RNA Backbone Validation, Correction, and Implications for RNA-Protein Interfaces

by

Gary Kapral

Department of Biochemistry
Duke University

Date:_____

Approved:

_____
Jane S. Richardson, Co-Supervisor

_____
David C. Richardson, Co-Supervisor

_____
Michael D. Been

_____
Arno L. Greenleaf

_____
Barbara R. Shaw

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Biochemistry in the Graduate School
of Duke University

2013

ABSTRACT

RNA Backbone Validation, Correction, and Implications for RNA-Protein Interfaces

by

Gary Kapral

Department of Biochemistry
Duke University

Date:_____

Approved:

_____
Jane S. Richardson, Co-Supervisor

_____
David C. Richardson, Co-Supervisor

_____
Michael D. Been

_____
Arno L. Greenleaf

_____
Barbara R. Shaw

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the Department of
Biochemistry in the Graduate School
of Duke University

2013

Abstract

RNA is the molecular workhorse of nature, capable of doing many cellular tasks, from genetic data storage and regulation, to enzymatic synthesis—even to the point of self-catalyzing its own replication. While RNA can act as a catalyst on its own, as in the hammerhead ribozyme, the added efficiency of proteins is often a necessity; the ribosome—the large ribozyme responsible for peptide chain formation, is aided by proteins which ensure correct assembly and structural stability. These complexes of RNA and proteins feature in many essential cellular processes, including the RISC silencing complex and in the spliceosome. Despite its enormous utility, structural determination of RNA is notoriously difficult—particularly in the backbone, since a nucleotide standardly has 12 torsion angles (including χ) and 12 non-hydrogen atoms, compared to 4 torsions (including χ_1) and 4 non-H atoms in a typical amino acid. The abundance of backbone atoms, their conformational flexibility, and experimental resolution limitations often result in systematic errors that can have a significant impact on the interpretation. False trails due to structural errors can lead to significant loss of time and effort, especially with such high-profile complexes as the ribosome, telomerase, or the RISC complex.

My research has focused on harnessing the recently discovered ribosome structures and the Richardsons' RNA dataset to find trends in RNA backbone

conformations and motifs that were then used to develop structural validation techniques and provide improved diagnosis and correction techniques for RNA backbone. Methods for fixing RNA structure have been developed herein for both NMR and X-ray crystallography. For NMR structures, a method for assigning RNA backbone structure based on NOE data was developed, leading to improved identification and building of RNA backbone conformation in NMR ensembles. For crystallography, our method of diagnosing the correct ribose pucker from clear observables allows reliable assessment of pucker in validation or refinement. Observed differences in bond-lengths, bond-angles, and dihedrals have been categorized by sugar pucker in the PHENIX refinement package. I have shown that this improves the refinement behavior of both pucker and geometry.

I have also made improvements in how structural motifs are identified in RNA. Many previously studied structural motifs have now been defined in terms of backbone suitestrings, a series of 2-character code divisions of RNA backbone that show the best clustering of dihedral angle correlations. Combined with our BLAST-like alignment program called SuiteAlign, these suitestrings were quickly and easily identified in a number of structures, eventually leading to the discovery of known motifs in new places, such as the multiple instances of TΨC-loop structures in the ribosome, as well as novel motifs, like the OHO pentaloop, whose poor sequence conservation had hidden them from traditional motif identification techniques.

To facilitate error diagnosis and corrections in RNA-protein complexes, as well as to expand the knowledge base of the scientific community as a whole, a dataset of model RNA-protein complexes has been determined, rooted in the quality-filtering, visualization, and analysis techniques of the Richardson lab, particularly those developed by Laura Murray specifically for RNA structures. From this dataset, key features of the RNA-protein interface have been identified, and a set of RNA backbone motifs along the interface has been described.

Taken together, the consensus RNA backbone conformers, sugar pucker diagnosis, and all-atom contacts have been combined to develop first manual and then automated tools for RNA structure correction; the newly defined structural motifs described here provide a new method of validating the structure. I have applied all these techniques to improve the accuracy of a number of important RNA and RNA/protein complex structures, including the *E. coli* ribosome, human 3′ exonuclease, and the HDV ribozyme, greatly improving the initial models and uncovering interesting information about RNA-protein interactions.

Dedication

For Amy and Gary and Christopher

Contents

List of Tables

List of Figures

List of Symbols and Abbreviations

Å	Ångstrom (1Å = $1x10^{-10}$ m)
α (alpha)	torsion (or dihedral angle), O3′-P-O5′-C5′ (see Figure 8 for backbone torsions)
β (beta)	torsion, P-O5′-C5′-C4′
γ (gamma)	torsion, O5′-C5′-C4′-C3′
δ (delta)	torsion, C5′-C4′-C3′-O3′
ϵ (epsilon)	torsion, C4′-C3′-O3′-P
ζ (zeta)	torsion, C3′-O3′-P-O5′
χ (chi)	torsion, O4′-C1′-N1-C2 for pyrimidine or O4′-C1′-N9-C4 for purine
ν_0	internal ribose ring torsion angle, (C4′-O4′-C1′-C2′)
ν_1	internal ribose ring torsion angle, (O4′-C1′-C2′-C3′)
ν_2	internal ribose ring torsion angle, (C1′-C2′-C3′-C4′)
ν_3	internal ribose ring torsion angle, (C2′-C3′-C4′-O4′)
ν_4	internal ribose ring torsion angle, (C3′-C4′-O4′-C1′)
δ_{n-1}	torsion for previous residue
θ (theta)	pseudotorsion, P_{n-1}-C4′$_{n-1}$-P-C4′
η (eta)	pseudotorsion, C4′$_{n-1}$-P-C4′-P_{n+1}
φ (phi)	protein backbone torsion, C-Cα-N-C
ψ (psi)	protein backbone torsion, N-C-Cα-N
Ψ (Psi)	pseudouridine (5-ribosyluracil)

CSD	Cambridge Structural Database
CTD	C-terminal domain
DNA	deoxyribonucleic acid
EF	elongation factor
ERRASER	Enumerated Real-space Refinement ASisted by Electron density under Rosetta
FT	Fourier transform
dsRNA	double-stranded RNA
miRNA	micro RNA
mRNA	messenger RNA
MSE	selenomethionine
NC	Non-canonical (non-Watson-Crick base pair)
ncRNA	noncoding RNA
NDB	Nucleic Acid Database
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
nps	non-pucker specific parameter
nt	nucleotide
P-perp	Perpendicular from a 3′ phosphate to the base plane or glycosidic bond vector
PDB	Protein Data Bank
PHENIX	Python-based Hierarchical Environment for Integrated Xtallography

ps	Pucker-specific parameter
RDC	residual dipolar coupling
RNA	ribonucleic acid
RNAi	RNA interference
RNP1	ribonucleoprotein binding domain 1 (octameric consensus sequence)
RNP2	ribonucleoprotein binding domain 2 (hexameric consensus sequence)
ROC	RNA Ontology Consortium
rRNA	ribosomal RNA
SCOR	Structural Classification of RNA database
SE	sugar-edge face of nucleic acid base
siRNA	small interfering RNA
SL	stem-loop
SLBP	stem-loop binding protein
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleic particle
TΨC	characteristic motif found in the T loop in tRNA
tRNA	transfer RNA
vdW	van der Waals
WC	Watson-Crick face of nucleic acid base
1MA	1-methyl adenine

Acknowledgements

I would like to acknowledge all of my family and friends who helped me get to this stage in my career. Of particular note are my wife and sons who provided motivation and necessary distraction, my parents and grandparents who would always provide encouragement, and the smallgroup for being there throughout my graduate career and reminding me that it's ok to be a scientist and still think Ecclesiastes 12 :12 is true.

I would also like to acknowledge my former professors who have fostered and encouraged my interest in biochemistry. Hubert Avery, who demanded constant attention to details with his draconian grading scheme that led to many nights "burnin' the midnight o'l", as he would say. Rebekah Owens, who opened my eyes to the wonder of chemistry and recognized I could put my talents to use at the North Carolina School of Science and Mathematics. Charles Roser, Noreen Naiman, and Sarah Allen, my chemistry instructors at NCSSM in advanced chemistry, biochemistry and polymer chemistry, respectively. Edward Tokas and Robert Gotwals, who mentored me in my first independent research projects. Marvin Illingsworth, mentor in inorganic and polyimide synthesis, and Suzanne O'Handley, whose excellent tutelage convinced me to pursue a graduate career in biochemistry.

I would also like to acknowledge the faculty at Duke, particularly Mike Been, Arno Greenleaf, and Tao-shih Hsieh for discussions on nucleic acids and Terry Oas for his guidance on statistics and thermodynamics.

Of course I would never have gotten this far without a lot of help from my current advisors Jane and David Richardson. Their willingness to work with me for the years it took is a testament both to their excellent mentoring abilities, not to mention their patience. Their model of mentorship will ever be the ideal which I strive to obtain in my academic career.

The ROC consortium and the PHENIX consortium have greatly influenced my graduate career, and I would like to particularly thank Jesse Stombaugh and Kevin Keating of ROC and Nigel Moriarty and Ralf Gross-Kunstleve of PHENIX for their support and aid through the years.

All the Rlab folks, thank you for the time we've spent working together. Not that things will change too much when I graduate, but still. Laura, thank you for setting the foundation for my own work, and for being someone I could talk to about all things RNA and non-RNA; for better or worse, I picked up most of my grad school habits from you. Swati, thank you for all your continuing work on RNA backbone—I know the future of the project is safe in your hands. Lizbeth, Bryan, Vince, Jeremy, and Ian, your presence in the lab is what made me decide to forgo my computational chemistry ways to venture into structural biology. Keedy, Christopher and Bradley, thank you for

helping me tackle the protein side of these complexes, and Bradley—watch your step, nucleic acid structures are a slippery slope into an entirely different world. Lindsay, it's awesome that you brought crystallography back to the lab, and thank you for teaching me how to actually do some of the experimental side of things.

Pat and Carrie, Vince and Stefanie, Ian and Katy, and Bob—it's been great gaming with you guys over the years. Though we're scattered at the moment, I hope time brings us back together to adventure again.

Chris and Joanna, Randy, Kyle, Nick, Steven—thank you for being there for me emotionally, scientifically, philosophically, and spiritually. You all have had a great influence on my life and I would never have made it this far without you.

All the departmental and organization staff who had a hand in my graduate school career, thank you. Amy Norfleet, you've been especially patient with me, and the same can be said of Marsha and Esther in Biochemistry and Carol Richardson with SBB and Annette with the Center for RNA Biology. Thank you for all the hard work you do, and for making sure the funding and departmental paperwork got handled without too much hassle.

Thanks to Duke Biochemistry, SBB, CRB, and the NIH for funding and for travel fellowships.

And finally, I would like to acknowledge my committee, who were always full of good ideas about how to take RNA backbone research to the next level.

1. Introduction

1.1 Where we are coming from: a brief history of RNA

Ribonucleic acid (RNA) is one of the most versatile biological polymers on earth, both in structure and in function. Structurally, RNA consists of a chain of ribonucleotides connected via phosphodiester bonds; each ribonucleotide contains a cyclic *d*-ribose with a phosphate covalently bound to C5′ and a base forming a glycosidic bond to C1′ (Figure 1). An RNA strand generally contains only four base types: Adenine, Guanine, Cytosine and Uracil, with occasional modified bases. Cytosine and Uracil are pyrimidine derivatives which bind to the ribose C1′ at the N1 position. Adenine and Guanine are purine derivatives, consisting of a fused pyrimidine ring and imidazole ring; the purines bind to the ribose C1′ at N9. From a primary structure point of view, RNA appears to be very similar to DNA (deoxyribonucleic acid), though RNA has a 2′ hydroxyl on the ribose and DNA uses thymine (5-methyl uracil) in place of uracil. Indeed, for some time DNA and RNA were both thought to be alternate versions of a single acid isolated from the cellular nuclei, with RNA residing in plants, and DNA in animals (Allen 1941)—thus their common moniker of nucleic acids. RNA's association with protein synthesis was discovered very early: RNA could be isolated from the cytoplasm and higher concentrations of RNA were present in the cytoplasm

during protein generation in rapidly growing cells (Caspersson 1939). But study of RNA was overshadowed for over a decade while DNA took the spotlight: the Avery-MacLeod-McCarty experiment showed that DNA was responsible for passing on genetic information (Avery 1944), Chargaff established the ratio of AT and GC pairs (Chargaff 1952), Hershey and Chase confirmed DNA as the vehicle for inheritance (Hershey and Chase 1952), and Watson and Crick solved the first structure of DNA (Watson and Crick 1953). Finally, in 1955, research once again began to focus on RNA, as Goldstein and Plaut showed that the RNA obtained from the cytoplasm had been synthesized in the nucleus (Goldstein and Plaut 1955). Based on this work, the Central Dogma of Molecular Biology was put forth by Crick (Crick 1958): DNA is transcribed to RNA through complementary pairing, which then is translated to protein (and never the reverse); RNA's nebulous role as a intermediary between DNA and protein was established.

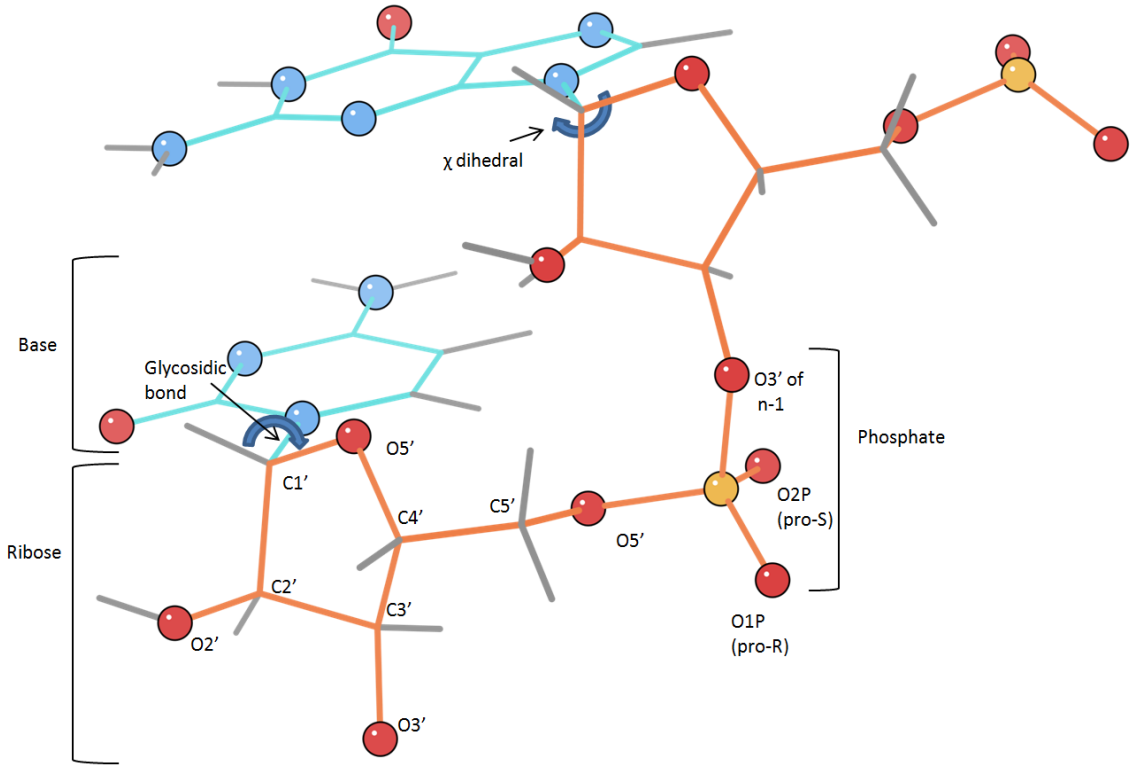


Figure 1: Overview of an RNA Residue.

The base, ribose, and phosphate that make up the nucleotide are labeled, as well as the major atoms of the RNA backbone. The glycosidic bond, and the χ dihedral around it, are also displayed.

Later studies classified RNA as a trans-nuclear messenger (mRNA), using its base complementarity with DNA to transcribe genetic data and sending the resulting RNA transcript out of the nucleus to the ribosome for translation (Crick 1970). The discovery of transfer RNA (tRNA) expanded RNA's role in translation (Holley 1965), and the clover-leaf secondary structure put forth by Holley was soon confirmed by

crystal structures (Ladner 1975, Kim 1973). Unlike mRNA, tRNA adopted a very specific L-shaped tertiary structure (Figure 2). This structure provided the answer for how the 3-nucleotide (nt) codon of mRNA (Nirenberg 1965) was translated to a single amino acid: tRNAs—each of which has a particular amino acid bound to the CCA stem—attach to mRNA via an anticodon found at the base of the tRNA structure, which is complementary to the correct mRNA codon. A third type of RNA was discovered in 1956 when George Palade discovered that the microsome organelle was RNA-rich, containing slightly more RNA than protein (Palade and Siekevitz 1956). By 1960, it was shown that the microsomes, renamed ribosomes, consisted of two subunits, each consisting of RNA and protein (Tissieres and Watson 1958), and were responsible for protein synthesis (Siekevitz and Palade 1958). The conflux of mRNA, tRNA, and ribosomal RNA (rRNA) in translation was shown by modification of the 16S ribosomal subunit (Noller and Chaires 1972). It was also discovered that the sequence of the 16S subunit was highly conserved among species; Carl Woese used this information to propose a new phylogenetic taxonomy based around 3 domains: Eukaryotes, Bacteria, and a new group called Archaea (Woese and Fox 1977).

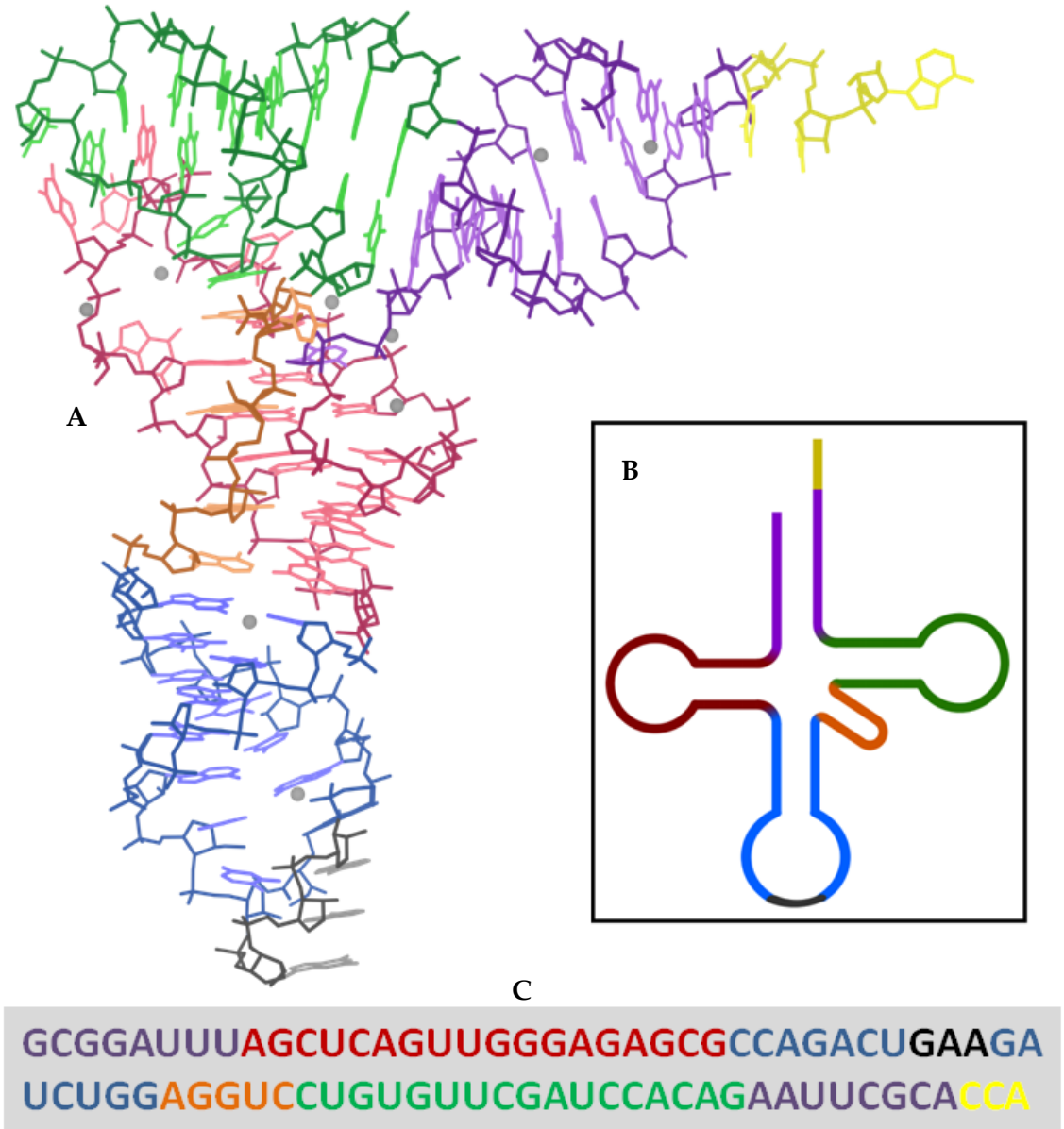


Figure 2: Primary, Secondary, and Tertiary structure of a tRNA.

The inverted L-shaped tertiary structure of the phenylalanine tRNA ($tRNA^{PHE}$) in (A) is colored to match its secondary structure (B), and primary sequence (C). The anticodon triplet is in black.

In addition to RNA's important roles in gene transcription and translation, it can also act as a catalytic agent. In 1967, Carl Woese suggested that the 2' hydroxyl could serve as a site for catalysis (Woese 1967). He focused on RNA's ability to store genetic information and putative catalytic ability as the basis for early life, an idea formalized and expanded on by Walter Gilbert in his *Nature* paper entitled *The RNA World* (Gilbert 1986). Thomas Cech and Sidney Altman discovered evidence of such catalytic "ribozymes" (Zaug 1986, Been 1986, Guerrier-Takada 1983) in *Tetrahymena* transcript self-splicing and RNase P catalysis. It has since come to light that RNA can catalyze the formation of RNA; a particularly good example is two artificially designed ribozymes that catalyze each other's formation indefinitely so long as resources are available (Lincoln 2009).

Ribozymes, tRNA, and rRNA belong to a group of RNAs called noncoding RNAs (ncRNAs) because, unlike mRNA, they do not code for amino acids. A number of uses have been found for ncRNA, particularly in gene splicing and regulation. Small, low-molecular weight RNAs found in the nucleus (snRNA) engage in complex interactions with particular proteins to form small nuclear ribonucleoproteins (snRNPs) (Lerner 1979, Lerner 1980). These snRNPs have very prominent roles in pre-mRNA processing, particularly in removing introns from the post-transcriptional pre-mRNA, as part of a large RNA-protein complex called the spliceosome (Jurica 2003). Some introns

bypass the need for a spliceosome: the group I and group II introns can self-catalyze, causing their own cleavage out of the transcript (Stahley 2005, Toor 2008).

RNA's role in gene regulation was expanded through the processes of SELEX and *in vitro* selection, which showed that particular mRNA transcripts could be engineered to form ligands, called aptamers, that bound to a particular target (Tuerk 1990; Ellington 1990). This set off a search for natural aptamer occurrences, resulting in the description of riboswitches (Nahvi 2002)—RNA sequences that bound directly to small metabolites, such as free guanine or thiamine pyrophosphate (TPP), forming a stable tertiary structure that prevented transcription of that region. Many riboswitches affect downstream translation of nearby genes, turning them on and off according to the presence of the riboswitch's target metabolite (Nahvi 2002). Some do so directly by folding to sequester the Shine-Dalgarno sequence, thus inhibiting translation (Winkler 2002).

A third, novel method of gene regulation was determined by Andrew Fire and Craig Mello (Fire 1998). Dubbed RNA interference, this method of regulation depends on short RNAs known as micro RNA (miRNA) and small interfering RNA (siRNA). Both begin as rather long double-stranded RNA (dsRNA) chains which are then converted to 21-nt dsRNA by the appropriately named enzyme Dicer (Bernstein 2001). The dsRNA is unwound into 2 single strands, one of which is degraded while the other binds to the RNA-induced silencing complex (RISC), a multi-protein complex with

RNase activity. RISC uses this bound miRNA or siRNA as a guide to bind with the target mRNA transcripts via base complementarity. The part of RISC with RNase activity, the Argonaute protein, then makes an endonucleolytic cut in the mRNA strand, cleaving the strand and preventing further expression. The main difference in siRNA and miRNA lies in how they accomplish their regulatory tasks. The siRNA exhibits perfect base pairing with its target and causes cleavage of its target in almost 100% of cases. On the other hand, miRNA is more lenient in where it binds to the mRNA, allowing some non-Watson-Crick base pairing. The miRNA-RISC complex still causes translation repression, but rather than degrading the target, the mRNA is moved to P-bodies (processing bodies) where other proteins determine whether the mRNA is degraded or stored until needed (Bashkirov 1997, Kulkarni 2010).

These myriad roles of RNA in controlling gene expression show how our understanding of RNA function has shifted in the past hundred years of study, from that of an intermediate, less efficient version of DNA to an integral part in cellular vitality. RNA has been shown to be a crucial factor in each step of protein synthesis, from providing the mRNA transcript to post-transcriptional regulation to peptide synthesis. Despite this, our understanding of RNA structure is woefully inadequate. RNA is difficult to crystallize, a fact borne out by the low number of RNA-containing structures in the Protein Data Bank (PDB): as of March, 2013, only 2522 RNA structures are in the PDB vs. 86,600 protein structures and 4177 DNA structures (Berman 2000).

Thus RNA accounts for just over a third of the total deposited nucleic acid structures. The relative lack of nucleic acid structures in general has a somewhat sinister effect in structural biology: because they differ by a single backbone atom and a single methyl group on a single base, many model building and prediction software packages treat both nucleic acids the same. This works poorly, because DNA has evolved for fidelity of information storage in double-helix form while RNA adopts very complex and diverse tertiary structures and functions, so that their properties are actually quite different.

In order to facilitate RNA study and capitalize on new results, structural biologists need a better way of looking at RNA structural interactions, and a nomenclature/classification system which reflects the uniqueness of RNA's backbone structure that strongly influences its ability to act as a genetic regulator and a catalyst. The development of a common classification system and tools with which to analyze the existing data and improve the sparse number of available models will go a long way to giving RNA biologists the insight into RNA structure they need to describe better experiments and determine new RNA structures. Because the amount of data available for study has recently reached the critical mass necessary to make an undertaking such as this feasible, the laying of the foundation for the new investigations of RNA backbone structure and interactions can now begin.

1.2 The chemical and structural importance of RNA backbone

The first steps to understanding how RNA performs such a wide variety of vital functions is to understand its structure. Given its important role, it is somewhat surprising, and disappointing, that the number of RNA structures deposited in the Nucleic Acid Database (NDB; Berman 1992) and PDB was very low at the turn of the millennium (236 RNA vs. 939 DNA and 9812 proteins). A part of this was simply due to lack of information—RNA primarily adopted A-form helix, which was structurally similar to A-form DNA; but when it did not, there were few methods of recourse to determine its structure. Even the refinement parameters used in X-PLOR and CNS (Parkinson 1996, Brunger 1998) were largely generalized to both nucleic acids, and tested almost exclusively on DNA structures. Thus, many early models of RNA structures resembled their DNA counterparts with only the very few high resolution structures working well to illustrate non-A-form RNA structure. Many early structures show clear distinction between the standardization and quality of A-form helix versus the intervening RNA "loop" regions that connect them (Duarte and Pyle 1998; Ferre-D'Amare 1998).

Fortunately, the publication of the ribosome structure in 2000, proving structurally that the ribosome is really a ribozyme (Ban 2000; Nissen 2000), set off an explosion of RNA structural studies that continues to this day. The revitalized focus on structure helped prove the existence of riboswitches (Winkler and Batey 2005), and

aptamer structures have been used to further medical advances in viral transcription inhibition (Eulberg and Klussman 2003). The structure of the RNA Induced Silencing Complex (RISC) has been solved, giving us more insight into RNAi and the siRNA and miRNA that drive the process and keep the body safe from invading RNA-based pathogens (Fire 1998; Tuschl 2001). Improved structures of the ribosome have elucidated translation even further, and led to a myriad of drugs designed to inhibit translation in pathogenic species (Matt 2012; Borovinskaya 2007; Bulkley 2010).

Traditionally, much of the study surrounding RNA has focused on the bases. The reasons for this are twofold. Chemically, they are variable pendant groups in a polymer of ribose-phosphodiester linkages, and with their strong tendency to form complementary base-pairs, changing the sequence of pendant groups is an obvious first step if one wishes to alter the polymer structure. Secondly, their positions are easy to determine using both X-ray crystallography and NMR spectroscopy. Base pairing and base stacking interactions stabilize the overall RNA structure, allowing them to be easily seen through NMR, and their electron-rich pseudo-aromatic properties make them easy to identify in the electron density. In addition, base positions can easily be inferred on the bases of Watson-Crick complementarity. Overall, the base sequence is easily changed, and the results of those changes are fairly easy to determine.

Yet much of the recent research, particularly concerning catalysis, has focused on the properties of the 2′ hydroxyl of the ribose. The 2′ hydroxyl has an extreme effect on

RNA structure. From a purely steric point of view, it prevents RNA from adopting the low-energy B-form helix common to most DNA structures; most RNA is in A-form helix (Murray 2003) which exposes the bases and the 2′ hydroxyl to solvent. The 2′ hydroxyl is responsible for self-splicing, as in the case of the Group I and Group II introns (Cate 1996; Toor 2008). It also can act as a proton shuttle, as in the peptidyl transferase site of the ribosome, catalyzing the binding of the amino-acyl amino from A-site tRNA to the amino-acyl carbonyl carbon of the P-site tRNA, thus lengthening the nascent peptide (Ban 2000). The 2′OH is also the center location of a wide variety of backbone modifications, particularly methylation (Kiss 2001), and can participate in hydrogen bond interactions (Bolton and Kearns 1978) or even ligand binding (Greenbaum 2001).

With so many important processes dependent on the 2′OH, and others dependent on the negatively charged phosphate, the RNA backbone is not to be ignored. The backbone of any given nucleotide has 12 torsion angles (including the χ angle around the glycosidic bond) and 19 atoms, compared to 4 torsions (including χ_1) and 7 atoms in an average amino acid backbone (Figure 1). While the huge number of torsions gives the RNA backbone the flexibility needed to effectively catalyze reactions, they wreak havoc on current structure building programs. As a result (Figure 3), errors along the backbone are frequent at the resolutions commonly attainable (2Å-3Å). The reason for this difficulty in determining backbone positions is evident from a look at the electron density (Figure 4): even at good resolutions (2.4Å), there is simply not enough

detail to accurately place C5′ or the sugar—even the phosphate oxygen positions are unclear. The problem is exacerbated because the crystallographically invisible but sterically important hydrogen atoms (Word, 1999a) are not modeled, while the phosphate's lack of hydrogens reduces the amount of information provided by ^{1}H NMR about the portion of the backbone between the ribose rings.

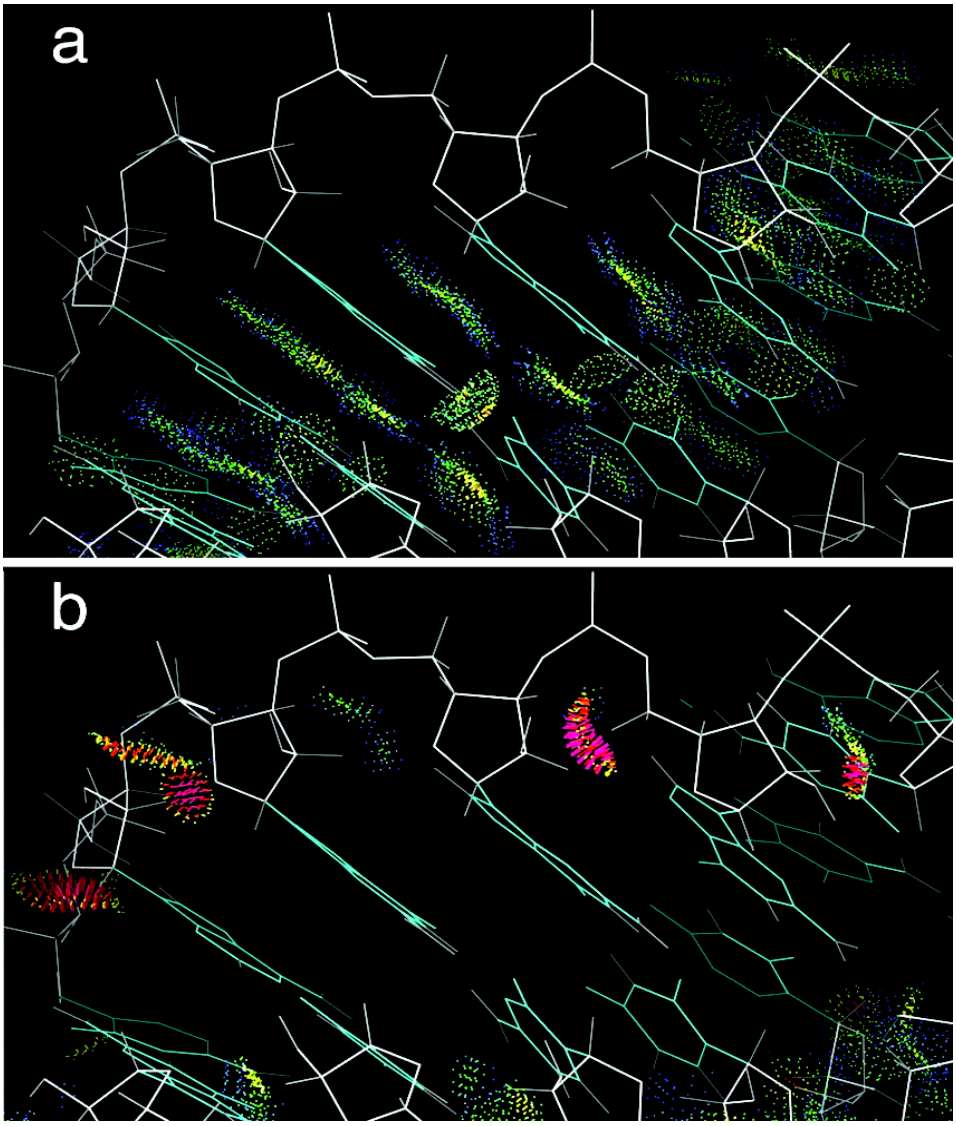


Figure 3: Base stacking vs. Backbone Clashes.

Well-fit base contacts in green (a) compared with clashing backbone contacts in red (b) for rr0033/1JJ2 23S rRNA (Klein 2001) at 2.4Å resolution

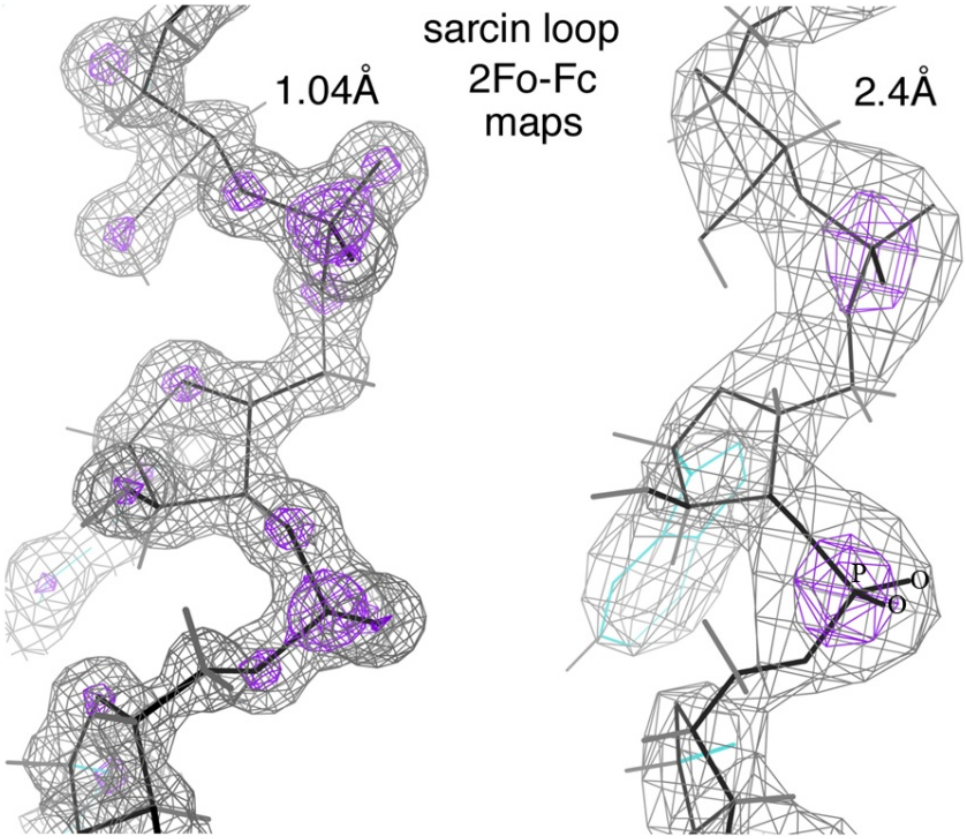


Figure 4: Density maps at different resolutions.

