RNA binding domains. RBD1 and RBD2, together with a linker peptide, form the binding site for the conserved nucleolin recognition element (NRE), a hairpin that displays a conserved loop sequence of six nucleotides. The second RNA-binding site is the evolutionary conserved motif (ECM), an eleven-nucleotide sequence, likely to be single-stranded. The NMR structure of nucleolin RBD1/2, complexed to a SELEX target that mimics NRE, showed that RBD1 contacted six nucleotides, while RBD2 contacted two

(Allain, Gilbert 2000). Perhaps more striking was the extensive involvement of the protein linker in sequence-specific recognition of the RNA. The linker is flexible in the free protein, but becomes ordered in the nucleolin-RNA complex (Figure 16, Allain, Bouvet 2000). Both RBDs and the linker region interact with the RNA. Similarly, the RNA hairpin loop is dynamic when free, but becomes fixed when in complex with nucleolin.

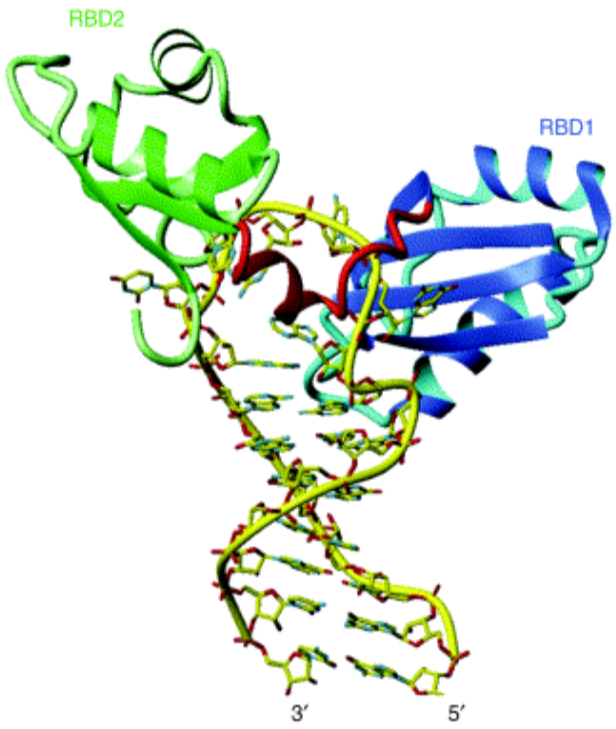


Figure 16 – NMR Structure of Hamster Nucleolin RBD1/2 Bound to SELEX NRE RNA (Allain, Bouvet 2000)

Hu Proteins

The 3' untranslated regions of mRNAs often contain binding sites for proteins that regulate translation. Among these classes of protein-binding sites are the adenosine-uridine-rich elements (AREs). Proteins that bind to AREs include the Hu proteins (e.g. HuD), which are speculated to regulate gene expression at the post-transcriptional level. Human Hu proteins contain three RBDs. Cocrystal structures of HuD RBD1/2 bound to an ARE sequence were solved (Wang 2001). The RNA strand runs from one RBD β sheet surfaces to the other, with the two protein surfaces parallel to each other. One RBD also makes several intraprotein contacts with residues in the linker regions. As in the nucleolin cocrystal, the linker region of HuD RBD1/2 is involved in contacts with the RNA.

p14 Protein

The spliceosome is a nuclear complex of over 100 different proteins and five small nuclear RNAs (snRNAs), where the proteins and snRNAs combine to create six small nuclear ribonucleoprotein particles (snRNPs) in the complex. The six snRNPs are labeled U1, U2, ..., U6, because the snRNAs are rich in the nucleotide uracil. This complex processes the pre-mRNAs in the nucleus by splicing out intronic nucleic acids, producing mature mRNA to be translated to protein by ribosomes in the cytosol. An evolutionary conserved 14 kDa protein (p14) of the U2 snRNP contains a canonical RBD, based on its RNP1 and RNP2 sequences, and has been cross-linked specifically to the branch-point adenosine at several points during the splicing process (Query 1996).

PTB Protein

The polypyrimidine tract binding protein (PTB) was first identified as part of the spliceosome, where its target RNA is the U/C-rich region of introns (Patton 1991). PTB contains four RBDs, and in its functionally relevant homodimeric form, it has eight. These RBDs have several unusual features, but PTB nonetheless is able to bind RNA sites in pre-mRNA introns and within the internal ribosome entry site (IRES) of picornaviruses (Witherall 1993).

U2AF Heterodimer

The association of the U2 snRNP at the 3' splice site is one of the initial events in pre-mRNA splicing. One of the proteins involved in recruiting this snRNP to the splice site is the U2 auxiliary factor (U2AF) (Zamore 1992). U2AF is a heterodimer, composed of $U2AF^{65}$ and $U2AF^{35}$. $U2AF^{65}$ consists of three RBDs and a serine-arginine (SR) region that binds to pre-mRNA. $U2AF^{35}$ contains an atypical RBD that is more likely involved in protein-protein interactions, as it interacts with a polyproline helix extracted from $U2AF^{65}$ (Kielkopf 2001).

Examples of Proteins with dsRBM

ADAR1 and ADAR2

The adenosine deaminases (ADARs) contain catalytic domains responsible for the deamination of adenine to inosine in RNA duplexes. ADAR1 contains three dsRBMs while ADAR2 contains two, and these motifs are responsible for binding to the enzymes' target RNA duplexes. ADARs will bind to any double-stranded RNA, but the patterns of deamination are dependent on the length, stability, and structural context of the RNA. The

extent of deamination depends on the length of the RNA duplex, with ADAR2 favoring shorter RNA duplexes over ADAR1 (Lehman 2000). This is likely a result of the larger binding site of ADAR1, suggesting its targets are longer than those of ADAR2. The mechanisms of these enzymes are largely unknown, such as how dsRBMs select their RNA target duplexes, if and how linker regions between the motifs contribute, and how the target RNA interacts with the catalytic domain of the ADARs.

PKR Kinase

PKR is a threonine/serine kinase with two dsRBM domains, and is activated by double-stranded RNA. The two dsRBMs have similar secondary and tertiary structure, but the linker between them is unstructured (Nanduri 1998). The RNA-binding properties of each PKR dsRBM separately and in tandem were studied using recombinant proteins (Tian 2001). dsRBMI and dsRBMII exhibited differences in RNA affinity, with dsRBMII alone having very little association with duplex RNA. A structural study of how PKR binds RNA was undertaken, by attaching a cleavage reagent (EDTA●Fe) at either of two engineered cysteine residues in dsRBMI (Spanggord 2001). The results indicate that dsRBMI sits at the loop/stem junction of the target RNA. The location of dsRBII is unknown from this study, but presumably sits on the stem (Figure 17).

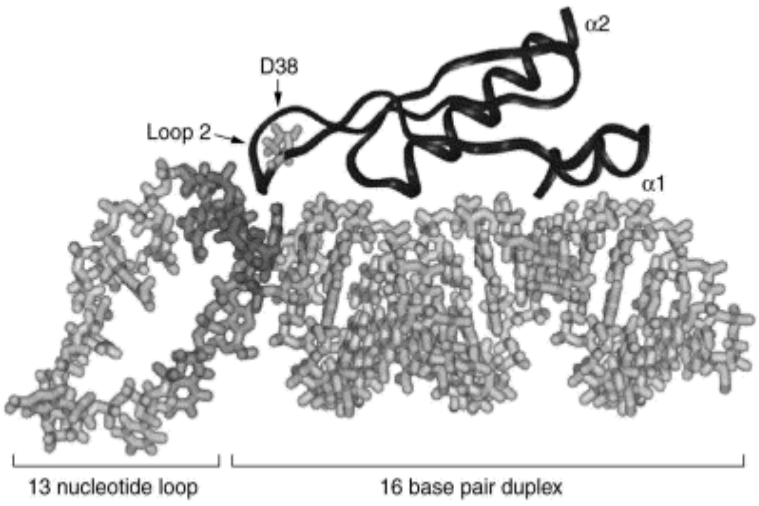


Figure 17 – Model of PKR dsRBMI Bound to Stem Loop RNA

D38 is the site of EDTA●Fe attachment to PKR dsRBM

Its position over the RNA loop is based on the cleavage pattern observed

(Spanggord 2001)

Chapter 3: RNA Backbone Conformation Clustering

With the rapidly growing body of knowledge in the field of RNA structure and function providing the impetus, the RNA Ontology Consortium (ROC) was born in order to create a shared vocabulary and system for describing, classifying, and comparing results of the RNA scientific community. The mission statement of this international group is to help scientists from various backgrounds communicate with one another and establish a basic framework for RNA structure. Part of this initial framework was the consensus set of RNA backbone conformations developed in this study in collaboration with other members of the ROC (Richardson 2008).

Research Approach

While RNA conformational space has been analyzed by different methods that mine the set of RNA structures (Duarte 2003; Hershkovitz 2003), this work focused on using the total torsion angle space of nucleotides as the means to understanding RNA conformation. Previously, Murray *et al*. (Murray 2003) used a ribose-to-ribose division to parse the RNA backbone into seven torsion angle repeats, from which 42 backbone conformations were discovered. Schneider *et al*. (Schneider 2004) clustered six combinations of three torsion angles, emphasizing those at the phosphodiester link, to discover 32 conformational families.

Data Gathering Methods

In this work, the NDB was mined for all X-ray crystal structures containing RNA. Sugar or phosphate modifications were excluded, and a three-angstrom resolution cutoff was

imposed. 232 structures remained in this initial data set. Then each RNA chain (of length *n*) was split into dinucleotide (di-nt) pairs, namely nucleotides *i* and *i*+1 for all nucleotides 1-*n*. The data was further filtered by crystallographic criteria, namely resolution and temperature factor, as well as stereochemical quality, mainly atom-atom close contacts (Murray 2003). 4148 di-nts remained. These di-nts were tabulated with the following information: structure ID, chain ID, residue number of the first nucleotide, residue ID of the first nucleotide, and the fourteen torsion angles of the two nucleotides. All di-nt lacking one or more torsion angle were removed, leaving 3751 di-nt from 101 different structures (Figure 18).

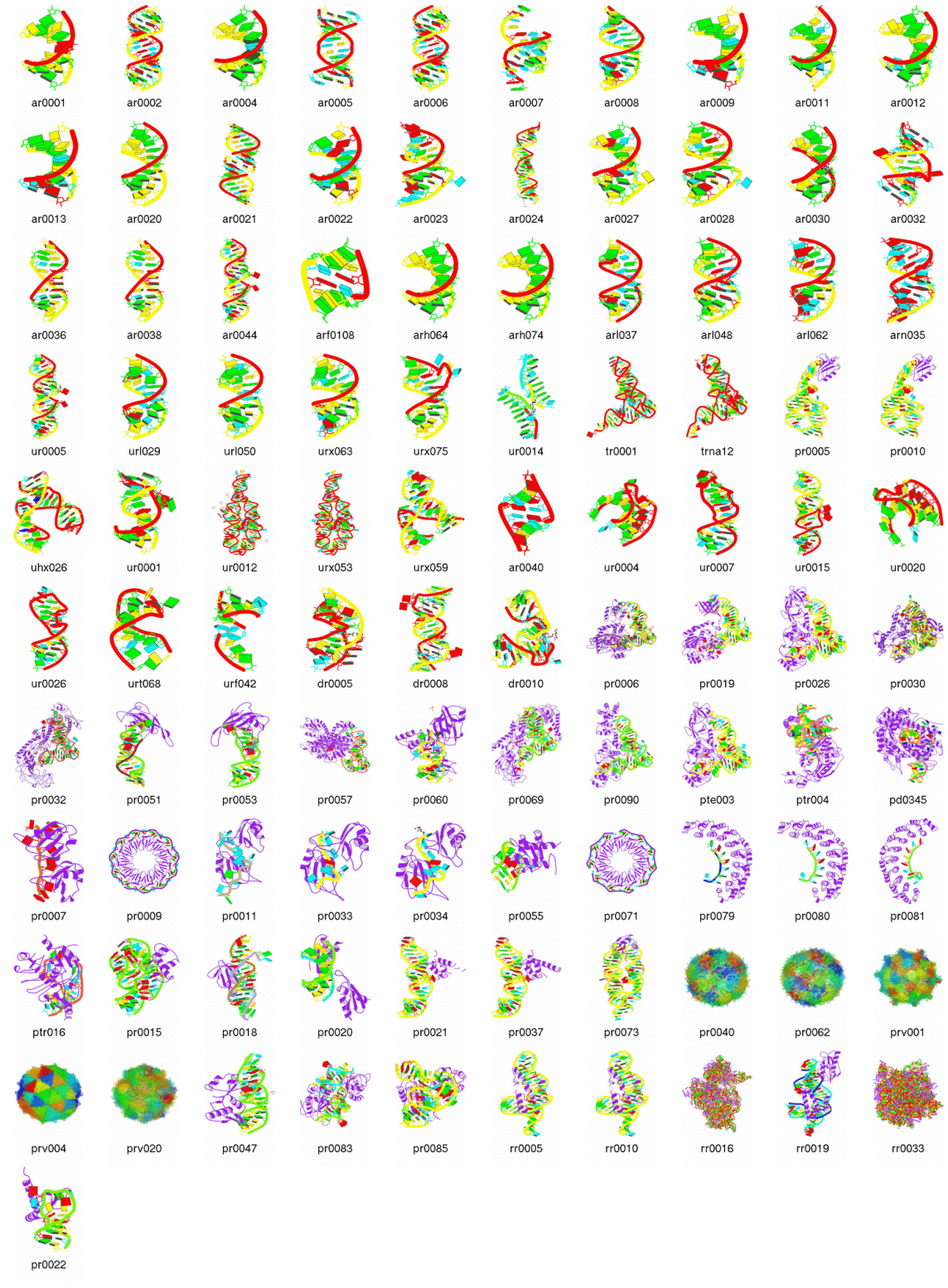


Figure 18 – The 101 NDB Structures used in Conformation Clustering

Each di-nt was assigned a unique number ID (1-3751) to distinguish between identical chain IDs across the different structures. Then, the data was partitioned into two sets, A-RNA versus non-A-RNA, by the phosphodiester linkage angles between the two nucleotides ($\zeta \approx 290°$; $\alpha+1 \approx 300°$ for canonical A-RNA). This resulted in 882 non-A-RNA di-nts, and 2869 A-RNA di-nts.

Classification Methodology

Torsion distributions of these fragments were analyzed in seventeen 3D scattergrams ("maps") by a Fourier averaging technique (Schneider 1993) to pinpoint areas of high torsion frequency. The rigid nucleotide concept served as the leading principle in lowering the torsion angle dimensionality problem, namely the fourteen torsion angles in a dinucleotide subunit. All angles spanning the backbone of a di-nt were represented by seventeen scattergrams of three torsion angles apiece (Figure 19).

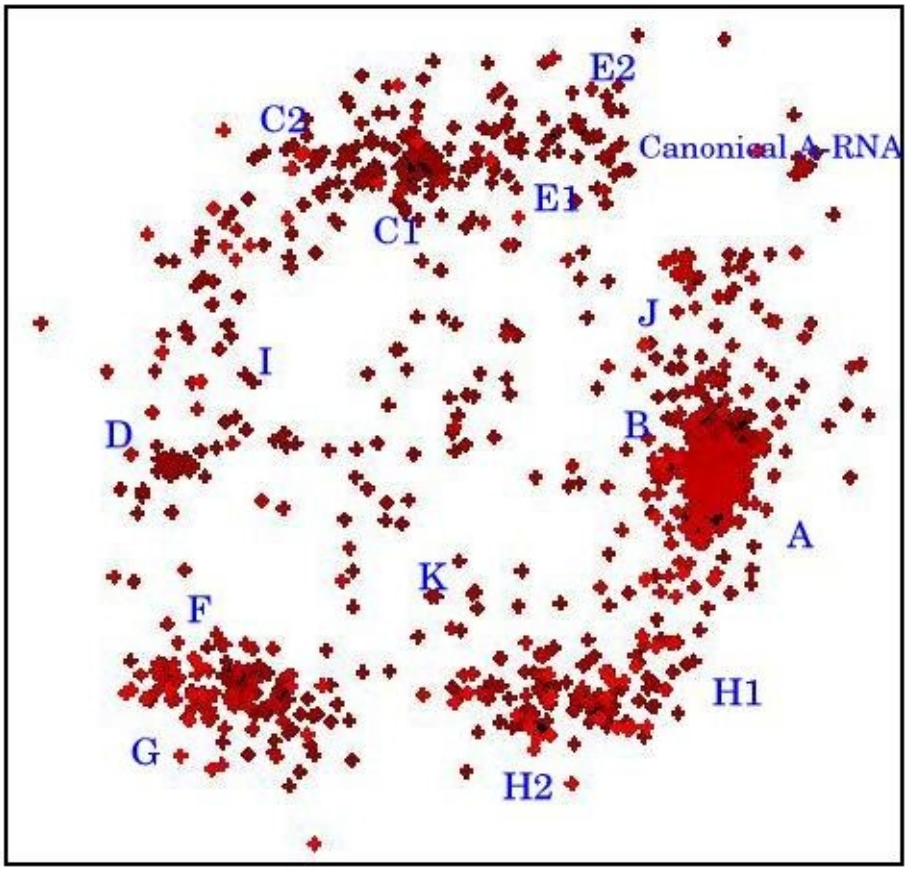


Figure 19 – Point Scattergram Example – ζ - α+1 - γ+1 3D Torsion Map

ζ = x-axis; α+1 = y-axis; γ+1 z-axis

The Fourier-averaging method was employed to create contours around the areas of high density in the scattergrams. Peak maxima were defined for each map, representing the preferred (FT-averaged) values for the three analyzed torsion angles of each map (Figure 20).

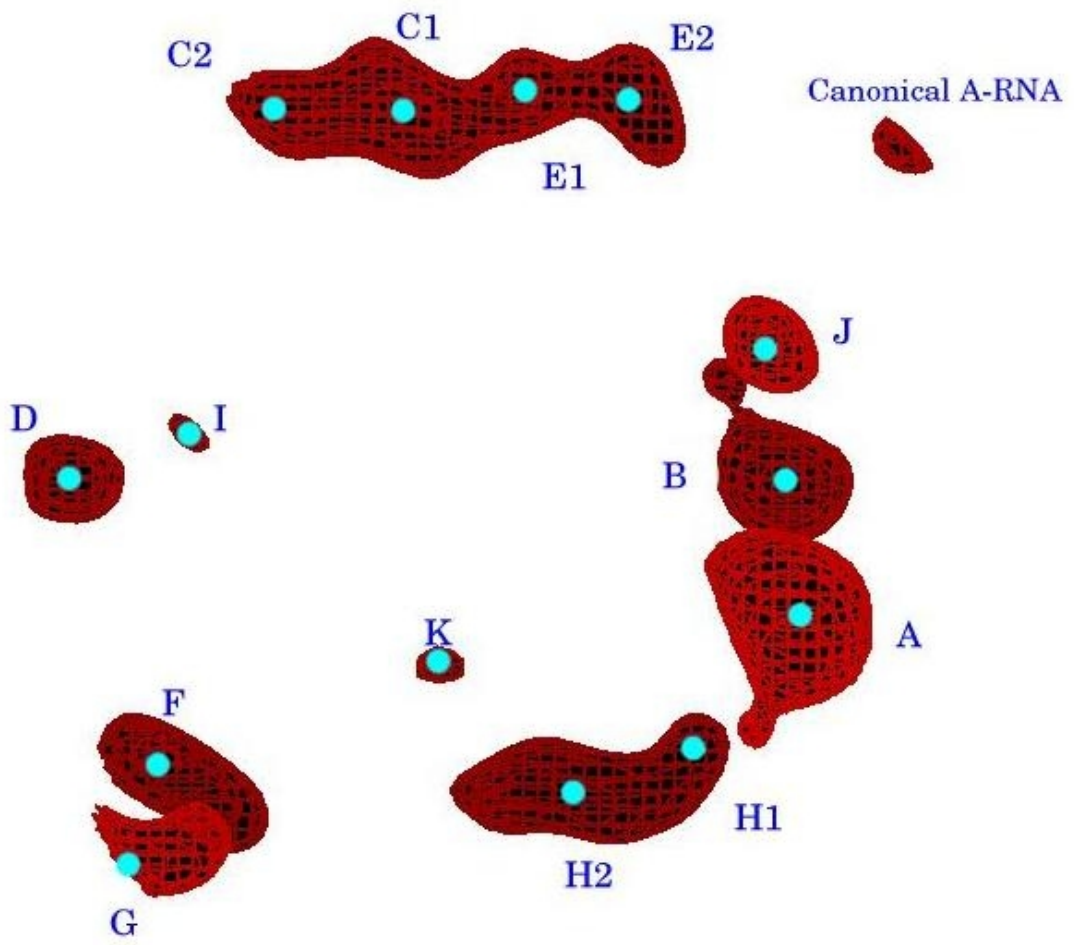


Figure 20 – Fourier-Averaged Representation of ζ - α+1 - γ+1 3D Torsion Map

Peak names labeled in blue

The di-nts were annotated with these peak names based on the distance between the projection of the di-nt's three torsion angles and the peak maxima of each map. Points within a variable sphere around each peak maximum were labeled accordingly; outliers were given an undefined peak name. The data was clustered lexicographically by the string of peak names created for each di-nt (Table 1).