The sarcin/ricin domain from *E. coli* 23S rRNA is shown at 1.04Å resolution (left) from ur0035/1Q9A (Correll 2003a), and from rr0033/1JJ2 (Klein 2001) at 2.4Å resolution (right). The 1.04Å density shows distinct atoms; at 2.4Å only a vague idea of the backbone shape remains, with no clues as to dihedral orientation within the density.

The problem is exacerbated further when one considers the nature of the ribose ring; the sugar is not flat, but its geometry causes it to have a distinctive pucker at one of its atoms, while the other four are close to coplanar. The most common puckers are C2'-endo and C3'-endo (Figure 5); if one considers the ring numbered clockwise as "face

up", then either the C2′ or the C3′ will be out of the plane defined by the other four atoms, called "endo" if on the face-up side (thus same-side, or endo, to the base; Egli and Saenger 1983). Trying to correctly model sugar pucker in a crystal structure can be very difficult, as the density often gives just enough information to resolve the general position of the sugar ring, but not enough detail to determine a specific pucker.

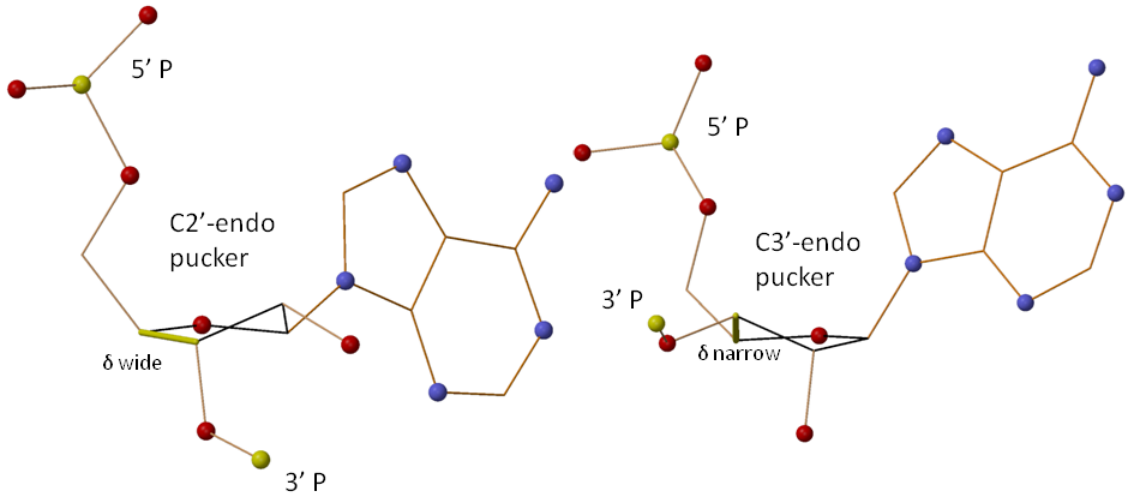


Figure 5: Ribose Pucker Conformations.

C2′-endo (left) and C3′-endo (right) sugar pucker; note the difference in δ dihedral that correlates to each, and the different positioning of the 3′ P relative to the glycosidic bond.

Overall, trying to fit the RNA backbone into the electron density at 2-3Å is rather like trying to make a model from tinker toys using a cardboard tube as a guide—it is difficult, time-consuming, and there is little guarantee of getting it perfect. Fortunately, we have more than just the electron density to help us. In chapter 2, I will present

research on discrete RNA backbone conformations and their implications for modeling. Chapter 3 will concern the common structural motifs we can identify that help us correctly fit backbone in the electron density. Validation and correction methods for RNA structures are presented in chapter 4, detailing how our methods for diagnosing problem areas of RNA backbone structure lead to better X-ray and NMR structures and have been incorporated into the PHENIX refinement package. The final results of these validation and correction methods will be presented in chapter 5, as we take on corrections of several RNA structures, including one of the largest RNA structures to date—two 70S ribosomes in one asymmetric unit.

1.3 All-atom contact analysis

Throughout this work, I have made significant use of the all-atom contact analysis developed in the Richardson lab, both for quality control and as an integral part of the research. The two principal components of all-atom contact analysis are the addition of hydrogens to the structure using REDUCE (Word 1999a), followed by analysis of the van der Waals (vdW) spheres of each atom with PROBE (Word 1999b).

Many crystallographic structures ignore hydrogens, an experimentally reasonable practice, as they are nearly impossible to detect. Yet implicit hydrogen models are often quite limited in their ability to elucidate H-bonds and vdW contacts, and are therefore inadequate as a method of quality control (Word 2000). REDUCE is a

program designed to automatically add hydrogens to a given structure, optimized by attention to polarity, sterics, and hydrogen-bond networks (cliques). The hydrogen positions are added directly to the PDB file, and thus are able to be used in most PDB analysis programs. REDUCE also includes the capability to perform automated corrections on sidechains that are easily fit incorrectly when explicit hydrogens are not being used to inform the initial build. For example, the N and O positions for the Asparagine and Glutamine amide can appear interchangeable in an implicit hydrogen model; adding hydrogens shows clearly that the NH2 is much larger physically than the O, since the N-H bond extends 0.6Å further than the bare oxygen, and incorrect positioning can have dire steric consequences for the structure. REDUCE alleviates the problem by performing "flips" for Asn, Gln, and His residues based on the steric positions and hydrogen bonding of the explicit hydrogens when they are added.

PROBE piggybacks on REDUCE's explicit hydrogen model, identifying any contacts between atoms in the model which can be tailored to user specifications. The program functions by rolling a probe of 0.25Å radius along the vdW surface of a given atom. When the probe touches the vdW surface of another atom, these atoms are considered to be in contact, and a dot is drawn on the vdW surfaces of both atoms; by default, PROBE will output 16 dots per square angstrom of surface contacts. PROBE uses a sliding color scheme to represent the distance between vdW surfaces. Blue and green dots are vdW contacts: they represent vdW surfaces that are between 0.25 - 0.5Å

(wide contacts) and 0.0 - 0.25Å (close contacts) apart, respectively. If the vdW surfaces overlap, PROBE checks to see if the overlap is between hydrogen bond donor-acceptor pairs or not. An overlap between valid hydrogen bond donor and acceptor pairs indicates an H-bond, which is represented by, sea-green dots that look rather like pillows (Figure 6). Overlaps between other atoms are drawn in increasingly warm colors from yellow to red as the overlap draws closer to 0.4Å, representing close contacts of increasingly unfavorable model tolerance. Overlaps between most atoms that are ≥ 0.4Å are drawn in hot pink; these hot pink spikes represent clashes—regions of the model that are sterically impossible in an actual biological macromolecule—thus representing errors in the model. Hydrogen bonds and salt bridges have a more forgiving clash limit since they require some amount of overlap already; clashes won't occur until an overlap of 0.6Å for standard H-bonds and 0.8Å for salt bridges. It is important to note that though PROBE does not calculate energies, the extremely large overlap of clashes makes them energetically impossible as well as sterically impossible.

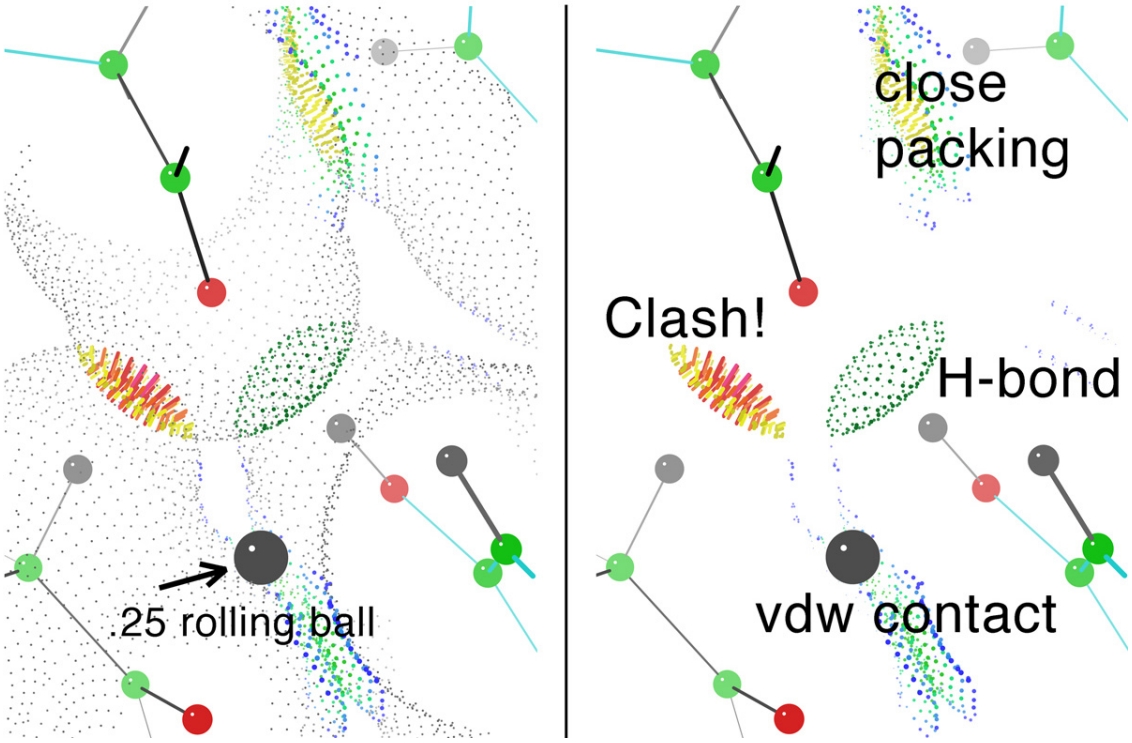


Figure 6: All-atom contact analysis dots.

On the left, the vdW spheres and the default 0.25Å-radius rolling probe are shown. The right shows the results drawn by the probe. Clashes are shown in orange and hot pink, H-bonds in sea green, and vdW contacts in green and blue. Slight overlaps, such as the yellow in the top right corner, can indicate close packing rather than outright steric hindrance.

1.4 MolProbity analysis and Kinemages

Throughout this work, I visualized the protein and nucleic acid models using kinemages, "kinetic images" designed for interactive scientific illustration (Richardson 1992). PDB files are converted to the kinemage format via programs like MolProbity and Prekin, and may be viewed with MAGE and KiNG (Richardson 2001; Chen 2009). These

viewers can display not only Cartesian coordinates, but markup produced by other programs, such as MolProbity, which can graphically indicate model quality and other features. Kinemages can contain multiple models at once, and the viewers have sophisticated tools for creating superpositions. More importantly, both MAGE and KiNG have tools for modeling protein sidechain rotations and mutations, for model improvement and new model design. KiNG also has methods for improving the protein and RNA backbone, allowing corrections to the original model that are difficult in most other programs.

The quality of protein and nucleic acid models was evaluated with MolProbity, a suite of programs oriented towards model validation and improvement (Davis 2004; Davis 2007; Chen 2010). In a typical MolProbity run, the input file is first stripped of hydrogens (a rarity for crystal structures); they are then added via REDUCE and the model is scanned for flippable amino acids as described in section 1.3. The whole model is then subjected to a series of validation programs; the results are presented in a sortable multi-criterion chart, and can be overlaid upon the model to create a multi-criterion kinemage.

Both protein and nucleic acid models have a Clashscore, derived from PROBE, which represents the number of clashes per 1000 atoms and for X-ray structures will display a rolling percentage based on how a given model compares to other models in its resolution range. Clashes are represented by pink spikes, as described above; H-

bonds and vdW contacts are also drawn (Figure 7). Backbone bond lengths and bond angles are also evaluated, and any that deviate by >4σ from the accepted parameters are flagged as bond or angle geometry outliers. For bond lengths, outliers are represented by springs; a red, elongated spring indicates the modeled bond length is too long, and a blue, compressed spring indicates it is too short. For angles, a fan-like series of angles approaching the correct ideal value are drawn; red for a modeled angle that is too large, and blue for too acute.

Proteins are analyzed for sidechain rotamericity (how well the sidechains fit into known rotamer positions), Ramachandran fit, and Cβ deviations. Poor sidechain rotamers indicate areas where protein sidechains have unusual χ dihedrals when compared to χ ranges determined in the Top8000, a database of the highest resolution and best overall protein chains in the PDB. These ranges have been defined for all twenty standard amino acids, and >99% of the protein sidechains throughout the PDB fall within these ranges, so any deviation is suspect; rotamer outliers are shown as golden sidechains. Even if the model is not wrong, it indicates an unusual conformation that is worthy of further investigation. Ramachandran outliers are regions where the protein backbone deviates from the standard positions on a Ramachandran plot, which represents accepted φ,ψ values of proteins; the offending dihedrals are marked in green (Lovell 2003). In addition to flagging outliers, MolProbity's Ramachandran analysis also evaluates how many amino acids are in favored vs. acceptable regions—a good model

should have >98% favored. Cβ deviations occur when the position of the Cβ is incorrect relative to the Cα on the protein backbone; these are marked by a purple sphere that grow larger relative to the deviation.

RNA specific evaluations are based on the sugar pucker and the backbone conformation, as described in Chapter 2. A sugar pucker outlier is marked by a perpendicular dropped from the 3′ P to the plane of the glycosidic bond vector, with a cross elongated parallel to the glycosidic bond in the direction of the base. The length of the perpendicular determines the sugar pucker, as described in Section 4.2.2, and is magenta when the modeled sugar should have C2′-endo pucker and dark purple when it should be C3′-endo. The backbone conformation analysis indicates the number of outliers; much like the sidechain rotamers in protein, this may indicate errors or residues of interest. There is no current markup for these conformations in MolProbity, but there are tools in KiNG that highlight RNA backbone outliers in green, much like the Ramachandran outliers, above. More details on the backbone conformation outliers will be found in Chapter 2.

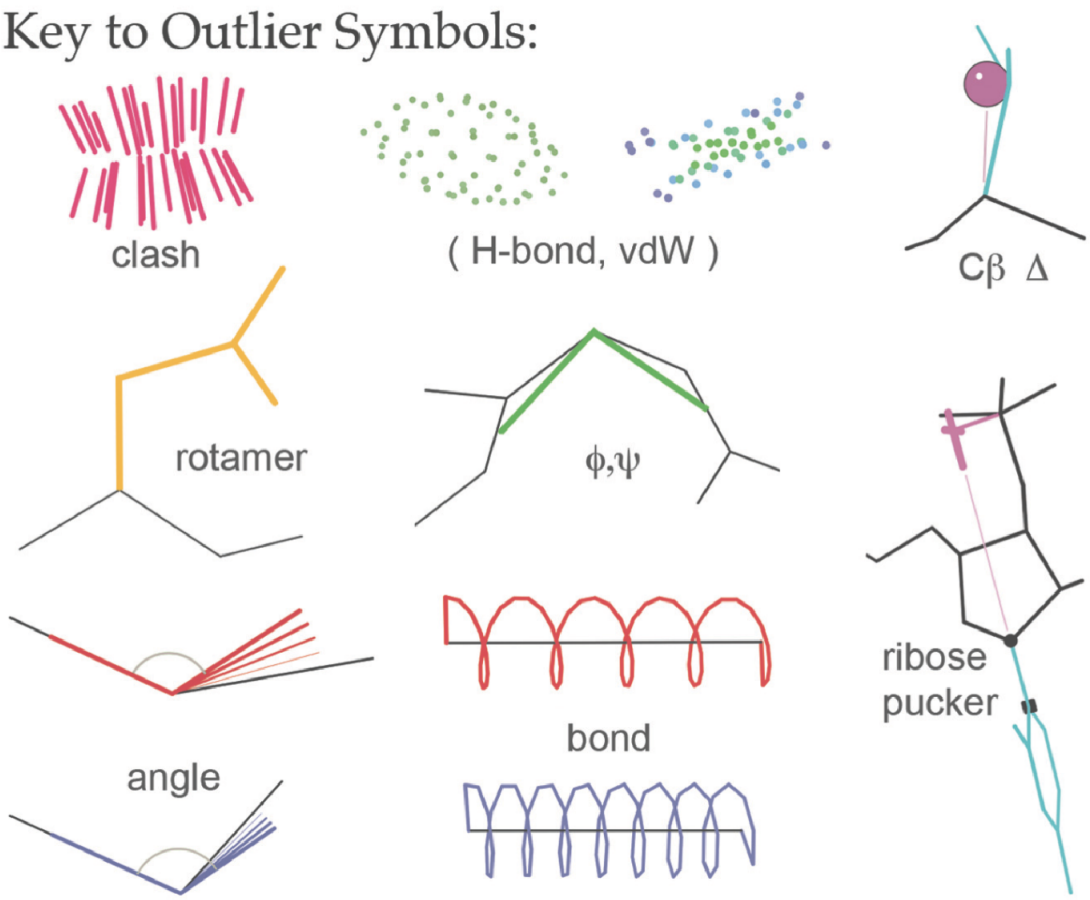


Figure 7: MolProbity Markup.

A list of the most common kinemage markups from MolProbity, as they would be displayed in KiNG.

2. The Backbone of RNA Structure Analysis

The RNA backbone is a very complex system to work with. While the bases are planar and have only one rotatable bond (χ), there are six (α-ζ) rotatable bonds for the RNA backbone, along with five (ν_0-ν_4) dihedrals within the sugar (Figure 8). There are physical boundaries that limit the amount of rotation around these dihedrals, but early research into the extent of these limits, and especially of their allowable combinations, was difficult due to the lack of structures. Small ribonucleotide structures found in the Cambridge structural database (CSD) (Allen 1979) were determined at very high resolution, but these were typically duplexes or single strands of RNA, and did not exhibit any of the complex structural behavior associated with most larger RNAs. Of just 253 structures as of the beginning of 2000, the largest were the 158-residue Group I intron P4-P6 domain (PDB: 1GID, Cate 1996) and several very similar 74- to 77-residue tRNA structures. Most X-ray and NMR structures of this era were duplex RNA on the order of 20-30 residues. It was not until crystal structures of the ribosomal subunits were solved (Ban 2000, Schluenzen 2000, Wimberly 2000) that enough data was present to give serious consideration to analysis of the RNA backbone.

This chapter outlines the work done to establish what conformations are adopted by the RNA backbone. This was a largely collaborative effort, with Jane Richardson, Laura Murray, and myself contributing the primary work in the Richardson Lab. Two conventions will be used throughout this chapter, and the rest of this dissertation. The

first concerns the magnitude of dihedral angles in RNA backbone; it can be difficult to automatically resolve clusters with torsions close to *trans*, since using the typical range of 180° to -180° can make torsions appear very far apart from each other, even if they are separated by only 20°, e.g., -174° and 166°. This problem is avoided by adding 360° to any torsion angle less the zero. Thus, 180° and -180°, would be the same value, and gauche minus would be 300° instead of -60°. The second convention is for the reader's benefit; all references to suites and suitestrings will be in **bold**.

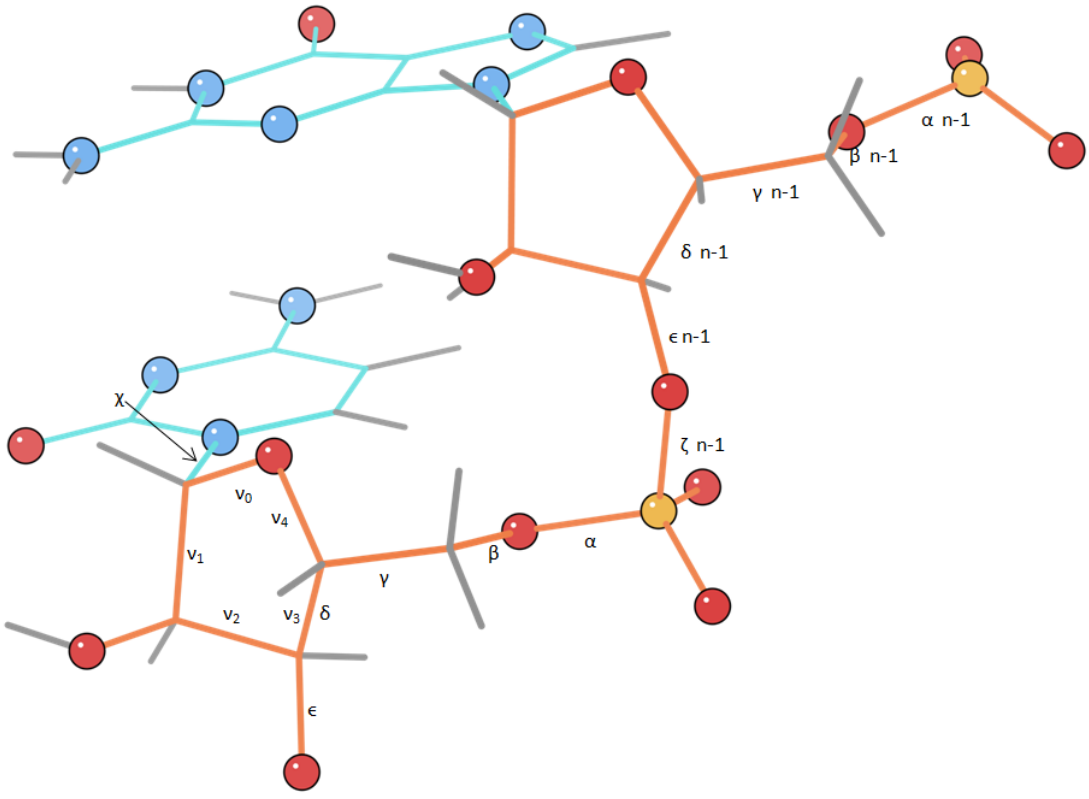


Figure 8: RNA backbone dihedrals.

2.1 Early backbone work

Even with the limited number of structures, early attempts were made to define the RNA backbone. Sasisekharan and Lakshminarayanan used hard-sphere models of dimethyl phosphate to search for allowed ζ-α pairs (Sasisekharan 1969), finding that having both gauche-plus or both gauche-minus was most favorable. They expanded their analysis to find that β should be near 180°, γ at 60°, 180°, or 300°, ϵ between 210-270°, and χ at 30° or 220° (Lakshminarayanan 1970). Later, George Rose's laboratory used hard-sphere models to determine allowed and forbidden regions for individual torsions and two-torsion combinations for α thru ζ (Murthy 1999); they also included both χ and also the phase and pucker amplitude of the ribose ring as formulated by Altona and Sundaralingam (Altona 1972). They were able to show that sterics forbid many dihedral combinations, but that many of the sterically allowed regions were not populated if one looked at the current available RNA structures (Murthy 1999). With such a small number of deposited structures, however, it was impossible to say whether the unpopulated regions represented undiscovered RNA, or were disallowed due to other reasons.

In contrast to hard-sphere models, Olson and Flory (Olson and Flory 1972) sought to elucidate RNA backbone by simplifying it into two pseudotorsions, creating virtual bonds from P_n to $C5'_n$ and $C5'_n$ to P_{n+1} and using the torsion around it to define RNA backbone restraints. Hoping to echo the success the Ramachandran plot showed in

clarifying protein backbone (Ramachandran 1963), they discovered that RNA does have particular clusters of "rotational states" defined by the pseudotorsions, particularly when dividing the RNA into C3'-endo and C2'-endo ribose pucker. The idea of using pseudotorsions was later taken up by the Pyle lab (using C4' instead of C5'), which found several more distinct clusters of η (P_n-$C4'_n$) and θ ($C4'_n$-P_{n+1}) pairs corresponding to distinct RNA structures (Duarte and Pyle 1998). This method was proven to be quite useful after the ribosome structures were solved, as it could correctly identify several structure motifs, such as the S-motif, in the RNA backbone (Duarte 2003).

2.2 The Big Three

With the publication of the ribosome structures, the total number of residues available for study in the PDB and NDB effectively doubled. More importantly, these structures were a far cry from the A-form duplex RNA common to the PDB and NDB at the time. While the early tRNA and 5S rRNA structures had some peculiar structural features that were non-A-form, the full ribosome contained numerous instances of triple helices, non-canonical base-pairing, single-nucleotide bulges, and other features that provided a wealth of information on what RNA backbone could do given its inherent flexibility. When a new structure of the 50S ribosomal subunit (Klein 2001) was published with 2.4Å data—very good for a structure its size—three major studies were quickly undertaken to establish the limits of RNA backbone structure conformation.

Using this 50S structure (PDB: 1JJ2) as their dataset, Loren Williams' laboratory developed a method for classifying RNA backbone by binning dihedrals α, γ, δ, and ζ for each residue. The γ and α dihedrals were found to have three peaks each, δ had two, with only one for ζ. A bin was created for each peak, with an extra bin for each dihedral to represent "miscellaneous" torsions—the total number of bins was $4\alpha*4\gamma*3\delta*2\zeta$, or 96 (18 if no miscellaneous torsions are counted). Their analysis showed 37 bins (13 non-miscellaneous) were populated, and that these bins could be used to identify helices and tetraloops (Hershkovitz 2003). It is important to note that the "missing" dihedrals β, ϵ, χ and the pucker phase described previously through hard-sphere analyses (Murthy 1999) went unused, not in an attempt at simplification, as with pseudotorsions, but because the dihedrals were assumed to be unimodal, and the pucker phase tracked well with δ; these variables were thus deemed to have minimal impact on the binning process and were not included in the analysis.

In contrast, a collaboration between Helen Berman, Bohdan Schneider, and Zdenek Moravek (Schneider 2004) attempted to capture as much backbone dihedral information as possible. Still using the structure 1JJ2 as their entire dataset, they analyzed dinucleotides, taking into account all dihedrals α through ζ as well as χ. This 14-variable set was split into six 3-D plots: $\zeta\alpha\delta$, $\zeta\alpha\gamma$, $\alpha\gamma\delta$, $\zeta\alpha\chi$, $\zeta\alpha\epsilon$, and $\zeta\delta\chi$; any dinucleotides that were roughly A-form (both ζ and α near 300°) were taken out to be analyzed separately, so as not to overwhelm the less populated regions of the plots. The

distributions were converted into a pseudo electron density, and the automated peak-picking functions of Xtalview (McRee 1999) were used to identify potential conformational clusters. They found 32 total clusters--14 peaks in the A-like data, and 18 in the non-A-like.

The third major approach to classify the RNA backbone using the new data implemented a radical departure from the others; rather than use the chemical nucleotide, the Richardson laboratory used a measure describing the dihedrals between adjacent sugars, dubbed the suite (Figure 9) (Murray 2003). The decision to use the suite as a unit of backbone division was influenced by observations on where most model errors lay according to all-atom contact analysis (see Section 1.3); most impossible steric clashes occurred between consecutive sugars, but comparatively few were within a residue, implying that the greatest conformational flexibility in the RNA backbone occurred between the sugars. Initial modeling of a seven-dihedral dataset (δ_{n-1},ϵ_{n-1},ζ_{n-1},α,β,γ,δ), showed several distinct clusters representing particular RNA backbone conformations. Plots were made of each individual dihedral, 2-D plots were made for most dihedral pairs, and two 3-D heminucleotide (δ,ϵ,ζ and α,β,γ) plots were generated to study the relationships between the torsions. Dependence on sugar pucker was very pronounced; δ values of 120°-175° corresponded to C2′-endo puckers, while values of 55°-110° corresponded to C3′-endo puckers. The α,β,γ plot was therefore split into two separate plots, one for each δ range. From these 3D plots, 18 $\alpha\beta\gamma$ peaks were found, 3 of

which only appear in sugars with C2′-endo pucker, and 3 only with C3′-endo pucker. Seven δεζ peaks were also identified, 3 for C2′-endo and 4 for C3′-endo puckers. To find how these dihedral sets related to each other, a new cross-picking function between kinemages was added to Mage (Murray 2005). To assure completeness, new αβγ plots were generated based on δεζ peaks for a total of 14 new plots (7 for each pucker). These plots showed a total of 42 peaks containing at least five suites in filtered data or constituting 10% of their particular αβγ distribution.

Unlike other studies, which used the ribosome as the entire dataset, the Richardson lab study included a broad range of structures. The total number of RNA structures from the NDB was filtered by resolution, keeping all structures with 3.0Å or better resolution. To minimize the effect of redundancy, common structures like the RNA duplexes were limited to one example in the final dataset, dubbed RNA03; in all, RNA03 contained 132 non-redundant RNA structures. A final filtering of the data to remove suites with internal clashes and B-factors greater than 60 led to a total of 4166 residues (3976 suites), increasing the amount of data drawn from the 1JJ2 structure (2876 residues) by nearly a third. Over 90% of these suites could be assigned to one of the forty-two 7-dihedral conformers.

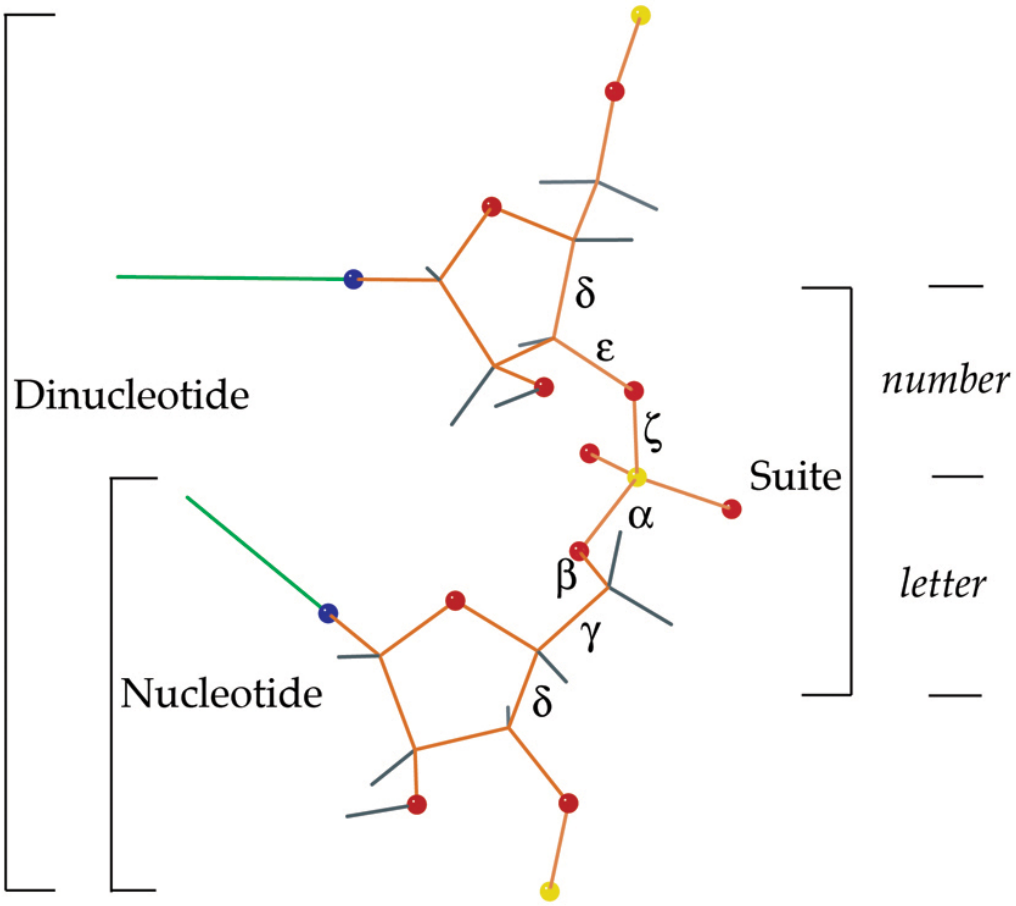


Figure 9: Three methods of defining RNA backbone.

The backbone is labeled according to the suite nomenclature. The suite is further divided into the heminucleotides corresponding to a number (δεζ) and a letter (αβγδ).

2.3 Consensus

With the publication of these three classifications of RNA backbone, those involved realized quickly that we would all benefit from closer collaboration and exchange of information. The first thing we noticed was that the results agreed overall, despite the vastly different methods used to achieve the several classification schemes. This vindicated our view that RNA could indeed be classified into discrete backbone conformations, akin to the way protein sidechains may be classified into discrete rotamers. The next thing we noticed was that each group had some conformations that were not present in the other groups' data, and that many of the differences were in difficult to resolve regions of torsion space where small clusters could be mistaken for noise (and vice-versa). It became clear that the three different approaches all had different merits and drawbacks, and overall the results of these methods were complementary to each other. Rather than choosing one method or attempting to independently resolve the differences in classification, we resolved to collaborate with the goal of forming a consensus set of RNA backbone rotamers that would combine the advantages of each group's methods and classification scheme.

Thus began the search for consensus, involving members of David and Jane Richardson's lab at Duke, Helen Berman's lab at Rutgers, Bohdan Schneider's lab in Prague, Anna Marie Pyle's lab at Yale University, and Loren Williams and Eli Hershkovitz from the Georgia Institute of Technology. The extensive collaboration,

involving five institutions across two continents, came together to seek a consensus that could describe RNA backbone structure. Much of this work was facilitated by the RNA Ontology Consortium (ROC) (Leontis 2006), an organization formed with the stated goal of creating an ontology that would fully describe RNA structure; RNA backbone classification was an integral part of this goal. The final results of this work were published in 2008 in the RNA Journal (Richardson 2008).

2.3.1 Initial Sharing and Updates

Initially, all the groups continued their individual research paths, making sure to inform the others of important updates. We provided the Georgia Tech group with a copy of our filtered version of the RNA03 dataset as they pursued a more multidimensional analysis on interdependence of dihedral angles. Influenced by our work and Bohdan's study, which found α and ζ to be interrelated, Eli Hershkovitz began adding cross-residue torsion combinations to the analysis, proving in the process that the rotamers were better classified using suites. The filtered RNA03 dataset was also run through the methods used by Helen Berman and Bohdan Schneider labs, which resulted in a slight expansion of their results.

As for our lab, Laura Murray updated the RNA03 dataset with more current structures, culminating in the release of RNA05. This dataset contained 171 files, including 1S72 (Klein 2004) and 1XMQ (Murphy 2004), newly revised structures of the 50S and 30S ribosomal subunits, respectively. RNA05 maintained the non-redundant

aspect of RNA03 but also added stricter filters, removing residues from consideration that had backbone-base and base-base clashes. The final, filtered version of RNA05 thus had only 4053 residues, but represented a wider range of structures.

Another update that came from the Richardson lab was the development of new tools for visualization of these multidimensional datasets. David Richardson encoded support for high-dimensional views into Mage, allowing selection of points in one set of axes, manipulation (such as recoloring) of these points, and then selection of new axes while the original points maintain their new color. The result was the ability to create N-dimensional kinemages, an easy way to track clusters of backbone rotamer through different combinations of backbone dihedrals. To add even more versatility (and easier identification), the ability to view parallel coordinates (Inselberg and Dimsdale 1990) was also incorporated into Mage. The parallel coordinate view shows all dimensions as axes, with color-coded lines tracing the values for these axes for all N-dimensional points. This coloring is retained from the cluster view, so the user can choose 3 axes, pick out and recolor a cluster of points, then activate parallel coordinate mode to view how well this 3-D cluster tracks across all seven dimensions representing the seven RNA backbone dihedrals of the suite (Murray 2005). Viewing and manipulation of the high-dimensional kinemages was also added to KiNG by Vincent Chen (Chen 2009), which enabled easy communication with our collaborators and the addition of web kinemages in the supplement to Richardson *et al.* 2008.

2.3.2 A Modular Heminucleotide Nomenclature

Each of the three analyses focused on different divisions of the RNA backbone (Figure 9). Trying to harmonize methods that used residues, suites, and dinucleotides proved to be quite difficult initially. While we could identify corresponding regions through discussion, even this proved difficult when each group used a different naming scheme to describe their conformations. Loren Williams' lab use single-character codes to identify the conformations they found, the Rutgers-Prague group used a six-character name based on the letters of the plots their clusters were derived from, and the Richardson lab used a 7-character system identifying dihedrals by sugar pucker (for δ only) and torsion value: e for eclipsed, t for trans, p for gauche plus, m for gauche minus. What is more, the Richardson lab indexed suites by the number of the second residue, since it contained four of the seven torsions and the phosphate; conversely the Rutgers-Prague group indexed by the first residue of a dinucleotide. These differences in nomenclature caused confusion at best, and outright mistakes at worst; it was apparent that a consensus dataset would only arise from a common nomenclature, but each group was adamant that their nomenclature was integral to their work.

To combat this problem, we devised a modular nomenclature based on heminucleotides, building on an idea from the Williams lab that their one-letter designation be expanded to two, and on Helen Berman's voice of experience that we were sure to soon regret staying with single-character names. This nomenclature

consists of a 2-character system, where the first character is a numeral, and the second character is a lowercase letter. Each character corresponds to a heminucleotide; as shown in Figure 10, numerals are associated with δϵζ heminucleotides, while letters refer to αβγδ heminucleotides. The nomenclature is further divided into C3′-endo and C2′-endo sugar puckers: odd numbers and letters **a,c-n** refer to C3′endo pucker, while even numbers and letters **b,o-z** refer to C2′ endo pucker, preserving suite 2b for B-form DNA (see below). Several nonstandard characters augment the 2-character heminucleotide system. The characters **&** and # refer to C3′-endo and C2′-endo puckers, respectively, for the δϵζ heminucleotide—they are used simply because we ran out of Arabic numerals. The bracket ("**[**") is used to represent a particular αβγδ heminucleotide that lends itself well to intercalation (Figure 11); the "**[**" character provides a visual approximation of the base positions in a such a conformation.