Unique ID	Structure ID	z-a1-d	z-a1-g1	a1-b1-g1	a1-g1-d1	d-e-a1	z-a1-c	z-b1-c1	b1-d1-c1
17	ar0036	AC	J.	E1	??	F.	A2	C.	D.
19	ar0036	AC	J.	E1	??	F.	A2	C.	D.
21	ar0038	AC	J.	E1	??	F.	A2	C.	D.
23	ar0038	AC	J.	E1	??	F.	A2	C.	D.
82	pr0018	AC	J.	E1	??	F.	A2	C.	D.
136	pr0051	AC	J.	E1	??	F.	A2	C.	D.
137	pr0051	AC	J.	E1	??	F.	A2	C.	D.
140	pr0053	AC	J.	E1	??	??	A2	C.	D.
276	rr0016	AC	J.	E1	??	F.	A2	C.	D.
347	rr0033	AC	J.	E1	??	F.	A2	C.	D.
415	rr0033	AC	J.	E1	??	F.	A2	C.	D.
671	rr0033	AC	J.	E1	??	F.	A2	C.	D.
778	rr0033	AC	J.	E1	??	F.	A2	C.	D.

Table 1 – Cluster Example Showing Eight First Peak Maxima used to Label Di-nts

The clusters were visually inspected by overlapping the di-nt structures to ensure comparable conformation. The canonical A-RNA conformation was used as the template on which to base each cluster overlap (Figure 21). Finally, each cluster was annotated with average torsion angles and biologically relevant motifs.

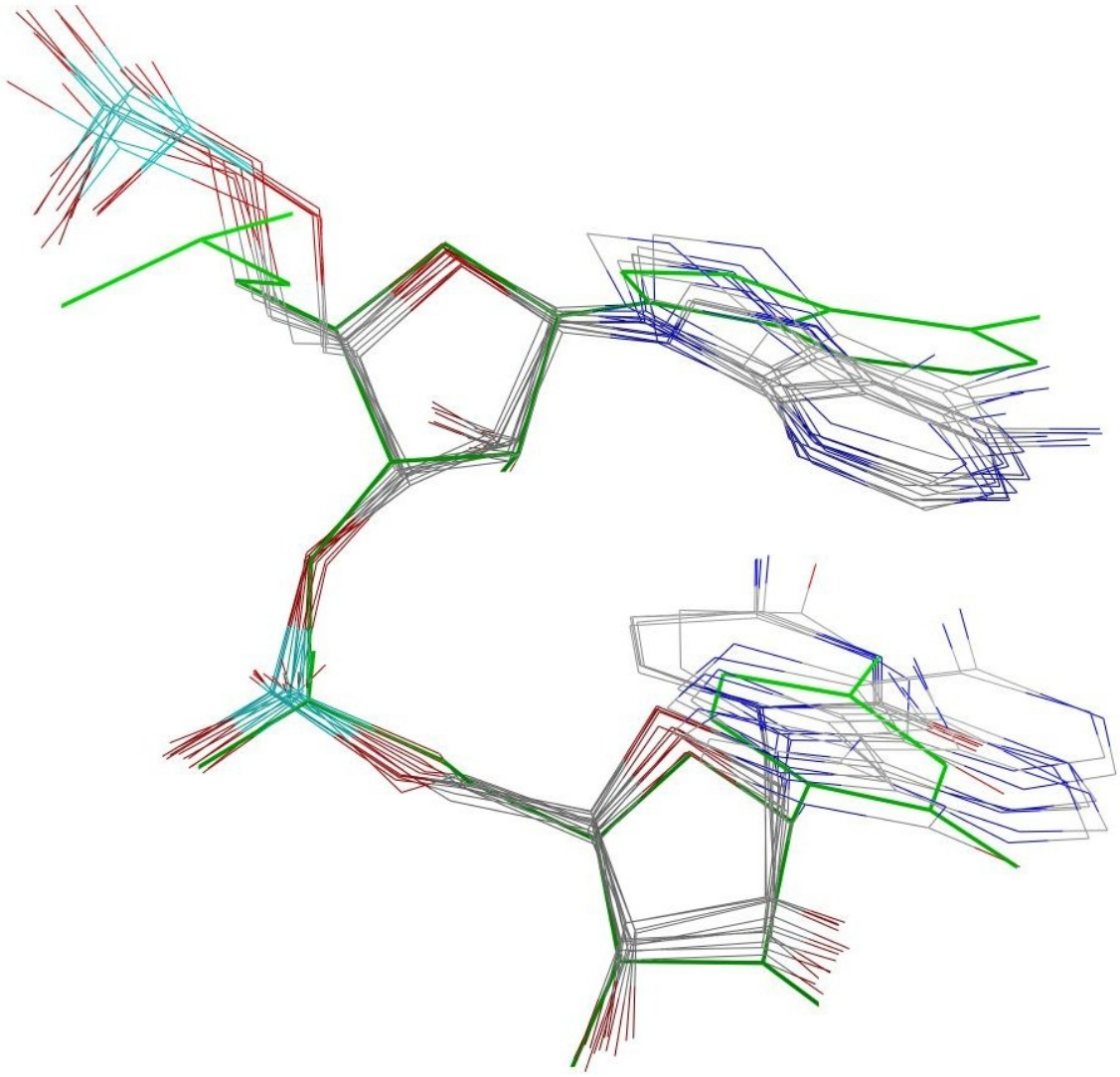


Figure 21 – “AC-J-E1” Cluster

Canonical A-RNA shown in green

Consensus Set Collaboration

Conformational clusters similar to those just described were found independently using protocols developed earlier (Murray 2003; Schneider 2004), with their geometries compared and validated. From these, a set of 46 different consensus RNA conformational families were finalized and presented to the ROC (Richardson 2008).

31 of the 46 backbone conformations uniquely identified in the individual studies compared very strongly, suggesting their validity as well as that of the overall research at hand. The other conformation sets were reconciled by various analytic methods between the collaborators, such as superposition of comparable conformation classes and expanding or decreasing torsion angle boundaries.

The torsion angle space that was settled on for the set of consensus conformations was the seven-angle torsions δ, ε, and ζ from the first nucleotide in the di-nucleotide pair; and α, β, γ, and δ from the second nucleotide. These dihedral angles accurately define the sugar-to-sugar backbone and shared the best fit comparisons between the different collaborators' work.

Concurrently while the set of consensus conformations were developed, a novel modular nomenclature was developed for each conformation class. Each conformation was given a two-character name comprised of a digit in the first position and a non-digit character in the second. The digit represents the torsion angles of the first nucleotide in the conformation (δ, ε, and ζ), while the non-digit represents those angles provided by the second nucleotide (α, β, γ, and δ).

This nomenclature was designed with bioinformatics string manipulation in mind, as the two-letter names can be used to model a string of RNA backbone conformations as a string of characters. The 46 total consensus conformations are listed below in Tables 2-5 with their two-letter names and torsion angles, as well as in graphical views of their structures in

Figures 22-25. Both sets of Tables and Figures are divided up into four distinct groups based on the puckers of the two sugars of each conformation class. The two δ angles in each conformation describe the sugar pucker for each sugar, and the nomenclature utilizes this in its naming pattern.

An odd digit in a conformation name indicates the first nucleotide's δ value is consistent with the C3'-endo pucker, namely between 74° and 94°. An even digit indicates a δ value consistent with the C2'-endo pucker, between 136° and 164°. For the second nucleotide in the conformation class, letters in the first half of the alphabet (a, c-n) indicate the 3' pucker, while (b, o-z) are used for the 2' sugar pucker. A di-nucleotide can have four different permutations of the first and second sugar puckers, which conveniently divides the consensus conformation classes into four distinct groups.

In the four tables below, the total number of di-nucleotides in each conformation consensus cluster is listed in the second column, next to that conformation's modular name. The next two columns provide comments on the conformation's structural role and a representative structural example listed by NDB ID and first nucleotide number. The final seven columns contain the ranges of the seven torsion angles that comprise each conformation's sugar-to-sugar torsion set. The first three angles belong to the first nucleotide, and are listed as δ1, ε1, and ζ1. The final four angles are from the second nucleotide and are listed as α2, β2, γ2, and δ2.

The most common conformation is listed in the first row for each of the four tables. For example, conformation 1a is listed first in the first table, which has both δ angles in the

C3'-endo pucker. 1a is the conformation of A-RNA, the most common conformation found in the study. For the rest of the conformations in each table, the particular torsion angle that most sets that conformation apart from the first conformation listed is shaded in light blue. Many of the conformations have multiple variations in torsion angles, but these initial shaded variations were utilized in the next step of this study in order to automatically assign the conformation nomenclature to novel, input RNA structures.

The four Figures 22-25 depict the conformations in the same order as listed in Tables 2-5. Conformations with similar torsion angle variations are grouped together on the same row. Each conformation is a uracil-adenine di-nucleotide, with the uracil base oriented in the foreground with the same orientation for all conformations. The uracil ribose ring is oriented such that the oxygen atom is pointing in the direction of the helix axis as in A-RNA (Sussman 1972).

Name	#dint	Comment	example	δ1	ε1	ζ1	α2	β2	Γ2	δ2
1a	4637	A-form	ur0020 11	77-85	202-222	282-296	287-303	166-182	48-60	78-84
1m	15	+B variation; some intercalate	rr0082 1940	79-89	202-234	277-307	276-308	210-234	48-68	79-93
1L	14	-B variation; overtwists base direction	rr0082 1460	82-90	239-251	255-281	296-312	134-142	52-72	74-84
&a	33	-Z variation; weak Hb O2'(-1)-O4'	pr0037 b163	77-87	184-198	259-271	287-305	172-192	44-58	77-87
7a	36	-Z variation; stack switch	ar0041 a6	79-87	194-240	208-236	294-312	146-176	43-55	79-85
3a	25	-Z variation; bases far	urb016 a2	81-89	192-240	159-187	277-301	148-180	39-53	79-91
9a	19	-Z variation; bases far; starts on ends loop	rr0082 2582	81-85	195-225	108-134	277-301	134-180	43-55	78-84
1g	78	-A variation; GNRA1-2; U-turn	rr0082 1864	78-84	211-227	282-300	159-175	144-176	46-56	82-88
7d	16	-A,-Z variation; bases far; can span 2 helices	rr0082 636	80-88	223-255	245-269	60-80	147-193	47-59	82-88
3d	20	-A,-Z variation; GNRA1-2; U-turn	rr0082 2118	81-89	229-259	189-219	47-85	158-204	49-61	82-90
5d	14	-A,-Z variation; P(-1) to P(+1) close; end or end+1 A-helix	ur0020 a9	76-84	195-209	49-77	56-80	113-173	43-57	81-85
1e	42	+G,-B,-A variation; S-motif strand2 “dent”; Hb O2(-1)-O4'	ur0035 2665	78-84	193-209	275-287	240-258	72-94	162-174	82-90
1c	275	+G,-A variation; GNRA 4-5; ttt “crankshaft” version of 1a	ur0020 a28	77-83	188-206	281-301	141-165	182-206	169-189	81-87
1f	20	+G,-B,-A variation; stack switch or intercalate	tr0001 22	79-83	189-217	283-305	161-183	126-152	166-186	81-87
5j	12	+G,-B,-A variation; bases far; 1-bulge return	ar0027 b17	80-94	201-247	65-95	58-76	99-119	170-182	80-88

Table 2 - First Quarter Conformations

name	#pts	Comment	example	δ1	ε1	ζ1	α2	β2	γ2	δ2
1b	168	leads into 2' suites; k-turn 0'; *syn* G Hb N2-OP2	pr0113 d208	80-88	205-225	279-299	291-309	165-189	51-65	138-152
1[	52	+B variation; best intercalation conformation	pr0019 b658	79-87	210-230	279-299	289-305	213-231	47-61	136-152
3b	14	-Z variation; bases far; ends A-helix	rr0082 904	82-88	208-244	151-185	278-308	156-200	44-54	145-151
1z	12	-A variation; UNCG 1-2; bulges	rr0082 1771	80-86	188-224	259-297	182-210	137-187	46-56	140-150
5z	42	-A,-Z variation; S-motif 1-2; Z32a dna; Hb OP2(-1)-O2'	ur0026 2654	80-86	201-211	46-60	159-169	138-158	45-55	144-152
7p	27	-A,-Z variation; bases far	pr0033 b8	81-87	213-261	205-235	56-80	170-230	47-61	140-152
1t	7	+G,-A variation; ttt version of 1b	pte003 b907	78-84	179-219	281-297	163-197	181-209	169-187	142-152
5q	6	+G,-B,-A,-Z variation; bases far	pte003 b973	74-90	199-211	55-83	54-72	98-132	170-182	142-150
1o	13	+G variation; starts t-bulge	rr0082 1108	80-88	200-234	272-302	290-304	186-264	287-301	138-164
7r	16	+G,-A variation; k-turn 1-2	rr0082 262	81-89	220-246	229-267	50-76	155-209	292-300	143-157

Table 3 - Second Quarter Conformations

name	#pts	Comment	example	δ1	ε1	ζ1	α2	β2	γ2	δ2
2a	126	leads out of 2' suites; 1-bulge return	rr0082 1711	137-153	248-272	271-307	275-301	176-210	46-60	79-89
4a	12	-Z variation; bases far	rr0082 2485	139-153	245-275	156-184	279-317	136-204	43-59	79-89
0a	29	-Z,-e variation; cross-stacked A-helix start; k-turn 4-5	rr0082 265	142-156	212-234	114-164	274-296	138-178	42-54	80-88
#a	16	-Z,-e variation; i to i+1 base pair; S-motif 3-4	rr0082 1371	145-151	187-197	140-152	282-296	139-163	38-46	82-88
4g	18	-A,-Z variation; i to i+1 base pair; non S-motif	ur0012 a226	140-156	243-271	144-186	191-219	150-180	42-56	79-87
6g	16	-A,-Z variation; sheared stack	pr0122 r151	138-152	245-281	64-96	175-233	167-213	53-63	78-92
8d	24	-A,-Z variation; some with Hb O2'(-1)-OP2(+1)	rr0009 c1062	143-155	261-281	224-258	52-72	153-199	50-58	84-90
4d	9	-A,-Z variation; tRNA 58-9; Hb O2'(-1)-OP2(+1)	tr0001 59	144-156	224-276	181-195	60-100	178-218	53-69	85-93
6d	18	-A,-Z variation; starts A-helix	rr0082 116	141-153	218-264	73-105	45-73	138-184	45-59	79-87
2h	17	+G variation; bases far	rr0082 2540	144-152	253-269	278-302	286-306	160-194	162-190	83-91
4n	9	+G,-A,-Z variation; stack or sheared stack	rr0082 767	137-151	213-241	190-218	62-86	197-237	185-203	78-84
0i	6	+G,-A,-Z variation; bases perpendicular	rr0082 940	147-151	255-295	87-113	70-92	236-260	179-185	81-85
6n	18	+G,-A,-Z variation; UNCG 3-4; Z23 dna; *syn* curled to base triple	rr0082 1773	144-156	257-279	77-93	59-69	183-199	168-186	81-91
6j	9	+G,-B,-A,-Z variation; bases far	pte003 975	134-150	216-272	51-81	64-80	100-144	176-188	81-87

Table 4 - Third Quarter Conformations

name	#pts	comment	example	δ1	ε1	ζ1	α2	β2	γ2	δ2
2[	40	UNCG 2-3; near B dna; k-turn 3-4	rr0082 264	138-154	243-275	274-308	280-304	189-231	47-61	141-155
4b	27	-Z variation; cross-stacked A-helix end	rr0082 247	138-152	225-265	150-176	288-300	158-186	40-52	140-152
0b	14	-Z variation; varied	rr0082 453	144-152	228-268	98-126	258-292	149-181	45-69	140-152
4p	13	-A variation; often starts 1-bulge; Hb O2'(-1)-N7(+1)	rr0096 873	140-160	244-276	195-233	59-85	181-235	43-71	144-152
6p	39	-A,-Z variation; k-turn 2-3	rr0082 1315	139-153	237-279	75-105	56-80	155-191	48-64	144-152
4s	8	+G,-B,-Z variation; S-motif 2-3	ur0026 2655	148-152	232-264	158-182	265-291	77-91	170-182	146-150
2o	12	+G variation; bases perpendicular, something between	pr0033 b5	141-153	241-271	280-312	283-291	169-221	287-301	147-153