A clever aspect of this modular nomenclature is that information about the structure is built into the character identifier. Besides classifying which δϵζ heminucleotide a structure belongs to, the starting numerals 1,3, and 5 as well as 2,4,and 6 represent the minus, trans, and plus ζ values, respectively. 7 and 9 or 8 and 0 represent eclipsed values, within +/- 25° of 120°(eclipsed) or 240° (negative eclipsed). A combination of low ϵ and ζ in the 240° range is represented by the "#". The seemingly out of place "**b**" from the second heminucleotide represents B-form DNA, whose conformation is sterically disallowed in RNA by the 2′ hydroxyl; aside from being an

homage to DNA, the decision for "**b**" to represent B-form structure gives a starting point for future expansion into categorizing suites for DNA backbone structures. Overall, out of all the possible number-letter combinations, only about 20% are used to represent clusters identified in the filtered RNA05 dataset, which leaves plenty of room to add conformations discovered in the future.

There are two other special characters used to identify specific situations rather than clusters. Incomplete torsions are identified by the double underscore, "__", a designation which appears mostly at chain ends or in gaps within the chain. Unusual conformations that are not in any of the clusters are denoted with exclamation points, "**!!**", a designation designed to draw the users' attention. Many times these outlier conformations will represent an error in the structure, but occasionally the **!!** will represent a genuine rare conformation as yet uncanonized or perhaps even an important strained conformation in a transition state or active site.

Using this modular system, it becomes possible to assign a 2-character identifier to a cluster of RNA conformations, regardless of which of the three earlier classification schemes was used to find it. A standard A-form suite would be identified by the number-letter combination **1a**. An A-form nucleotide would also consist of a "**1**" and an "**a**" heminucleotide, but the name would be reversed to **a1**, to show that the first heminucleotide in a residue is the $\alpha\beta\gamma\delta$ and the second is the $\delta\epsilon\zeta$ heminucleotide. When assigning residue numbers in the heminucleotide nomenclature, a chemical nucleotide is

referred to by its normal residue number; a suite, however, is referred to by the number of the second residue, which reflects the preferred reference to the backbone dihedrals as $\delta_{n-1}\epsilon_{n-1}\zeta_{n-1}\alpha\beta\gamma\delta$.

It should be noted that since the $\alpha\beta\gamma\delta$ and $\delta\epsilon\zeta$ heminucleotides share the same δ dihedral measure, there are certain logical restrictions on the successive suites. Using the **1a** example, the "**a**" indicates a C3'-endo pucker, so logically, the following suite should start off with an odd number or "**&**". Overall, **a,c-n** should be followed by odd numbered heminucleotides and **b,o-z,[** should be followed by even numbered nucleotides.

For δεζ heminucleotides:

C3′-endo puckers: Odd numbers: Code angles δεζ	C2′-endo puckers: Even numbers: Code angles δεζ
1 = 3′-em	**2** = 2′-em
3 = 3′-et	**4** = 2′-et
5 = 3′-ep	**6** = 2′-ep
7 = 3′-e-e	**8** = 2′-e-e
9 = 3′-ee	**0** = 2′-ee
& = 3′t-e	**#** = 2′te

For αβγδ heminucleotides:

Code angles αβγδ	Mnemonic
For C3′-endo pucker:	
a = mtp3′	**1a** is A-form
c = ttt3′	**1c** is "crankshaft" variant of A-form
d = ptp3′	inverted "p"; see below
e = -ept3′	**1e** is stack-shift dent; only eclipsed α
f = tet3′	
g = ttp3′	**1g** is suite 1–2 of GNRA tetraloop
h = mtt3′	
i = p-et3′	
j = pet3′	
L = mep3′	
m = m-ep3′	Minor 1a shoulder
n = ptt3′	**6n** is 2′3′ Z-form; "N" is rotated form of "Z"
For C2′-endo pucker:	
b = mtp2′	**2b** would be B-form DNA
o = mtm2′	**1o** and **2o** both put bases opposite each other
p = ptp2′	Most p angles in 2′ set
q = pet2′	
r = ptm2′	Rare reverse order of common m t p
s = mpt2′	**4s** is commonest suite 2–3 of S-motif
t = ttt2′	All-*trans*
z = ttp2′	**5z** is 3′2′ Z-form
[= m-ep2′	**1[** is commonest intercalation conformation

For all heminucleotides: (_) Suites with any undefined dihedrals (chain ends or disordered loops). "**L**" is used here for clarity, but would be lower case in computations.
(**!**) Unusual conformations: suites or heminucleotides not in the list, bad ε, etc. So, **!** denotes something that is either wrong or interesting.
Note: In the δεζ list, the "code" is a number (meaning a symbol in the 002–003 range of Unicode) for the first characters of modular consensus conformer names; in the αβγδ lists describing the second characters of conformer names, the "code" is a letter (a symbol >005A in Unicode). For the mean dihedral angles, m signifies −60° (minus); t, 180° (*trans*); p, +60° (plus); e, 120° ± 25°; and -e, −120° ± 25°.

Figure 10: Components of the modular consensus nomenclature.

Each heminucleotide has a number or letter identifier. As much as possible, letters were assigned with mnemonics in mind, such as g for the GNRA tetraloop, and z for Z-form structure.

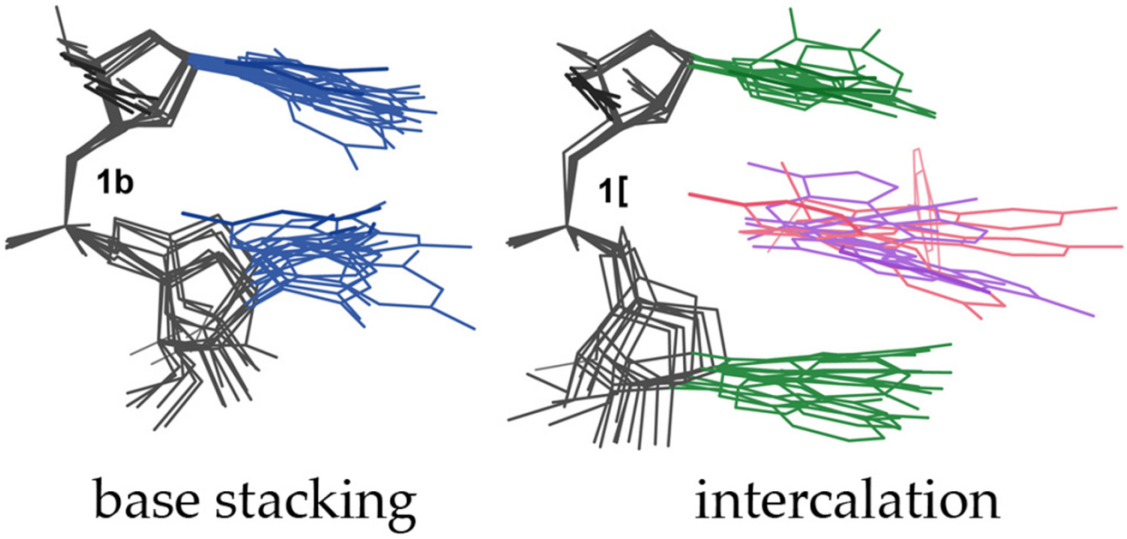


Figure 11: Two sample backbone suites.

1b and 1[look similar, and in fact differ only in β. 1b suite shows significant base stacking, while the 1[bases are far enough apart to allow intercalation.

2.3.3 RNA05 and Consensus

When we updated our dataset to RNA05, we had augmented Laura's original RNA03 dataset with many new RNA structures, as well as adding higher-resolution versions of some structures that had already existed in RNA03. Armed with the new modular consensus nomenclature, we revisited the RNA05 dataset to determine what the new rotamer distribution would be. Using the RNA05 dataset in conjunction with the 7-dimensional kinemages, Jane Richardson identified a new set of conformational rotamers. With the 7-D kinemages, she could show conclusively that the δ and γ dihedrals had the clearest divisions. The bimodal δ and trimodal γ led to the creation of a $\delta_{n-1}\delta\gamma$ plot, which defines 12 bins—two for each δ and three for γ (Figure 12). This plot provided an effective first pass filter through which to examine candidate suites; we eliminated any δ that did not fit into the rather generous ranges for C3'-endo (55°-110°) and C2'-endo (120°-175°), and any ϵ that was not between 155° and 310°, since fewer than 3.5% of the total unfiltered ϵ data is outside that range. The out of range torsion values are related to common mistakes, such as incorrect sugar pucker (see Section 4.2.2), though it is possible that in some cases they represent legitimate conformations that are strained (e.g., part of an active site) and do not have enough similarity with other points to be identified as a separate conformer.

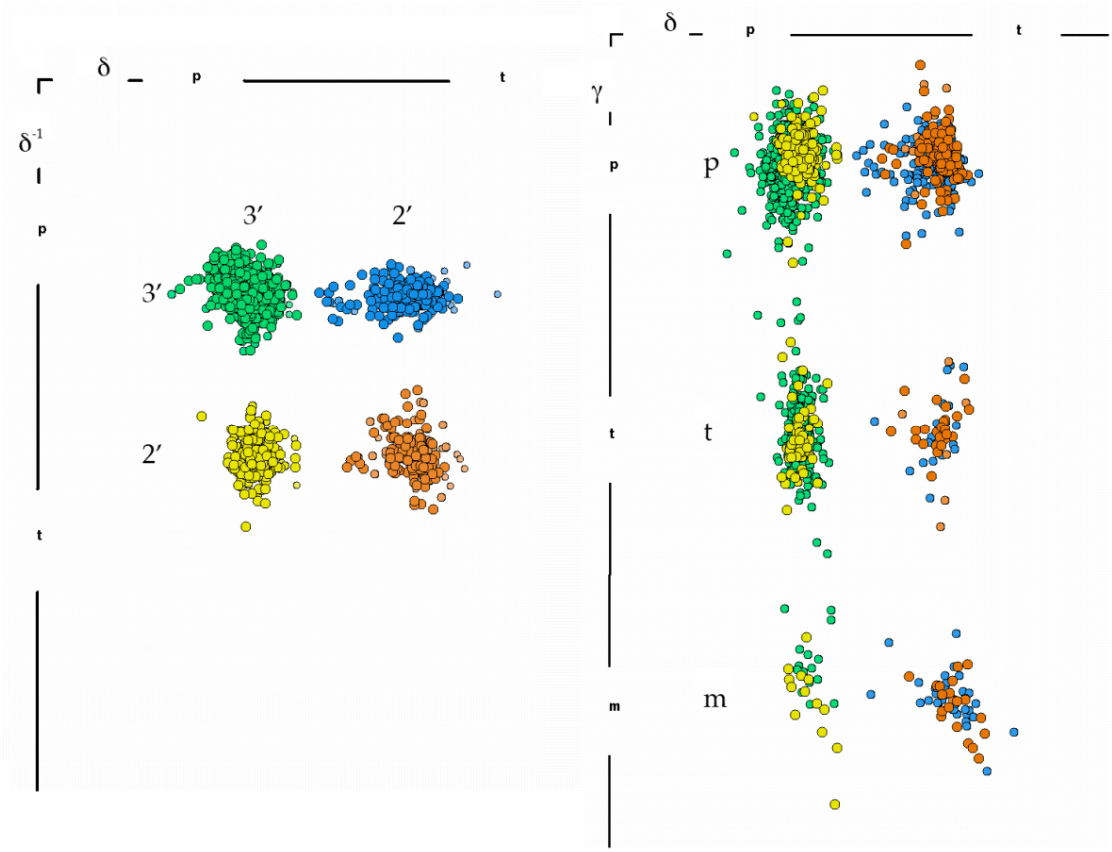


Figure 12: $\delta_{n-1}\delta\gamma$ division.

On the left is $\delta_{n-1}\delta$, which corresponds to binning by sugar pucker. On the right is $\delta\gamma$, showing the 6 distinct bins.

Each of the 12 bins were then examined for clustering among the α,β,ϵ, and ζ dihedrals, and the clusters were confirmed by Laura Murray using the 7-D kinemages. We then reexamined small clusters for conserved structural roles and consistency in Cartesian space; we identified 50 suite clusters, though nine were considered marginal.

With our new 7-D suite cluster analysis, we compared our results with those of the Rutgers-Prague group's Fourier transform (FT) analysis to build the consensus

conformer set. Three cases arose: both groups agreed fully on a cluster, both groups agreed on the cluster region, but disagreed on the number of clusters represented (e.g., two suite clusters, but only 1 FT peak), and in some cases, one group had a cluster that the other did not.

In the first case, we simply added the cluster to the consensus dataset, since both labs already agreed on it. Thirty-one clusters were agreed upon in this manner.

The second case was more difficult to resolve. To determine whether a cluster should remain independent or be merged with another, we examined the internal consistency of the cluster in Cartesian space, searched for unique structural roles fulfilled by the cluster, and determined if their $\zeta\alpha\beta$ clustering was similar to the $\zeta\alpha\beta$ clusters in other $\delta_{n-1}\delta\gamma$ bins. We also made use of the Pyle lab's η/θ torsions (Duarte and Pyle 1998) to help distinguish similar rotamers, though we altered their analysis to use θ/η instead, since this measure corresponded better to the definition of a suite.

In this way, we better defined the clusters we had found, and were able to distinguish between disputed regions that contained two genuine clusters, and regions where only one cluster was justified. The **2b**/**2[** and **1b**/**1[** clusters exemplify this process. Through our analysis, **2b** was identified rarely, but appeared at first to be a separate suite. Upon further analysis, the **2b** cluster was very similar to the **2[** cluster in all respects, so they were merged into the **2[** cluster. **1b** and **1[**, on the other hand, have a

distinct β shift that separates them (see Figure 11); while there are a few examples that are hard to distinguish at the edges, the cluster means are clearly distinct.

In the third and last case, some clusters were only identified in one analysis or the other. These clusters were analyzed using similar methods to case 2: for acceptance, the cluster must either have 5 examples or 3 with one at high resolution and must have either a defined structural role or be consistent when superimposed in Cartesian space. Two clusters now in the consensus dataset are **9a** and **1t**, found only in the FT analysis and the 7D cluster analysis, respectively.

Overall, 46 clusters were identified that now make up the consensus conformer dataset. There were also some clusters that seemed promising but were eliminated by having too few examples, or being found only in low-resolution structures and high B-factor. These examples have been included as "wannabes"; they are not listed in Table 1 or Figure 13, but in practice have usually been included in conformational analysis or validation and are the most likely clusters to become conformations as new and better RNA structures are developed. Each of the 46 primary consensus clusters can be seen in Tables 1 and 2, along with their heminucleotide nomenclature names and their original identifiers in the three earlier studies (columns 6-8). Column 3 lists the total number of points in each cluster and Column 5 lists the structural exemplar, identified by NDB code and residue number (and chain where necessary). The dihedral means and standard deviations for each cluster can be found in Table 2. Finally, Figure 13 shows 2-

dimensional $\zeta_{n-1}\alpha$ plots of each cluster, separated by $\delta_{n-1}\delta\gamma$ bin, with further plots of select overlapping clusters that are separable only in β or ϵ.

Table 1: Consensus conformers with comparison to original assignments.

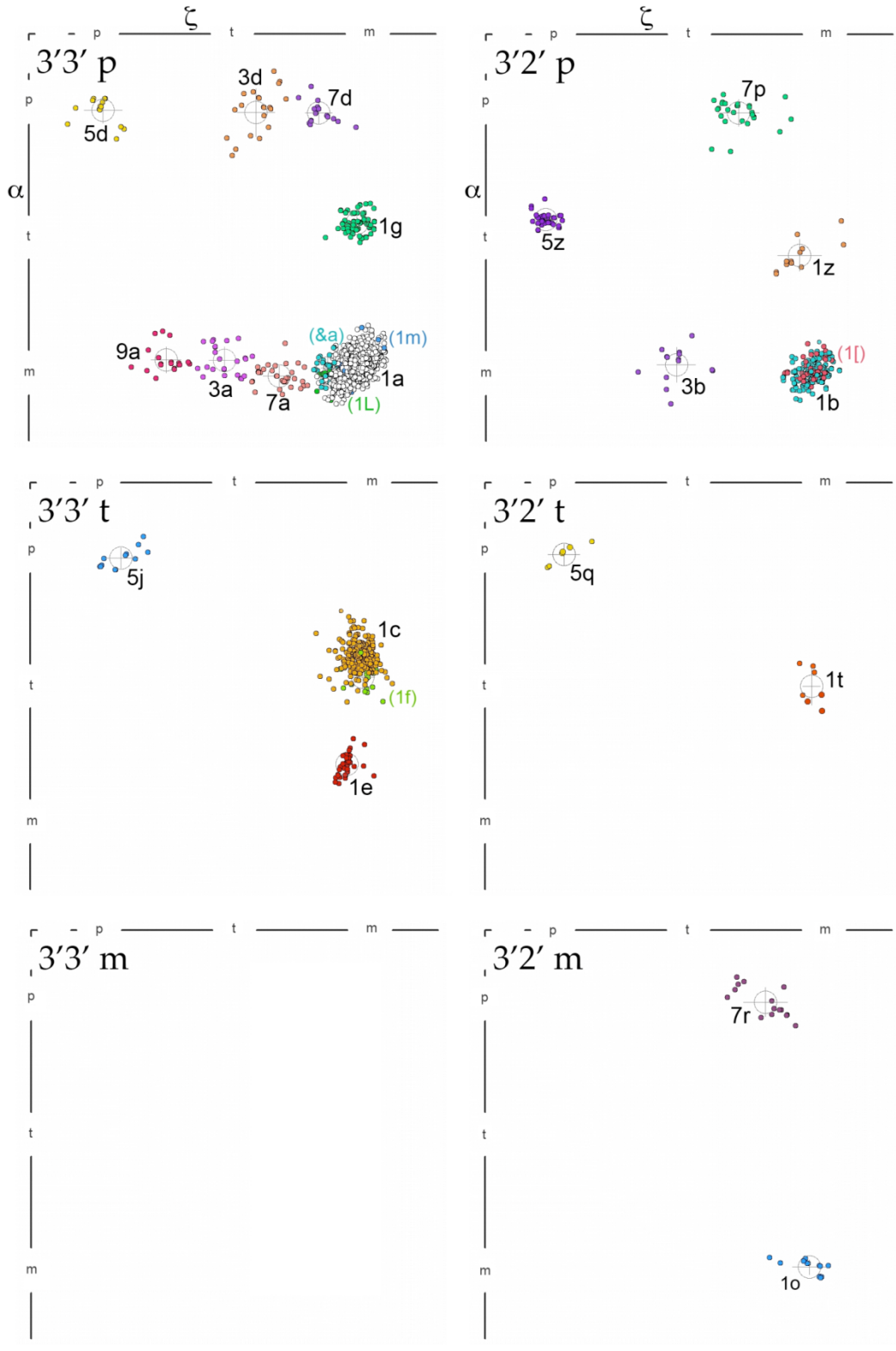
$\delta_{-1}\delta\gamma$	name	# pts	comment	example	=dinuc	=suite	=bin
33p	**1a**	4637	A-form	ur0020 11	BD-1	3'emmtp3'	a
	1m	15	-β shoulder on 1a; some intercalate	rr0082 1940		3'emm-135p3'	a
	1l	14	+β shoulder on 1a; overtwists base direction	rr0082 1460			a
	&a	33	εζ shoulder on 1a; weak Hb O2'$_{-1}$-O4'	pr0037 b163			a
	7a	36	stack switch	ar0041 a6	16,17	3'e-140mtp3'	e
	3a	25	bases far; 7a, 3a, 9a all touch in ζ	urb016 a2	BD-9	3'etmtp3'	e
	9a	19	bases far; starts or ends loops	rr0082 2582	BD-15		e
	1g	78	GNRA 1-2; U-turn	rr0082 1864	BD-18	3'emttp3'	o
	7d	16	bases far; can span 2 helices	rr0082 636	BD-26	3'emptp3'	T
	3d	20	bases far; starts or ends A-helix	rr0082 2118	BD-27	3'e-140ptp3'	t
	5d	14	P$_{-1}$ to P$_{+1}$ close; end or end+1 A-helix	ur0020 a9	BD-24	3'epptp3'	t
33t	**1e**	42	S-motif strand 2 "dent"; Hb 2'O$_{-1}$-O4'; low β	ur0035 2665	BD-7	3'em-11080t3'	u
	1c	275	GNRA 405; ttt "crankshaft" from 1a	ur0020 a28	BD-2	3'emttt3'	i
	1f	20	+β shoulder on 1c; stack switch/intercalate	tr0001 22	BD-6	3'emt135t3'	i
	5j	12	bases far; 1-bulge return	ar0027 b17	BD-25	3'epp110t3'	L
32p	**1b**	168	k-turn 0', syn G Hb N2-O2P	pr0113 d208	BD-4	3'emmtp2'	n
	1[	52	best intercalation conformation	pr0019 b658	BD-5	3'emm-135p2'	n
	3b	14	bases far; ends A-helix	rr0082 904	BD-12	3'etmtp2'	E
	1z	12	UNCG 1-2; bulges	rr0082 1771	BD-19	3'emttp2'	g
	5z	42	S-motif 1-2; Z32a DNA; Hb O2P$_{-1}$-2'O	ur0026 2654	BD-20	3'epttp2'	s
	7p	27	bases far	pr0033 b8	32,33	3'e-140ptp2'	m
32t	**1t**	7	ttt variation from 1b	pte003 b907		3'emttt2'	
	5q	6	bases far	pte003 b973	BD-22	3'epp110t2'	
32m	**1o**	13	starts 1-bulge; wide in β	rr0082 1108	BD-34	3'emmtm2'	
	7r	16	k-turn 1-2	rr0082 1314	BD-13	3'e-140ptm2'	d
23p	**2a**	126	1-bulge return	rr0082 1711	BD-37	2'emmtp3'	F
	4a	12	bases far	rr0082 2485	BD-8	2'etmtp3'	A
	0a	29	cross-stacked A-helix start; k-turn 4-5	rr0082 265	BD-14		A
	#a	16	base platform; S-motif 3-4; low ε	rr0082 1371		2'etmtp3'	A
	4g	18	base platform, non S-motif	ur0012 a226		2'etttp3'	b
	6g	16	sheared stack	pr0122 r151		2'epttp2'	b
	8d	24	some with Hb 2'O$_{-1}$-O2P$_{+1}$	rr0009c1062	BD-28	2'emptp3'	
	4d	9	tRNA 58-9; Hb 2'O$_{-1}$-O2P$_{+1}$	tr0001 59		2'etptp3'	f
	6d	18	starts A-helix	rr0082 116		2'epptp3'	f
23t	**2h**	17	bases far	rr0082 2540	BD-30	2'emmttt3'	
	4n	9	~stack or sheared stack	rr0082 767		2'etptt3'	l
	0i	6	-β next to 6n; bases perpendicular	rr0082 940			l
	6n	18	UNCG 3-4; Z23 DNA; syn base triple	rr0082 1773	BD-36	2'epptt3'	l
	6j	9	+β next to 6n; bases far	pte003 975			l
22p	**2[**	40	UNCG 2-3; near B-DNA; k-turn 3-4	rr0082 264	BD-38	2'emm-135p2'	r
	4b	27	cross-stacked A-helix end	rr0082 247	BD-10	2'etmtp2'	R
	0b	14	varied	rr0082 453	BD-11	2'epmtp2'	R
	4p	13	often starts 1-bulge, Hb O2'$_{-1}$-N7$_{+1}$	rr0082 873		2'etptp2'	c
	6p	39	k-turn 2-3	rr0082 1315	BD-21	2'epptp2'	c
22t	**4s**	8	S-motif 2-3; low β	ur0026 2655			h
22m	**2o**	12	bases perpendicular, something between	pr0033 b5	BD-23	2'emmtm2'	

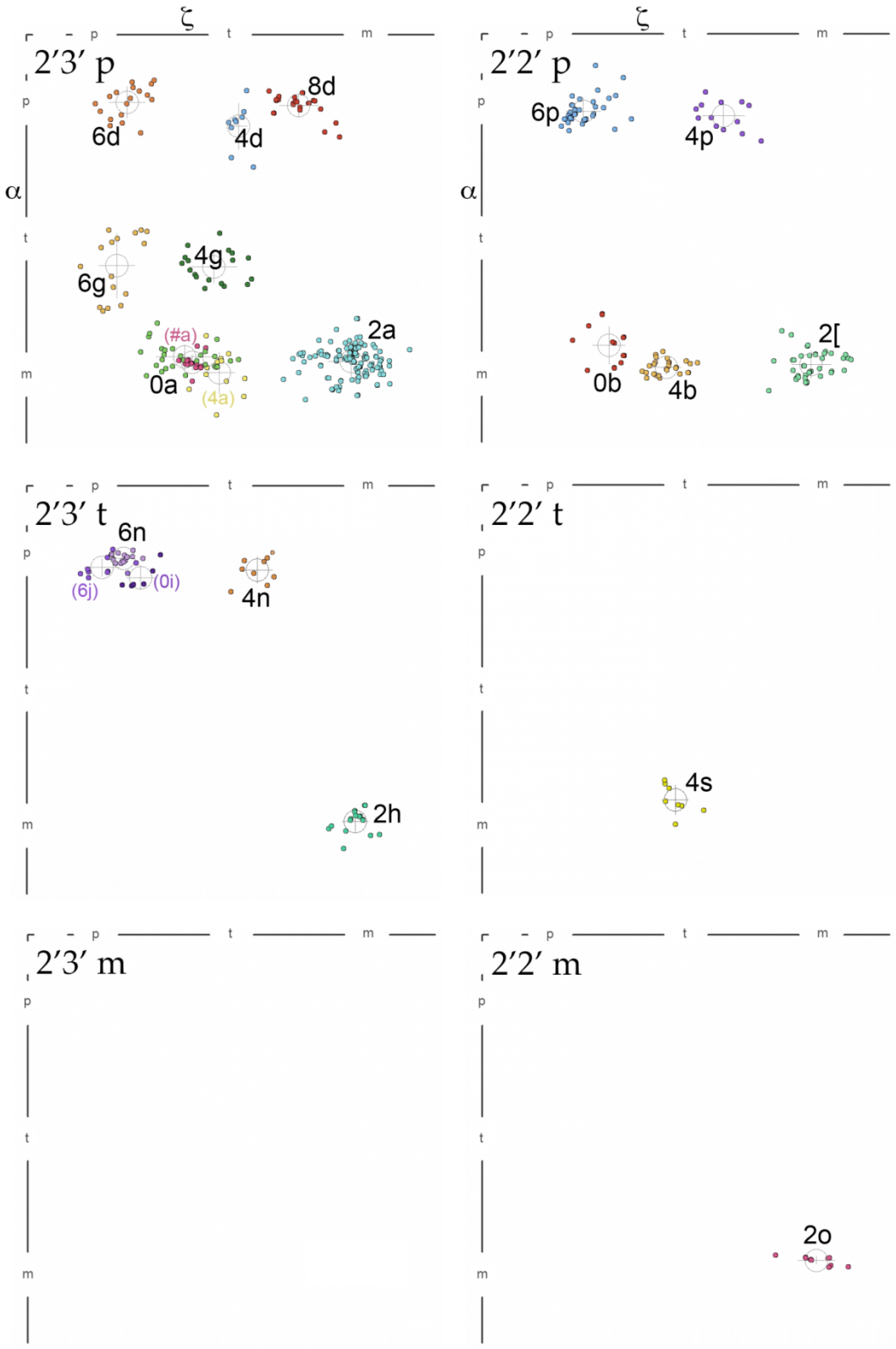
Notes:
List is sorted by δ_{-1}, then δ, then γ, then α, then ζ, starting from the A-form or most common value.
"name" is the 2-character modular consensus cluster name.
Cluster points include those from both FT and 7D suite analyses.
Examples are numbered according to the central P of the suite (2nd base)
"=dinuc" is the updated equivalents to Schneider et al. 2004; "=suite" is taken from Murray et al. 2003; "=bin is the suite-binned equivalent updated from Hershkovitz et al. 2003. See text.

Table 2: Consensus conformer dihedral means and standard deviation

$\delta_{-1}\delta\gamma$	name	δ-1	ε-1	ζ-1	α	β	γ	δ
33p	**1a**	81 (4)	-148 (10)	-71 (7)	-65 (8)	174 (8)	54 (6)	81 (3)
	1m	84 (5)	-142 (16)	-68 (15)	-68 (16)	-138 (12)	58 (10)	86 (7)
	1l	86 (4)	-115 (6)	-92 (13)	-56 (8)	138 (4)	62 (10)	79 (5)
	&a	82 (5)	-169 (7)	-95 (6)	-64 (9)	-178 (10)	51 (7)	82 (5)
	7a	83 (4)	-143 (23)	-138 (14)	-57 (9)	161 (15)	49 (6)	82 (3)
	3a	85 (4)	-144 (24)	173 (14)	-71 (12)	164 (16)	46 (7)	85 (6)
	9a	83 (2)	-150 (15)	121 (13)	-71 (12)	157 (23)	49 (6)	81 (3)
	1g	81 (3)	-141 (8)	-69 (9)	167 (8)	160 (16)	51 (5)	85 (3)
	7d	84 (4)	-121 (16)	-103 (12)	70 (10)	170 (23)	53 (6)	85 (3)
	3d	85 (4)	-116 (15)	-156 (15)	66 (19)	-179 (23)	55 (6)	86 (4)
	5d	80 (4)	-158 (7)	63 (14)	68 (12)	143 (30)	50 (7)	83 (2)
33t	**1e**	81 (3)	-159 (8)	-79 (6)	-111 (9)	83 (11)	168 (6)	86 (4)
	1c	80 (3)	-163 (9)	-69 (10)	153 (12)	-166 (12)	179 (10)	84 (3)
	1f	81 (2)	-157 (14)	-66 (11)	172 (11)	139 (13)	176 (10)	84 (3)
	5j	87 (7)	-136 (23)	80 (15)	67 (9)	109 (10)	176 (6)	84 (4)
32p	**1b**	84 (4)	-145 (10)	-71 (10)	-60 (9)	177 (12)	58 (7)	145 (7)
	1[	83 (4)	-140 (10)	-71 (10)	-63 (8)	-138 (9)	54 (7)	144 (8)
	3b	85 (3)	-134 (18)	168 (17)	-67 (15)	178 (22)	49 (5)	148 (3)
	1z	83 (3)	-154 (18)	-82 (19)	-164 (14)	162 (25)	51 (5)	145 (5)
	5z	83 (3)	-154 (5)	53 (7)	164 (5)	148 (10)	50 (5)	148 (4)
	7p	84 (3)	-123 (24)	-140 (15)	68 (12)	-160 (30)	54 (7)	146 (6)
32t	**1t**	81 (3)	-161 (20)	-71 (8)	180 (17)	-165 (14)	178 (9)	147 (5)
	5q	82 (8)	-155 (6)	69 (14)	63 (9)	115 (17)	176 (6)	146 (4)
32m	**1o**	84 (4)	-143 (17)	-73 (15)	-63 (7)	-135 (39)	-66 (7)	151 (13)
	7r	85 (4)	-127 (13)	-112 (19)	63 (13)	-178 (27)	-64 (4)	150 (7)
23p	**2a**	145 (8)	-100 (12)	-71 (18)	-72 (13)	-167 (17)	53 (7)	84 (5)
	4a	146 (7)	-100 (15)	170 (14)	-62 (19)	170 (34)	51 (8)	84 (5)
	0a	149 (7)	-137 (11)	139 (25)	-75 (11)	158 (20)	48 (6)	84 (4)
	#a	148 (3)	-168 (5)	146 (6)	-71 (7)	151 (12)	42 (4)	85 (3)
	4g	148 (8)	-103 (14)	165 (21)	-155 (14)	165 (15)	49 (7)	83 (4)
	6g	145 (7)	-97 (18)	80 (16)	-156 (29)	-170 (23)	58 (5)	85 (7)
	8d	149 (6)	-89 (10)	-119 (17)	62 (10)	176 (23)	54 (4)	87 (3)
	4d	150 (6)	-110 (26)	-172 (7)	80 (20)	-162 (20)	61 (8)	89 (4)
	6d	147 (6)	-119 (23)	89 (16)	59 (14)	161 (23)	52 (7)	83 (4)
23t	**2h**	148 (4)	-99 (8)	-70 (12)	-64 (10)	177 (17)	176 (14)	87 (4)
	4n	144 (7)	-133 (14)	-156 (14)	74 (12)	-143 (20)	-166 (9)	81 (3)
	0i	149 (2)	-85 (20)	100 (13)	81 (11)	-112 (12)	-178 (3)	83 (2)
	6n	150 (6)	-92 (11)	85 (8)	64 (5)	-169 (8)	177 (9)	86 (5)
	6j	142 (8)	-116 (28)	66 (15)	72 (8)	122 (22)	-178 (6)	84 (3)
22p	**2[**	146 (8)	-101 (16)	-69 (17)	-68 (12)	-150 (21)	54 (7)	148 (7)
	4b	145 (7)	-115 (20)	163 (13)	-66 (6)	172 (14)	46 (6)	146 (6)
	0b	148 (4)	-112 (20)	112 (14)	-85 (17)	165 (16)	57 (12)	146 (6)
	4p	150 (10)	-100 (26)	-146 (19)	72 (13)	-152 (27)	57 (14)	148 (4)
	6p	146 (7)	-102 (21)	90 (15)	68 (12)	173 (18)	56 (8)	148 (4)
22t	**4s**	150 (2)	-112 (16)	170 (12)	-82 (13)	84 (7)	176 (6)	148 (2)
22m	**2o**	147 (6)	-104 (15)	-64 (16)	-73 (4)	-165 (26)	-66 (7)	150 (3)

Note: Cluster torsion angle means are given in degrees, with standard deviations in parentheses.


ζ
p
t
m
3′3′ p
α
5d
3d
7d
1g
9a
3a
7a
(&a)
(1m)
1a
(1L)
3′2′ p
7p
5z
1z
3b
(1[)
1b
3′3′ t
5j
1c
(1f)
1e
3′2′ t
5q
1t
3′3′ m
3′2′ m
7r
1o


ζ
p
t
m
2′3′ p
6d
8d
4d
α
6g
4g
(#a)
0a
(4a)
2a
2′2′ p
6p
4p
0b
4b
2[
2′3′ t
6n
(6j)
(0i)
4n
2h
2′2′ t
4s
2′3′ m
2′2′ m
2o

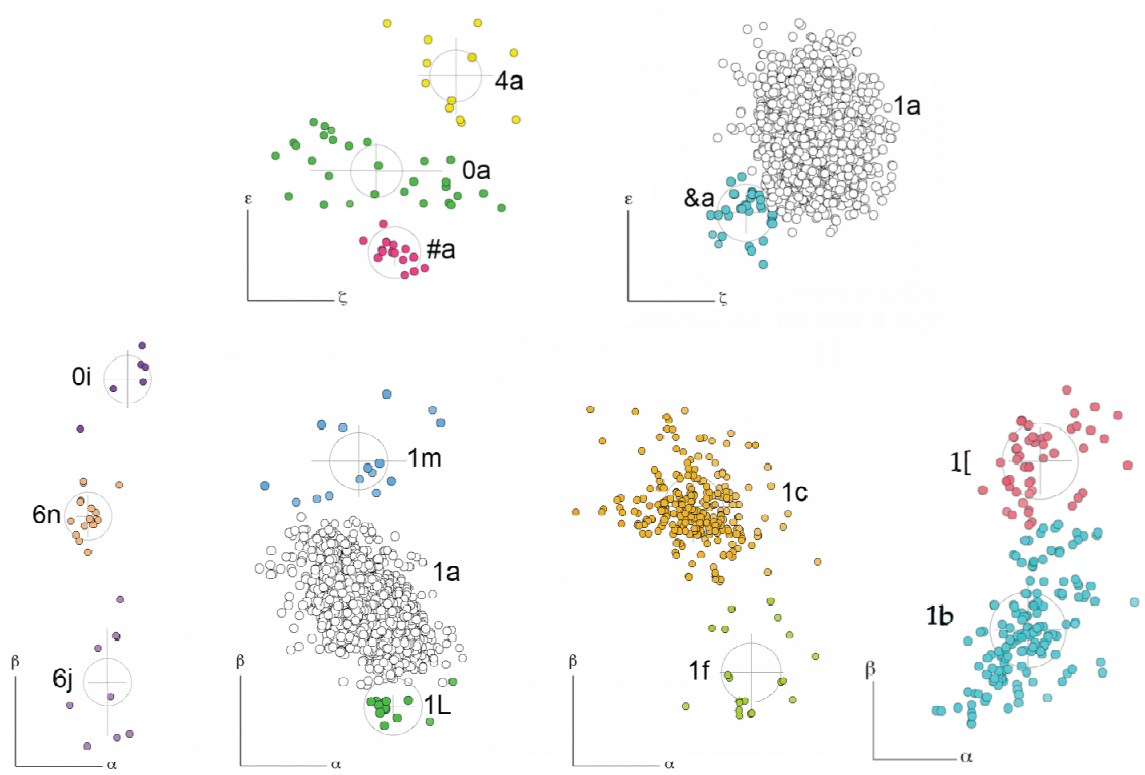


Figure 13: Panels of all 46 suite conformers.

Panels 1 and 2 (preceding pages) show the $\zeta_{n-1}\alpha$ plots of the $\delta_{n-1}\delta\gamma$ bins. The above illustration shows 2 $\epsilon_{n-1}\zeta_{n-1}$ divisions and 4 $\beta\alpha$ divisions, further differentiating each suite conformer. Note: Figure 10, Tables 1 & 2, and Figure 13 are reproduced here from Richardson, *et al.*, 2008.

2.4 Automatic Suitename Assignment

Almost all the work put into developing the consensus conformers was done by hand, thus depending on the expertise of the scientists involved to assign each suite to a particular cluster. With seven dihedrals for a given suite, manual assignment of a suite to a particular cluster is time-consuming, and scientists unfamiliar with RNA backbone or high-dimensional datasets may find the assignment process overwhelming. To ameliorate these aspects of the analysis, and to preclude erroneous assignments due to bias or unfamiliarity, we developed a new software package called Suitename. Written in C by David Richardson, Suitename takes as input a set of $\delta_{n-1}\epsilon_{n-1}\zeta_{n-1}\alpha\beta\gamma\delta$ torsions generated by DANGLE, an earlier program developed by Ian Davis and Daniel Keedy to measure dihedral angles. Suitename then uses an algorithm to assign these suite torsions to clusters using the modular consensus nomenclature, and also gives a measure of how well a particular suite fits into its assigned cluster.

As mentioned above, $\delta_{n-1}\delta\gamma$ plots provide a good starting point for assignment, since δ is bimodal and γ is trimodal, producing 12 well-defined bins for first-pass sorting (see Figure 12). For Suitename, these bins are defined by δ=55°-110° (C3′) or 120°-175° (C2′), and γ= 20°-95° (p), 140°-215° (t), or 260°-335° (m). Suitename then finds all clusters in ϵ_{n-1},ζ_{n-1},α, or β torsion space to which the suite could be assigned. This is done by generating an axially-oriented ellipsoid centered on the mean of a given cluster, whose semi-axis in each coordinate direction is $3\langle\sigma\rangle + 15°$, where $\langle\sigma\rangle$ is the average of all

cluster deviations in that dimension. If the ϵ_{n-1},ζ_{n-1},α, and β torsions lie within the boundary of a clusters' ellipsoid, a 4-dimensional distance is calculated from the cluster mean to the suite. This distance is scaled from 0-1, with 0.0 at the cluster mean, and 1.0 at the ellipsoid surface. A lower number, therefore, means a suite is closer to the mean of that cluster; if a suite fits within the ellipsoids of multiple clusters, it is assigned to the cluster with the smallest mean-to-suite distance.

We soon found that the ellipsoids were very useful, but were sometimes foiled by the significant diagonal spread in clusters with many data points, such as **1a**, **1c**, and **1g**. To correctly assign these cases, we expanded Suitename to use superellipsoids (Gielis 2003), mathematical constructions which start as an ellipse, but move progressively towards the corners of a superscribed rectangle as the exponent increases. The superellipse used by Suitename has the following equation:

$$|\varepsilon / a|^n + |\zeta / b|^n + |\alpha / c|^n + |\beta / d|^n = 1$$

where a,b,c,d are halfwidths of torsion angles ϵ, ζ, α, and β, respectively, and exponent n is 3. The superellipsoid greatly improves the coverage of the above clusters and many others as well.

Not all of the clusters are cleanly divided. **1a**, **1c**, **1b**, **0a**, and **6n** are dominant clusters that each have satellite clusters with >50% overlap, even using superellipsoids, which can lead to significant assignment problems. For example, in the case of **1a** vs. **1L**, the boundary plane that would divide the suite assignment is 4 times farther from the

mean of **1a** than from the mean for **1L**, causing many false **1a** assignments that should have gone to **1L**. To decide membership between these and other cluster pairs with similar problems, the 4D distance is scaled in the relevant dimension by the same ratio as of the two distances from the cluster means to the boundary. If the suite can be potentially assigned to more than two clusters, its closest non-dominant cluster is found using the general algorithm with default scaling, and then the asymmetric comparison is made with the dominant cluster.

There are two special cases in which suites will not be assigned to a valid conformation. Suites at the beginning of a chain or immediately following a chain break will not have an n-1 residue to measure from, making assignment impossible; these are assigned the a null conformation marker: __. If a suite simply does not fit into any cluster, on the other hand, they are assigned the outlier conformation of **!!**. To help identify *why* a particular suite is a **!!**, certain common error diagnostics have been built into Suitename. Near-zero values of δ signify incorrect ribose stereochemistry, and ϵ lower than 155° or higher than 310° (or a mismatch of δ with 3'P perpendicular) indicates a sugar pucker outlier. Other single-angle outliers include β outside 50°-290° or α or γ outside 25°-335° (Murray 2007). Since these are almost certainly errors (as opposed to rare or new conformations), they are marked with "**trig**" (for triaged) and the angle name to help identification and correction.

Now that the suites have their correct assignments, it is useful to evaluate how well the suites fit into their clusters (if they have one). To do this, a scaled superellipsoid distance is computed in all seven dimensions, including δ_{n-1}, δ, and γ, and is converted into a "suiteness" value using the equation

$$S = (\cos(\pi d) + 1) / 2$$

where d is the 7D superellipsoid distance, and S is the suiteness. Note that this suiteness value is 1.0 at the cluster mean and moves to zero at the surface of the superellipsoid. To avoid rounding errors, the lowest a non-outlier can go is .01; all outliers have a suiteness of 0. A higher suiteness indicates that a particular suite fits its assigned conformational cluster better, but it is important to keep in mind that other validation metrics such as clashes and geometry should take precedence when building a model.

2.5 Structure and Properties of Individual Suite Conformers

With the demarcation of each collection of backbone dihedrals into a suite conformer, we discovered that certain suites were more suited for particular structural tasks than others. In our **1[** example above, the distance between the bases allowed for intercalation with distant RNA structure, with small molecule aromatic rings, or with aromatic protein sidechains. In contrast, the base stacking in **1a** does not lend itself well to such interactions. In this section, I describe some of the differences in particular suites and elucidate some of their general roles in RNA structure.

2.5.1 H-bonding patterns in suites

To better determine the relationship between the consensus conformers and their surroundings, particularly in regions with RNA-protein contacts, I developed a script to identify H-bonding patterns for each suite. I ran this script on each of the structures from the RNA05 dataset, and tabulated the results. For each atom in the suite that was involved in an H-bond, I listed the donor and acceptor atoms, how far apart the atoms were in sequence space (if they were in same chain), and what the preceding and following suites were. This provided information on the locale of each suite, and suites were flagged if there were at least two instances of a conserved hydrogen bond with the same set of parameters (donor-acceptor pair, sequence distance, and surrounding suites). The results gave insight into local conditions surrounding the structure of a given conformation, and could be used to help establish the role of a given suite conformer within the overall structure. For example, many **#a** conformations contained conserved hydrogen bonds between base$_n$ and base$_{n-1}$, indicating that this suite can potentially make dinucleotide platforms; looking at the **#a** in context of surrounding structure reveals that it is responsible for the dinucleotide platform between G and U of the S-motif (Correll 2003a). Another common set of base$_n$ and base $_{n-1}$ H-bonds were found for the 4g conformation, this time found to be part of the Adenine platform (Cate 1996).

The most valuable results from the H-bond analysis came from suites H-bonding with non-adjacent residues. In the tRNA structures, there are several repeated occurrences of a backbone hydrogen bond between a **1[** suite and a **1b** suite that skipped a suite, **4d**, between them. Upon examination of other occurrences, we were surprised to find examples of these conserved bonds in the ribosome; this led to the discovery of an entire tRNA-like TΨC domain that was replicated between 618-642 of the 23S subunit of the *H. marismortui*, complete with a methyl-adenine at 628 and similar neighboring chain-interactions, e.g., intercalation of an outside base between the bases of the **1[** suite. This large motif is discussed and generalized further in Chapter 3.

For the **1g** conformation, I observed three different conserved H-bonds. One was a base interaction between H21 of the n-1 residue of the **1g** suite to the N7 of the n+2 residue; this indicated a non-Watson-Crick base pair between a G and another purine, which upon further analysis proved to be an A in every case. The second conserved H-bond was between the backbone 2′ hydroxyl of n-1 and the N7 of n+1, showing that the **1g** backbone conformation has a specific interaction with a neighboring purine. The last bond was between H22 of n-1 and O2P of n+2, another specific base-backbone interaction. These three conserved H-bonds turn out to be those that define the G position in the GNRA tetraloop (Heus and Pardi 1991; Correll 2003b), as seen in Figure 14; the **1g** conformation is a rotation of the standard A-form structure around α by

roughly 120° that is necessary to make the loop structure. The fact that all three H-bonds are conserved for many **1g** suites shows just how stable the GNRA tetraloop motif is.

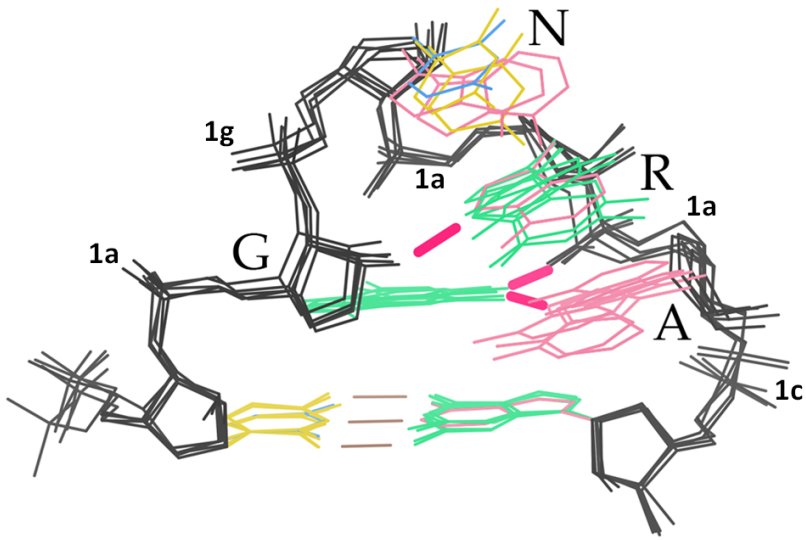


Figure 14: GNRA tetraloop with three conserved Guanine H-bonds highlighted in hot pink.

Other H-bond patterns are not as conserved, but still lead to useful insights into the structure. I found several instances of **1m** suites with conserved H-bonds between $base_n$ to $base_{n-42}$—a surprisingly long sequence distance, but one that implied a long helix with an internal loop. Indeed, these interactions were identified in Helix 30-31 and Helix 68 of the 50S ribosome structure. Upon closer inspection, it was clear that the **1m** suite was forming an intercalation interaction with the neighboring strand, and the second base of **1m** was involved in a base triple. In Helix 30, the **1m** suite takes part in the base triple that divides Helix 30 and Helix 31; in Helix 68, the **1m** suite forms part of a base

triple that sets off a loop region between Helix 68 and Helix 69. In both cases, the **1m** was the last suite in the helix, providing an easily identifiable interruption to the secondary structure with one base participating in the last helical base-pair and the other base beginning the transition to the next structural motif.

2.5.2 Discovery of a new suite

While we worked towards consensus, we continued to examine the suite conformations from the RNA05 dataset in hopes of finding new ones to add to our repertoire. The first new suite to be discovered was a variant from the TΨC loop in tRNA. In this alternate structure, there is a pucker change in the trailing residue of the TΨC turn, where the C2′-endo pucker of suite **1b** is replaced by a C3′-endo pucker, resulting in suite **1a**. The suite following this is normally **2a**, after which standard A-form helix occurs; with the new C3′ pucker, there was no known suite that could make the transition, and the result was marked as a **!!**. The only remarkable feature of this suite was a hydrogen bond between the second base of the suite to the phosphate of the previous residue. In an attempt to find other examples, I searched for other $base_n$ to $phosphate_{n-1}$ H-bonds, and observed several examples in **!!** suites in the ribosome. There were only a few examples, but they were similar enough to each other to be classified as a new "wannabe" suite, **3g**. Our later updates of the RNA dataset to RNA11 reveals that **3g** is indeed a new conformation, and it has since been promoted from "wannabe" status to a fully recognized conformation.

2.6 Suite conformations along RNA-protein interfaces

The discovery of suite conformations now gives us a new perspective on the role of the backbone in RNA-protein interactions. Even though the 53 non-A-form conformers only account for 30% of the total RNA suites, most RNA-protein interfaces contain at least one non-A-form conformation. To see whether this was due to suite distribution, or to more specific interaction between RNA and protein, I analyzed each suite along the RNA-protein interfaces in our RNA_Prot2011 dataset (section 4.1.2). By comparing the distribution of suites along the interfaces to the distribution among the overall RNA structure, I were able to discover some insight into how RNA backbone interacts with protein. To determine if there was a significant difference between free RNA and the RNA-protein complexes, the suite distribution in RNA_Prot2011 was also compared to the suite distribution of RNA09, an updated version of RNA05.

The first interesting result is that the standard A-form RNA backbone conformation, **1a**, is more common in non-interacting RNA. A-form RNA accounts for 73.4% of non-outlier suite conformations in RNA09, but only 70.0% for RNA_Prot2011. Furthermore, along the RNA-protein interfaces, **1a** accounts for only 67.8% of the suites (Table 3). This means that compared to general RNA structures, we expect to find 3.4 fewer **1a** suites per hundred non-**!!** suites in RNA that interacts with protein, and we expect to find 5.6 fewer **1a** suites out of every 100 non-outlier suites along the RNA-protein interface itself. This provides the first solid evidence to confirm that protein

generally interacts with non-standard, well-defined backbone conformations in the RNA, rather than simply recognizing A-form RNA.

The next interesting result is that the 53 non-A-form conformations did not simply increase in relative abundance along RNA-protein interfaces. In fact, half of the 16 non-A suites with two C3′-endo sugar puckers (3′3′) decreased, including almost all of the A-form satellite clusters. However, only 4 of the 16 3′2′ clusters decreased, while only 2 of the 12 2′3′ and none of the 2′2′ decreased, showing that RNA-protein interfaces have a relative preference for C2′-endo sugar puckers over C3′-endo. Quantitatively, the relative abundance of 2′2′ suites increased the most at RNA-protein interfaces, followed closely by 3′2′ suites and further behind the 2′3′ suites.

This led us to investigate how individual suites were distributed across the RNA backbone in RNA-protein complexes. Overall, just over a third (34.46%) of RNA suites in the RNA_Prot2011 dataset had contacts with protein according to PROBE; if the distribution of suites showed no preference for general RNA structure vs. RNA-protein interfaces, we would expect 34.36% of the instances of each suite to be along an interface. Instead, we found significant differences between the number of interacting suites vs. the non-interacting suites. Four conformations, including **1a**, were significantly underrepresented along the interface (by at least 3σ, or 99.73% confidence). This demonstrates that RNA-protein interactions generally involve more specific backbone interactions, rather than the non-specific binding we find in many DNA-protein

complexes. **1L** and **1m** are satellite clusters of **1a**, differing only in their β dihedral, so their absence from RNA-protein interactions helps bolster the evidence that proteins bind to non-standard RNA backbone structure. The last significantly underrepresented conformation was **7a**, which is one of the major stack-switch conformations. It is interesting that the more compact, 3′3′ stack switches (**7a** and **1e**) are rare along the interface, but the 2′-containing stack switch conformations, **0a** and **4b**, are some of the most overrepresented.

This analysis also revealed 15 conformers with a significant preference for RNA-protein interfaces (Table 3). Some of these form recognized motifs that will be discussed in further detail in the next chapter: **4s, 6p, 2[, 0a, 6g,** and **1b** represent suites that are part of S-motifs, kink-turns, U1A binding sites, and even TΨC motifs from tRNA. Other suites are good at interacting with protein individually; the **2o** and **2z** are not part of any recognized motifs, but they are ideal binding sites because their backbone dihedrals orient the bases perpendicular to each other, which encourages stacking or intercalation from either direction. A particularly poignant example of this versatility can be seen in the $tRNA^{ARG}$ bound to its cognate synthetase (Figure 15)—one base from a **2o** suite in the D-loop stacks on a tryptophan from the synthetase, while the other base in the suite is intercalated into the TΨC loop (see Figure 27), helping stabilize the tRNA structure.

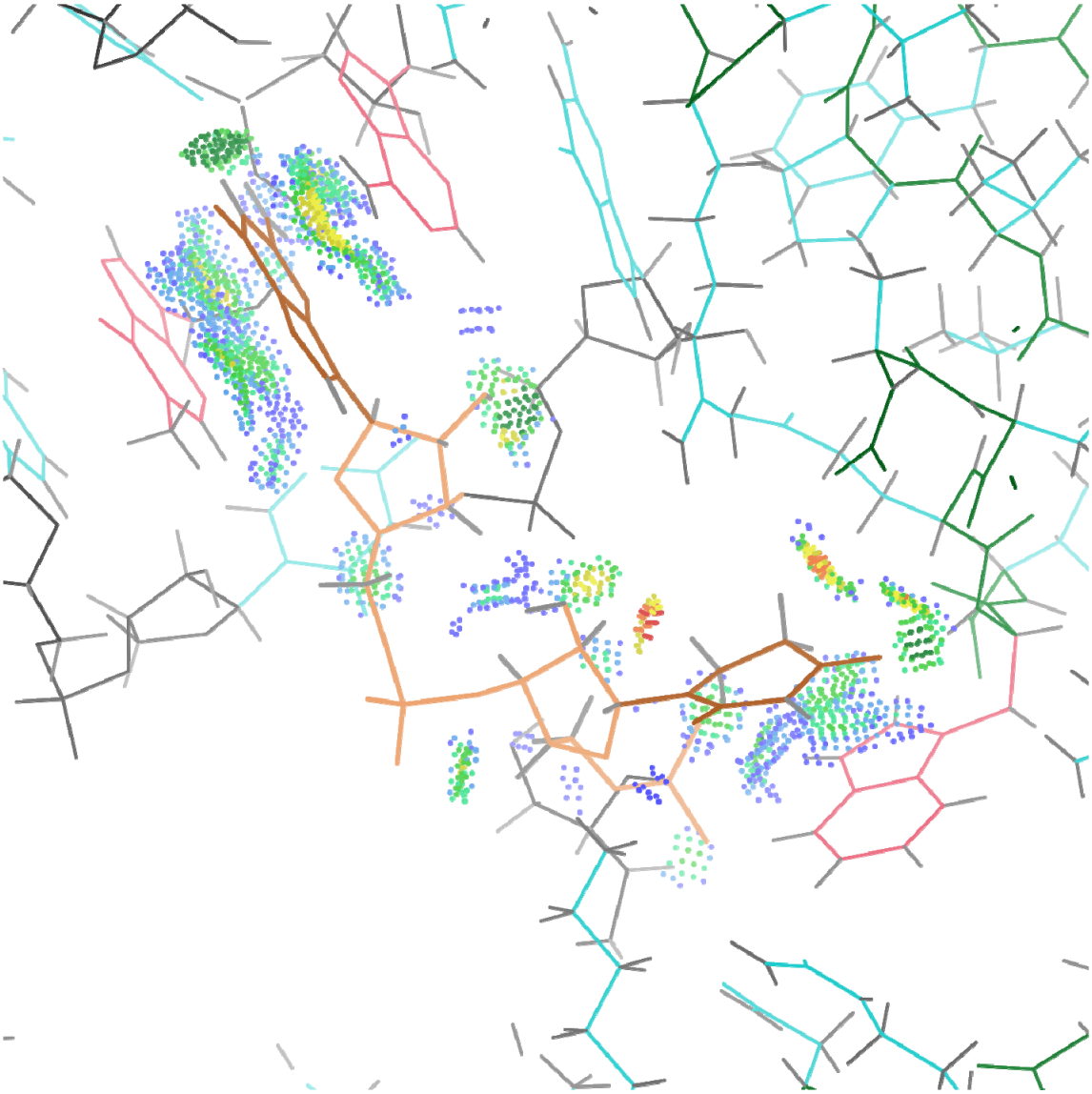


Figure 15: 2o from tRNA^{ARG}.

The 2o suite (peach) has bases at 90°, allowing stacking on both faces of each base. Residues 916-917 of tRNA^{ARG} (PDB: 1F7U; Delagoutte 2000) have a 2o suite; the modified base (H2U) of residue 916 interacts with W60 of the tRNA^{ARG} synthetase, while G917 intercalates between A957 and 1MA958 in the tRNA TΨC loop (see Figure 27). Residues 916-917 are in peach (lighter for backbone, darker for bases), while their stacking partners are in pink; contact dots are generated from PROBE (see section 1.3)

Among the interaction-preferred suites, **5j** is especially interesting; not only is it one of the few 3'3' structures that prefers RNA-protein interfaces, but it beats out all the other conformations save **6p** for strength of preference. This appears to be due to its extended backbone structure, resulting in two bases pointing away from each other, and providing a nearly flat surface which facilitates binding with either 2'OH or stacking with bases (Figure 16). Additionally, the **5j** suite appears almost exclusively as the closing suite of single-nucleotide bulges, which make excellent binding sites for protein (Hermann 2000); one particularly conserved example is rRNA's interaction with ribosomal protein L6.

Besides examining the propensity of a given suite to appear along a binding site, we also wanted to determine which parts of the suites were most likely to interact. For example, in A-form RNA, the bases primarily pair with each other, occupying most of the hydrogen-bond donor/acceptor sites and stacking with each other; conversely, the backbone faces the solvent, leaving the phosphate, and to a lesser extent the 2'OH, exposed for protein binding. Thus, there are many examples of the **1a** conformation along the RNA-protein interface in which the protein interacts with the backbone, but not the bases. Of our 54 conformations, 13 have a significant preference for interacting via the backbone rather than bases, including **1L, 1m, 5j** and **1b** mentioned above. It is interesting to note that while **1a**, **1L**, and **1m** are less common along an interface, when

they do interact, it is much more likely to be via their backbones rather than bases; **5j** and **1b**, which are overrepresented along interfaces, also interact via their backbones. Numerous other suites (Table 3) show no preference for RNA-protein interfaces or general RNA structure, but still mostly do interact with protein via their backbones.

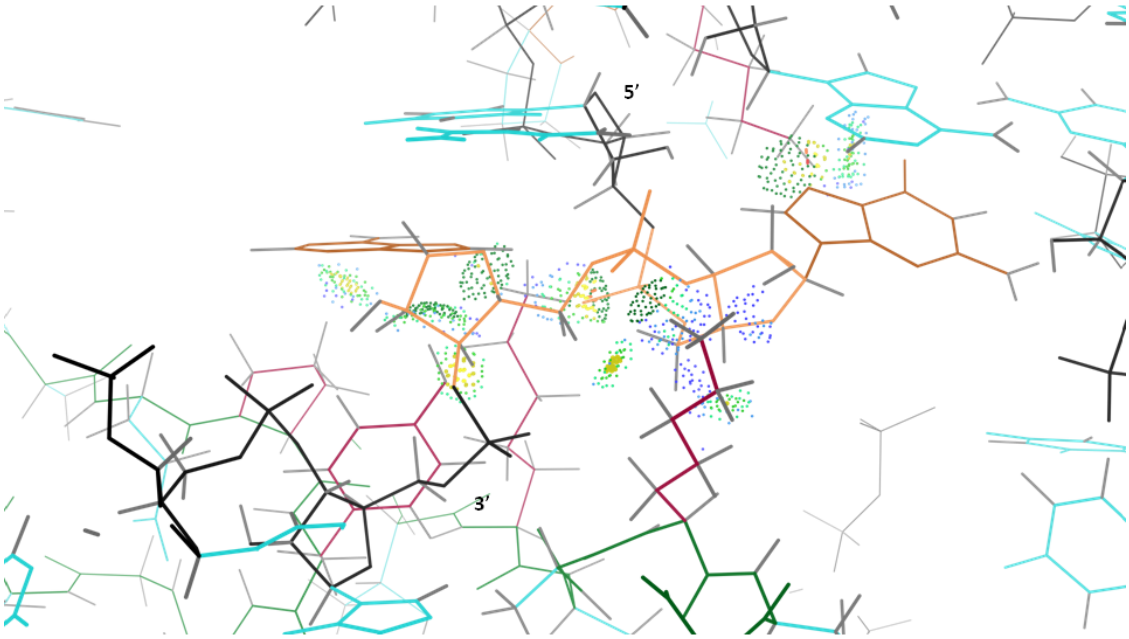


Figure 16: 5j from 23S rRNA

The 5j suite has bases pointing in opposite directions, presenting a broad surface area that accommodates extensive protein interactions. Residues 2529-2530 from the 23S rRNA (PDB: 3R8S; Dunkle 2011) interact closely with ribosomal protein L6, contacting P155, Y156, K171, and K174. Residue 2529 also interacts with K32 of L36. Residues 2529-2530 are in peach, with the interacting protein sidechains in maroon.

The tendency for some suites to interact via backbone is quite intriguing, not least because it confirms both that the negatively charged phosphate and 2'OH have very important roles in binding protein and that we can gain some insight into these roles by using suites, which in effect describe the physical positions of these functional groups in relation to each other and the bases. To gain some insight into the protein side of these interactions, I conducted a survey of RNA-protein interactions identified via PROBE across the RNA09 dataset (see section 4.1.1), searching for which types of sidechains interacted with the RNA most often; I found ~21100 amino acids interacting with RNA, roughly an average of ~1000 per amino acid type. As expected from our abundance of interactions between the negatively charged RNA backbone and protein, the most common sidechains along RNA-protein interfaces were the positively charged Arginine and Lysine, interacting 4x and 3x more than the average sidechain, respectively. Glycine is the third most common sidechain, and is recognizable from the well-characterized RGG domain (Kiledjian and Dreyfuss 1992). Cysteine interacts with RNA the least, an order of magnitude smaller than the average. Suprisingly, the next least common is tryptophan; phenylalanine and tyrosine are only slightly below average, but the aromatic with the most surface area is rarely found along RNA-protein interfaces. Taken together, these data support our earlier results that the RNA backbone plays a significant role in RNA-protein interactions.

Table 3: Suite behavior along RNA-protein interfaces. Underlined values are >3σ lower than expected values, bold values are >3σ higher than expected.

Suite	Along RNA-prot Interface	Not along RNA-prot interface	Interactions via backbone	Interactions via base
#a	23	40	**23**	8
&a	49	127	43	30
0a	**75**	82	53	58
0b	17	19	16	14
0i	26	33	26	15
0k	8	11	7	3
1L	159	**424**	**144**	81
1[	70	102	53	46
1a	6066	**12087**	**5628**	2915
1b	**244**	299	**227**	152
1c	600	1240	**544**	318
1e	41	103	33	22
1f	80	136	**77**	47
1g	165	268	145	122
1m	78	**278**	**62**	41
1o	15	15	14	8
1t	**46**	42	**43**	23
1z	28	46	23	20
2[	**78**	63	70	53
2a	108	219	**99**	62
2g	11	17	9	10
2h	**27**	29	26	25
2o	**15**	10	14	15
2u	9	14	7	6
2z	**23**	16	19	21
3a	60	104	52	43
3b	26	31	23	18
3d	43	78	40	37
3g	16	24	16	10
4a	28	37	22	18
4b	**50**	38	46	37
4d	12	26	12	7
4g	35	56	32	28
4n	15	19	**15**	7

4p	**38**	27	35	34
4s	**17**	9	15	17
5d	20	39	14	18
5j	**44**	25	**38**	20
5n	10	21	10	6
5p	**14**	12	14	7
5q	4	12	4	1
5r	8	13	5	6
5z	51	72	**47**	21
6d	16	60	13	14
6g	**44**	47	37	35
6j	15	22	15	11
6n	35	45	**34**	20
6p	**74**	49	69	66
7a	49	**150**	41	31
7d	47	59	43	42
7p	**35**	31	31	33
7r	15	41	13	11
8d	19	37	18	13
9a	47	71	41	38

2.7 Discussion

The study of RNA backbone has taken many twists and turns, but the most important step was the discovery that there were indeed backbone rotamers analogous to those found in protein sidechains. Working with other labs with similar interests, we were able to establish a set of 46 suite conformers that were universally recognized by nucleic acid structural biologists (Richardson 2008). As more RNA structures have been solved at ever improving resolution, 6 of the 8 wannabe conformers, including the newly established **3g** suite, have now been upgraded to full suite status, giving a grand total of 52 RNA backbone conformers. These snippets of RNA structure are useful for motif identification, as described in the following chapter, as well as RNA structure building and modeling, as described in chapter 5. What is more, they also give us the ability to describe the RNA backbone structure independently from the sequence, in a way that had never been possible before. The suite analysis from Suitename has been incorporated into MolProbity, and is currently used as an additional model evaluation; it has seen much use in the recent flood of ribosome structures.

Taken altogether, this work also confirms our suspicions that the RNA backbone is particularly important to RNA-protein interactions, and emphasizes the need for correct structural models—Chapter 4 addresses the work we have done to facilitate the creation and improvement of such models, while Chapter 5 shows how we have implemented these tools in newly solved structures.

3. RNA Motifs

RNA is known to form a variety of stable local structures in 3D, as determined by the overall base sequence; such structural motifs can be found in many folded RNAs (Tamura 2004), providing insight into both secondary and tertiary structure in a given region. Identification of motifs within a given RNA structure makes it easier to correctly model the RNA during initial building, and to identify functionally relevant binding sites or active sites. RNA motifs can also represent small, stable units of RNA structure that can form natural building blocks; the rapidly growing field of RNA tectonics (Westhof 1996; Grabow 2011) takes advantage of this stability to construct new nanoparticles from RNA. Given the importance of RNA structural motifs, there are surprisingly few methods available for identifying them in 3D Cartesian space—many programs used for motif identification rely on determining secondary structure motifs from the primary sequence and chemical probes like SHAPE (Merino 2005). While RNA secondary structure motifs are indeed useful, they largely ignore the vast differences in local interactions and chemical environment that influence the ultimate 3D structure. Furthermore, the RNA backbone is almost entirely ignored in these traditional secondary structure motif definitions, save for 5′-3′ connectivity and the occasional 2′ hydroxyl sugar-edge interaction. In this chapter, I present a survey of RNA structural motifs of varying sizes, how they were originally defined, and how their traditional definitions can be augmented by the inclusion of RNA backbone conformations. I will

also show how the use of RNA backbone conformational suitestrings can be used to identify structural motifs where the sequence is poorly conserved, something nearly impossible to do by sequence and secondary structure modeling alone.

3.1 Early RNA motifs

Early RNA motif definitions relied almost exclusively on the identification of particular sequences that were repeatedly found in primary and secondary structures. Through observation of primary sequence and chemical determination of hydrogen bonding, with confirmation provided by the few available RNA structures, many motifs had already been tentatively identified before the release of the ribosome structures in 2000 and 2001; the original Structural Classification of RNA (SCOR) database (Klosterman 2002) consisted almost entirely of structural motifs that used sequence and secondary structure predictions as their major determinants. Even by 2003, only the hook-turn (Szep 2003), A-minor motif (Nissen 2001), and Kink-turn (Klein 2001), all three of which required tertiary structure knowledge from the ribosome, could be considered "new" motifs (Duarte 2003). Unfortunately, even with improvements to the SCOR database (Tamura 2004), we found these traditional motif definitions, done in terms of primary sequence and secondary structure, to be too imprecise, as many times a search for a particular motif would turn up multiple models that differed wildly from canonical examples in their 3D structure; such bad models could make up almost 50% of

the results from a motif search. Armed with our set of RNA backbone rotamers (Murray 2003), I investigated a set of known RNA motifs (See Table 4) to determine if they had conserved backbone structure; this would show whether they were truly conserved in 3D Cartesian space or merely possessed similar primary sequence or secondary structure. In addition, I sought to determine whether conserved backbone was found consistently in regions with putative binding sites for proteins. Four known RNA backbone motifs—the kink-turn, the S-motif, the dinucleotide platform, and the tetraloop, are described briefly below; each of these motifs are known to associate with proteins, with binding involving the backbone atoms.

The kink-turn, or K-turn, is a known RNA-protein interaction motif that involves unusual RNA backbone conformations (Klein 2001). It consists of a helix/internal loop/helix, and includes a sharp kink in the phosphodiester backbone that bends the RNA helix axis by ~120°. In the *Haloarcula marismortui* 50S ribosomal subunit, K-turns interact with nine of the ribosomal proteins: L4, L7Ae, L10, L15e, L19e, L24, L29, L32e, and L37Ae. Small ribosomal proteins S11 and S17 also interact with K-turns in the *Thermus thermophilus* 30S structure. From its inception, the kink-turn was recognized to contain a highly conserved backbone structure, but diverse sequence, making a consensus sequence difficult to obtain. In fact, no two of the eight K-turns in the ribosome have the same sequence, yet the average backbone RMSD is only 1.7Å. The K-turn consists of two stems flanking an internal loop as seen in Figure 17: a helical stem

with canonical WC base pairs (the canonical, or C-stem), which ends at the internal loop, and a second helical stem that follows the internal loop and starts with two non-canonical base pairs (the NC-stem). The internal loop, the eponymous kink itself, is asymmetrical, with three unpaired nucleotides on one strand and none on the other. The 5′ nucleotide in the long strand of the loop stacks on the C-stem, the second on the NC-stem, with the third protruding into the solution. In many K-turns, a Hoogsteen-Sugar Edge (SE) interaction occurs between the starting G of the C-stem and the closing A of the NC-stem. There is considerable variability in how the K-turns interact with proteins and no recognized protein structure motifs are consistently involved, as demonstrated in Figure 18. Even so, four principal surface features of the RNA are recognized: the widened major groove of the C- stem, the flattened minor groove of the RNA NC-stem, the sharply kinked sugar-phosphate backbone and the protruded nucleotide, and the exposed base planes. The method of interaction presumably depends on the protein's surface features (most of which are sidechains) in relation to the unique surface of the K-turn motif (much of which is backbone). The K-turns all appear at or near the surface of the ribosome, further solidifying their importance to protein-RNA binding, but identification and prediction of K-turns in terms of binding location and consensus sequence has met with only limited success.

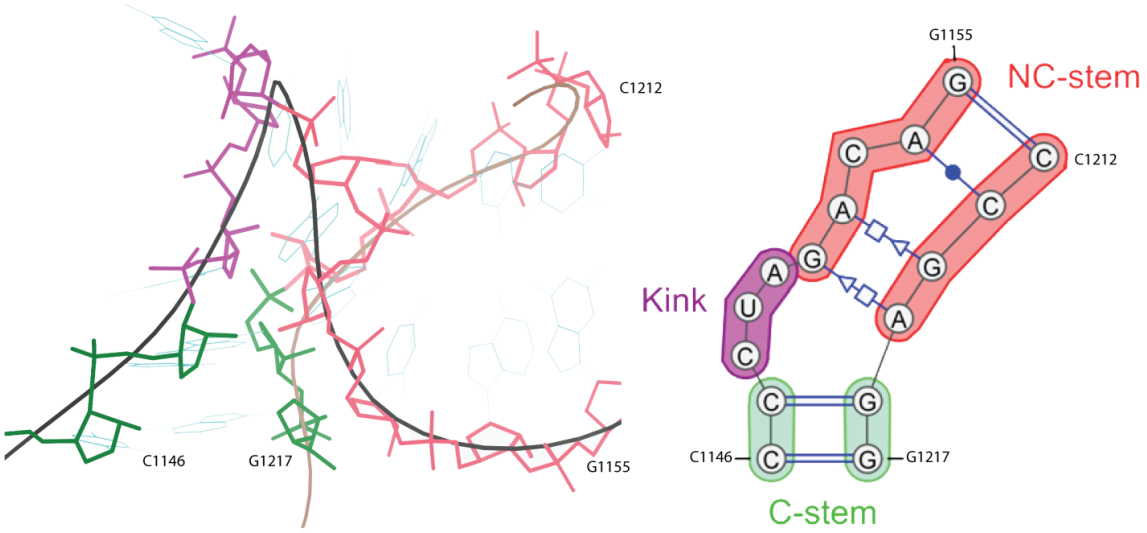


Figure 17: Secondary structure of Kink-turn.

The canonical stem (green) has standard WC pairing, and precedes the kink. After the kink (purple), the stem re-forms, but with non-canonical pairing . K-turn numbering taken from Klein, *et al.*, 2000, using the coordinates from the more up to date *H. marismortui* 50S subunit structure, 3CC2 (Blaha 2008). The secondary structure diagram was made with VARNA (Darty 2009). See Figure 24 for suitestring.