Table 5 - Fourth Quarter Conformations

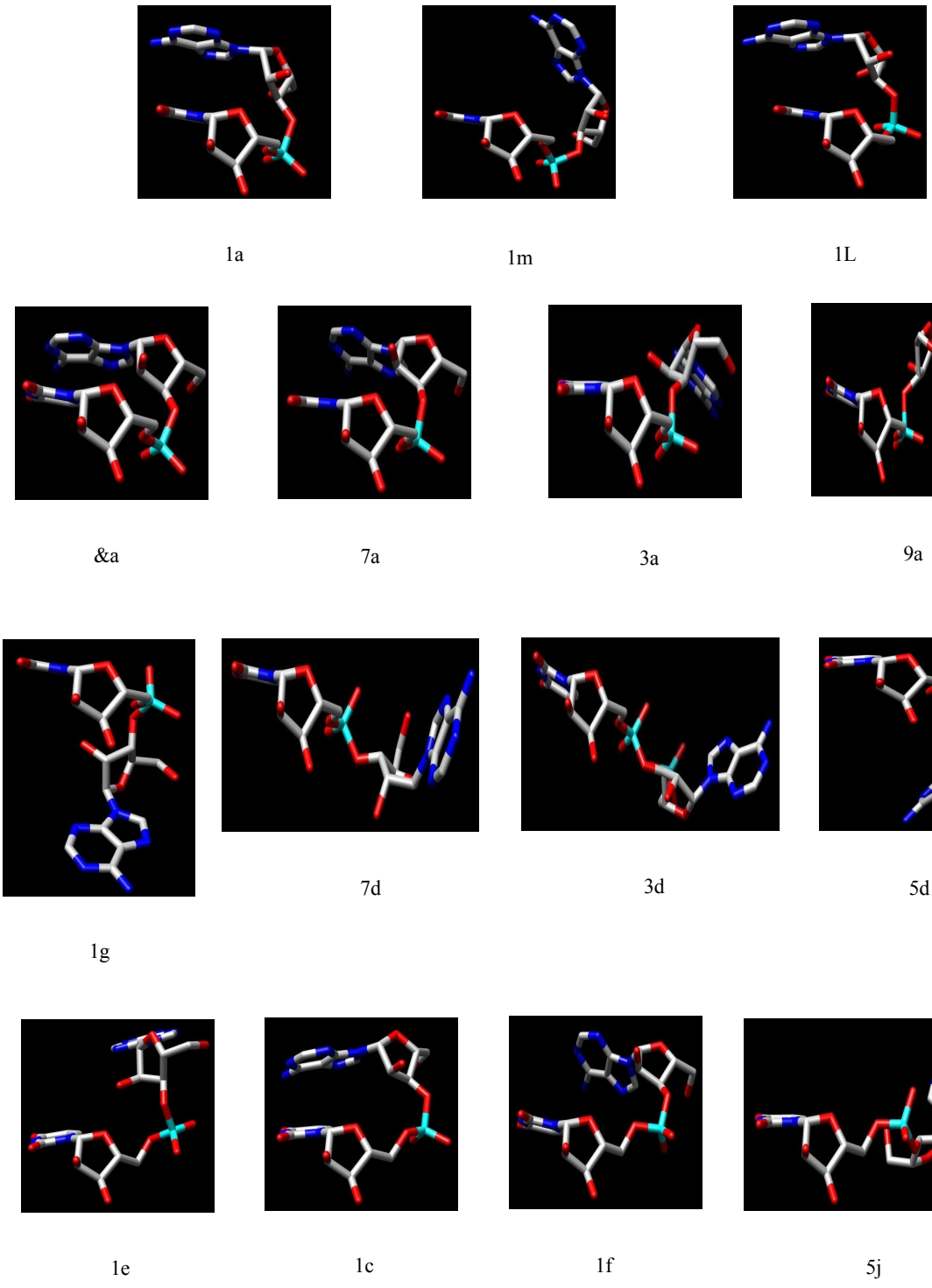


Figure 22 – First Quarter Conformations

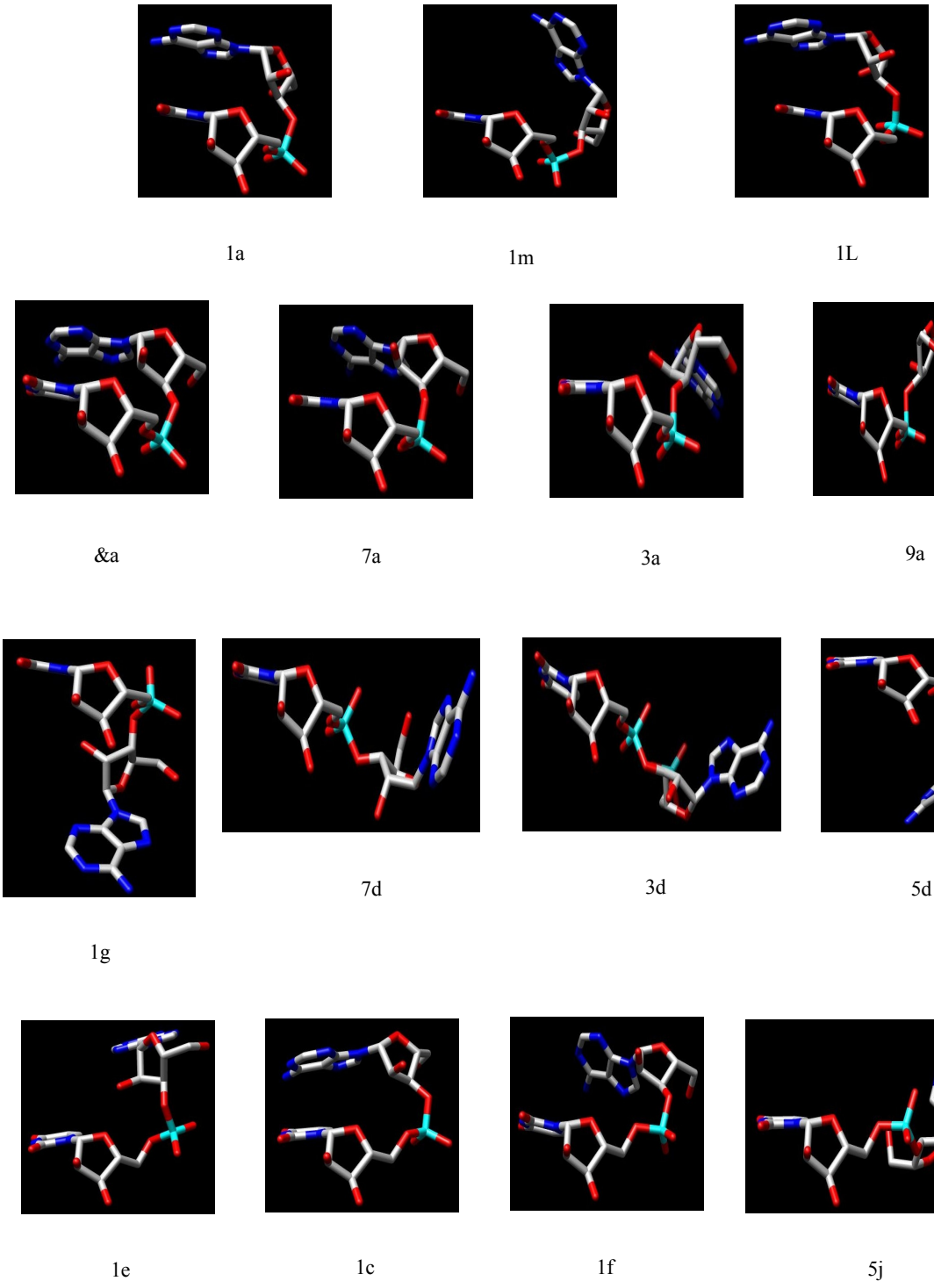
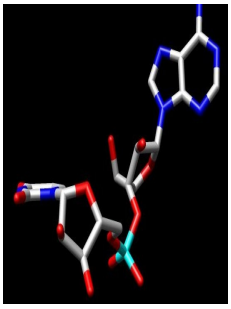
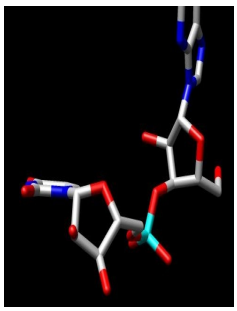


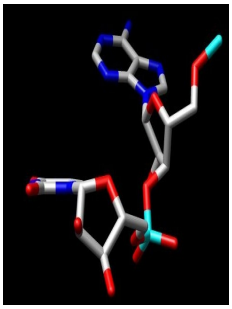
Figure 23 – Second Quarter Conformations



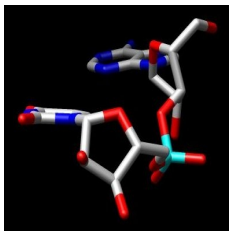
2a



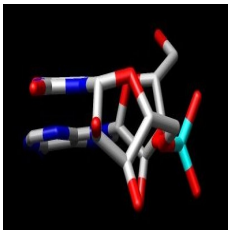
4a



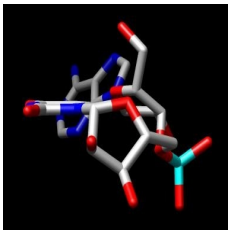
0a



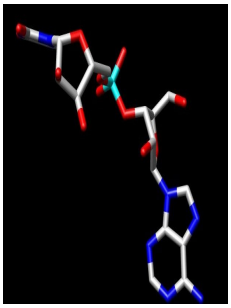
#a



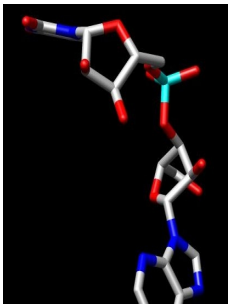
4g



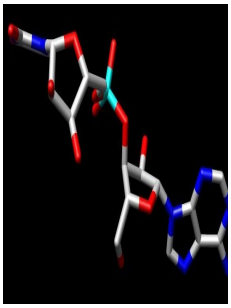
6g



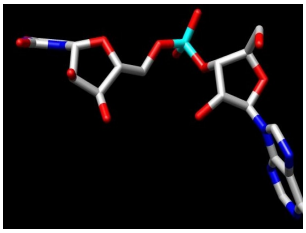
8d



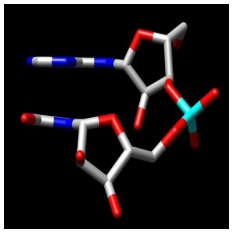
4d



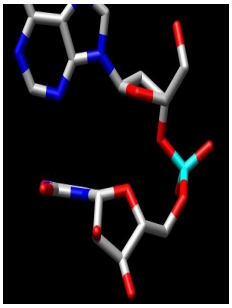
6d



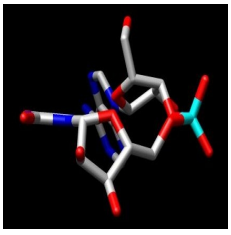
2h



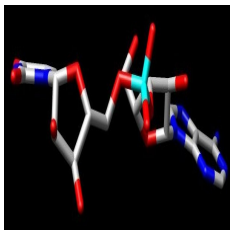
4n



0i

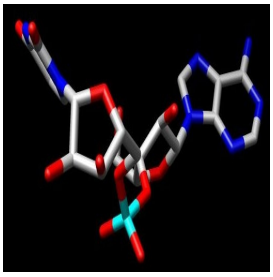


6n

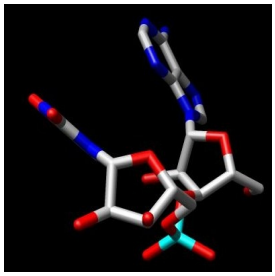


6j

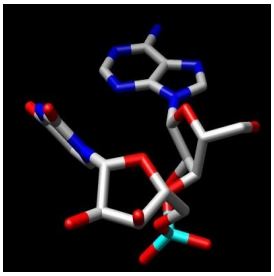
Figure 24 – Third Quarter Conformations



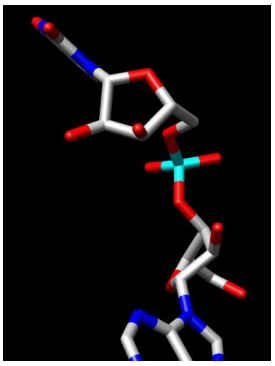
2[



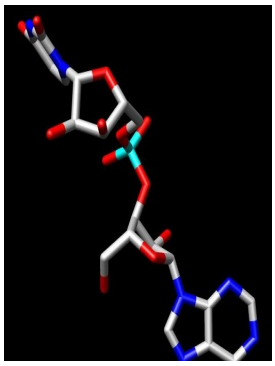
4b



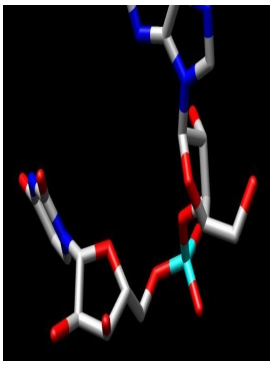
0b



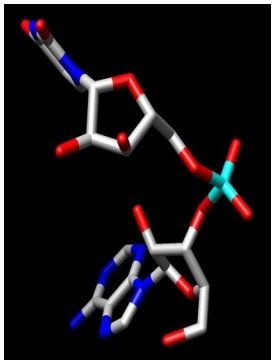
4p



6p



4s



2o

Figure 25 – Fourth Quarter Conformations

Chapter 4: DiCAT Software Tool

Design

In order to utilize the set of RNA conformations in a practical way, this study designed and implemented a software tool to automatically assign input RNA structures with the novel RNA conformation nomenclature. The tool was named DiCAT – an acronym for Dinucleotide Conformation Assignment Tool.

The program was based on a decision tree algorithm. Each level of the decision tree represents one of the seven torsion angles that define the set of RNA conformations. Within each level, sets of angle ranges that are tested on the input RNA to determine what path it should take down the decision tree.