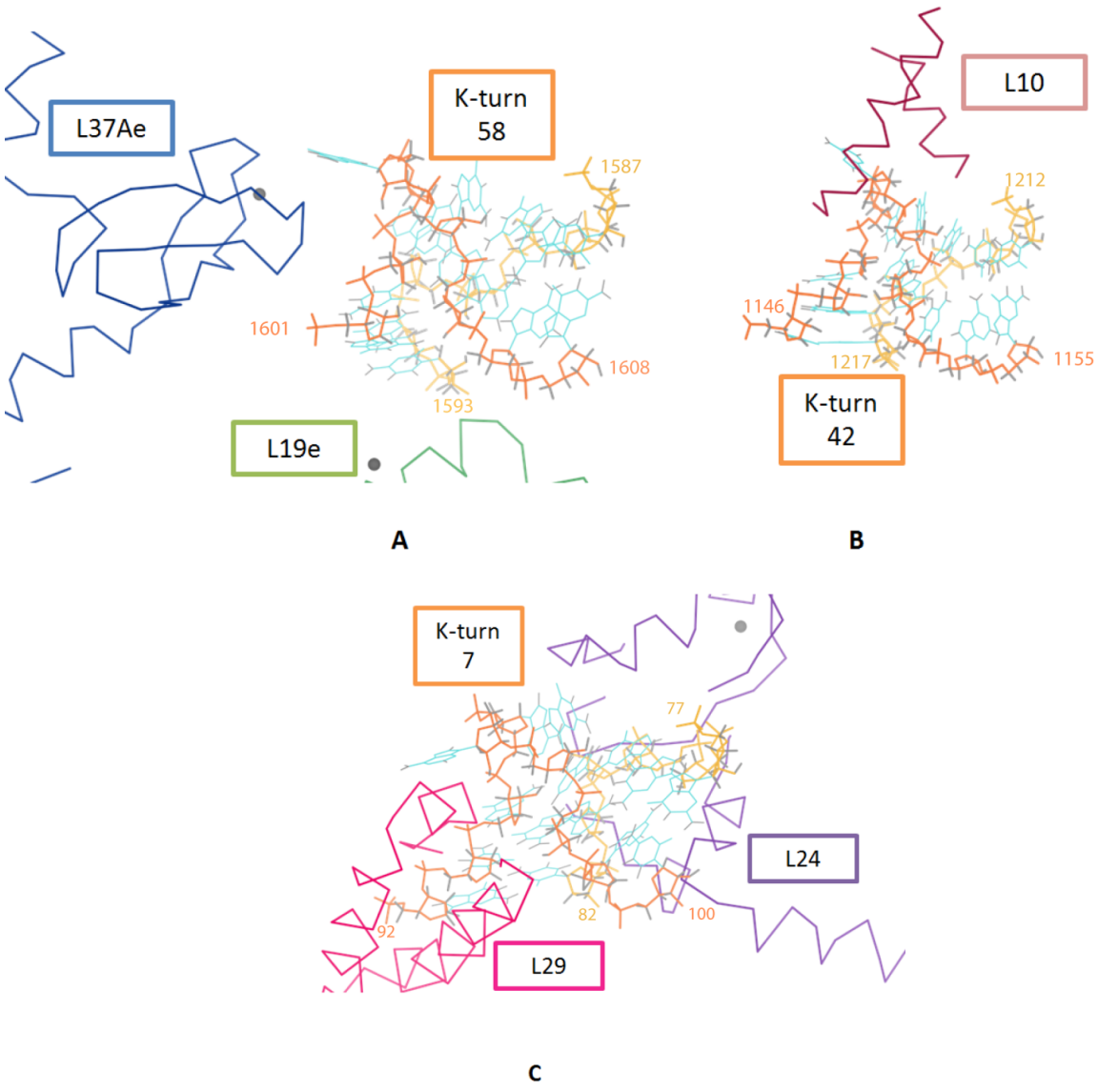


Figure 18: (A), (B), and (C) show the diversity of K-turn interactions; the K-turn itself is in orange for each case, with the back strand in gold.

(A) shows KT-58, which makes limited contacts with proteins L37Ae and L19e. (B) depicts a helical segment of L10 contacting the bulged nucleotide in KT-42, and (C) shows how L24 and L29 bind to opposite sides of KT-7. K-turn numbering taken from Klein, *et al.*, 2000, using the coordinates from the more up to date *H. marismortui* 50S subunit structure, 3CC2.

The S-motif (or sarcin-, S-turn-, bulged G-) is a distinctive and highly structured, asymmetric internal loop within an A-form RNA double helix, especially common in ribosomal RNAs; an example is shown in Figure 19. It includes several non-canonical base pairs and a base triple, and the backbone forms a pronounced S-shape on the primary strand with a stack switch on the secondary strand. The S-motif is named for its distinctive shape, but was first discovered as part of the larger loop E motif of the 5S ribosomal RNA (Branch 1985) and in the highly conserved sarcin/ricin loop of the large ribosomal subunit, which binds essential translation factors (Leontis 1998). Biochemical studies have shown that the middle suite (**4s**) is protected when Elongation Factor (EF) II (EF-G in bacteria) binds (Correll 2003a); EFIα/EF-Tu binding protects the bend in the back strand, colored pink and labeled **1e** (Szewczak 1993). Toxins like sarcin, ricin, and restrictocin inactivate ribosomes by cleaving the sarcin loop; the S-motif is at the toxin binding site (Correll 2003a). Classic S-motifs and variants also occur elsewhere in ribosomal and other RNAs, so there are many similar but not identical examples in our RNA05 structural database, including a few at very high resolution, e.g., ur0035/1Q9A at 1.04 Å resolution (Correll 2003a), a part of whose remarkable electron density is shown in Figure 4.

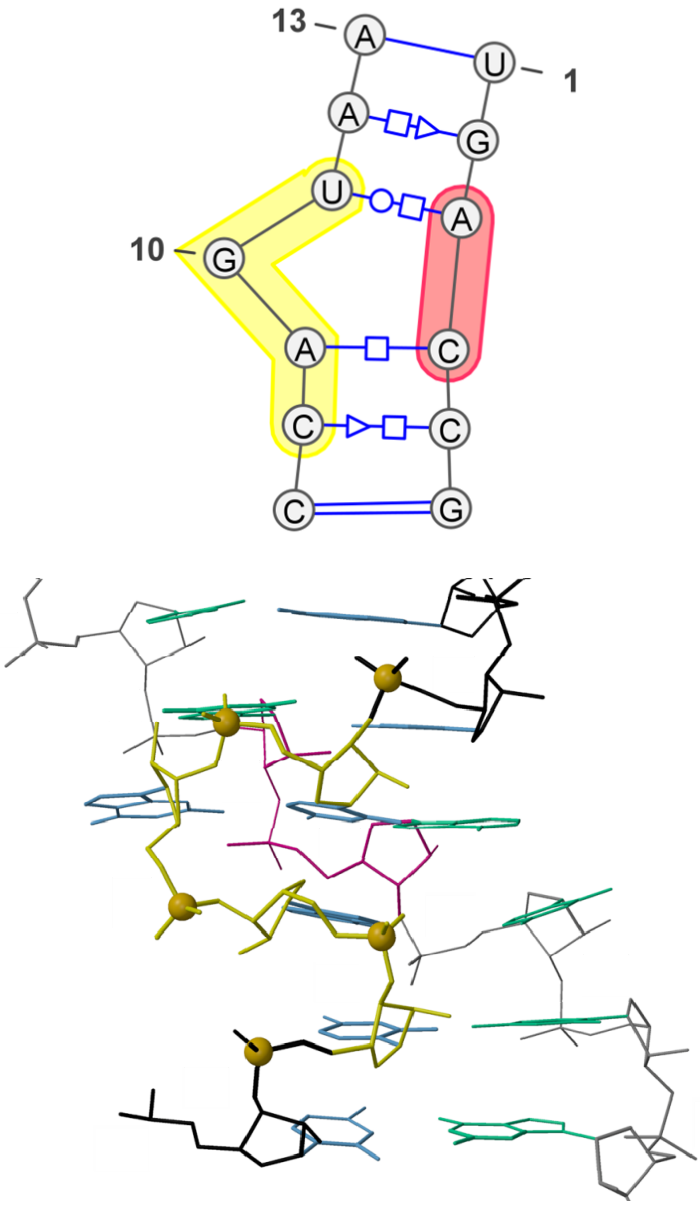


Figure 19: Secondary and 3D structure of an S-motif.

Both the secondary structure (top) and 3D structure (bottom) show the S-motif primary strand sequence in yellow and the conserved back strand in pink. The 3D structure is the S-motif from PDB code 1S72 (Klein 2004), residues 585-592 (front strand, black and yellow), and residues 566-572 (back strand, grey and pink). See Figure 25 for details on the suitestring.

The dinucleotide platform is an important motif first discovered in the Group I intron at three sites of long-range contacts in the tertiary structure, forming critical stacking interactions with the GAAA tetraloop as part of the GAAA tetraloop receptor (Cate 1996). Every dinucleotide platform consists of a coplanar base-pairing between two sequence-adjacent residues. This motif is most commonly found in the midst of standard A-form helix as a mediator for long-range tertiary interactions; the adjacent base pair results in a large planar surface that makes an ideal location on which bases or aromatic protein sidechains can stack. There are two major classes of dinucleotide platforms: adenosine platforms, in which the N3 5′ A forms a hydrogen bond with the N6 of the 3′A (Cate 1996), and GpU platforms, in which N2 of the 5′ G forms a hydrogen bond with the O4 of the 3′ uracil (Lu 2010). The GpU platform does double duty, since it makes up the last two residues in the primary strand of the S-motif as well as appearing on its own in various RNA structures. Despite the relatively simple characteristics of the dinucleotide platform –adjacent base pairing and coplanarity—many examples in the SCOR database that are identified as such do not have these characteristics, making it an ideal candidate for classification through our backbone system; if our system proves robust, we expect to identify only the subset of true dinucleotide platforms, and provide definitive evidence that the others deviate significantly from this motif. Examples of the two most common dinucleotide platforms are found in Figure 20.

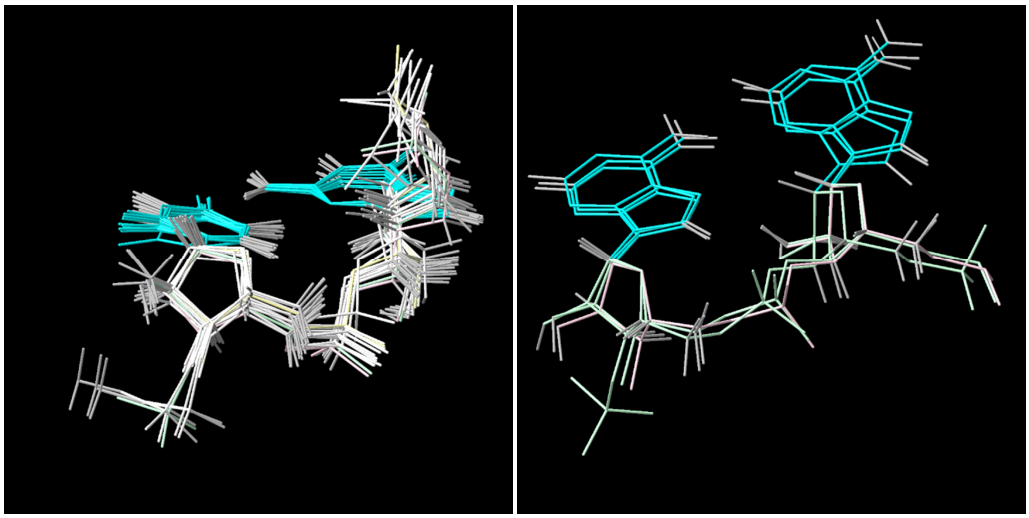


Figure 20: Dinucleotide platforms.

Superimposed GpU platforms (left) and adenosine platforms (right).

The last major RNA motif I searched for is the tetraloop. Two major classes of tetraloops, GNRA and UNCG, were early on identified by sequence; the N means any base can occupy that sequence space, and the R indicates any purine (pyrimidines are indicated by Y). The GNRA tetraloop accounts for over 50% of all hairpins, and is a very prominent motif in ribosomal structure. The G is in the 5′ stack and makes 3 H-bonds while the N-R-A are approximately in the 3′ stack. The UNCG tetraloop is the second most common tetraloop and contains U in the 5′ stack and the C and G in the 3′ stack, with the N base looped out (Holbrook 1991; Cheong 1990). Figure 21 shows examples of GNRA and UNCG tetraloops. These tetraloops join the two strands of a double helical region, and are closed off by at least one Watson Crick base pair (Heus 1991). Tetraloops

provide a means for RNA to fold back upon itself, reversing the direction of the phosphodiester backbone, and they often form tertiary interactions with other regions of RNA, establishing and reinforcing the tertiary structure. GNRA tetraloops are also important in binding proteins: the ribotoxins sarcin and ricin are site-specific enzymes that target the GAGA tetraloop (a GNRA), inhibiting protein synthesis on the same sarcin/ricin loop that also contains an S-motif (Wool 1992). One study showed that eight GNRA tetraloops share a similar backbone geometry but have an unexpectedly large variation in orientation of the last three bases (Correll 2003b). Thus, both the backbone conformation and the nucleic acid sequence may play a role in RNA-protein binding specificity.

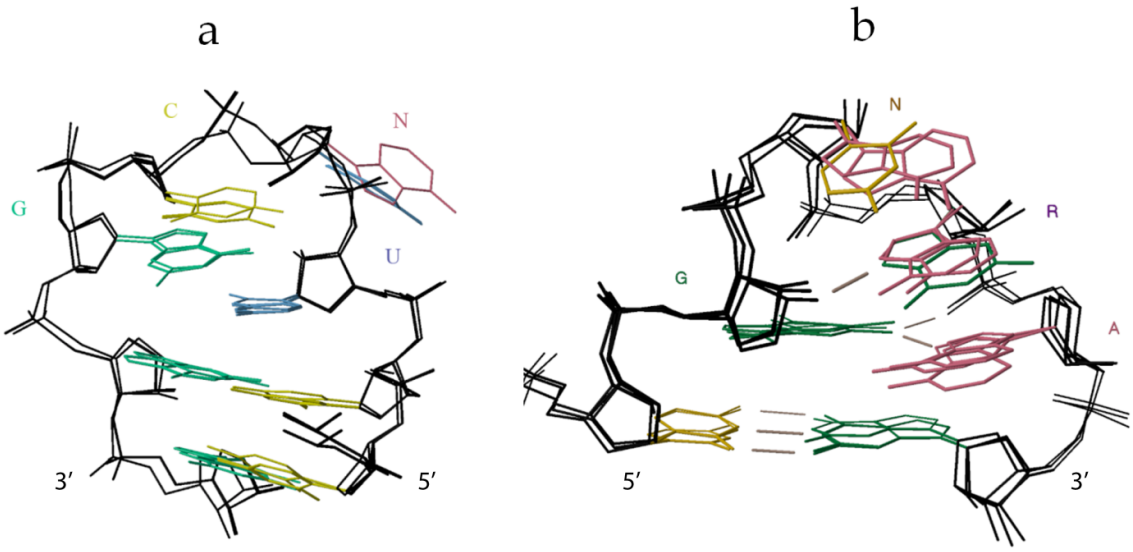


Figure 21: Tetraloops.

The UNCG (a) and GNRA (b) tetraloops are shown, with Watson-Crick base pairs closing the bottom of the loop. Note the bases extending away from the rest of the structure at the peak of the loop; these are often recognition sites for protein binding. Base types are color-coded: G green, C yellow, A pink, and U blue

While these motifs have already been established through biochemical means, they do not include descriptions of the RNA backbone. As established in chapter 2, the consensus modular nomenclature uses a 2-character name to describe the RNA backbone conformation. Adding this information to the existing motifs will aid in characterizing them; in addition, it will provide a check on how accurately motifs are defined when characterized solely by primary sequence and secondary structure.

3.2 Suitestrings

An advantage to using primary sequence to define motifs is the ease with which primary sequence strings can be searched. BLAST (Altschul 1990) takes advantage of the 1-dimensional nature of the primary sequence by treating it as a string, and efficiently finding many similar strings, corresponding to similar sequences. BLAST and ClustalW, a multiple sequence alignment tool founded on similar principles (Thompson 1997), are rooted in the premise that the sequence determines the overall tertiary conformation, and can be used to search large numbers of primary sequences for conserved regions. Yet these alignments only include information on the primary sequence—it is assumed that the tertiary structure is similar when the primary sequence is similar. This assumption can cause three major difficulties for those trying to predict structure accurately. The first occurs when small changes in sequence result in similar sequence alignments, but yield different RNA tertiary structures. This is true in particular for C→U mutations, as sometimes the wobble GU pair will maintain the same geometry, yet other times the loss of one H-bond will destabilize the tertiary structure, causing it to change dramatically; this is particularly noticeable for internal loops in which the G-C triple H-bond is crucial for maintaining stability. The second difficulty arises when two structures contain regions of similar tertiary structure, but the sequences are not well-conserved; both the TΨC-loop and the S2-motif have trouble with this problem, making it hard to tell from sequence alone which ones are real and which are false positives. The

third case—more insidious than the first two—results in a mismatch between the primary and tertiary structure alignments: the best scoring primary sequence alignment and the best scoring tertiary structure alignment place the aligned residues and gaps differently. This third case is more likely to occur in large structures, where such a mismatch could be easily overlooked in the overabundance of data; one example can be found in Figure 22.

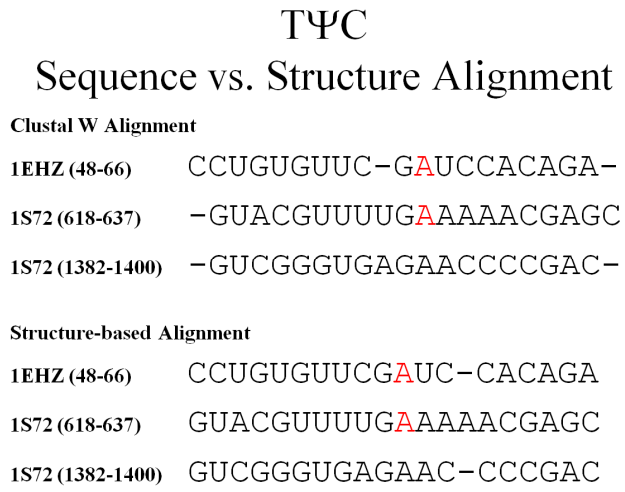


Figure 22: Sequence vs. structure alignments.

The sequence alignment done in ClustalW shows gaps in the 1EHZ (tRNA^{PHE}) sequence and none in the others. The structure-based alignment accurately depicts the inserted Adenine in the TΨC loop at position 631 of 1S72. 1-methyl Adenine modifications are in red.

The modular consensus nomenclature (see section 2.3) addresses these problems by providing a way to include the tertiary structure information directly in the sequence string. By making a string out of the identifiers of consecutive suite conformers, henceforth referred to as a suitestring, the tertiary structure of the RNA backbone can be seamlessly incorporated into the primary sequence, vastly increasing the available structural information. For example, the first residue of the GNRA tetraloop is G; the suite that follows G is in the **1g** conformation, which is then followed by base N. This can be written as **G1gN**, indicating that base G is part of the heminucleotide **1** (the δϵζ of the G residue), which is followed by heminucleotide **g** (the αβγδ of the N residue). Alternatively, the heminucleotide nomenclature can be used to refer to the residue—the G residue becomes **aG1**, where **a** is the αβγδ heminucloetide of residue G, with the **1** still representing the δϵζ heminucleotide of G. In this way, sequence alignments can be extended to include tertiary structure, whether it be defined in terms of residues or suites. The combined "sequence suitestring" for the GNRA tetraloop is **N1aG1gN1aR1aA1cN1a**: the sequence is NGNRAN, and the suitestring is **1a1g1a1a1c1a**. Any structure that matches this sequence suitestring is almost guaranteed to satisfy the motif requirements for a GNRA tetraloop listed in section 3.1. Note that the suitestring will always have a similar 3D structure, as each suite represents a particular 3D conformation. However, since there is some variation in each suite, the most robust suite definitions are given by using the sequence and suitestring in concert whenever

possible, rather than either one separately. The definition is even more robust if it also specifies the key H-bonding, but that cannot yet be expressed or manipulated as a simple string.

3.3 Alignment with suitestrings

Like primary sequence, suitestrings are represented by strings of individual terms that are easily parsed and easily searchable for patterns. Unlike sequence, which is a 1D string representing 1D structural information, the suitestring is a way of compressing 3D tertiary structure information into a 1D string. Suitestrings thus have the potential to revolutionize the world of structure alignments and homology modeling by allowing structural biologists to search for conserved 3D structure as easily as they would a BLAST search. This capability was not lost on our lab, and I immediately set to work implementing it.

3.3.1 SuiteBlast

My initial attempt to get a working search function for tertiary structure was called SuiteBlast, named thus as an homage to BLAST. Users could input a series of structures and then specify the reference suitestring they would use in the search. SuiteBlast then ran the structures through Suitename to get their suitestrings, and parsed all the suitestrings to have the length of the reference suitestring for each structure. The total number of unique suitestrings for each structure was then counted, and ranked by

a very basic scoring function that listed how many of the conformations in the suitestring matched those in the reference suitestring, using Chain ID and residue numbers to identify the appropriate region. In lieu of a reference suitestring, a length could be input instead, and the output would be an ordered list of suitestrings of the input length, ranked from most common to least. An additional search mode for SuiteBlast allows the user to query a particular suite conformer and specify the number of suites on either side of the central (queried) suite; the program then sorts the output based on the queried conformer followed by the 5′ suite, then 3′ suite.

This straightforward method allowed for easy search and comparison of suitestrings, but did not include common pseudo-alignment factors, such as gaps or substitution, that would allow more accurate ranking for inexact alignments. Furthermore, it did not deal with **!!** outlier conformations; any **!!** in the searched suitestrings was treated like a normal suite conformation, automatically penalizing that particular suitestring (since a **!!** would never match the reference suitestring). Nonetheless, this initial program was useful for identifying new suitestrings and keeping track of how often they appeared, giving an effective jumping off point for further investigation. For example, after seeing a particular suitestring of six conformations show up in several different structures, I was able to identify it as a conserved suitestring in the RNA-protein binding region of the U1 Hairpin II backbone (see Table 4). The alignment feature, as rudimentary as it was, also helped us discover a

rare instance of the TΨC loop outside tRNA—in this case, in the center of the 50S ribosomal subunit!

Table 4: Selected motifs and their defining sequences and suitestrings.

Motif	Sequence	Suitestring	Function
UNCG tetraloop	UNCG	1a:1z:2[:6n:3′	Tertiary structure contacts; protein binding site
GNRA tetraloop	GNRA	1a:1g:1a:1a:1c	Tertiary structure contacts; kissing loops
S-motif	YAGUA YA_AG	1a:5z:4s:#a:1a:3′ 1a:1a:1e:1a:1a	Binding site of elongation factors and other proteins
S2-motif	not conserved	1a:5z:6p:8d:1a	Protein binding site
Kink-turn	GC[loop]GAAC CG_____AGGG	1a:1a:7r:6p:2[:0a:1a:1a Not conserved	Protein binding site; structural scaffold
TΨC	not conserved	1a:1a:1a:1a:1a:1g:1a:1[:4d:1b:2a:1a:1a:1a	Protein binding site; structural scaffold
U1 hairpin II	AUUGCAC	1a:1a:1[:6g:1[:0a:1a	Protein binding site
Dinuclotide platform	GU or AA	2′:#a:3′ or 2′:4g:3′	Stacking interactions; minor groove accessibility

If the motif contains multiple strands, each strand is written on a separate line. The base sequences for the S-motif and Kink-turn have been shifted to show how they align; the second strand is continuous despite the underscore. The kink-turn has been highlighted to reflect the submotifs shown in Figure 17. The 3′ label refers to a poorly conserved suite with a C3′-endo pucker; similarly, 2′.

3.3.2 SuiteAlign

With our initial tests successful, we set about trying to create a new suitestring alignment program that acted more like BLAST and ClustalW; we wanted a program that allowed gaps, and made some determination on which suites were closely related to each other and so might substitute (or be confused with) each other. To this end, the lab brought in Parawee Lekprasert, a graduate student in the Computational Biology and Bioinformatics program. She developed new alignment software using mine as a basis, resulting in SuiteAlign, a program that can be used to align suitestrings of many structures, and account for variations in conformation and gaps in the alignment. A substitution matrix was generated using a scoring factor inversely proportional to the RMSD of each ideal suite to each other ideal suite. The greater the RMSD, the lower the score on the matrix, and the less likely SuiteAlign will choose that particular suite as a substitute for the other in an alignment.

SuiteAlign's ability to look for inexact suitestring matches bore fruit right away. A search for the suitestring associated with the TΨC loop in tRNA uncovered another similar loop with several minor differences (Figure 23). This loop was also in the ribosome and included a very close match to the TΨC loop and, unlike the other ribosomal TΨC loop, extended to largely match the CCA stem as well (RMSD = 3.37Å for loop+stem).

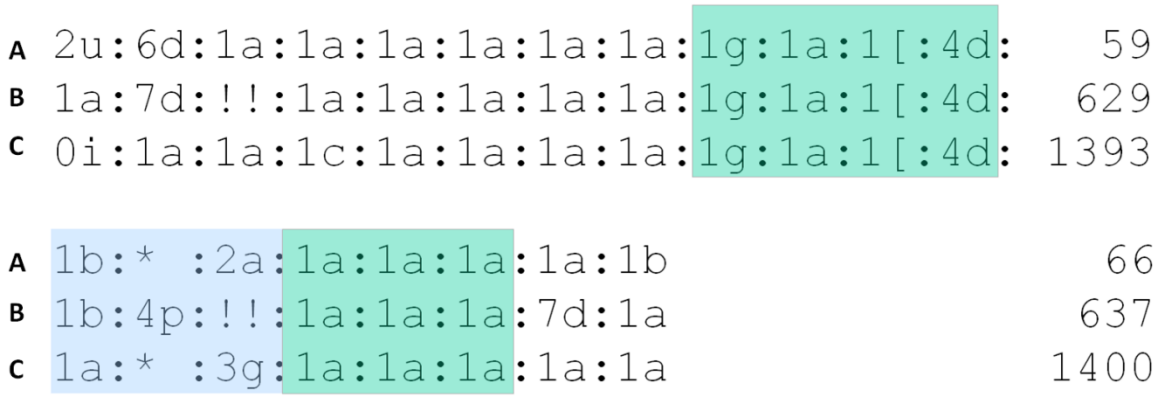


Figure 23: Suitestring Alignments of TΨC loop.

Suitestring A is from the tRNA^{PHE} (PDB: 1EHZ; Shi 2000) and B and C are both from the 3CC2 archaeal 50S ribosomal subunit. SuiteBlast identified the similarity between A and B, while the extra versatility of SuiteAlign allowed it to recognize C as an additional candidate. See Figure 28 for structures.

3.4 Redefining motifs with suitestrings

By redefining known RNA motifs in terms of backbone as well as sequence, we can define RNA in a manner that is consistent in both sequence and structure. The RNA motifs described in the following subsections are presented in Table 4, with both their sequence and suitestring definitions.

3.4.1 Results from backbone redefinitions

Examples for each of the four motif types discussed in section 3.1 were found by using the SCOR database, a text search of the PDB, and a more general literature search on PubMed, thus ensuring that the bulk of the definitive motif examples were analyzed.

Once a set of structures was compiled, the structures were downloaded from the PDB and run through DANGLE to obtain dihedral angles for each suite. This data was then used to assign 2-character conformation names using Suitename. The results for each motif are discussed below.

From a backbone point of view, the kink-turn and S-motifs are the most uniform of our four motif examples. While a particular kink-turn may vary slightly in terms of sequence, the backbone suitestring is consistently **1a:1a:7r:6p:2[:0a:1a:1a**, and generally occurs in the midst of A-form helix (Figure 24). The C-stem, being made up of WC base pairs, is A-form, but the kink is characterized by the **7r** and **6p**, as it changes from C3'-endo pucker to C2'-endo; the bases from **6p** are oriented opposite each other, with the second one sticking prominently away from the structure, allowing it to easily make contact with proteins or intercalate with other RNA structure. The first two residues of the NC-stem are also C2'-endo pucker, resulting in the **2[** and **0a** suites, the latter of which begins the transition back to A-form helix in the NC-stem. The back strand is a little more varied; it is A-form in the C-stem and part of the NC-stem, but the transition between the C-stem and NC-stem results in at least two residues with C2'-endo pucker, which interact with their counterpart **2[** and **0a** suites from the main strand. Both the sequence and the suites for this transition on the back strand are quite variable, implying that it is the kink in the kink-turn that defines the motif, while the back strand simply adapts to fill in the space.

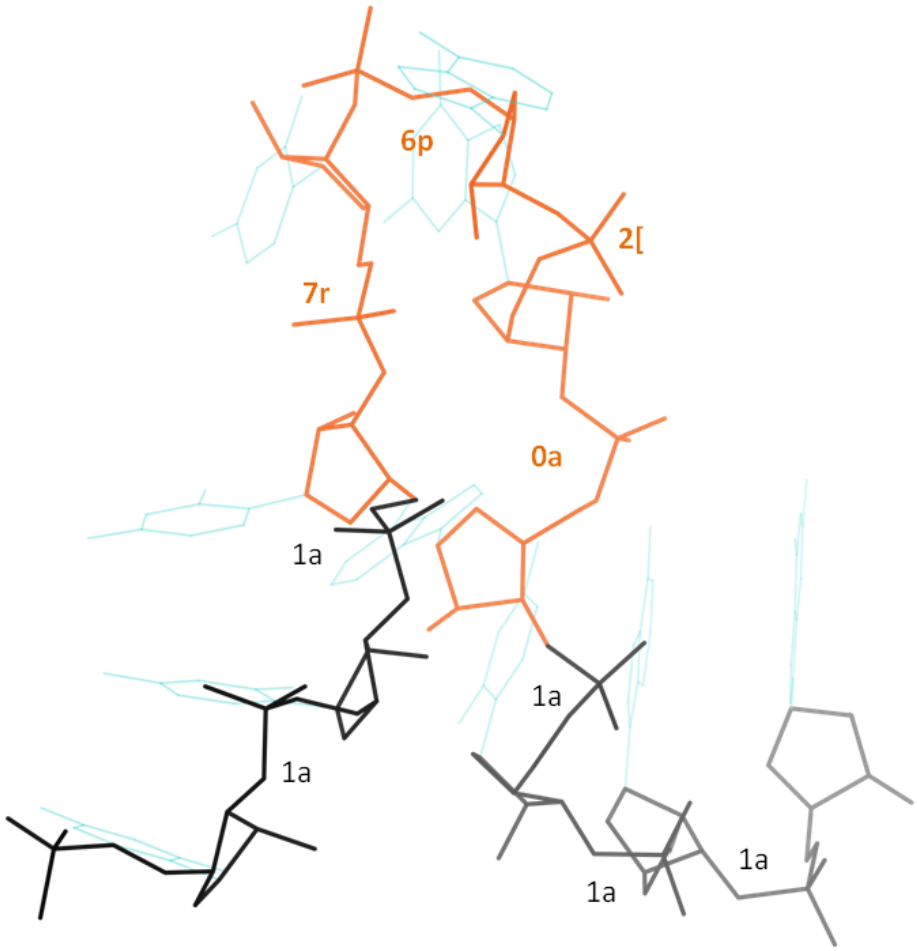


Figure 24: Kink-turn primary strand with suitestring labels.

The primary strand of Kinkturn-42 (residues 1146-1155)from *H. marismortui* (PDB: 3CC2) is shown; with the consensus suitestring in orange and the surrounding A-form sequence (1a suites) in black. Refer to Figure 17 for more detail on the sequence.

The S-motif also occurs as an interruption of A-form helix, with the suitestring **1a5z4s#a1a** on the primary strand. The distinctive S-shape is due to the changes in ζ_{n-1},α,β, and γ dihedral; they undergo an inversion from pttp to tmpt, a rare occasion where β is close to the gauche plus range. The **5z** conformation is stabilized by a characteristic backbone H-bond between the $O2P_{n-1}$ and the $2'OH_n$. The **4s** conformation is quite strained, and only occurs naturally within the context of an S-motif; it forms part of the binding site of EF-II. The #**a** is a GpU dinucleotide platform that sets off the stack switch with the back strand. Unlike in the kink-turn, the back strand disruption of the S-motif is conserved, and follows a **1a1a1e1a1a** suitestring (Figure 25); the **1e**, shown in pink, is the only other suite besides **4s** with a low positive β value, and forms as a response to the #**a** dinucleotide platform. The result is an extended base triple, where the U from the #**a** dinucleotide platform interacts with the Hoogsteen edge of the A from **1e**. The **1e** conformation also facilitates the stack switch between the front and back strands of the S-motif. Because the suitestrings for the S-motif and the kink-turn are so readily identifiable, it is an easy task to find all of their occurrences, and they occur frequently in larger structures, such as the *H. marismortui* 50S ribosomal subunit (Figure 26).

It is also worth noting that the consistency of the S-motif suitestring helped confirm a distinct variant form, known as the S2 motif (Wadley and Pyle 2004). The S2 motif had been identified through the Pyle lab's η/θ analysis, and applying our suite analysis yields a backbone suitestring of **5z6p8d**, in contrast to the standard **5z4s#a**.

Unlike **4s**, the **6p** is largely similar to **5z**, except for a 120° rotation around α. Instead of forming a dinucleotide platform, the base is instead pointed in the exact opposite direction, flipped out into the solvent where it can make tertiary stacking interactions with the rRNA and bind to ribosomal proteins; such an example can be found in residues 893-895 of the *H. marismortui* 23S rRNA (PDB: 3CC2). Thus, there is no base triple and the back strand can have **7a** or **1[** to facilitate the stack switch, though **1e** is still often used.

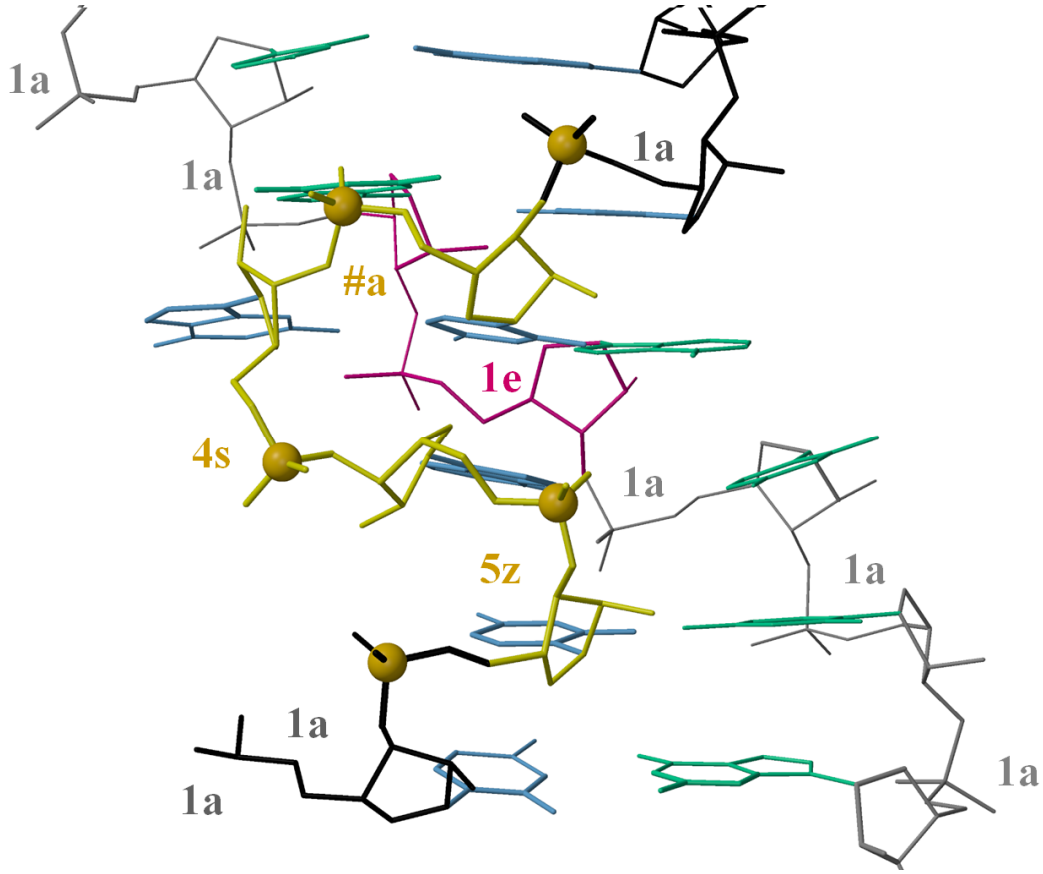


Figure 25: S-motif with suitestring labels.

The front strand (yellow) has a dinucleotide platform at #a, and the back strand shows the 1e (pink) stack switch conformation. The 3D structure is the S-motif from the *H. marismortui* 50S (PDB: 1S72; Klein 2004), residues 585-592 (front strand, black and yellow), and residues 566-572 (back strand, grey and pink).

rr0082 / 1S72 rRNA partial suitestring, with kink-turns & S-motifs

```
__U!!A1aU1aG1aC1aC1aA1aG1aC1aU1aG1aG1aU1aG1bG0aA1aU1aU1aG1aC {1S72:0:  29:C}
1bU!!C2aG1cG1aC1aU1aC1aA1aG1aG1aC1aG1eC1aU1aG7rA6pU2[G0aA1aA {1S72:0:  49:A}
1aG1aG1aA1aC1cG1aU1[G2gC1aC1[A!!A1aG1aC1aU1aG1aC1bG!!A6pU!!A {1S72:0:  69:A}
!!A!!G1aC1aC1aA1aU1aG1aG1aG1aG1bA0iG1aC1aC9aG1aC5zA!!C!!G1aG {1S72:0:  89:G}
1aA1aG1aG1aC1aG7rA6pA2[G0aA1aA1aC1aC1aA1aU1aG1aG1aA1aU1aU3aU {1S72:0: 109:U}
1aC1aC1aG1mA!!A!!U6dG1aA1aG1aA!!A3aU5nC1aU1aC1aU__A1aA!!C3aA {1S72:0: 131:A}
1aA1aU1cU1aG1aC1aU1gU!!C6gG!!C4gG1aC1cA1aA1aU1bG2aA1aG1aG!!A {1S72:0: 151:A}
1LA1aC1aC1aC1aC1cG1aA1aG1eA1aA1aC1aU1aG1aA!!A2aA1aC1bA!!U1aC {1S72:0: 171:C}
1aU1aC5zA4sG#aU1aA1aU1aC1aG1aG1aG&aA1aG1bG!!A1cA1aC1aA1aG!!A {1S72:0: 191:A}
!!A2hA9aA1aC1bG0kC5rA6nA!!U2aG1aU1bG4dA1aU1cG1aU1aC1aG1aU1aU {1S72:0: 211:U}
5zA4sG#aU1aA1aA1aC1aC!!G5rC6nG1aA1cG1aU1aG1eA1aA1aC1aG1aC1aG {1S72:0: 231:G}
1cA1aU1aA1cC1zA!!G1cC1aC1cC1aA1aA1aA1aC1cC1bG4bA6nA1aG1aC1aC {1S72:0: 251:C}
1aC1aU1gC1aA1aC1aG1aG1aG1aC1aA7rA6pU2[G0aU1aG1aG1aG1bU!!C {1S72:0: 271:C}
!!A!!G1aG1aG1aC1aU1aA1cC1aC1aU1aC!!U!!C!!A1aU1aC1aA1aG1aC1aC {1S72:0: 291:C}
1aG1eA1aC1aC1aG1aU1aC1aU1aC1aG1aA1aC1aG7aA1aA1oG!!U!!C1aU1aC {1S72:0: 311:C}
1aU1aU1aG1gG1aA1bA!!C2aA1aG1aA1aG1aC1aG1aU1aG1aA7pU2[A2oC0kA {1S72:0: 331:A}
1aG1aG1aG1aU!!G!!A2zC2[A8dA1aC!!C1aC1aC!!G1aU1aA1aC1aU1aC1aG {1S72:0: 351:G}
1aA1aG1aA1aC1aC5zA!!G#aU1aA1aC1aG1aA1aC1aG1aU1bG4aC1aG1aG1aU {1S72:0: 371:U}
1aA1aG1aU7dG!!C1aC1aA3bG2[A!!G#aU1aA1aG1aC1aG1aG1aG1cG1aG1aU {1S72:0: 391:U}
1aU1gG1aG5nA5pU!!A3dU1aC1aC1aC1aU1aC1aG1aC1aG&aA1aA1bU4bA6pA {1S72:0: 411:A}
```

Figure 26: Kink-turn and S-motif suitestrings in 50S.

Kink-turn (blue) and S-motif suitestrings are highlighted in this segment of Suitename results for the *H. marismortui* 50S ribosomal subunit (PDB: 1S72).

Unlike the S-motif and kink-turn, which have very recognizable sequence and 3D conformations, dinucleotide platforms and tetraloops are more nebulous. Much e-ink has been spilled, for example, on whether a dinucleotide platform qualifies as such if the n to n+1 bases are coplanar but not H-bonded, or whether a 5-residue loop containing a GNRA sequence might qualify as a GNRA tetraloop; such ambivalence is reflected in the

examples contained within the SCOR database. It behooves us, then, to apply our more stringent classification of the backbone to each of these examples, and thus establish a more rigorous definition of these general motifs and their subclasses.

The dinucleotide platforms are somewhat of a special case since only one backbone suite is necessary to span the n to n+1 base pairing needed to establish a platform. Thus it comes as no surprise that 25 of our 54 suite conformers (46%) have one example such that the bases interact to create an n to n+1 H-bond; the problem is that the bases are often skewed relative to each other, and a true dinucleotide platform should have a strong degree of coplanarity. Across over one hundred examples, a significant majority of platforms, defined as being coplanar and having at least one n to n+1 H-bond, (61%) were associated with only two backbone conformations: #**a** and **4g**. Each requires a C2′ endo sugar pucker to be followed by a C3′ endo sugar pucker, as well as trans ζ and β values, which results in the bases being coplanar rather than stacking. Interestingly, the #**a** conformation occurs only 50% more often than the **4g** conformation, even though GpU platforms occur three times more frequently than Adenosine platforms. This indicates that, besides the many GpU platforms that are present as part of S-motifs, and therefore #**a**, independent (non-S-motif) GpU and Adenosine platforms can adopt either conformation. We observed this indeed to be the case, though Adenosine platforms show a preference for the **4g** conformation and non-S-motif GpU for the #**a** (about 2:1 ratio for both cases).

Tetraloops, like dinucleotide platforms, can be created by multiple sequences. The most common sequence corresponds to the GNRA tetraloop; its backbone turns out to be very well-defined, with a standard suitestring of **1a:1g:1a:1a:1c (**Figure 27). The **1g** conformation describes the rotation around α that forms the loop bend of the tetraloop, including its peak, while the subsequent **1a** conformations form stacking interactions that make up the rest of the loop. **1c** contains a compensating α and γ rotation to reincorporate the loop into the A-form helix, with the second base of **1c** forming a WC base pair with the base preceding the G. The **1a** conformations on either end denote a standard A-form helix both preceding and following the GNRA tetraloop, a shared aspect of all tetraloop motifs.

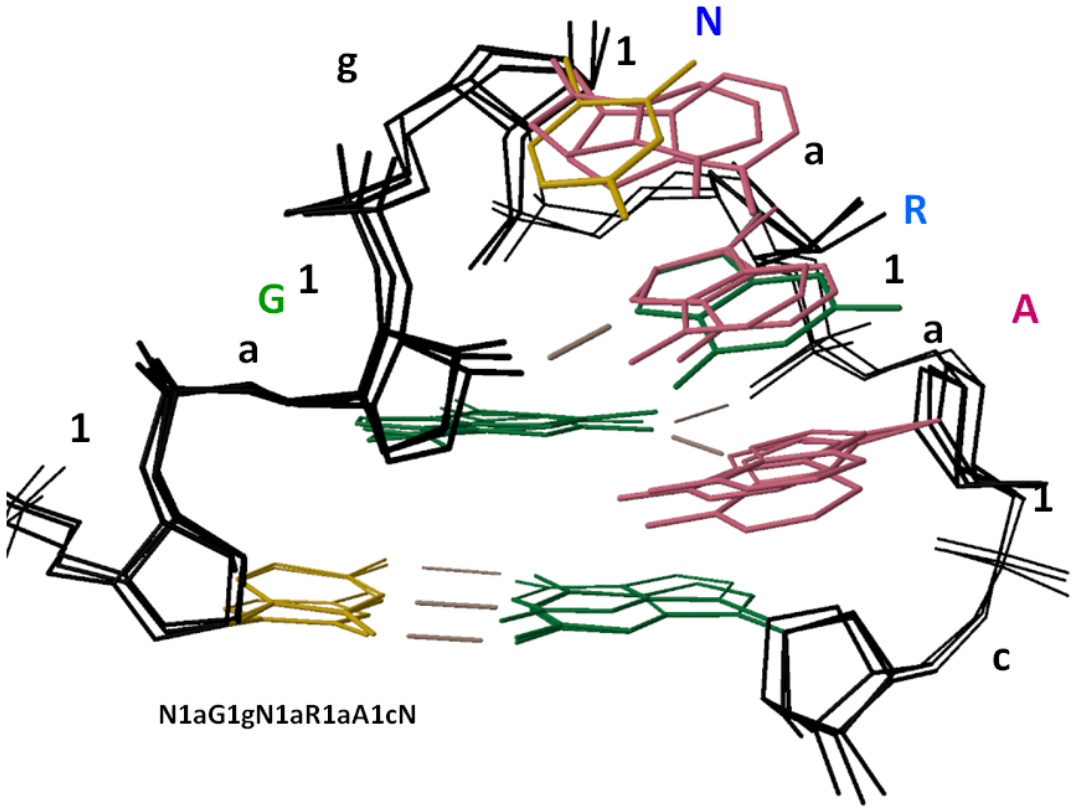


Figure 27: 3 superimposed GNRA tetraloop examples with suitestring labels.

GNRA tetraloops colored by sequence, with suitestring labels at each heminucleotide. Examples are taken from superimposed structures of residues 2658-2663 of 483D (Correll 1999), 153-158 of 1HQ1 (Batey 2001), and 576-581 of 1S72 (Klein 2004).

The second most common tetraloop is the UNCG tetraloop (see Figure 21). Unlike the GNRA tetraloop, whose sugars are all C3'-endo, the UNCG tetraloop contains 2 residues with a C2'-endo sugar pucker. The tetraloop itself is defined by **1aU1zN2[C6nG1a**, a significant departure from the GNRA tetraloop, which is mostly A-form. The **1z** and **2[** conformations allow the N residue to be flipped out into the solvent,

while the C and G are stacked with each other. The **2[** conformation also places its phosphate out into the solvent where it would be readily accessible for binding. The starting residue of **1z** and the second residue of **6n**—the U and the G, respectively—participate in a SE-WC base pair, before the structure returns to standard A-form helix.

Overall, we find that informing the RNA structure motifs with RNA backbone is very useful. It has allowed us to provide a solid backbone definition to the kink-turn and S-motifs, and helped us distinguish the main S-motif from its variant S2 motif; it has also been useful as a tool for improving our ability to identify and correct errors in the model. Our analysis of the tetraloops and dinucleotide platforms goes further by allowing us to establish well-defined submotifs and the backbone conformations that correspond to them. We also have shown that within tetraloops, the local backbone structure is correlated to the sequence that creates the secondary structure; different base sequences may form the same overall secondary structure, but do so using correspondingly different preferred backbone structural motifs, as seen most readily in the GNRA and UNCG tetraloops.

3.4.2 Old friends in new places: rediscovering motifs via suitestrings.

These strong correlations between the backbone and existing motifs was encouraging, but what convinced us of the practicality of suitestrings was our re-discovery of several motifs we were not intentionally seeking. Using a perl script

precursor to SuiteBlast, I searched the suitestrings of every structure found in the RNA05 dataset, enumerating the most common suitestrings. I uncovered two suitestrings not already studied that were common enough to warrant further investigation. These common suitestrings were later discovered via SuiteBlast to be associated with the TΨC loop and the U1A snRNP.

3.4.2.1 TΨC loops in the ribosome

While examining the suitestrings of all RNA05 structures, one of the most conserved suitestrings was also one of the longest. Spanning 14 suites (15 bases), the **1a:1a:1a:1a:1a:1g:1a:1[:4d:1b:2a:1a:1a:1a** suitestring was highly conserved across many structures. We quickly identified this suitestring with the TΨC loop found in tRNA, which explained both its prevalence and length—the TΨC loop is the most structurally stable of the loops in tRNA, and tRNA structures are common in RNA05. We were able to further identify a conserved backbone H-bond between the OP2 of the **1b** and the 2HO' of the first sugar of the **4d** conformation (which is the second sugar of the **1[** conformation). Furthermore, almost every case had a conserved 1-methyl adenenine (1MA) modification on the first residue of **4d**. This well-conserved motif was used to test the SuiteBlast program, with great success—every tRNA structure in the dataset was found. To our surprise, an additional match was also found—not from a tRNA, but from the center of the *H. marismortui* 50S ribosomal subunit!

The ribosome's own version of the tRNA TΨC loop spanned residues 622-635, and even included the characteristic 1MA and backbone H-bond between residues 628 and 630. The modified base is especially interesting, as this particular TΨC loop is 15Å from the closest nucleic acid surface of the 50S subunit, and thus has no interactions with proteins or RNA other than the 23S, and no accessibility to enzymes once the 23S is folded; it is also interesting that the rRNA lacks the characteristic pseudouridine (Ψ) which gives the tRNA TΨC loop its name. The interactions with the surrounding nucleotides also parallel tRNA; A1081 hydrogen bonds with U626 to form an analogue of the characteristic hydrogen bond between tRNA residues 19 and residue 56. Furthermore, the **1[** conformation in the ribosome allows for intercalation of C2071 between G627 and 1MA 628; in tRNA, the same intercalation is accomplished by residue 18 betwen residue 57 and 1MA 58.

Improvements in the code for SuiteAlign led to the discovery of a second TΨC loop in the 23S rRNA (Figure 28).This second TΨC loop is on the surface of the ribosome as part of helix 53, and has several key differences from the other TΨC loops, including the other ribosomal example. While the backbone-backbone hydrogen bond is conserved, the adenine of the **4d** conformation is unmodified. Furthermore, the conserved tertiary interactions are missing; though it does include an intercalation at the **1[** conformation, the intercalating base enters from the opposite side of the loop when compared to the standard motif, and no analogue to the tRNA res19-res57 H-bond

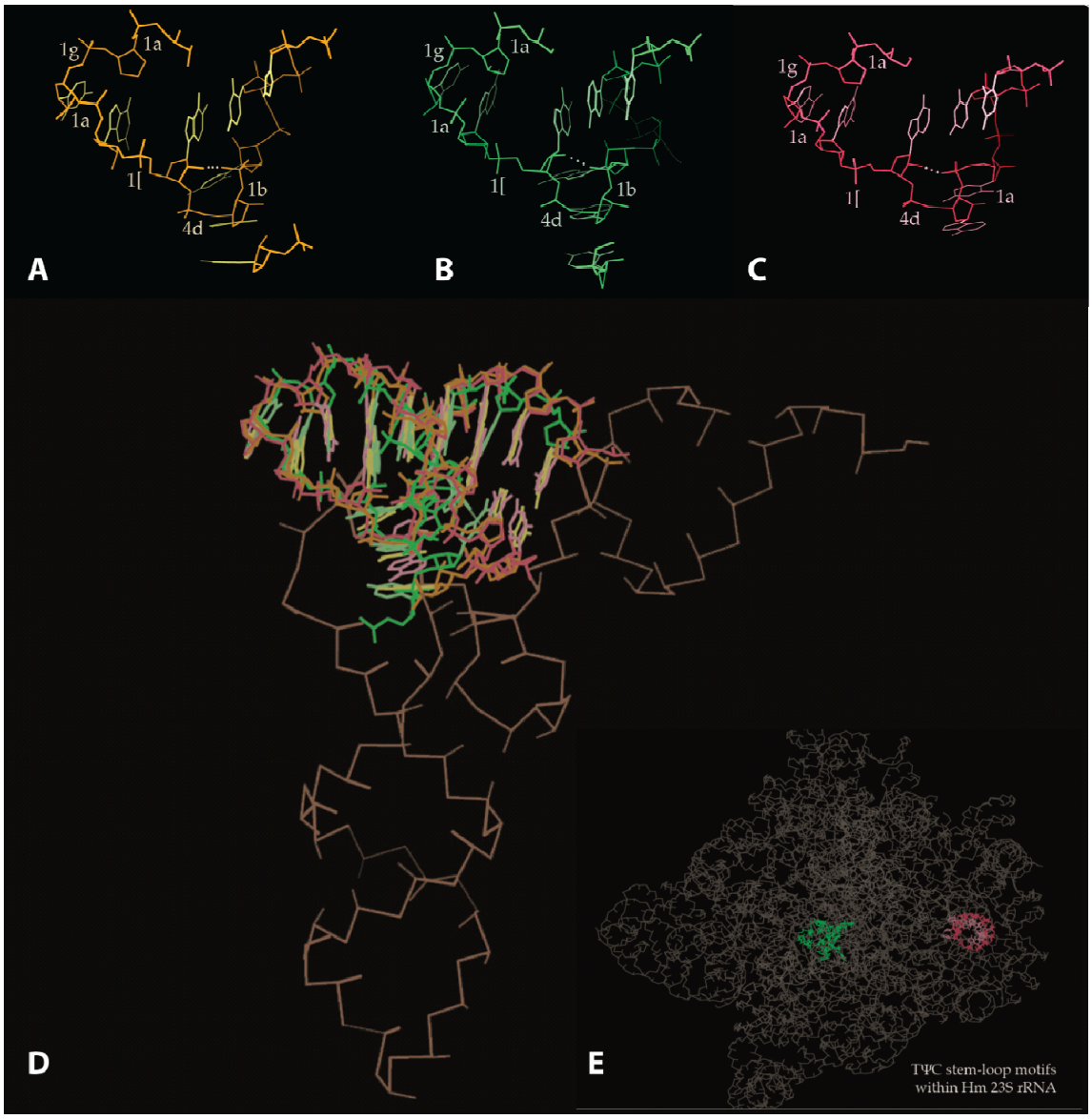


Figure 28: TΨC loops and locations.

The labeled TΨC loop from tRNA^{PHE} (A) is nearly indistinguishable from the TΨC loops from the 23S rRNA (B-C). Superimposed on the tRNA (D), their remarkable similarity emphasizes the accuracy of the suitestring method of identification. Their actual locations in the ribosome are in panel E.

exists. This may be because, unlike the other ribosomal TΨC-loop, this example interacts extensively with helices from L19e and L31e, with the proteins binding opposite each other to the A-form helix preceding the loop itself, and preventing the tertiary RNA-RNA contacts that one finds in the tRNA. It is also worth noting that this second example also contains an extended region of RNA that corresponds closely to the loop region leading into the CCA-stem; the motif can be extended by an additional five residues without diverging significantly from tRNA-like structure.

Both of the ribosomal TΨC loops were discovered through their suitestrings, while the other conserved features, such as the backbone hydrogen bond and base intercalation, were observed as we investigated each model in KiNG. The strong conservation of sequence and hydrogen bond patterns made us wonder if this structure had been described before; it seemed unlikely that such a well-defined motif would have been overlooked by the early motif hunters. Our search led us to two instances, the U-turn motif and the T-loop, both of which describe the reversal of phosphate direction that occurs between residues 55 and 57 of the standard $tRNA^{PHE}$ example. The U-turn is a generalized motif that describes a backbone reversal following a single-stranded UNR sequence (Gutell 2000); a "Type E" U-turn is a U-turn with a flanking Y:R base pair. This definition perfectly describes 5MU54,Ψ55, C56, G57,1MA58 of $tRNA^{PHE}$ and U624,U625, U626, G627,1MA628 in the 23S subunit. However, this sequence-derived description falls apart for U1388, G1389,A1390,G1391,A1392, since the G1389 is not a U or one of its

derivatives. Accordingly, the second TΨC loop in the 23S subunit is not a canonical U-turn.

This brings us to the T-loop, an author-described upgrade of the U-turn to include a more robust definition for finding tRNA-like loops in ribosomal RNA (Nagaswamy and Fox 2002). The T-loop consists of four to five nucleotides, starting at a U or Ψ, and has a UA *trans* WC/HG interaction that stacks with a WC base pair. Furthermore, it must contain a U-turn as well as a 2-3 nucleotide bulge 3′ of the final A. Finally, there must be a stabilizing base-sugar bond between n+1 and n+4 (the "U" and "R" from the U-turn's "UNR" sequence). The T-loop definition succeeds in more completely describing the TΨC loops in 48-66 of $tRNA^{PHE}$ and 618-637 of the 23S subunit, but strictly speaking still rules out the one from 1382-1400 of the 23S, due to the G1389 not making a canonical U-turn. The authors include this third example anyway, since the other aspects of a T-loop are there, and categorize it as a "variant" T-loop, thus opening the door for examples from Ribonuclease P as well as more in the ribosome (Krasilnikov and Mondragon 2003).

What neither of these motif definitions possess is a robust backbone definition that describes the TΨC loop in its entirety. The U-turn found a small portion, the T-loop a larger amount, but the key residues in the bulge are not defined by either in any practical way—the bulged nucleotides are just that, with no description as to what they are actually doing. Yet these bulged nucleotides have a very well defined backbone

structure in our suitestring definition of the full TΨC-loop. Once again, the suitestring for the full loop is as follows: **1a:1a:1a:1a:1a:1g:1a:1[:4d:1b:2a:1a:1a:1a**

The beginning of the loop is A-form, which is not unusual, but the U-turn is accommodated by the $\alpha\gamma$ crankshaft that is characteristic of **1g** (much like the GNRA tetraloop). The turn is completed with another **1a**, which is immediately followed by a **1[** that allows a base intercalation (not mentioned in the U-turn and T-loop definitions) between what the T-loop would call the n+4 and n+5 residues. This base stacks with these two residues as well as forming a WC/SE basepair with the first residue of the U-turn. The **4d:1b** forms a two-nucleotide bulge that eases the transition into canonical WC base-pairing with the residues that began the 5′ end of this motif. This is a departure from other instances of the T-loop, as the bulged nucleotides rarely lead back to WC pairing with the incoming strand, but rather go on to take part in long-range tertiary interactions. The final A-form residues form WC base-pairs with the initial A-form residues, bringing the TΨC loop to a close.

This more strict definition of the TΨC loop provides interesting insights into the development of ribosomal structures. While U-turns and T-loops are found in ribosomes of all observed species, the ribosomal TΨC loop is found in *H. marismortui* and in *S. cerevisiae*, but not in *E. coli*, *T. thermophilus*, or in *D. Radiodurans*. This may indicate that this tRNA-like structure may be unique to archaea and eukaryotes, while bacteria retain at best a truncated portion of the motif.

3.4.2.2 U1 snRNP shows up everywhere

Another common suitestring identified in RNA05 was **1a:1[:6g:1[:0a:1a**. We found this suitestring to be associated with U1 snRNA hairpin II, specifically the region that is bound by the U1A spliceosomal protein (Oubridge 1994). Unlike the tRNA TΨC loop, which appeared in unexpected places naturally, the reason for our abundance of U1 hairpin II motifs is due to the U1 snRNP's use as a crystallization agent; many ribozymes and other independent RNA structures are difficult to crystallize, but engineering a binding site for U1A allows crystallization of the RNA-protein complex with minimal disruption to the main RNA structure (Figure 29). Thus, we see the U1A:U1 hairpin II complex appear in structures ranging from the actual U1 snRNP to the c-di-GMP-I riboswitch (Smith 2011).

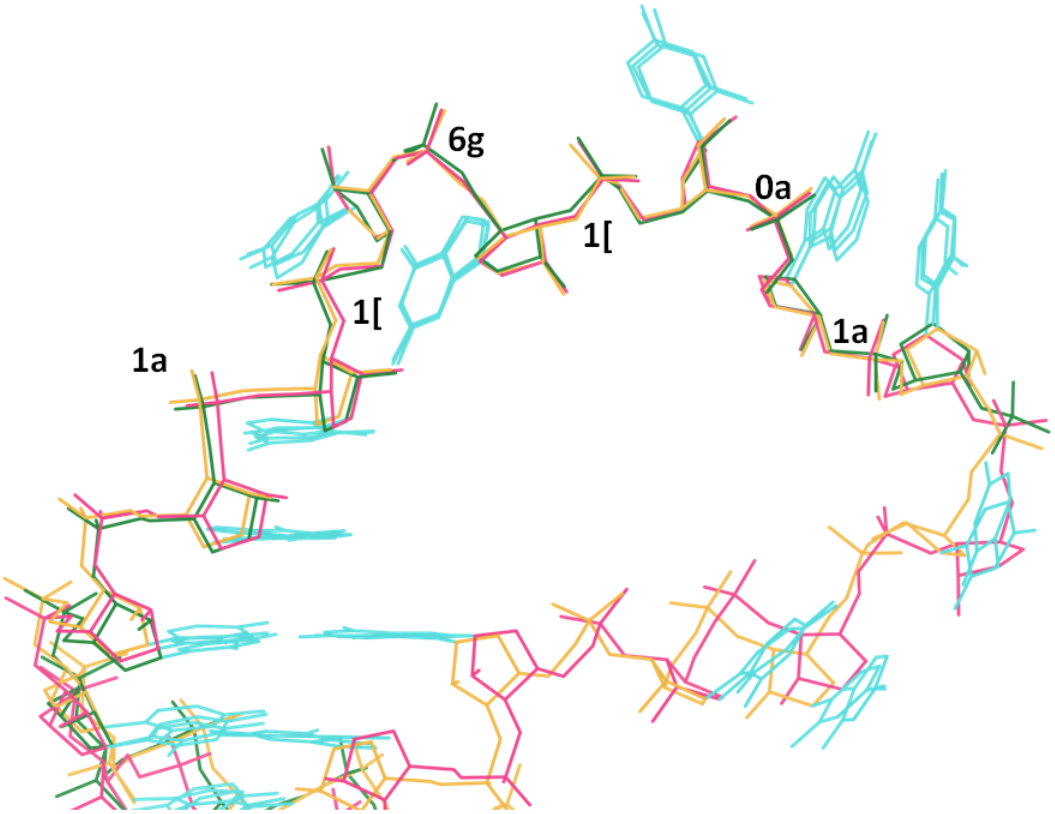


Figure 29: U1 hairpin II superimposed.

U1 hairpins are from a U1A to U1 hp II complex (green), a U1hpII engineered into the hairpin ribozyme (pink) and the HDV ribozyme (yellow). Structures are from PDB IDs 1URN (Oubridge 1994), 1M5O (Rupert 2002), and 1CX0 (Ferre-D'Amare 1998) respectively.

While we know that the complex is present in many different structures, the consistency of the hairpin's suitestring demonstrates that across all of these different environments, the RNA backbone is binding to the U1A protein by adopting the exact same conformations. Of the ten unpaired nucleotides in the U1 hairpin II sequence, the

first seven, AUUGCAC, bind to the protein's RNA-binding domains RNP1 and RNP2 (Bandziulis 1989), as well as the C-terminal domain (CTD) of U1A (Oubridge 1994). The first three suites, **1a:1[:6g**, interact with the protein via base-sidechain hydrogen bonding to the CTD (Figure 30). The next two suites, **1[:0a**, are more interesting. The first residue of **1[**, G, is closely packed with Q54 of the RNP1 domain while also forming a 2'OH backbone hydrogen bond with MSE 51. Meanwhile, the second residue, C, stacks on Y13 from RNP2 and makes a backbone H-bond between its 2'OH and K88 of the CTD. Finally, the second residue of **0a**, A, stacks with F56 from RNP1. The direct interactions of the RNA backbone coupled with the stacking interactions makes this interaction motif particularly stable, and thus easily identified by suitestring. It is interesting that one of the earliest high-resolution structures had a pucker error that resulted in a suitestring of **1a:1[:6g:1m:!!**, the **!!** indicating a non-recognized conformer. A simple rotation of the **!!** conformer around the α and γ dihedrals leads to a corrected **1[:0a** conformation, adhering to the observed structural motif; the details of this correction will be discussed further in chapters 5, but the fact remains that the U1 snRNA hairpin II suitestring is conserved enough to be used to aid model validation and correction.

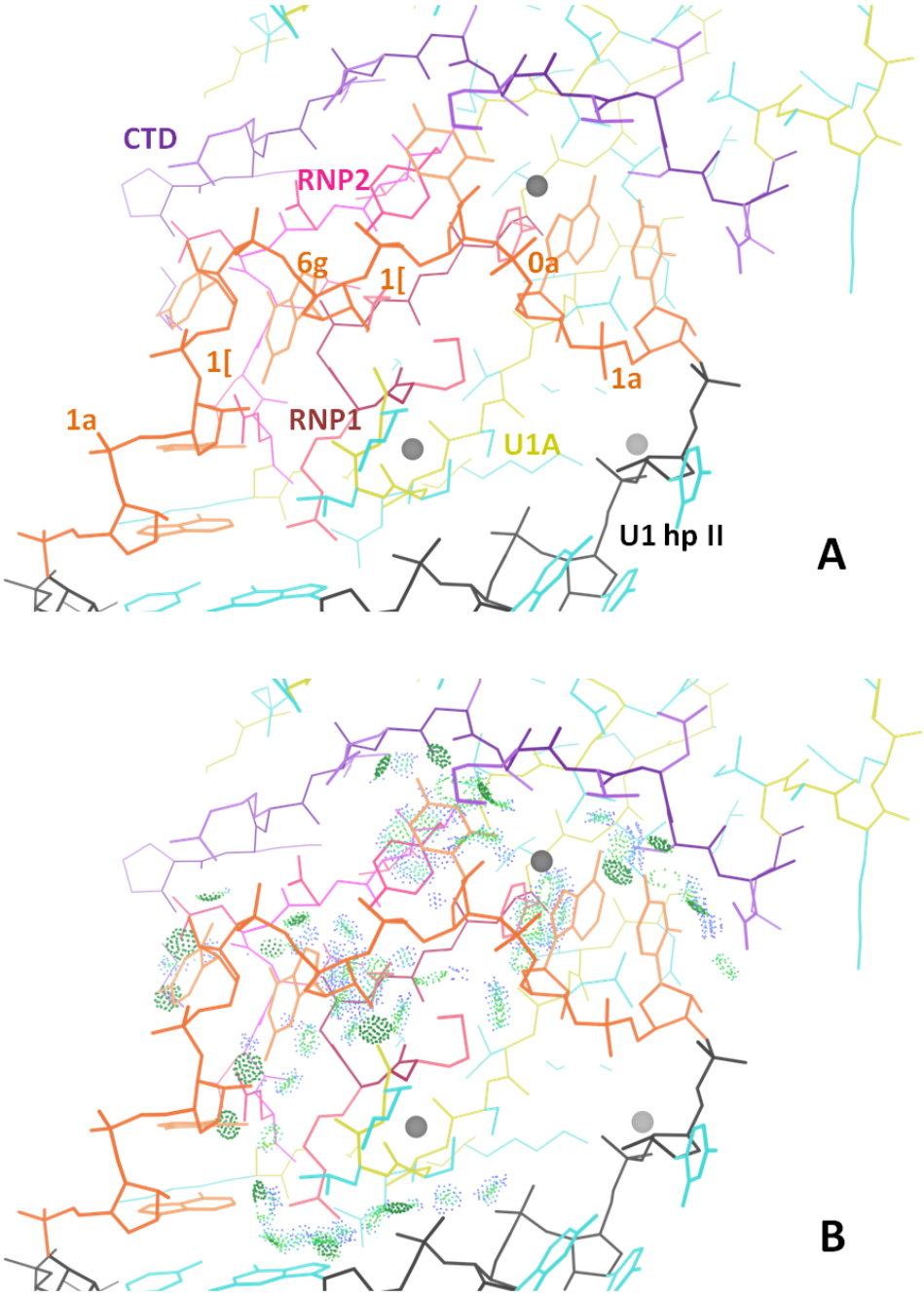


Figure 30: U1A binding to U1 hairpin II.

Panel A shows the RNP1, RNP2, and CTD regions of U1A, while Panel B shows the contacts between the RNA and protein.

3.5 The OHO pentaloop—a novel backbone motif

Redefining motifs in terms of suitestrings led to the rediscovery of many motifs in unusual places, but the true test of the suitestring method would be the discovery of a new motif which has never heretofore been described. Using SuiteBlast to analyze all suitestrings 3-suites long in RNA-protein interactions from RNA_Prot2011 (see section 4.1.2), I discovered four instances of the suitestring **4b6p2a**, which, upon inspection via KiNG, described a pentaloop with conserved suitestring **1b4b6p2a**. This pentaloop has a highly conserved backbone structure and a conserved H-bond between the H of the 2'OH of residue n and the O of the 2'OH of residue n+4, earning it the name of the OHO pentaloop (Figure 31). A second conserved H-bond also occurs at the end of the pentaloop, but appears to have two alternatives: it is either between the $2'OH_n$ and $base_{n+4}$ or it may be between $base_n$ and $2'OH_{n+4}$. Stacking between pentaloop bases can stabilize the 3D structure, but varies greatly between instances, as there is little sequence conservation.

By running this same procedure on the RNA11 dataset (section 4.1.1), I was able to expand the initial four instances of this pentaloop to eight, and all but two instances contained the **1b4b6p2a** suitestrings; one had a **1[** in place of the **1b**, which is fine, as the two clusters are very similar (see Figure 11, section 2.3.2). The other example, however, began with a **!!** outlier conformation. To address this error, I used ERRASER (section 4.5.3, where an OHO is illustrated) to correct the offending suite, which resulted in the **!!**

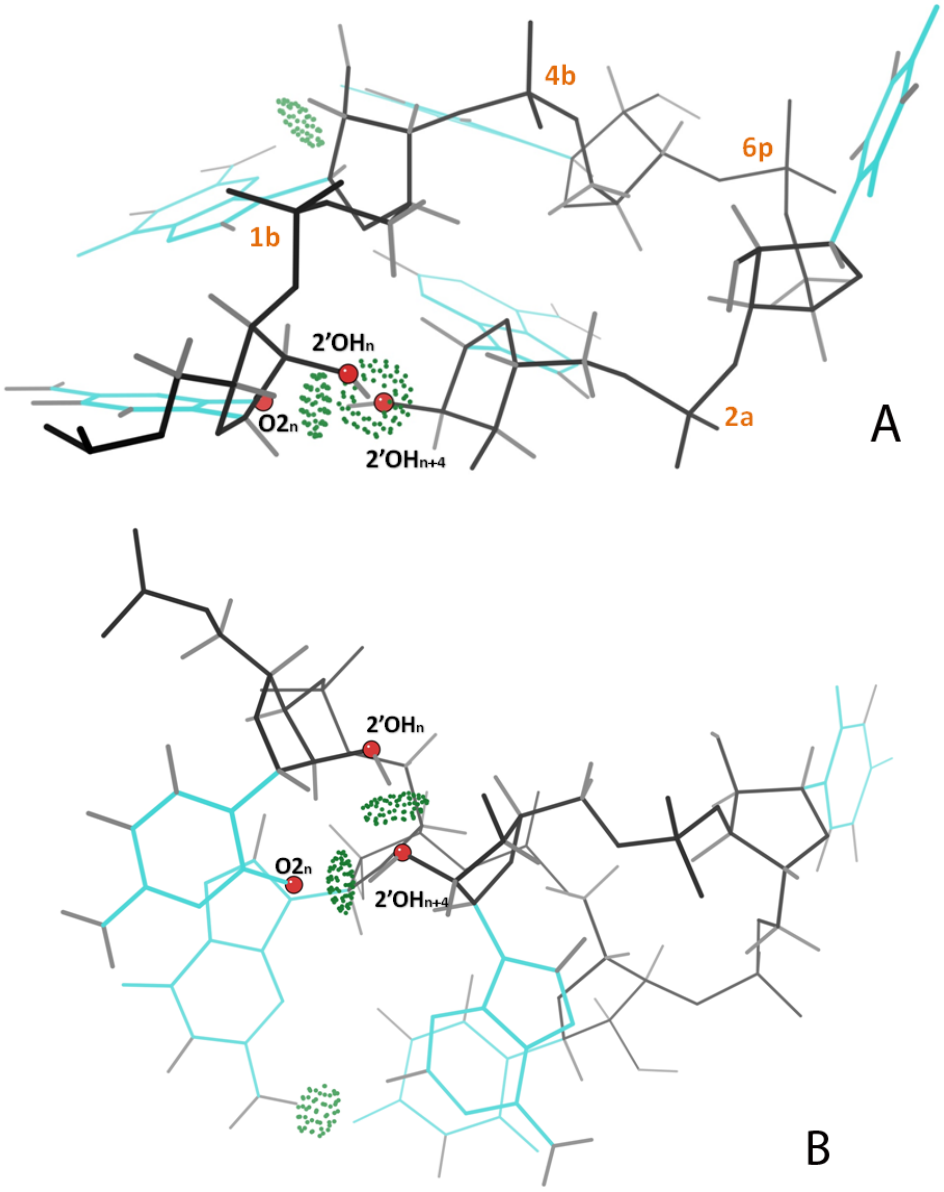


Figure 31: OHO pentaloop overview

Panel (A) shows the best overall view of the backbone of the OHO pentaloop from the GlmS ribozyme, residues 89-93 (PDB 2Z75, Klein 2007). The suitestring is in orange, and the H-bonds in dark green dots. The atoms contributing to the primary H-bonds are also labeled. Panel (B) shows the same pentaloop from the best view to clearly show two primary H-bonds that close the loop.

being corrected to the expected **1b**. Spurred on by this success, I decided to search for candidate OHO pentaloops that may contain outlier suite conformations that could be corrected via ERRASER. I used AMIGOS II (Wadley 2007) from the Pyle lab to do a pseudotorsion search on the RNA_Prot2011 and RNA11 datasets; the pseudotorsions were defined based on the OHO motif formed by residues 34-37 from an A-riboswitch crystal structure (PDB: 1Y26, Serganov 2004). The seven new candidate motifs found by AMIGOS II each had at least two **!!** outlier conformations, which is why they did not show up in our initial search. Each of these candidate OHO motifs was run through ERRASER by separating the motif and the 10 adjacent upstream residues and 10 adjacent downstream residues from the rest of the structure. This new, shorter model was run through ERRASER rather than the original structure—this was particularly necessary for the OHO motif candidates found in the ribosome. ERRASER was able to find good candidates for 9 of the 17 **!!** conformations. After these corrections, five of the seven candidates fit the overall OHO motif shape (2.67Å RMSD or better) and had similar (though not exact) suitestrings.

Altogether, we found 13 high-quality instances of the OHO motif in six different types of structures (Figure 32). In the A-riboswitch, G-riboswitch, and luteoviral pseudoknot, the OHO pentaloop acts as a normal loop ending a stem. In the GlmS ribozyme, the OHO pentaloop acts as a helix junction, interrupting an A-form helix to facilitate the strand's transfer to a second helix. The *H. marismortui* and *E. coli* ribosomes

each have several OHO motifs, some as stem-loops and some as junctions. Yet despite these diverse functions, the OHO backbone structures are remarkably similar.