The current order of torsion angles down the tree will be delta-1, delta, alpha, beta, gamma, zeta-1, and epsilon-1. This order allows for easy decision-making between the various conformation groups. The greatest delimiters are the two delta torsion angles, and thus they are the first two decision-making levels of the tree traversal. When traversing the tree, the input RNA flows down the tree based on the ranges of the cluster conformations. The traversal continues until a conformation cluster name is found, and if not, will return a best-fit conformation that most closely matches the input set of angles.

The assignment method thus traverses the decision tree using input RNA torsion angles to find the best conformation match. The tree structure allows for an output option to return

multiple possible conformations for input RNA. This may be important for ambiguous input. Figure 26 depicts the decision tree logic used by the DiCAT algorithm.

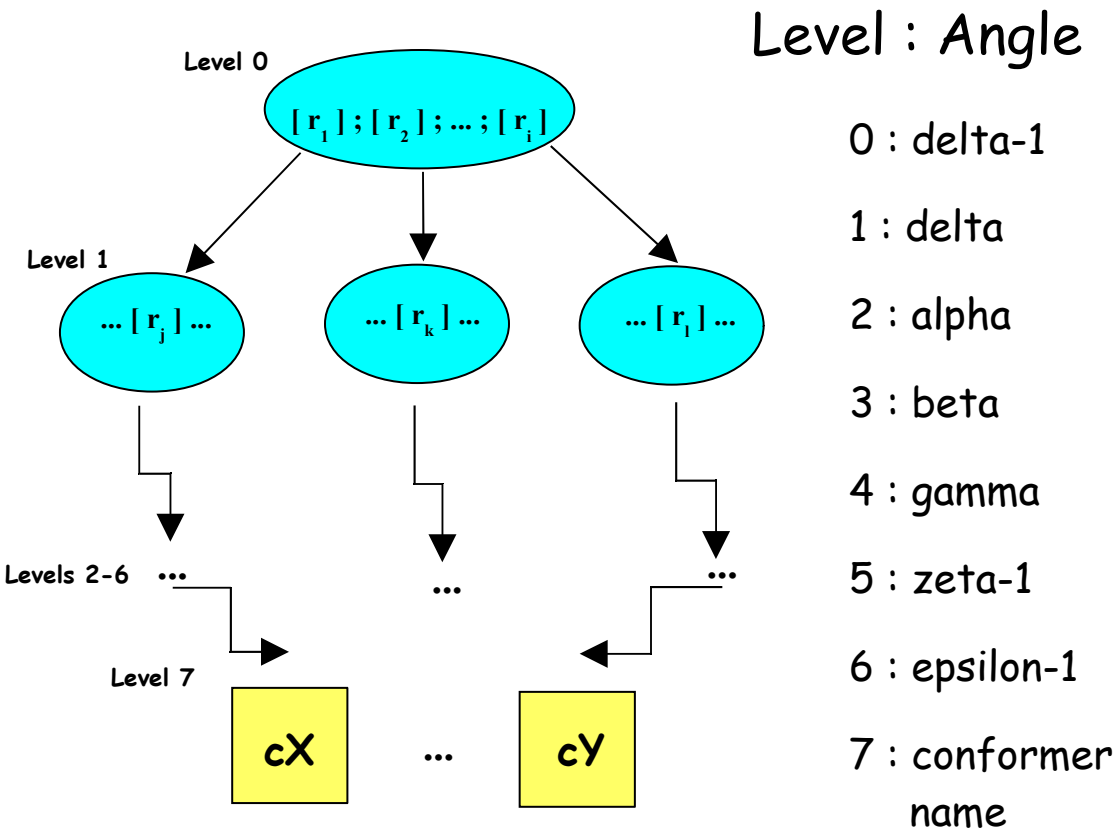


Figure 26 – DiCAT Decision Tree Logic

Development

The DiCAT software tool runs in a series of steps fueled by shell and Perl scripts. The primary input is a list of PDB IDs, namely the RNA structures of interest that are to be annotated with the conformation names.

First, this PDB ID list is used in a shell script that downloads the corresponding PDB files.

Second, the PDB ID list is again used in another shell script that utilizes the RNAView program (Yang 2003) to produce files containing the torsion angles for each di-nt in each RNA structure. Output from the RNAView program is processed so that the torsion files produced can be easily uploaded into a spreadsheet program like Microsoft Excel.

Finally, the third and last shell script is run that calls the DiCAT algorithm. This script outputs a conformation file for each structure. Each file contains the di-nt torsions in the same spreadsheet format as before, but now also contains a conformation name as assigned by DiCAT.

Testing

To test the accuracy of the DiCAT algorithm, the original set of 3,751 RNA dinucleotides used to form the consensus conformation set were run through the program. 2,663 di-nt in this set were assigned to conformation groups in the clustering process. Initially, 2,497 of the di-nucleotides were assigned the appropriate conformation name by the DiCAT algorithm. Thus 166, or 6.2% of the assigned di-nts, were assigned different conformation names than that found in the initial pilot study.

140 of the 166 mislabeled dinucleotides were originally classified as 1a conformation, namely A-RNA. The DiCAT algorithm assigned this group into the following sets – 90 as &a, 30 as 1L, and 20 as 1m. All three classes are single torsion variations of the 1a conformation. Random inspection of representatives of each set of outliers showed that each di-nt torsion angle set fell closer to the conformation labeled by DiCAT more so than that of 1a.

Similarly, the other 26 mislabeled dinucleotides all fell in a range closer to the conformation assigned by DiCAT. For example, two di-nts originally classified as 0a were assigned the #a conformation by DiCAT. Both di-nts should actually be #a because the ε-1 and ζ-1 torsion angles were closer to the value in the #a conformation than that of the 0a. This was the case with the remaining 24 di-nt that did not match up to their originally classified conformations. Thus, the DiCAT tool has proven to be accurate in properly assigned RNA dinucleotides with the RNA torsion conformation classification.

Chapter 5: DiCAT Case Study

RRE RNA Structures

As described above, the Rev protein of HIV-1 viral RNA is an α-helix and binds the major groove of its target RRE RNA. Binding of the protein induces conformational change in the RRE RNA. This Rev-RRE complex controls the export of HIV-1 from the nuclei of infected cells. Six RRE RNA structures were selected from the PDB to put the DiCAT tool into use and determine if any conformation patterns could be discerned. Two of the structures are free of protein (1duq and 1csl), while the other four are bound to a polypeptide (1i9f, 1g70, 1etf, and 1etg).

While DNA-binding proteins often use α-helices to target specific bases in the major groove of their target DNA, the major groove of A-form RNA helices are too deep and narrow to provide access to a protein α-helix. In the Rev-RRE complex, non-Watson-Crick interactions widen the RNA's major groove. Two purine-purine base pairs (G-A and G-G) are responsible for this widening, which distorts the RNA backbone into an S-shaped architecture to the backbone from nucleotides G70 to A73, an undertwisting of the base pairs in the internal loop, and an opening of the major groove by roughly 5 Å (Battiste 1996).

Figure 26 shows the RRE RNA hairpin from structure 1g70 (Gosser 2001). Wild-type nucleotides are in uppercase, while non-wild-type are lowercase. The bulge in the RNA helix created by the two purine-purine base pairs is clearly visible. Bases G46-A52 and

bases U66-C74 were common to all six RRE RNA structures and were used in the DiCAT analysis of their di-nucleotide backbone conformations.

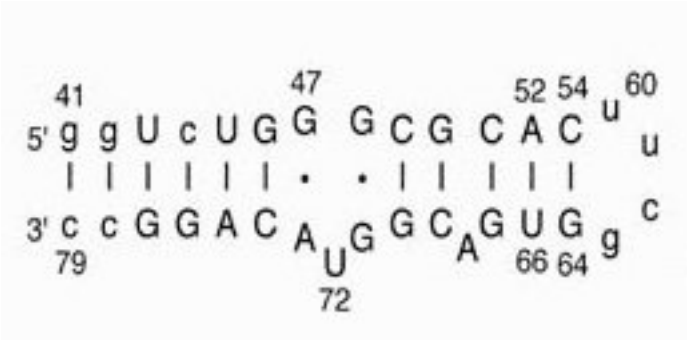


Figure 27 – RRE RNA Hairpin

Data Analysis

Table 6 shows the backbone conformations for the 14 di-nucleotides common to all six RRE RNA structures and known to play a role in Rev protein binding. The first column lists the PDB ID for each of the six structures. The four structures with bound protein are listed first. An empty row then separates the four RRE-REV complex RNA structures followed by the two unbound RRE RNA structures. Each column shows the backbone conformation output by DiCAT for each structure. An empty column separates the two distinct sets of bases G46-A52 and bases U66-C74 from opposite sides of the strands. The A68 and U72 bulge nucleotides add two extra di-nucleotides to the second set of backbone conformations.

PDB ID	GG	GG	GC	CG	GC	CA	UG	GA	AC	CG	GG	GU	UA	AC
1etf	4a	1b	2a	1a	1a	1a	1a	1o	2h	1a	7p	6p	2a	1a
1etg	2h	1t	2a	1a	1a	1c	1a	1t	6j	1a	5z	4s	2a	7d
1i9f	4a	1b	2a	1a	1c	1a	1a	1e	1c	1a	7p	6p	2a	1a
1g70	1[	2h	1c	1a	1c	1c	5j	1g	1e	1a	1z	6n	1c	1c
1duq	1a	1a	1a	1a	1a	1a	1a	1g	5j	1a	1b	8d	5j	1a
1csl	1a	1a	1a	1a	1a	1a	1a	1g	5j	1a	1b	2[	2h	1a

Table 6 – RRE RNA Backbone Conformations

As shown in Table 6, the protein-free RNA was found to have mostly canonical A-RNA conformations, especially on the 5' side of the protein binding site. The protein-bound RNA has non-canonical conformations in most of the known binding sequence, with notable exceptions at both CG di-nucleotides.

While no clear-cut pattern is apparent in the RRE RNA in complex with the Rev protein, this study does demonstrate the utility of the DiCAT tool to search for patterns in RNA backbone conformation.

Chapter 6: Conclusions

Summary

RNA structure is a complex yet exciting area of study. Thousands of RNA structures exist in the PDB, a ripe source of bioinformatics data ready to be harvested and analyzed. RNA backbone conformations are one component of RNA structure that can be utilized to better understand patterns of RNA structure. This study has helped to define common backbone conformations in RNA structures by using the sugar-to-sugar torsion angles found in di-nucleotide sequences. Lastly, a tool was developed to assign the novel RNA backbone conformations to target input RNA.