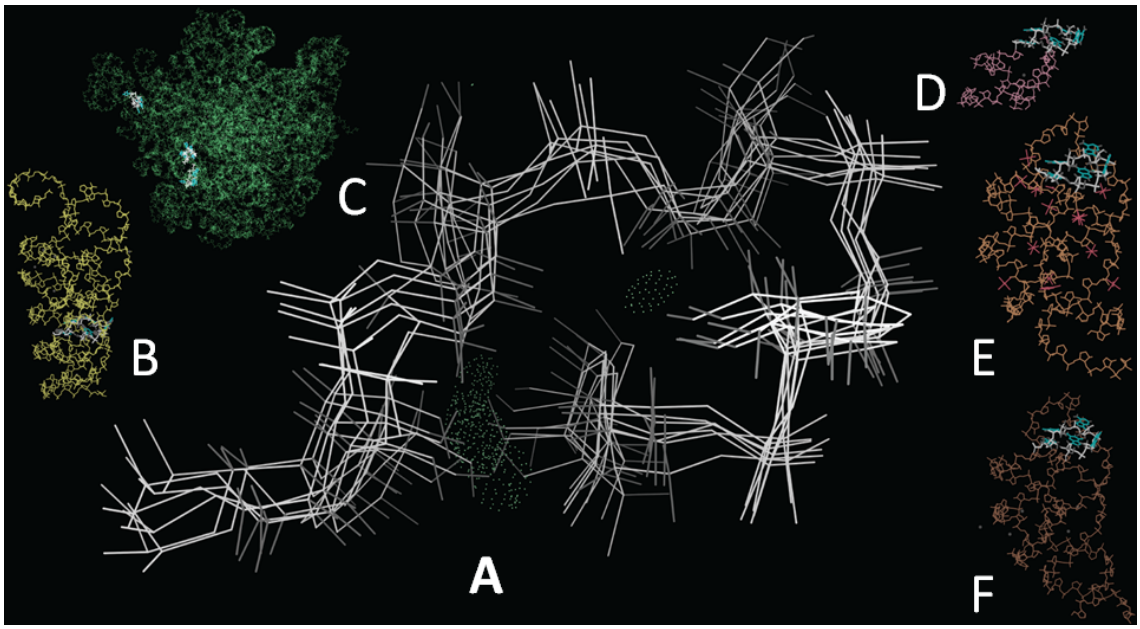


Figure 32: OHO pentaloop structures

Panel A (center) shows a backbone view of the OHO pentaloops superimposed on each other. Clockwise starting with B, we show the pentaloop from the GlmS ribozyme (PDB: 2Z75, Klein 2007), three OHO pentaloops in the *E. coli* 50S ribosomal subunit in C (PDB: 3R8T, Dunkle 2011), then D-F show pentaloops from the luteoviral pseudoknot (PDB: 2A43, Pallan 2005), G-riboswitch (PDB: 3GER, Gilbert 2009), and A-riboswitch (PDB: 1Y26, Serganov 2004), respectively.

It is especially interesting then that despite their structural similarity, these pentaloops come in a great variety of base sequences (Table 5). Out of 13 examples of the motif, one sequence occurs 3 times (AUAUG), and one occurs twice CGAUA. The other 8 examples each have a different base sequence. None of the positions are well conserved, save for the last position, which is almost always a purine. The only other commonality in the sequence is that the n and n+4 residues that make up the beginning and closing portions of the pentaloop are never potential Watson-Crick pairs, perhaps to ensure that the closing pair is facilitated by backbone H-bonds and won't be disrupted by n to n+4 base-pair interactions.

Table 5: OHO loop sequences

PDB ID	Residue range	Suitestring	Base Sequence	Description
3CC2	408-412	1b4b6p2a	AUAAC	*H. marismortui* 50S
1Y26	33-37	1b4b6p2a	AUAUG	A-riboswitch
3GER	33-37	1b4b6p2a	AUAUG	G-riboswitch
3R8T	402-406	7r4b6p2a	AUAUG	*E. coli* 50S
3R8T	445-449	1b4b6p2a	CGAUA	*E. coli* 50S
3R8T	69-73	1b!!6p2a	CGAUA	*E. coli* 50S
2Z75	89-93	1b4b6p2a	CGUUA	GlmS ribozyme
3R8T	1140-1144	1[!!4p2a	CUAAA	*E. coli* 50S
2A43	16-20	1b4b6p2a	CUCAA	Luteoviral pseudoknot
3CC2	1873-1877	1[4b6p2a	GUACG	*H. marismortui* 50S
3R8T	1817-1821	1b4b6p2a	GUAUA	*E. coli* 50S
3R8T	1798-1802	1[!!!2a	UGCAA	*E. coli* 50S
3R8T	773-777	1[!!!6n	UGGGG	*E. coli* 50S

3.6 Discussion and Conclusions

As is evident from the examples in this chapter, the suitestring analysis provides a facile method for identification of RNA tertiary structure, and, when combined with primary sequence, also gives a robust method for motif identification. Considering that the number of RNA structures has increased by an order of magnitude in the past decade, with no sign of slowing down, our easy way to indentify motifs promises to be very useful as a way to guide model building and identify future research targets.

One drawback of using suitestrings is that they are rather sensitive to errors of backbone modeling that produces **!!** outlier suites or incorrect suites. Our lab's program of improving RNA crystal structure accuracy helps address this. A second drawback is that they cannot be done directly from base sequence but must rely on a known structure in order to identify the suitestring. Several approaches are in development to address this issue. The first uses RNA prediction methods from Rosetta and related programs to computationally determine the RNA tertiary structure from a given sequence and follow up with suite assignment. The results are then compared to the suitestrings of known structures. Unfortunately, at this stage the experiment is as much a test of the robustness of Rosetta's computational methods as it is a test of suitestrings' utility in structure prediction. Until reliable RNA structure can be predicted, suitestrings as a method of motif identification will be used less for prediction and more as an aid to structure building, correction, and classification.

A second, related ongoing project seeks to relate the suitestrings to experimental data generated by SHAPE analysis. SHAPE (Selective 2′-Hydroxyl Acylation analyzed by Primer Extension) is a method developed in Kevin Weeks' lab for studying structural flexibility in difficult-to-determine RNA structures (Wilkinson 2006). Because the SHAPE reagents react with the 2′-hydroxyl at different rates, the speed and amount of reactivity between a SHAPE reagent and an RNA residue gives an indication of how flexible that residue is in relation to the rest of its structure. Recently, it was shown that time-differential SHAPE analysis using multiple reagents that act on different time-scales can be used to infer whether a particular residue has C2′-endo sugar pucker or is part of a single-nucleotide bulge. I have been working with Kady-Ann Steen-Burrell and Jennifer McGinnis-Merkel of the Weeks lab to correlate this information with suitestring data with the goal of tying suitestrings to SHAPE data, thus allowing improved motif and tertiary structure prediction before a solved structure is available.

4. Validation and Correction of RNA Structures

RNA structures have traditionally suffered greatly from a lack of validation methods. Forced to rely heavily on parameters tuned for the structurally quite different DNA (Parkinson 1996), RNA crystallographers relied on their individual skill to correctly build the RNA backbone—the main point of difference between the two nucleic acids. In the Richardson lab, we have pioneered many methods for improving RNA structure building and validation, from automatic backbone correction to introducing pucker-specific parameters in the PHENIX refinement suite to developing some of the initial validation tools for NMR structures. This chapter details the work we have done to make improved RNA structure a reality, by developing structural parameters and implementing them as part of publicly available software packages.

4.1 Datasets

The foremost problem plaguing RNA structural bioinformaticists, including parameter choice, is the extremely limited amount of data available. There are only 2522 RNA-containing structures in the PDB at this stage, a pittance compared to the 86600 structures containing proteins. A little over half of the RNA structures also contain protein, but only three hundred of these structures are at a resolution better than 3Å. The difficulty in working with such sparse data has prompted us to painstakingly create curated, non-redundant datasets of RNA-containing structures filtered by resolution,

clashes, and B-factor. These filters ensure that we are using only the highest quality data for our analyses.

4.1.1 RNA09/RNA11

As mentioned in chapter 2, work on RNA began with a dataset called RNA03, developed by Laura Murray, which contained the filtered, best resolution structures up to 2003. RNA03 was superseded by RNA05, which was used to establish the consensus modular nomenclature, i.e., number-letter heminucleotide method of describing the suites and suitestrings. Structures included in the RNA05 dataset had to match the following criteria: resolution ≤ 3Å at the file level, and B-factor < 60 with no steric clashes in the backbone at the residue level. As new structures were deposited in the PDB, and with our lab excited to search for new suitestring motifs and better data, we soon began another update, this time to RNA09. Swati Jain and Laura Murray worked closely together to find new and improved structures with interesting RNA backbone. The RNA09 dataset contained a greater variety of RNA samples, but the structure list no longer had limitations on B-factor or internal clashes; in practice, these criteria were applied as a second layer of filtering to each residue, rather than determining whether a structure was included in the dataset or not. It was also important to control for redundancy; a representative PDB file for a given RNA molecule was chosen based on resolution and MolProbity statistics. If a given PDB file contained multiple copies of the

model in the asymmetric unit, only the best copy was retained, unless the other copies had significantly different conformations.

The RNA09 dataset contained 287 structures vs. 171 in RNA05. Of these, 101 stayed the same, but many structures from RNA05 were dropped in favor of higher resolution or more complete examples. In a shift of focus, RNA09 dropped many of the simple duplex RNA structures in favor of duplex RNA with mismatches and wobble pairs, or with different drugs bound to them. Individual rRNA pieces (e.g., 16S rRNA) were also discarded in favor of whole ribosomal subunits and full 70S structures. Notable additions to RNA09 include the first sub 3.2Å *E. coli* 70S, the full Group I intron, and the Puf protein/RNA complexes.

Even more recently, we have developed a newer dataset version, RNA11; drawing from the NDB and PDB up through 11/11/2011, this dataset contains 311 structures, while still following the criteria from RNA09. New additions include more aptamers and riboswitches, including the recent c-di-GMP-II riboswitch (Smith 2011), the first solved CRISPR structures (Haurwitz 2010), and a new *E. coli* 70S at 3Å resolution.

4.1.2 RNA_Prot2011

While a dataset of RNA structures was a good starting point, not all of these structures were suitable for a study of RNA-protein complexes. Furthermore, many of

the most interesting RNA-protein complexes were of such poor resolution that they did not meet the criteria for RNA09. Working with Ben Lewis from Drena Dobbs' lab, I reconciled our dataset of RNA with their set of RNA-protein complexes used to determine protein motifs. This hybrid dataset contained many results which were of lower quality than the structures admitted to RNA09, but the wider variety of RNA-protein interfaces were considered essential. As we updated the RNA09 to the RNA11 dataset, I also revised the initial RNA-protein dataset into RNA_Prot2011 with the addition of more recent structures, and better resolution copies of structures that were already present. The resulting dataset of 201 structures contained over 33000 suites, with more than 11000 of them occurring along an RNA-protein interface. Of these, only 21 files are not in RNA09 or RNA11, with 12 between 3.0Å and 3.5Å resolution, hailing from difficult to crystallize RNA-protein complexes, like tRNA-MnmA (Numata 2006) and 16S rRNA-IF1 (Carter 2001). The other 9 are better than 3Å resolution, but represent common RNA structures bound to different proteins, producing different RNA-protein interfaces; they would have been cut from RNA11 for redundancy due to similar RNA sequence, but for RNA_Prot2011 they were considered necessary as representatives of distinct RNA-protein interfaces.

4.2 Error Diagnosis

Armed with the best models and the best data for RNA and RNA-protein complex structures, we could begin to develop methods for improving RNA structures. The first step on this path is to identify commonplace errors in the models, particularly in the RNA backbone, and use that to inform the construction of new models. But when we began, there were precious few resources for evaluating RNA models, and none that dealt specifically with RNA backbone. The discovery that RNA backbone could be clustered into rotamers (Murray 2003), while we could not see the same for DNA, impressed on us that the current, generalized methods of nucleic acid evaluation would be at best inadequate, and at worst, misleading. Thus, we embarked on a quest to develop robust methods to diagnose errors in RNA backbone and make them available to the greater RNA community. To do so, we used the all-atom contact analysis developed in the Richardsons' lab to evaluate steric clashes, and introduced three new metrics to determine sugar pucker, bond and angle geometry, and backbone conformation, incorporating each metric into MolProbity. The sugar pucker and backbone conformation metrics were first based on the RNA05 dataset, and the geometry parameters were based on analysis of small, high-resolution structures (section 4.2.3); we have continued to improve these methods over the years as our datasets and our understanding improved.

4.2.1 All atom contact analysis

As mentioned in detail in Chapter 1, the all-atom contact analysis developed by the Richardson lab can identify vdW interactions, H-bonds, and steric overlaps by running a spherical probe of .25Å radius over the vdW shell of an atom and drawing a dot whenever it also contacts a different vdW sphere (excluding adjacently connected atoms). If the vdW spheres overlap and are not a H-bond donor-acceptor pair, then the probe will draw spikes growing from yellow to red to hot pink. At overlaps of .4Å, the hot pink spikes represent impossible steric overlaps where the model cannot possibly be correct; most of these model errors accrue in regions of high flexibility and low signal. Unfortunately, the RNA backbone between sugars, the region encompassed by the suite (see Figure 9, section 2.2), is particularly prone to misfit steric overlaps because the backbone is flexible and has little scattering power save for the phosphate (in crystallography), and little chemical shift signal save the C5' hydrogens (in NMR). When generating our ideal datasets and suite definitions, we removed suites with internal clashes from consideration, allowing us to use suite conformer validity as a way of correcting the RNA backbone structures that have these problems. The all-atom contact analysis is available online via the MolProbity webservice, and generates an entire-structure clashscore representing the average number of clashes per thousand atoms. Each individual residue is listed in a sortable multi-criterion chart along with the most extreme clash within that residue, making it easy to find regions of sequence that have

large steric clashes within suites. Furthermore, the clash markup is generated for each model for display in a multi-criterion kinemage, so it is easy to find steric errors in RNA backbone by looking at the 3D structure in KiNG or Mage.

4.2.2 Sugar Pucker and Base-Phosphate Perpendiculars

Beyond steric overlaps, a second evaluation for RNA structure is correct sugar pucker. Analysis both of the Cambridge Structural Database (CSD) for small molecules and of the high resolution, B-factor filtered residues in our RNA structure dataset reveals that RNA sugar pucker is highly two-state: the ribose has a C3′-endo pucker or a C2′-endo pucker. Alternative puckers which occur regularly in DNA, like C4′-exo, are virtually nonexistent in RNA except for seemingly strained conformations and active sites. The sugar pucker is correlated with the δ dihedral, with a δ of 55°-110° indicating a C3′-endo pucker and δ of 120°-175° corresponding to a C2′-endo pucker. Any sugar modeled outside these δ ranges is highly suspect, as the C2′-endo and C3′-endo pucker regions cluster very tightly, and non-standard sugar puckers are also often accompanied by other error indicators, like steric clashes.

However, at low or medium resolution it is quite possible to fit a C3′-endo sugar pucker in a residue that should be C2′-endo, causing steric clashes and resulting in ϵ outliers to compensate. As a way to distinguish such cases of mistaken identity, Jane Richardson noticed a correlation between the 3′ phosphorus and the plane of the base: if

the sugar was C2′-endo, the phosphorus was close to the base plane, while if it was C3′-endo, the phosphorus was further away. Furthermore, any C3′-endo sugar with a 3′ phosphorus close to the base plane also had steric clashes and poor bond length and bond angle geometry, as well as nonstandard ϵ dihedral values. This indicated a correlation between 3′ phosphorus distance from the base plane and correct sugar pucker (Figure 33). We evaluated this correlation by measuring the perpendicular distance from the 3′ phosphorus to the base plane, and plotting it against the δ dihedral. Upon filtering for clashes, geometry, high-resolution structure and B-factors < 60, two clusters remained, corresponding to the C3′-endo pucker and C2′-endo pucker; a perpendicular distance of ≥2.9Å indicated C3′-endo, while a distance <2.9 Å indicated C2′-endo (Figure34).

This metric worked well initially, until it was discovered that a base with a *syn* or high-*anti* χ dihedral gave artificially low phosphate-perpendicular (P-perp) to base plane values, causing misassigned sugar puckers. To avoid this, we began using the perpendicular distance between the 3′ phosphorus and the glycosidic bond vector, thus eliminating the influence of χ on the perpendicular distance. Without filtering, this results in 97.9% of C3′-endo sugar puckers with a P-perp to line distance of 2.9Å and within the δ range of 55°-110, and 99.5% of C2′-endo sugar puckers falling below 2.9Å and within the δ range 120°-175°. With filtering, these numbers go up to 99.8% and 99.99%, respectively. We also noticed that 81.4% of the unfiltered C3′-endo outliers are

also ϵ outliers (<155°). As a low ϵ orients the 2′ hydroxyl and phosphate to promote nucleophilic attack and cleavage, this may explain why such conformations are strongly disfavored but also why a few of these outliers persist and may be real.

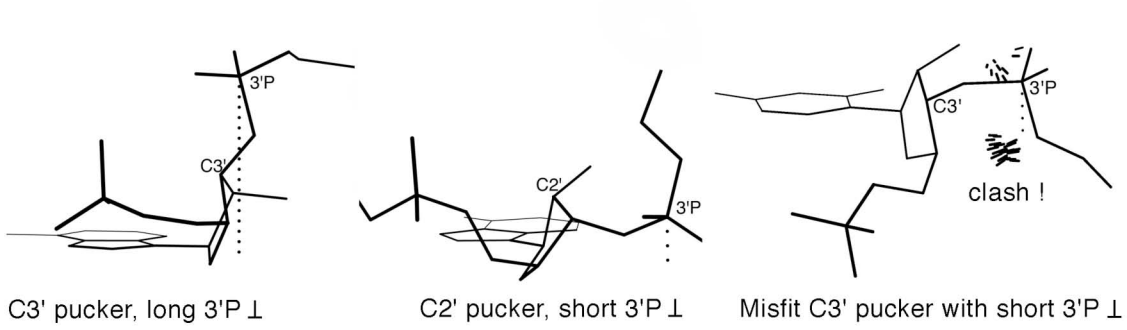


Figure 33: Ribose sugar pucker vs. 3′ phosphorus perpendiculars.

On the left is a long perpendicular indicating C3′-endo, center is C2′-endo with a short perpendicular. On the right, the short perpendicular indicates a C2′-endo pucker, but it is fit as C3′-endo, resulting in a steric clash.

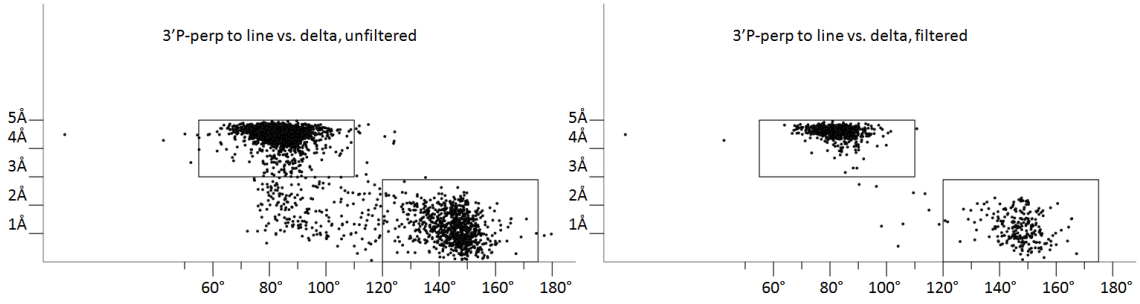


Figure 34: 3′ phosphorus perpendicular length vs. δ .

On the left is the unfiltered data from RNA05. When filtered for clashes, geometry, and B-factor, almost all the outliers disappear (right). The boxes represent the accepted δ ranges and P-perp to line distances for standard C3′-endo and C2′-endo pucker.

The P-perp to line calculation was introduced into the program PREKIN and is now part of the standard evaluation of RNA backbone by MolProbity. It is also now utilized in PHENIX refinement to allow implementation of pucker-specific target parameters (discussed in section 4.3).The multi-criterion chart has a column for P-perp to line outliers, and will also specify whether the ϵ dihedral is also an outlier. Visually, the markup in KiNG is represented by a magenta or dark purple cross for C2′-endo and C3′-endo outliers, respectively. In general, we find that the most common pucker errors are C2′-endo sugars misfit with C3′-endo pucker, due to the overwhelming dominance of C3′-endo sugars in RNA structure.

4.2.3 RNA backbone geometry

The nucleic acid bond length and bond angle parameters from Parkinson *et al.* provided the first comprehensive set of structure-derived parameters for RNA. Unfortunately, the number of nucleic acid structures at the time were very limited, and most of the solved crystals were of DNA, further limiting the available data. To obtain our set of parameters, Lizbeth Videau and I looked at the bond length and bond angles of 459 ultra-high resolution ribose structures in the Cambridge Structural Database (CSD), with R-factors < .1. These were further filtered to 399 by eliminating structures that had noticeably strained conformation due to interactions with ligands. Our initial update resulted in bond lengths that were by and large similar to Parkinson's, though

we included a standard C1′-N1 that they did not, and our mean C5′-O5′ bond length differed by .013Å, an order of magnitude greater change than for any of the other bonds (see Table 6). Due to our expanded number of structures, the standard deviation on each measure increased relative to the Parkinson numbers. Our updated bond angle means typically differed from the Parkinson set only by about 0.1° to 0.2° (see Table 7). The largest changes in mean were in C3′-C2′-O2′ and C4′-C5′-O5′, showing differences in mean of 1.101° and .699°, respectively. Overall, these differences are not particularly large, which demonstrates the care with which Parkinson, *et al.* made their initial parameters; with our larger dataset, we have confirmed that their parameters, commonly in use by XPLOR, CNS, and PHENIX, do indeed reflect the behavior of RNA structure fairly well.

Later work by Gelbin *et al.* attempted to use the 80 ribose structures in the CSD at the time to define separate parameters for C2′-endo and C3′-endo residues (Gelbin 1996). Notably, the phosphodiester linkage was measured for only eight structures—all the phosphodiester linkages available for RNA in the CSD as of 1996. Realizing the need for updated pucker-specific parameters, we re-evaluated our 399 structures accordingly, though once again, we did not find any significant differences in bond length and bond angle compared to those suggested by Gelbin, *et al.*

For the new phosphodiester linkage values, we assembled a set of 43 CSD structures and 36 high resolution (< 1.5Å) duplex RNA structures from the PDB. The

Gelbin analysis showed that C3′-endo and C2′-endo puckers had similar phosphodiester linkage parameters, and we confirmed this to be the case, though our standard deviations were somewhat more broad. With our larger dataset, we determined new values for the O3′-P bond length and the angles around the O3′ and P. Of particular note is the O3′/O5′→P→O1P/O2P angles. In the Gelbin analysis, these separated into two main categories—a "high" value of ~110.5° and a "low" value of 105° (note: for clarity, the 105° value is not listed on Table 7). With our larger dataset, the middle ground is more populated, and it is clear that, while there is a slight preference for O3′/O5′→P→O1P to be 109° and O3′/O5′→P→O2P to be 107.5°, we decided to average them for a value of 108° for all four angles. A summary of the changes from the Parkinson and Gelbin values can be found in Tables 6-7; the phosphodiester linkages are separated from the other values by a thick black line.

These bond length and bond angle parameters have been included in MolProbity as part of its covalent geometry analysis (Chen 2010). If the bond length or bond angle is greater than 4σ from the mean value, then it is marked as an outlier. For bond length outliers, a spring of 6 turns is drawn along the bond axis. If the length was too long, a stretched red spring is drawn; if it is too short, it is represented by a compressed blue spring (in analogy with astronomical red-shift or blue-shift). Bond angles outliers are represented by bold lines along the ideal, with a fan of increasingly thin lines fading towards the model. Again, red is used to indicate that the modeled angle is too large,

and blue for too small. On the multi-criterion chart, bond length and bond angle outliers each have their own separate, sortable column—observations based on the multi-criterion chart make it clear that many geometry outliers accompany sugar pucker outliers, as the model attempts to compensate for the incorrect sugar while maintaining connectivity to the base and the phosphates on either side.

Table 6: RNA Pucker Specific Parameters—Bond Lengths. All bond lengths are in Å.

			C3'				C2'			
Bond Length	Parkinson $\bar{x}$	Parkinson σ	Gelbin $\bar{x}$	Gelbin σ	Richardson $\bar{x}$	Richardson σ	Gelbin $\bar{x}$	Gelbin σ	Richardson $\bar{x}$	Richardson σ
O4' -> C1'	1.414	0.012	1.412	0.013	1.413	0.018	1.415	0.012	1.415	0.022
O4' -> C4'	1.453	0.012	1.451	0.013	1.449	0.016	1.454	0.01	1.455	0.022
C4' -> C3'	1.524	0.011	1.521	0.01	1.520	0.016	1.527	0.011	1.528	0.021
C4' -> C5'	1.51	0.013	1.508	0.007	1.506	0.017	1.509	0.012	1.512	0.021
C3' -> O3'	1.423	0.014	1.417	0.014	1.416	0.015	1.427	0.012	1.423	0.023
C3' -> C2'	1.525	0.011	1.523	0.011	1.527	0.019	1.525	0.011	1.529	0.019
C2' -> C1'	1.528	0.01	1.529	0.011	1.529	0.018	1.526	0.008	1.521	0.027
C2' -> O2'	1.413	0.013	1.42	0.01	1.422	0.016	1.412	0.013	1.407	0.020
C1' -> N1			1.483	0.015	1.476	0.020	1.464	0.014	1.463	0.023
C5' -> O5'	1.44	0.016	1.42	0.009	1.427	0.030	1.424	0.016	1.426	0.039
P-> O5′			1.593	.01	1.593	0.015	1.593	.015	1.593	0.015
P -> O3′			1.607	.012	1.602	.062	1.607	.012	1.602	.062
P -> O1P/O2P			1.485	.017	1.485	.015	1.485	.017	1.485	.015

Table 7: RNA Pucker Specific Parameters—Bond Angles.

Bond angle	Parkinson $\bar{x}$	Parkinson σ	C3' Gelbin $\bar{x}$	C3' Gelbin σ	C3' Richardson $\bar{x}$	C3' Richardson σ	C2' Gelbin $\bar{x}$	C2' Gelbin σ	C2' Richardson $\bar{x}$	C2' Richardson σ
O4' -> C1' -> C2'	106.4	1.4	107.6	0.9	107.3	1.3	105.8	1	105.9	1.8
O4' -> C1' -> N1	108.2	1	108.5	0.7	108.6	1.2	108.2	0.8	108.0	1.8
C4' -> O4' -> C1'	109.6	0.9	109.9	0.8	109.9	1.5	109.7	0.7	109.4	2.1
C3' -> C4' -> O4'	105.5	1.4	104	1	104.3	1.3	106.1	0.8	106.1	1.6
C5' -> C4' -> O4'	109.2	1.4	109.8	0.9	109.5	2.1	109.1	1.2	108.8	1.6
C3' -> C4' -> C5'	115.5	1.5	116	1.6	115.8	2.3	115.2	1.4	115.0	1.8
O3' -> C3' -> C4'	110.6	2.6	113	2	112.3	2.4	109.4	2.1	109.3	2.3
C2' -> C3' -> C4'	102.7	1	102.6	1	102.5	1.2	102.6	1	102.6	1.1
C4' -> C5' -> O5'	110.2	1.4	111.5	1.6	110.5	2.7	111.7	1.9	111.3	2.4
C2' -> C3' -> O3'	111	2.8	113.7	1.6	113.6	2.1	109.5	2.2	109.4	2.4
C1' -> C2' -> C3'	101.5	0.9	101.3	0.7	101.5	1.4	101.5	0.8	101.2	1.3
C3' -> C2' -> O2'	113.3	2.9	110.7	2.1	109.8	2.1	114.6	2.2	114.3	2.2
C1' -> C2' -> O2'	110.6	3	108.4	2.4	108.0	2.1	111.8	2.6	112.0	2.4
C2' -> C1' -> N1	113.4	1.6	112.6	1.1	112.6	1.7	114	1.3	114.5	2.0

C5′ -> O5′ -> P			120.9	1.6	120.9	1.5	120.9	1.6	120.9	1.5
O5′ ->P ->O1P/O2P			110.7 105	1.2	108.0	3.0	110.7 105	1.2	108.0	3.0
O1P -> P -> O2P			119.6	1.5	119.6	3.0	119.6	1.5	119.6	3.0
C3′ -> O3′ -> P			119.7	1.2	120.3	4.8	119.7	1.2	120.3	4.8
O3′ ->P-> O1P/O2P			110.5	1.1	108.0	3.0	110.5	1.1	108.0	3.0

4.3 Pucker Specific Parameters for Xray Crystallography

The refinement parameters (Parkinson 1996) used in X-PLOR and CNS (Brunger 1998), the major crystallographic model-building software at the time, were determined using the small nucleotide structures from the CSD (Allen 1979), with RNA and DNA structures taken from the NDB (Berman 1992) to account for the phosphodiester linkage. These parameters, which included bond lengths, bond angles, and dihedral angles, were rigorously tested in X-PLOR refinements of various DNA structures, and the few RNA structures available with moderate success. However, before the invention of our P-perp to line method of determining correct sugar pucker, sugar pucker errors were difficult to diagnose and correct; as such, pucker-specific parameters such as those found by Gelbin, *et al.*, (Gelbin 1996) were sparsely implemented or nonexistent in these model building and refinement packages. This was unfortunate, because it was established even in 1996 that sugar pucker was a significant factor in nucleic acid structure, and had a large impact on the structure of the overall nucleotide; for example, the backbone dihedrals for a C3′-endo DNA are more similar to that of a C3′-endo RNA than to a C2′-endo DNA, despite the fact that RNA has a 2′-hydroxyl on the sugar, which introduces additional steric constraints to the sugar dihedrals. As part of the PHENIX consortium, we were able to implement the P-perp to line test as part of the standard software design

and directly test how these old parameters fared against new, pucker-specific parameters derived from the above work and our dataset. This section focuses on how we incorporated those parameters into PHENIX and the results of these tests.

4.3.1 PHENIX introduction

PHENIX—or Python-based Hierarchical ENvironment for Integrated Xtallography—is a comprehensive Python-based system for macromolecular structure solution, and one of the foremost crystal structure refinement programs (Adams 2010). Consisting of a system of tightly integrated scripts and compiled software modules, PHENIX allows users to do substructure determination, phasing, and molecular replacement, as well as model building, refinement, and validation. PHENIX can be run from the command line or via GUI, and is available for Windows, OSX, and Linux. The procedures can be highly automated, from phase determination to homology modeling, and the work environment is extremely detailed and customizable, allowing the crystallographer to specify any unusual aspects of the crystal as well as control which refinement methods are used on a residue by residue bases. Furthermore, PHENIX is integrated with both Coot (Emsley and Cowtan 2004) and PyMOL (DeLano 2002; Schrodinger 2010), allowing both to be run and updated simultaneously as the

refinement runs, and incorporating any changes made in these programs into the next refinement.

Over the years, our lab has made major contributions to the validation and refinement packages, in part by including all MolProbity validation measures within the PHENIX software, ensuring that all structures are analyzed after refinement and any errors reported by MolProbity are highlighted in KiNG and Coot for more convenient correction. We have also worked with Tom Terwilliger, among others, to include RNA model building in the PHENIX autobuild procedure based on the well-determined phosphate and base positions; the autobuild program is now poised to take suites into account, attempting to build with recognized RNA backbone conformers if possible.

When I joined the PHENIX project, proteins and ligands were handled fairly well, but nucleic acid parameters were generalized such that both DNA and RNA were lumped together with the same bond length, bond angle, and dihedral parameters. I spearheaded the incorporation of RNA-specific parameters, and, as our dataset grew, pucker-specific parameters for refinement of the C3'-endo and C2'-endo sugars, thus making PHENIX the first refinement package to consider RNA puckers separately.

4.3.2 Pucker-specific parameters for the PHENIX software package

Putting pucker-specific parameters into PHENIX started out as a fairly difficult task. There was no infrastructure to handle two separate parameter files for each residue, and each nucleotide was represented by one set of values, meaning all DNA and RNA bases were the same. Working with Ralph Gross-Kunstleve, we adjusted the PHENIX code to accept multiple files for the same residue, and split the nucleotides into base definitions and backbone definitions. Furthermore, we split the backbone definitions into DNA and RNA specific parameters using the values in Parkinson *et al.*

We now had separate nucleic acid parameters, which improved refinement to an extent; this remained the apogee of our nucleic acid parameters until we were able to distinguish sugar puckers by including the P-perp to line method for identifying sugar pucker in PHENIX validation. Now that we could reliably tell what a ribose's pucker should be, we could use pucker-specific parameters to correctly model them. We started by splitting the RNA backbone parameters into four files—a pair for the C2'-endo and C3'-endo parameters for the backbone of a single nucleotide, and a pair for pucker-specific parameters for the phosphodiester linkages between residues. To obtain these pucker-specific parameters, we revisited the dataset used for the bond length and bond angle parameters (Section 4.2.3). As mentioned above, we had found only slight differences in bond lengths and bond angles, with each of their values very similar to the

Gelbin *et al.* parameters that had already seen use in other programs. Because of this, we opted to continue using the Gelbin *et al.* parameters for the means of all bond lengths and bond angles with one significant difference. However, rather than use the Gelbin "high" and "low" values for O5′→P→O1P/O2P and O3′→P→O1P/O2P, we opted to use the average of the more moderate values we had obtained, so both of these angles are listed in PHENIX as 108°.

With bond lengths and bond angles taken care of, we turned to dihedrals (see Figure 8, chapter 2). For these measures, we examined RNA05 (and later confirmed our results with RNA09). This rewarded us with the greatest difference in pucker-specific values. The most affected dihedrals were not just those around the sugar ring, but also those involving the relative position of the O3′; δ (defined as O3′→C3′→C4′→C5′) and ν_2 (C4′-C3′-C2′-C1′) changed by >71°! Dihedrals O3′→C3′→C2′→O2′, O3′→C3′→C4′→O4′, ν_3, and ν_1 all had rotations between 60° and 66°. Meanwhile, on the opposite side of the sugar, dihedrals ν_0, ν_4, and C4′→O4′→C1′→N1 (which includes the glycosidic bond) only changed by ~23°. As seen in Table 8, these dihedral parameters had values very close to those reported in Gelbin, *et al*.

However, our analysis of the non-sugar backbone dihedrals resulted in noticeable changes between our parameters and those from Gelbin *et al.* (Table 9). To begin, α has three peaks rather than two, at 65°, 165°, and -66°; these peaks are very

similar for C2′-endo and C3′-endo puckers. Rather than two peaks for β about 20° apart, we found the wide swath of β to include a range spanning 180° +\- 60°, with one small peak around 83°. Our γ and δ values mostly agreed with those in Gelbin, *et al.*, along with the observation that γ is overall the same for both puckers, and δ changes significantly from 81° (C3′-endo) to 147° (C2′-endo). The ϵ dihedral also depends on pucker: the C3′-endo value of -150° agrees with that from Gelbin *et al.*, but we found an additional peak at -100 for C2′-endo puckers. The strangest is ζ—it is trimodal, like α and γ, but the largest peak (-71) remains constant for both puckers, while its two smaller peaks shift: 172° and 52° for C3′-endo and 145° and 78° for C2′-endo. Because of this offset, if ζ is plotted without pucker specificity, it appears to have a constant low occurrence across most of the range. A list of these parameters can be found in Table 8-9.

Table 8: RNA Pucker Specific Parameters—Interior Sugar Dihedral Angles.

	C3'				C2'			
Dihedral	Gelbin $\bar{x}$	Gelbin σ	Richardson $\bar{x}$	Richardson σ	Gelbin $\bar{x}$	Gelbin σ	Richardson $\bar{x}$	Richardson σ
C4' -> O4' -> C1' -> C2' (ν0)	2.8	6.1	2.6	5.0	-20.8	5.2	-22.2	7.6
O4' -> C1' -> C2' -> C3' (ν1)	-24.6	4.9	-24.5	5.2	35.2	3.4	35.9	5.6
C1' -> C2' -> C3' -> C4' (ν2)	35.9	2.8	34.9	4.8	-35.4	2.8	-35.2	5.0
C2' -> C3' -> C4' -> O4' (ν3)	-35.3	3.1	-33.9	4.7	24.2	4.4	23.3	6.0
C3' -> C4' -> O4' -> C1' (ν4)	20.5	5.1	19.3	5.1	-2.3	5.7	-1.0	8.1
O3' -> C3' -> C4' -> O4'	-158.2	4.2	-156.2	6.7	-91.9	5.3	-92.8	6.4
C5' -> C4' -> C3' -> C2'	-156	3.1	-154.2	5.3	-96.6	4.1	-97.1	5.7
C4' -> O4' -> C1' -> N1	-118.6	6.5	-122.0	10.1	-143.4	5.5	-145.6	7.3
O2' -> C2' -> C3' -> O3'	44.3	4.5	38.9	8.5	-40.3	4.2	-42.4	5.1

Table 9: RNA Pucker Specific Parameters—non-Sugar Backbone Dihedral Angles.

	C3'				C2'			
Dihedral	Gelbin $\bar{x}$	Gelbin σ	Richardson $\bar{x}$	Richardson σ	Gelbin $\bar{x}$	Gelbin σ	Richardson $\bar{x}$	Richardson σ
O3' -> P -> O5' -> C5' (α) (m)	-74.7	9.8	-66	9.4	-74.7	9.8	-66	12.5
O3' -> P -> O5' -> C5' (α) (t)			165	20.3			172.6	17.0
O3' -> P -> O5' -> C5' (α) (p)	81	12.1	70.4	16.5	81	12.1	70.4	19.4
P -> O5' -> C5' -> C4' (β)	-176.5	13	180	60.0	-176.5	13	180	60.0
P -> O5' -> C5' -> C4' (β) p	163.8	23	83	10.0	163.8	23	83	10.0
O5' -> C5' -> C4' -> C3' (γ) (m)	-67.1	12.3	-66	15.0	-67.1	12.3	-66	15.0
O5' -> C5' -> C4' -> C3' (γ) (t)	179.4	6.4	178	15.0	179.4	6.4	178	15.0
O5' -> C5' -> C4' -> C3' (γ) (p)	52.5	5.7	52	10.0	52.5	5.7	52	10.0
C5' -> C4' -> C3' -> O3' (δ)	**81**	4.4	**83.6**	8.0	**147.3**	4.9	**146.8**	8.0
C4' -> C3' -> O3' -> P (ϵ)	-146	8.6	**-150**	35.0	-146	8.6	**-100**	35.0

C3' -> O3' -> P -> O5' (ζ) (m)	-70.8	4.8	-71	14.2	-70.8	4.8	-71	15.0
C3' -> O3' -> P -> O5' (ζ) (t)	163.1	0.6	**172**	30.0	163.1	0.6	**145**	30.0
C3' -> O3' -> P -> O5' (ζ) (p)	80.7	14.3	**52**	40.0	80.7	14.3	**78**	30.0

Key: Underlined values: large change between Gelbin value and Richardson value

Bold values: large change between C3′ pucker value and C2′ pucker value

4.3.3 Pucker-specific parameters results

These pucker-specific parameters were incorporated into the PHENIX database for use in the autobuild, refinement, and validation protocols. To see the effect of these changes in PHENIX, I ran a series of tests on 20 diverse structures from the RNA09 dataset, chosen for variability in resolution, size, and ligands. Each structure was run through PHENIX refinement with the original, non-pucker specific (nps) parameters, the nps parameters with tight standard deviation constraints (to see if limiting the model's search space would improve the structure), pucker specific (ps) parameters, and pucker specific parameters with tight standard deviations. The resulting models were scored based on R and R_{free} values, as well as on improvements in clashscore, pucker and geometry outliers, and suite outliers.