Future Work

Discover RNA Conformations Prevalent in RNA-Protein Interactions

With the RNA conformation set and DiCAT tool in place, all RNA structures can be annotated with the conformation nomenclature. RNA structures can then be modeled as sequences of conformation names. These conformation sequences can then be probed with conformation sequences of interest, such as that common to RNA known to interact with protein. Patterns of the conformations can be collected and compared among the RNA structures to determine if the di-nt conformations can be used as building blocks for larger motifs when in complex with protein.

A separate method can model the conformation gallery and probes as discrete three-dimensional objects, and use a block matching algorithm similar to that applied to the human face recognition problem (Podilchuk 2005). Thus the probes can be compared to the

conformations by their shape and volume. The first test for the method can be a full scan of the large ribosomal subunit in order to take classification snapshots of every RNA dinucleotide in the structure. Then, RNA-protein structures can be scanned in order to find those RNA conformations that are common to the RNA-protein interface.

Determine Protein Motifs Associated with Known Protein-Binding RNA

The RNA-protein structures used above can be reanalyzed, this time determining what protein motifs associate with the developed set of RNA conformation patterns. First, those structures supported by literature and annotation can be used to group the proteins by known RNA-binding motifs, such as the RBD and dsRBM. RNA conformations commonly associated with such motifs can be noted. Next, the RNA conformation patterns previously tabulated can cluster all other proteins in complex with RNA. Proteins grouped together can then be compared for structural motifs to determine common modes of RNA binding.

BIBLIOGRAPHY

Aboul-ela, F., J. Karn and G. Varani (1995). "The structure of the human immunodeficiency virus type-1 TAR RNA reveals principles of RNA recognition by Tat protein." J Mol Biol **253**: 313-332.

Aboul-ela, F., J. Karn and G. Varani (1996). "The 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.

Allain, F.H.-T., C.C. Gubser, P.W.A. Howe, K. Nagai, D. Neuhaus and G. Varani (1996). "Specificity of ribonucleoprotein interaction determined by RNA folding during complex formulation." Nature **380**: 646-650.

Allain, F.H.-T., D.E. Gilbert, P. Bouvet and J. Feigon (2000). "Solution structure of the two N-terminal RNA-binding domains determine the RNA binding specificity of nucleolin. J Biol Chem **303**: 227-241.

Allain, F.H.-T., P. Bouvet, T. Dieckmann and J. Feigon (2000). "Molecular basis of sequence-specific recognition of pre-ribosomal RNA by nucleolin." EMBO J **19**: 6870-6881.

Antson, A.A. (2000). "Single stranded RNA binding proteins." Curr Opinion Struct Biol **10**: 87-94.

Argaman, L., R. Hershberg, J. Vogel, G. Bejerano, E.G.H. Wagner, H. Margalit and S. Altuvia (2001). "Novel small RNA-encoding genes in the intergenic regions of *E. coli*." Curr Biol **11**: 941-950.

Avis, J., F.H.-T. Allain, P.W.A. Howe, G. Varani, D. Neuhaus and K. Nagai (1996). "Solution structure of the N-terminal RNP domain of U1A protein: the role of C-terminal residues in structure stability and RNA binding." J Mol Biol **257**: 398-411.

Ban, N., P. Nissen, J. Hansen, P.B. Moore and T.A. Steitz (2000). "The complete atomic structure of the large ribosomal subunit at 2.4 A resolution." Science **289**: 905-920.

Bass, B.L. (2000). "Double-stranded RNA as a template for gene silencing." Cell **101**: 235-238.

Battiste, J.L., H. Mao, N.S. Rao, R. Tan, D.R. Muhandriam, L.E. Kay, A.D. Frankel and J.R. Williamson (1996). "Alpha helix-RNA major groove recognition in an HIV-1 rev peptide-RRE RNA complex." Science **273**: 1547-1551.

Berman, H.M., W.K. Olson, D.L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.-H. Hsieh, A.R. Srinivasan and B. Schneider (1992). "The Nucleic Acid Database: a comprehensive relational database of three-dimensional structures of nucleic acids." Biophys J **63**(3): 751-759.

Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-242.

Caprara, M.G., G. Mohr, and A.M. Lambowitz (1996). "A tyrosyl-tRNA synthetase protein induces tertiary folding of the group I intron catalytic core." J Mol Biol **257**: 512-531.

Colegrove-Otero, L.J., N. Minshall and N. Standart (2005). "RNA-binding proteins in early development." Crit Rev Biochem Mol Biol **40**: 21-73.

Crowder, S.M., R. Kanaar, D.C. Rio and T. Alber (1999). "Absence of interdomain contacts in the crystal structure of the RNA recognition motifs of Sex-lethal." Proc Natl Acad Sci USA **96**: 4892-4897.

Cusack, S., A. Yaremchuk and M. Tukalo (1996). "The crystal structure of *T. thermophilus* lysyl-tRNA synthetase complexed with *E. coli* $tRNA^{Lys}$ and a *T. thermophilus* tRNA(Lys) transcript: anticodon recognition and conformational changes upon binding of a lysyl-adenylate analogue." EMBO J **15**: 6321-6334.

Cusack, S., A. Yaremchuk, I. Krikliviy, and M. Tukalo (1998). "tRNA(Pro) anticodon recognition by *Thermus thermophilus* prolyl-tRNA synthetase." Structure **6**: 101-108.

DeJong, E.S., B. Luy and J.P. Marino (2002). "RNA and RNA-protein complexes as targets for therapeutic intervention." Curr Topic Med Chem **2**: 289-302.

Dock-Bregeon, A.C., B. Chevrier, A. Podjarny, D. Moras, J.S. deBear, G.R. Gough, P.T. Gilham and J.E. Johnson (1988). "High resolution structure of the RNA duplex $[U(U\text{-}A)_6A]_2$." Nature **335**: 375-378.

Duarte, C.M., L.M. Wadley and A.M. Pyle (2003). "RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space." Nucl Acids Res **31**(16): 4755-4761.

Ferre-D'Amare, A.R., K.H. Zhou and J.A. Doudna (1998). "Crystal structure of a hepatitis delta virus ribozyme." Nature **395**(6702): 567-574.

Fierro-Monti, I. and M.B. Mathews (2000). "Proteins binding to duplexes RNA: one motif, multiple functions." Trends Biochem Sci **25**: 241-245.

Fischer, U., J. Huber, W.C. Boelens, I.W. Mattaj and R. Luhrmann (1995). "The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs." Cell **82**: 495-506.

Frankel, A.D. and C.A. Smith (1998). "Induced folding in RNA-protein recognition: more than a simple molecular handshake." Cell **92**: 149-151.

Gait, M.J. And J. Karn (1993). "RNA recognition by the human immunodeficiency virus Tat and Rev proteins." Trends Biochem Sci **18**: 255-259.

Ginisty, H., F. Amalric and P. Bouvet (2001). "Two different combinations of RNA-binding domains determine the RNA binding specificity of nucleolin." J Biol Chem **276**: 14338-14343.

Goldgur, Y., L. Mosyak, L. Reshetnikova, V. Ankilova, O. Lavrik, S. Khodyreva, and M. Safro (1997). "The crystal structure of phenylalanyl-tRNA synthetase from *Thermus thermophilus* complexed with cognate tRNA(Phe)." Structure **5**: 59-68.

Gosser, Y., Hermann, T., Majumdar, A., Hu, W., Frederick, R., Jiang, F., Xu, W. and D.J. Patel (2001). "Peptide-triggered conformational switch in HIV-1 RRE RNA complexes." Nature Structural Biology **8**: 146-150.

Grishok, A., A.E. Pasquinelli, D. Conte, N. Li, S. Parrish, I. Ha, D.L. Baillie, A. Fire, G. Ruvkun and C.C. Mello (2001). "Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing." Cell **106**: 23-34.

Gubser, C.C. and G. Varani (1996). "Structure of the polyadenylation regulatory element of the human U1A pre-mRNA 3'-untranslated region and interaction with the U1A protein." Biochemistry **35**: 2253-2267.

Hall, K.B. (2002). "RNA-protein interactions." Curr Opinion Struct Biol **12**: 283-288.

Handa, N., O. Nureki, K. Kurimoto, I. Kim, H. Sakamoto, Y. Shimura, Y. Muto and S. Yokoyama (1999). "Structural basis for recognition of the *tra* mRNA precursor by the sex-lethal protein." Nature **398**: 579-585.

Hershkovitz, E., E. Tannenbaum, S.B. Howerton, A. Sheth, A. Tannenbaum and L.D. Williams (2003). "Automated identification of RNA conformational motifs: Theory and application to the HM LSU 23S rRNA." Nucl Acids Res **31**(21): 6249-6257.

Inoue, K., K. Hoshijima, H. Sakamoto and Y. Shimura (1990). "Binding of the *Drosophila sex-lethal* gene-product to the alternative splice site of *transformer* primary transcript." Nature **344**: 461-463.

Izaurralde, E., J. Stepinski, E. Darzynkiewicz and I. Mattaj (1992). "A cap binding protein may mediate nuclear export of RNA polymerase II-transcribed RNAs." J Cell Biol **118**: 1287-1295.

Jones, K. and B. Peterlin (1994). "Control of RNA initiation and elongation at the HIV-1 promoter." Annu Rev Biochem **63**: 717-743.

Kanaar, R., A.L. Lee, D.Z. Rudner, D.E. Wemmer and D.C. Rio (1995). "Interaction of the sex-lethal RNA-binding domains with RNA." EMBO J **14**: 4530-4539.

Kielkopf C.L., N.A. Rodionova, M.R. Green and S.K. Burley (2001). "A novel peptide recognition mode revealed by the x-ray structure of a core $U2AF^{35}/U2AF^{65}$ heterodimer." Cell **106**: 595-605.

Kim, S.-H., H.M. Berman, N.C. Seeman and M.D. Newton (1973). "Seven basic conformations of nucleic acid structural units." Acta Crystallogr **B29**: 703-710.

Kranz, J.K. and K.B. Hall (1999). "RNA recognition by the human U1A protein is mediated by a network of local cooperative interactions that create the optimal binding surface." J Mol Biol **285**: 215-231.

Lagos-Quintana M., R. Rauhut, W. Lendeckel and T. Tuschl (2001). "Identification of novel genes coding for small expressed RNAs." Science **294**: 853-858.

Le Hir, H., D. Gatfield, E. Izaurralde and M.J. Moore (2001). "The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay." EMBO J **20**: 4987-4997.

Lehman, K.A. and B.L. Bass (2000). "Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities." Biochemistry **39**: 12875-12884.

Leulliot, N. and G. Varani (2001). "Current topics in RNA-protein recognition: control of specificity and biological function through induced fit and conformational capture." Biochemistry **40**: 7947-7956.

Long, K.S. and D.M. Crothers (1995). "Interaction of human immunodeficiency virus type 1 Tat-derived peptides with TAR RNA." Biochemistry **34**: 8885-8895.

Mittermaier, A., L. Varani, D.R. Muhandiram, L.E. Kay and G. Varani (1999). "Changes in side-chain and backbone dynamics identify determinants of specificity in RNA recognition by human U1A protein." J Mol Biol **294**: 967-979.

Moore, P.B. (1999). "Structural Motifs in RNA." Annu Rev Biochem **68**: 287-300.

Murray, L.J.W., W.B. Arendall III, D.C. Richardson and J.S. Richardson (2003). "RNA backbone is rotameric." Proc Natl Acad Sci USA **100**(24): 13904-13909.

Murthy, V.L. and G.D. Rose (2003). "RNABase: an annotated database of RNA structures." Nucleic Acids Res **31**(1): 502-504.

Nanduri, S., B.W. Carprick, Y. Yang, B.R.G. Williams and J. Qin (1998). "Structure of the double-stranded RNA-binding domain of the protein kinase PKR reveals the molecular basis of its dsRNA-mediated activation." EMBO J **17**: 5458-5465.

Neidle, S. (2002). Nucleic Acid Structure and Recognition. New York: Oxford University Press Inc.

Olson, W.K. (1982). In Topics in Nucleic Acid Structure, Part 2 (ed. Neidle, S.), pp. 1-79. Macmillan Press, London.

Orengo, C.A. and J.M. Thorton (1993). "Alpha plus beta folds revisited: some favoured motifs." Structure **1**: 105-120.

Oubridge, C., N. Ito, P.R. Evans, C.H. Teo and K. Nagai (1994). "Crystal-structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin." Nature **372**: 432-438.

Packer, M.J. and C.A. Hunter (1998). "Sequence-dependent DNA structure: the role of the sugar-phosphate backbone." J Mol Biol **280**(3): 407-420.

Patton, J.G., S.A. Mayer, P. Tempst and B. Nadalginard (1991). "Characterization and molecular cloning of polypyrimidine tract binding protein. A component of a complex necessary for pre-mRNA splicing." Genes Dev **5**: 1237-1251.

Peterson, R.D., D.P. Bartel, J.W. Szostak, S.H. Horwath and J. Feigon (1994). "1H NMR studies of the high-affinity Rev binding site of the Rev responsive element of HIV-1 mRNA: base pairing in the core binding element." Biochemistry **33**: 1026-1033.

Peterson, R.D. and J. Feigon (1996). "Structural change in Rev responsive element RNA of HIV-1 on binding Rev peptide." J Mol Biol **264**: 863-877.

Podilchuk, C., A. Patel, A. Harthattu, S. Anand and R. Mammone (2005). "A new face recognition algorithm using bijective mappings." In Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on, p 165.

Price, S.R., P.R. Evans and K. Nagai (1998). "Crystal structure of the spliceosomal U2B''-U2A' protein complex bound to a fragment of U2 small nuclear RNA. Nature **394**: 645-650.

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

Query, C.C., S.A. Strobel and P.A. Sharp (1996). "Three recognition events at the branch point adenine." EMBO J **15**: 1392-1402.

Richardson J.S., Schneider B., Murray L.W., Kapral G.J., Immormino R.M., Headd J.J., Richardson D.C., Ham D., Hershkovits E., Williams L.D., Keating K.S., Pyle A.M., Micallef D., Westbrook J., and H.M. Berman (2008). "RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution)." RNA **14**: 465-481.

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

Rould, M.A., J.J. Perona and T.A. Steitz (1991). "Structural basis of anticodon loop recognition by glutaminyl-transfer RNA-synthetase." Nature **352**: 213-218.

Ruvkun, G. (2001). "Glimpses of a tiny RNA World." Science **294**: 797-799.

Ryter, J.M. and S.C. Schultz (1998). "Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA." EMBO J **17**: 7505-7513.

Saenger, W. (1984). Principles of Nucleic Acid Structure. New York: Springer-Verlag.

Sankaranarayanan, R., A.C. Dock-Bregeon, P. Romby, J. Caillet, M. Springer, B. Reesc, C. Ehresmann, B. Ehresmann and D. Moras (1999). "The structure of threonyl-tRNA synthetase-tRNA(Thr) complex enlightens its repressor activity and reveals an essential zinc ion in the active site." Cell **97**: 371-381.

Scherly, D., C. Kambach, W. Boelens, W.J. van Venrooij and I.W. Mattaj (1991). "Conserved amino acid residues within and outside of the N-terminal ribonucleoprotein motif of U1A small nuclear ribonucleoprotein involved in U1 RNA binding." J Mol Biol **219**: 577-584.

Schneider, B., D.M. Cohen, L. Schleifer, A.R. Srinivasan, W.K. Olson and H.M. Berman (1993). "A systematic method to study the spatial distribution of water molecules around nucleic acid bases." Biophys J **65**: 2291-2303.

Schneider, B., S. Neidle and H.M. Berman (1997). "Conformations of the sugar-phosphate backbone in helical DNA crystal structures." Biopolymers **42**: 113-124.

Schneider, B., Z. Morávek and H.M. Berman (2004). "RNA conformational classes." Nucleic Acids Res **32**(5): 1666-1677.

Shen, L.X., Z. Cai and I. Tinoco, Jr. (1995). "RNA structure at high resolution." The FASEB Journal **9**: 1023-1033.

Sosnowski, B.A., J.M. Belote and M. McKeown (1989). "Sex-specific alternative splicing of RNA from the *transformer* gene results from sequence-dependent splice site blockage." Cell **58**: 449-459.

Spanggord, R.J. and P.A. Beal (2001). "Selective binding by the RNA binding domain of PKR revealed by affinity cleavage." Biochemistry **40**: 4272-4280.

Stutz, F., M. Neville and M. Rosbash (1995). "Identification of a novel nuclear pore-associated protein as a functional target of the HIV-1 Rev protein in yeast." Cell **82**: 495-506.

Sundaralingam, M. (1969). "Stereochemistry of nucleic acids and their constituents." Biopolymers **7**: 821-860.

Sussman, J.L., Kim, S.-H., and H.M. Berman (1972). "Crystal structure of a naturally occurring dinucleoside phosphate: uridylyl 3',5'-adenosine phosphate model for RNA chain folding." J Mol Biol **66**: 403-421.

Tan, R. and A.D. Frankel (1994). "Costabilization of peptide and RNA structure in an HIV Rev peptide-RRE complex." Biochemistry **33**: 14579-14585.

Tanaka, Y., S. Fujii, H. Hiroaki, T. Sakata, T. Tanaka, S. Uesugi, K. Tomita and Y. Kyogoku (1999). "A'-form RNA double helix in the single crystal structure of r(UGAGCUUCGGCUC)." Nucl Acids Res **27**(4): 949-955.

ten Dam, E., K. Pleij and D. Draper (1992). "Structural and functional aspects of RNA pseudoknots." Biochemistry **214**: 437-453.

Tian, B. and M.B. Mathews (2001). "Functional characterization of and cooperation between the double-stranded RNA-binding motifs of the protein kinase PKR." J Biol Chem **276**: 9936-9944.

Tinoco, I.J. and C. Bustamente (1999). "How RNA folds." J Mol Biol **293**: 271-281.

Tocilj, A., F. Schlunzen, D. Janell, M. Gluhmann, H.A.S. Hansen, J. Harms, A. Bashan, H. Bartels, I. Agmon, F. Franceschi, *et al.* (1999). "The small ribosomal subunit from *Thermus thermophilus* at 4.5 A resolution: pattern fittings and the identification of a functional site. Proc Natl Acad Sci USA **96**: 14252-14257.

Valegard, K., J.B. Murray, P.G. Stockley, N.J. Stonehouse and L. Liljas (1994). "Crystal structure of an RNA bacteriophage coat protein-operator complex." Nature **37**: 623-626.

Varani, G. (1997). "RNA-protein intermolecular recognition." Acc Chem Res **30**: 189-195.

Varani, G. and K. Nagai (1998). "RNA recognition by RNP proteins during RNA processing." Annu Rev Biophys Biomol Struct **27**: 407-445.

Wang, X. and T.M. Tanaka Hall (2001). "Structural basis for recognition of AU-rich element RNA by the HuD protein." Nat Struct Biol **8**: 141-145.

Weeks, K.M. and T.R. Cech (1996). "Assembly of a ribonucleoprotein catalyst by tertiary structure capture." Science **271**: 345-348.

Williamson, J.R. (2000). "Induced fit in RNA-protein recognition." Nat Struct Biol **7**: 834-837.

Wimberly, B.T., D.E. Brodersen, W.M. Clemons, Jr., R.J. Morgan-Warren, A.P. Carter, C. Vonrhein, T. Hartsch and V.R. Ramakrishnan (2000). "Structure of the 30S ribosomal subunit." Nature **407**: 327-332.

Witherall, G.W., A. Gil and E. Wimmer (1993). "Interaction of polypyrimidine tract binding protein with the encephalomyocarditis virus mRNA internal ribosomal entry site." Biochemistry **32**: 8268-8275.

Yang, H., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H.M., Westhof, E. (2003). "Tools for the automatic identification and classification of RNA base pairs." Nucleic Acids Research **31.13**: 3450-3460.

Ye, X., A. Gorin, A.D. Ellington and D.J. Patel (1996). "Deep penetration of an alpha-helix into a widened RNA major groove in the HIV-1 rev peptide-RNA aptamer complex." Nat Struct Biol **3**: 1026-1033.

Yusupova, G.Z., M.M. Yusupov, J.H.D. Cate and H.F. Noller (2001). "The path of messenger RNA through the ribosome." Cell **106**(2): 233-241.

Zacharias, M. and P.J. Hagerman (1995). "The bend in RNA created by the trans-activation response element bulge of human immunodeficiency virus is straightened by arginine and by Tat-derived peptide." Proc Natl Acad Sci USA **92**: 6052-6056.

Zamore, P.D., J.G. Patton and M.R. Green (1992). "Cloning and domain structure of the mammalian splicing factor U2AF." Nature **355**: 609-614.

Zeng, Q. and K.B. Hall (1997). "Contribution of the C-terminal tail of U1A RBD1 to RNA recognition and protein stability." RNA **3**: 303-314.