Of the 20 structures, 4 showed improved R/R_{free} values when run with the pucker specific parameters vs. nps (Figure 35). In 12 of the structures, the R value got worse while the R_{free} improved, indicating that there was less overfitting and that the new model derived with the ps parameters improves the fit to the data. There were four cases where both the R and R_{free} got worse, but the change was less than .4% in each case, and the clashscore improved while the number of bad bonds, angles, puckers, and suites were reduced tremendously—the angles in particular showed an average improvement of 68% for these structures by using the ps parameters instead of the nps ones. Across

the other sixteen structures, the percentage of bond and angle outliers dropped by 4.93% and 26.1%, respectively, resulting in greatly improved geometry. Geometry terms are highly weighted in PHENIX, and thus any change to the parameters will have huge implications for any model—for our new models to improve by so much indicates that the pucker-specific parameters are indeed having a positive impact on refinement. This is further borne out through our reduction in pucker and suite outliers—our pucker outliers are reduced by 11.5% on average, while our suite outliers are reduced by 8.0%. Clashscore, on the other hand, only dropped in half the studied cases. There are two reasons for this apparent retention of steric errors in the model. The first is that in several cases, these clashes are due to corrected residues clashing with waters; in such cases, it is likely that the water, which is typically difficult to see, should only have a partial occupancy, or is merely incorrectly placed. The second source of clashes is the order of operations of PHENIX refinement: in general, the terms with the worst outliers are corrected ahead of the general refinement. Since geometry and pucker errors have such a significant impact on the final model, these are among the first categories corrected, with the result that clashscore is given a lower priority. The contribution of this second source of error could be evaluated by loosening the estimated standard deviations in the pucker-specific parameters, thus lowering the weight of any given bond length or bond angle outlier. As expected, loosening the restraints on geometry

resulted in slightly fewer clashes, but it also greatly increased the number of pucker and suite outliers, as well as resulting in worse R/R_{free} values. As such, we decided to keep the pucker-specific parameters with tight standard deviations as the best compromise between geometry and steric contributions to model accuracy.

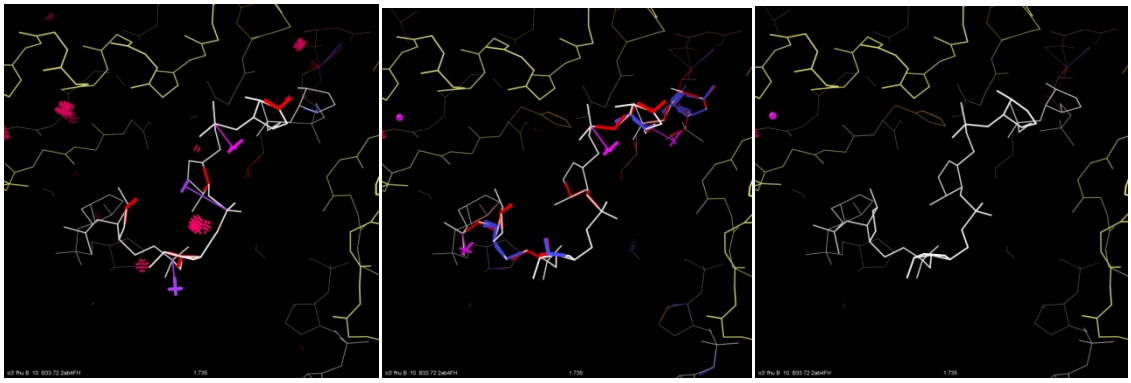


Figure 35: Pucker-specific parameters vs. Non-Pucker-Specific (NPS).

The original structure of pseudouridine 55 synthase complexed with its RNA substrate (left) contains severe clashes, as well as geometry and pucker outliers, within the RNA residues 11-15 (PDB: 2AB4, Phannachet 2005). Refinement with NPS parameters fixes the clashes but causes more geometry outliers and pucker problems (due to poor ϵ dihedrals). Using pucker-specific parameters produces a model with no pucker or geometry outliers (right).

4.4 Sugar Pucker and Suites in NMR

With our success at improving RNA-containing crystal structures through the use of our new parameters, we set our sights on doing the same for NMR structures. NMR structures of RNA were quite dominant before the turn of the millennium, but

crystallography of RNA has been surging forward. There are now over five times as many solved RNA structures presently in the PDB as there were at the end of 2000, with the increased rate mainly due to x-ray crystallography.

The usual way to determine nucleic acid NMR structures relies heavily on NOE (Nuclear Overhauser Effect) through-space distance constraints, allowing the placement of hydrogen atoms ~5 Å or less apart. But determining the detailed conformation of RNA backbone using NMR is quite tricky, since the density of observable and useful proton-proton distances is much lower than for proteins, and the interesting RNA structures tend to be the most difficult to analyze (Varani 1996). In practice, it is easy to identify the regions of A-form RNA structure vs. the rest. In addition to their role in refinement, the NOE constraints are used to guide early model building, such as determining helices and hairpin loops in RNA, and in some cases structural features as small as single-base bulges. One place where NMR has an advantage over crystallography is that sugar pucker can be determined by J-coupling measurements that reflect individual torsion angles in the ring. Late in the refinement process, RDC (residual dipolar coupling) orientation measurements are sometimes added, especially to determine long-range shape (Wang and Donald 2004).

My work on NMR, done in collaboration with Jeremy Block, maps out all the expected interatomic distances between hydrogen atoms (and thus the potentially

observable NOEs) along RNA backbone in each of the suite conformers, allowing the user to display them both in tabular form and in parallel coordinates. Viewing the possible NOE distances in parallel coordinates is an especially revealing way to identify patterns corresponding to particular backbone conformations. As with suitestrings, series of such conformations have their own identifiable multi-residue patterns; thus, an S-motif and a GNRA tetraloop imply distinct, repeatable NOE patterns. Even if the base is facing the solvent, and thus has little NOE data to identify the base position, the observed NOEs can be used to systematically pare down the possible backbone conformers at that suite; these can then be combined to identify common suitestrings and thus the local RNA 3D structure. Taken together, patterns of distance constraints and the system of suite conformers should provide a powerful tool to help elucidate the 3D conformations of RNA backbone in NMR structures.

4.4.1 Methods for RNA determination in NMR

As detailed in Chapter 3, RNA structural motifs such as the tetraloop or the S-motif can be specified by the corresponding strings of suites that describe their backbone dihedrals, as shown in the following two figures. The GNRA tetraloop (Figure 36) has suitestring **1a1g1a1a1c**; bases can be included to fully describe the RNA structure: **1aG1gN1aR1aA1c**. Strings of non-A-form RNA indicate large deviations from helices,

usually in the form of stem-loops, internal loops, or junctions. The distinctive S-shape of the primary strand of the aptly named S-motif, for example, has a suitestring of **1a5z4s#a1a** (Figure 37). The corresponding back strand of the S-motif has suitestring **1a1a1e1a1a**, the **1e** conformation being necessary to make the stack switch that accommodates the primary strand's distinctive shape.

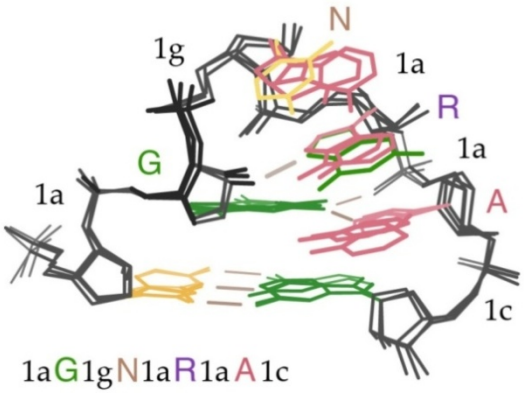


Figure 36: RNA Backbone Nomenclature for GNRA tetraloop

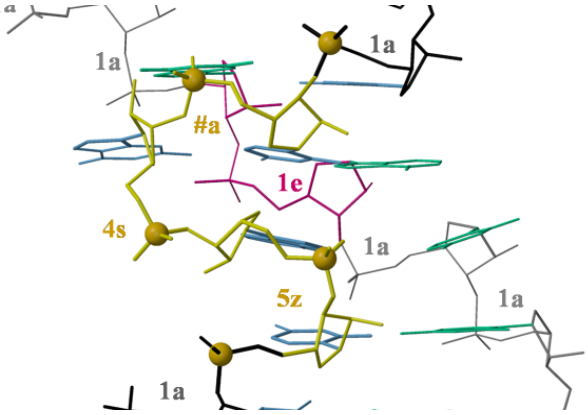


Figure 37: RNA Backbone Nomenclature for S-motif

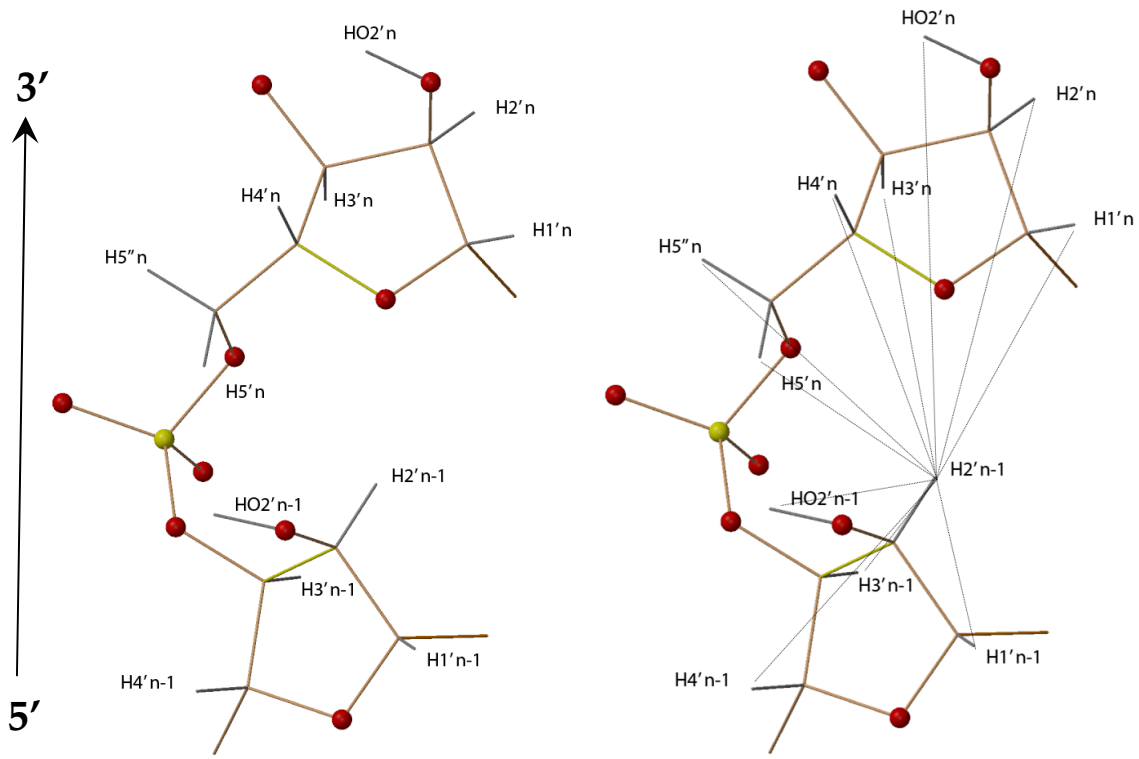


Figure 38: An RNA backbone suite with potential H2′n-1 backbone NOE's shown

The RNA backbone within a suite has 12 available hydrogens whose 3D arrangement is determined by the suite conformation. Figure 38 shows an RNA backbone suite with all the hydrogen atoms named (on the left of the image pair), and lines for the eleven potential backbone NOE's involving the H2′(n-1) atom (on the right of the image pair). Since NOEs are based on the distance between hydrogens, one can calculate the expected NOE's for each RNA backbone conformation based on their

relative hydrogen positions. The resulting calculations can be used as a lookup table for identification of RNA backbone structure during the 3D structure determination and building process.

4.4.1.1 NOE's by CRMA

NOE constraints are often determined by calculating distance between atoms via complete relaxation matrix analysis (CRMA). The relaxation matrix itself is related to the intensity of the NOE peaks (Cavanagh 2006). This method is much more complete and exact than the semi-quantitative distance approach. In practice, not all the intensities are known, resulting in an incomplete intensity matrix and thus interatomic distances cannot be computed exactly. A general solution for this problem is to build a model, often idealized A-form helix for RNA (Boelens 1989; Schmitz 1995), from which a model relaxation matrix is calculated. Observed NOEs are substituted for the theoretical values where possible, and a hybrid matrix is constructed from which distances are calculated. This process of substitution and back-calculation is iterated several times until the calculated and experimental distances are within an acceptable range (Wijmenga 1998). Many alternative methods have been proposed for building a complete relaxation matrix (Borgias 1990a; Kaluarachchi1991; van de Ven 1991).

A severe disadvantage of using CRMA is overdependence on starting models (Borgias 1989; Borgias 1990b). Using A-form RNA in the starting model yields an A-form-based model intensity matrix. A-form RNA accounts for >63% of all the residues in RNA structures, and >73% of the non-outlier conformations, (Murray 2003) and thus will almost always give a good overall match to the data. However, this prejudice will severely hamper the ability to model non-A-form structure when looking at interesting areas, particularly binding and active sites.

To avoid such bias, using NOE lookup tables could enable non-A-form structure to be determined for particular stretches of residues, allowing for better starting models with less A-form bias. Combining these lookup tables with known suitestrings, whole motifs could be handled in CRMA without resorting to A-form placeholders in the initial structure.

4.4.1.2 Semi-Quantitative Distances

The semi-quantitative distance approach relies on a large number of NOE constraints rather than precision of a given NOE. The NOE distances are estimated based on their relative intensities with respect to generally observed reference NOEs as measured on the same sample, and are classified only into bins as strong, medium, or weak. These reference atoms are often the H5-H6 distance (2.43Å) for strong, H1′-H3′

(~3.5Å) for medium, and H6/H8 to H1′ for weak (>5Å) (Varani 1996); note that the strong and weak references contain atoms from the base as well as the backbone. Using a relative scale based on reference atoms is important for offsetting the influence of mixing time on the intensity of the peaks when spin diffusion skews the distance measurements over longer mixing times.

In the semi-quantitative approach, each bin corresponds to a rough distance estimate, less than the upper bound of the constraint for that bin. Due to the difficulty in determining the lower bound accurately, most constraints have no lower bound. Looking at the upper bounds provides an invaluable resource for comparing observed NOEs to NOE constraints calculated from ideal RNA conformers because more NOEs in each bin means more cross-references to RNA backbone conformations; thus one can attempt to create a more precise estimate of which conformation is present.

4.4.2 Novel NOE method of suite identification and correction

The NOE method of identification for errors in RNA backbone was developed on the basis of RNA conformations possessing different NOE patterns. Using the 54 defined RNA backbone conformations, we created a series of twelve tables, one for each of the twelve hydrogens in the suite considered as one end of the potential NOE pair. Names of the 54 rotamers are listed vertically and the twelve hydrogens for the other end of the

pair are along the horizontal direction. Each table contains shaded-out cells (in purple) where the interatomic NOE distances are long enough (>5.2 Å) that they are very unlikely to be observed experimentally (Figure 39).

H2' n-1 to:	H3' n-1	H4' n-1	H1' n-1	H2' n-1	HO2' n-1	H5' n	H5'' n	H4' n	H3' n	H2' n	HO2' n	H1' n
1a	2.459	3.773	2.770	0.000	2.597	2.951	4.295	4.232	4.335	5.389	6.030	4.039
1c	2.458	3.765	2.773	0.000	2.615	4.358	3.356	3.733	4.956	5.735	5.910	3.840
1e	2.459	3.773	2.770	0.000	2.252	5.125	4.498	4.822	6.127	6.958	6.959	4.964
1f	2.459	3.773	2.771	0.000	2.621	4.307	3.280	4.717	5.516	6.891	6.969	5.466
1g	2.459	3.774	2.771	0.000	2.650	4.865	3.804	5.206	2.504	4.547	5.883	6.037
1L	2.468	3.813	2.756	0.000	2.546	3.801	5.041	4.215	4.679	5.100	5.693	3.153
1m	2.463	3.796	2.761	0.000	2.604	2.469	3.802	4.738	5.220	6.882	7.306	5.835
3a	2.466	3.804	2.759	0.000	2.650	5.851	6.692	7.972	6.780	8.463	9.819	8.146
3d	2.466	3.803	2.758	0.000	2.638	4.682	4.543	6.759	7.009	9.332	9.109	9.235
3g	2.466	3.804	2.759	0.000	2.864	6.077	5.891	8.097	6.905	9.216	9.854	9.562
5d	2.458	3.765	2.773	0.000	2.614	5.277	4.759	6.954	5.495	7.879	8.647	8.577
5j	2.468	3.820	2.754	0.000	2.570	5.566	5.615	6.813	8.065	9.938	9.407	8.888
5n	2.468	3.813	2.756	0.000	2.620	6.810	5.531	7.555	7.903	9.313	9.879	7.887
7a	2.463	3.789	2.764	0.000	2.137	5.060	6.019	6.560	5.819	7.004	8.276	6.084
7d	2.464	3.797	2.762	0.000	2.597	3.288	2.145	4.123	4.913	6.982	6.086	7.232
9a	2.463	3.789	2.765	0.000	2.528	5.561	6.633	7.824	7.446	9.163	9.985	8.294
&a	2.460	3.781	2.768	0.000	2.290	4.020	5.008	5.635	4.784	6.168	7.365	5.477
1b	2.464	3.797	2.762	0.000	2.627	2.598	4.142	4.147	5.318	4.586	6.504	4.746
1o	2.464	3.796	2.762	0.000	2.644	2.699	3.726	5.484	4.501	4.589	7.064	6.951
1t	2.459	3.773	2.770	0.000	2.287	4.557	3.657	4.034	5.693	5.064	6.580	4.576
1[	2.463	3.790	2.765	0.000	2.627	2.545	3.866	4.875	5.612	5.304	7.536	6.168
1z	2.462	3.789	2.765	0.000	2.683	5.078	4.301	5.939	3.873	3.461	5.904	6.413
3b	2.466	3.804	2.759	0.000	2.620	5.968	6.465	8.215	7.429	6.548	9.381	8.932
5p	2.464	3.797	2.762	0.000	2.573	5.757	5.260	7.276	5.630	5.301	7.911	8.142
5q	2.461	3.780	2.768	0.000	2.591	5.236	5.286	6.583	7.610	8.067	10.047	8.782
5r	2.463	3.788	2.764	0.000	2.626	5.876	5.862	7.181	6.693	8.001	9.881	9.827
5z	2.463	3.788	2.764	0.000	2.584	5.562	6.156	6.997	6.140	4.193	6.896	6.430
7p	2.464	3.797	2.762	0.000	2.609	3.478	3.257	5.214	5.588	6.892	8.542	7.950
7r	2.466	3.804	2.759	0.000	2.585	3.555	2.336	4.477	5.956	6.070	7.883	6.371
0a	2.524	3.887	3.084	0.000	2.490	6.087	6.812	8.516	7.988	10.006	10.738	9.516
0i	2.523	3.888	3.083	0.000	2.782	7.480	6.508	6.330	7.980	8.157	7.984	5.587
0k	2.524	3.887	3.084	0.000	2.846	5.905	5.236	7.520	8.232	10.357	10.155	9.709
2a	2.506	3.894	3.085	0.000	2.834	4.171	5.663	6.354	6.891	8.329	8.618	6.903
2g	2.490	3.896	3.086	0.000	2.769	6.884	6.535	7.743	5.010	6.562	8.386	7.940
2h	2.520	3.888	3.086	0.000	2.412	3.519	5.127	5.266	5.750	7.814	6.808	8.457
4a	2.506	3.897	3.079	0.000	2.842	5.746	7.248	7.464	7.607	8.543	9.456	6.896
4d	2.527	3.886	3.084	0.000	2.802	6.598	5.023	7.352	5.433	7.631	8.246	9.228
4g	2.521	3.888	3.086	0.000	2.818	7.099	7.609	9.063	7.556	9.317	10.705	9.293
4n	2.501	3.894	3.087	0.000	2.864	6.354	4.733	7.129	6.938	8.316	9.142	7.114
6d	2.514	3.891	3.085	0.000	2.831	6.411	5.741	8.187	7.195	9.640	10.009	10.265
6g	2.505	3.894	3.086	0.000	2.791	6.106	7.220	7.807	7.401	8.515	9.639	7.297
6j	2.491	3.897	3.086	0.000	2.862	5.799	6.031	7.546	8.397	10.523	10.113	9.888
6n	2.528	3.886	3.084	0.000	2.802	6.819	5.418	7.677	7.354	8.724	9.642	7.614
8d	2.523	3.887	3.084	0.000	2.807	5.522	4.278	6.596	6.233	8.568	8.207	9.432
#a	2.521	3.887	3.086	0.000	2.864	5.682	6.710	8.154	7.785	9.621	10.323	8.878
0b	2.520	3.888	3.086	0.000	2.843	5.549	6.638	8.020	8.413	7.791	10.354	9.082
2o	2.513	3.890	3.085	0.000	2.845	4.501	6.031	6.079	3.488	4.274	5.627	6.848
2u	2.502	3.895	3.086	0.000	2.819	4.863	6.235	5.690	7.246	8.762	9.697	8.656
2[	2.506	3.898	3.079	0.000	2.808	3.880	4.874	6.347	7.091	7.076	9.366	8.126
2z	2.495	3.896	3.086	0.000	2.864	6.699	5.777	7.509	5.164	4.988	7.164	7.989
4b	2.505	3.895	3.086	0.000	2.798	5.869	7.033	8.291	8.326	7.237	9.939	8.790
4p	2.528	3.886	3.083	0.000	2.863	6.488	4.810	7.042	4.600	5.156	6.687	7.919
4s	2.528	3.886	3.083	0.000	2.821	6.792	7.523	5.225	7.904	8.914	8.772	7.088
6p	2.507	3.897	3.079	0.000	2.841	6.774	5.569	7.965	6.002	6.110	8.393	8.994

Figure 39: RNA backbone NOE Lookup Table for distances from the H2′(n-1) hydrogen

Each suite conformer is identified by its appropriate number/letter combination describing the dihedral-angle values. Cross-referencing the observed NOE with the calculated NOEs in the table allows the spectroscopist to determine what subset of the 54 backbone conformations are possible given the data. For example, a spectroscopist observing an NOE of 3.2 Å between H2′ and H1′ of the following residue uses the H2′$_{n-1}$ table, and will see that the only possible calculated NOE in the H1′$_n$ column that fits is 3.153 Å, belonging to the **1L** conformation. Incidentally, the **1L** conformation differs from A-form in the β and ε dihedrals resulting in a more twisted base positioning.

If, on the other hand, a medium strength NOE of 4.0 Å is observed, the **1a** and **1c** (as well as **1L**) conformations constitute the reasonable subset of the 54 rotamer choices. For a longer-distance NOE of 4.7 Å or so, the best match would be **1b** or **1t** conformations (both 3′2′ puckers). Importantly, the observed NOEs in the semi-quantitative method are only representing upper bounds; conformations with lower NOE constraints (**1a**,**1c**, and **1L** in this case) may still be candidates.

Not all observed NOEs will be deterministic—many will result in ten or so conformations that match the data. This can be refined further by referencing multiple sets of observed NOEs and paring down the subset of backbone conformations plausible; using the 4.7 Å NOE example above, a second observed NOE between H2′$_{n-1}$ and H5′′$_n$ of 3.6 Å further pares the choices down to **1c** or **1t** , which can be further

distinguished based on sugar pucker. Sugar pucker can be determined independently with J-coupling or ^{31}P chemical shifts (Varani 1996). Pucker combinations can be identified on the table by color: 3′3′ is highlighted in blue, 3′2′ in green, 2′3′ in yellow, and 2′2′ in pink.

Ultimately, the strength of the approach is that even though multiple RNA backbone conformations are possible, the number of plausible ones that match the data will be reduced significantly compared to the 54 known RNA backbone rotamers and allows the spectroscopist to begin with more reasonable starting models for relaxation matrix generation.

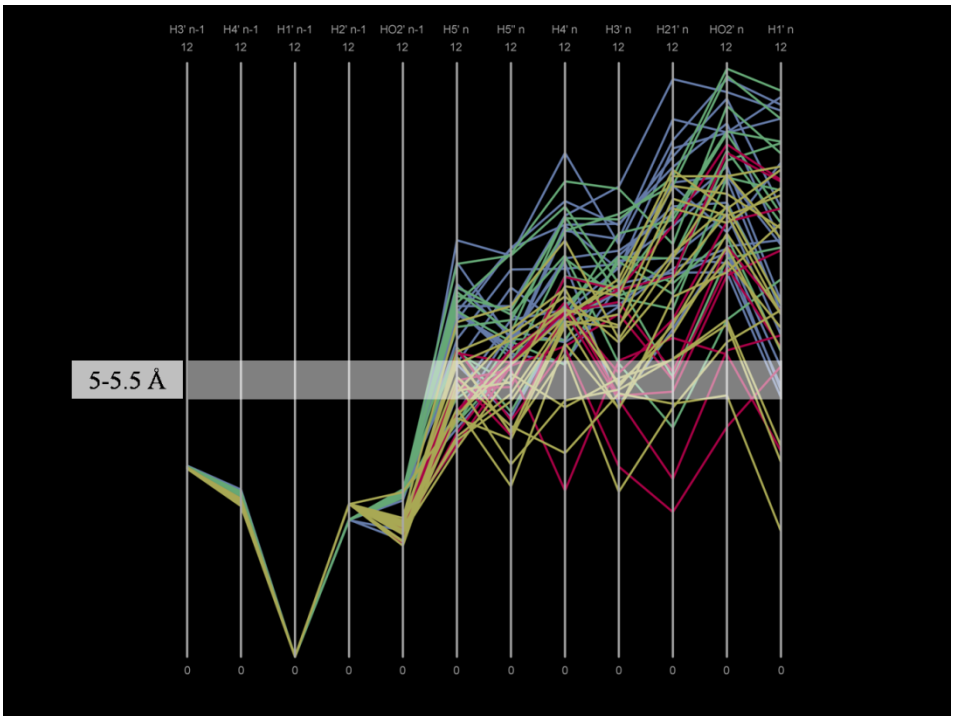


Figure 40: Parallel Coordinate Plot of RNA Backbone Rotamer NOE distances to H1′(n-1)

Complementary to the tables are parallel coordinate plots (Figure 40) where each axis in the plot represents a NOE distance between two hydrogens within the suite and is populated by the distances for each one of the 54 RNA backbone conformations. Twelve parallel coordinate plots are available, one for each hydrogen in the suite. These twelve plots provide a visual representation of the NOEs found in the tables. A dividing bar is shown at 5-5.5 Å, indicating where it becomes difficult to experimentally determine the NOE. The color scheme for pucker combinations is maintained (blue for 3′3′, green for 3′2′, yellow for 2′3′, pink for 2′2′). Each conformation has a unique

polyline representing the ideal NOE constraints. The parallel coordinate view can act similarly to the lookup table, and each can also give the spectroscopist insight into what other NOEs should be observed for any given conformation, and thus aid in the interpretation of the experimental data.

4.4.3 NOE suite assignment testing

A series of test structures culled from the PDB were used to assess the limits of the NOE suite assignment method. We started with using the Advanced Search function at www.pdb.org, with a first query of experimental method being solution NMR, with data present, followed by a second query for molecule type with the restriction of RNA. A total of 271 structure hits returned from the late 1990's until 2011. Ten of the 271 were selected for an initial test of the method, based on the number of NOEs in the restraints file (.mr file), and representing a variety of non-A-form structure and functions as well as a varying number of residues (Table 10).

Table 10: RNA structures used for NOE test

PDB ID	Reference	Year	Description	Number of Residues
1F9L	Rüdisser	2000	Analogue of P5ABC region of Group I ribozyme	22
1I3Y	Blanchard	2001	A-loop of 23S rRNA	19
1IKD	Ramos	1997	$tRNA^{ALA}$ acceptor stem	22
1LDZ	Hoogstraten	1998	Lead-dependent ribozyme	30
1T4X	Popenda	2004	Z-RNA	12
1UUU	Sich	1997	18S rRNA hairpin with UUU	19
1YMO	Theimer	2005	Human telomerase P2b-P3 pseudoknot	47
2B7G	Johnson	2006	Smaug recognition element	19
2HGH	Lee	2006	5S rRNA bound to TFIIIA	55
2JYM	Schwalbe	2008	SLα of HepB PTRE	22

Each of the test structures was run through the program Suitename to get their suite conformer assignments (Richardson 2008). The .mr data files containing the NMR restraints used to calculate the structures were downloaded and non-NOE data spliced out; we also ignored data from residues marked as outliers (**!!**) in the string of suites assigned by Suitename. Using the tables and parallel coordinate plots, each NOE distance restraint observed within a suite was used to look up the assigned identity on the table, indexing down to the value and assigning the appropriate subset of 3D backbone conformers. Where more than one NOE was present in the same suite, each NOE was used to generate subsets of possible conformers, and the union of these subsets was then taken as the result, as seen in Figure 41.

				Conformer Subsets
NOE1:	17 RGUA H2'	18 RCYT H5'	2.70	1a,1c,1e,1f,1L,&a,1b,1t,
NOE2:	17 RGUA H2'	18 RCYT H1'	5.00	1a ,1b,1m,1o,1[,

Intersection Conformer Subset:	1a, 1b
Union Conformer Subset:	1a,1b,1c,1e,1f,1L,1m,1o,1t,1[,&a

Figure 41: Two NOEs, their union and intersection.

NOE1, between H2′ and H5′ is 2.7Å; the conformer subsets that fit this distance are to the right. NOE 2, H2′-H1′, is 5.00Å and fits a different set of conformers. The intersection of these subsets is suites 1a and 1b, while their union is more expansive.

As can be seen, taking the union of the subsets often resulted in many more possible conformers than the intersection, but we still found the union to be more satisfactory for assigning possible suites. The main reason for this is the variable quality of deposited restraints for many NMR structures; often, at least one NOE is incompatible with the others, resulting in no possible conformers at the intersection. Even worse, sometimes multiple NOEs are at odds to such an extent that no single model can satisfy each of them simultaneously. Fortunately, the union of all NOEs still results in a substantial reduction from the 54 possibilities.

We began with 2HGH, a 54-residue portion of the 5S RNA found in the ribosome. Only 39 of these have backbone NOEs, though several of the residues contained multiple NOE constraints, bringing the total to 44 observed backbone-backbone NOEs. For each backbone NOE, the set of possible suite conformers was determined from the lookup tables. This suiteset was then compared to the suitestring generated by Suitename. Out of the 39 residues, 34 contained the Suitename-generated conformer in their lookup table suite conformer set. Nearly all of these suitesets consisted of only 6 conformers (out of the 54 possible), meaning that the amount of possible conformers to search before reaching a correct suite conformer was reduced by 89%. The average reduction of conformational search space by using the NOE lookup

tables was calculated to be 84%; this would result in substantial savings in calculation time when generating the structures.

Because most building programs assume A-form RNA structure, we decided to look specifically at non-A-form RNA suite conformers to see how lookup tables affected them. Out of 8 non-A-form suites, 7 contained the correct suite conformer in their lookup table suitesets, and the reduction in conformation space is 77%. This is extremely valuable as an addition to the previous only way to effectively deal with non-A-form structure is by multiple refinements of the hybrid matrix, which is very time-intensive, and may still not give the correct answer.

To further investigate non-A-form RNA structure in NMR, we analyzed 1T4X, one of the first structures of Z-form RNA to be determined by NMR. This structure is especially interesting as it contains no A-form structure, which means it cannot rely on the structural replacement methods mentioned earlier to create even a hybrid relaxation matrix. Indeed, the structure has many NOEs, and they are necessary to capture the structural details in the absence of a hybrid relaxation matrix. Unlike 2HGH, 1T4X contains many backbone NOEs per residue. It consist of 12 residues in two strands, with 48 total backbone NOEs between adjacent residues—which is more than found in the 54-residue 5S structure! With so many NOEs, it seemed that using the union of suitesets would become a problem. To further complicate things, the suitestring for this structure

was also hard to determine, with the consensus among the multiple deposited models being 6n!!6n1z6n on each strand.

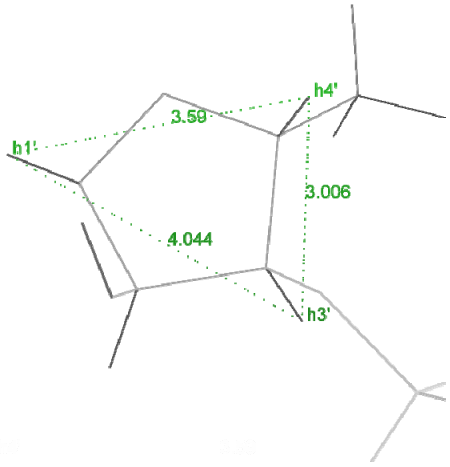
We found that for the **6n** suites, which were well-determined in the structure, we had an average of 11 suite conformers per suiteset (80% reduction from 54), despite this being the union of 4 suitesets per residue. This is even better than the results found in the non-A-form suite conformers from 2HGH, which only had 1-2 NOE restraint per residue, and demonstrates the usefulness of the lookup tables as a way to determine sets of suite conformers when dealing with non-A-form RNA.

With 80% of the suite conformers eliminated, it becomes much easier to determine which structure should be present, particularly since one can use outside information to pare down the possibilities further. The largest suiteset generated among the 6n suites was that between residues 2 and 3. The possible conformers according to the union of NOE suitesets are: **0i,2z,4d,4n,4p,4s,5n,6d,6g,6n,6p,8d**. Knowing that Z-form RNA, like Z-form DNA, alternates between residues with C2′-endo pucker and those with C3′-endo pucker (which can be confirmed experimentally) means that we can eliminate suites **2z,4p,4s,6p**, and **5n** (the first four are 2′2′, the last one is 3′3′) from the list of possible conformers, thus leaving us with only 7 to choose from and 87% of the suite conformers eliminated from the search. Thus, the NOE lookup tables combined

with outside biochemical and structural knowledge can greatly enhance the model building process by systematically eliminating suite conformers that do not fit the data.

The NOE lookup tables can also be used to diagnose areas where the structure should be refit. When looking at the second conformation in the Z-form RNA—the **1z** or **!!**, depending on the region, one can see that there is some ambiguity as to what suite is present even in the suitestring. When looking at the NOE-derived suitesets, we find that great disparity among them—so much so that the union of suitesets for any given residue contains 49 suite conformer possibilities. It is apparent, therefore, that the ambiguous NOE constraints have led to an overabundance of possible conformers and an incorrect fit to the structure. As mentioned earlier we know from Suitename that the suitestring is **6n!!6n1z6n**. However, looking at the NOEs, we find that the conformation **5z**, which is near **1z**, is contained in 12 of the 22 NOE suitesets describing these residues, while **1z** is only identified 4 times. By comparing glycosidic bond vectors, it is clear that those represented in 1T4X more closely match those of a **5z** conformation rather than a **1z** conformation. Substituting both ideal **5z** and **1z** suite conformers into the structure shows that the **5z** is indeed a better fit to both the data and the structure. This is further bolstered by the fact that ideal Z-form DNA also contains the **5z** suite conformer. Thus, we have shown that the NOE lookup table method has application for model error and identification as well as its main purpose of speeding up initial builds.

In the 1F9L structure, it is clear that rather arbitrary standard restraint values are used for NOE's that are close in the covalent structure (Figure 42). Due to these uncertainties in scaling in the NOE data, we removed 1F9L from the study. Unfortunately, such oddly standardized restraints are present in a number of other structures in the late 1990's and can be found even today. In some cases, the values have large discrepancies between their deposited NOE distance restraint value and the value measured on the structure itself. The simplest potential explanation for this discrepancy is that the bounds for semi-quantitative distance bins are not considered to matter much and are therefore not assigned very carefully.




```
 {constraints for residue 5 from expt}.
assign (residue 5 and name h3')(residue 5 and name h1') 2.0 0.0 2.0.
assign (residue 5 and name h4')(residue 5 and name h1') 2.0 0.0 2.0.
assign (residue 5 and name h3')(residue 5 and name h4') 2.0 0.0 2.0.
```

Figure 42: 1F9L residue 5—differences in NOE restraint distances and final model distances, showing the problem with scaling

The actual distance between H3′ and H1′ is 4.044Å, while the ranges from the NOE in the restraints file are 0.0Å-2.0Å. Yet H3′-H4′ is 3.006Å in the model, but has the same NOE restraints, indicating a standardization of the NOEs. Because these NOEs are unreasonably short, 1F9L was removed from the study.

4.4.4 NOE suite assignment discussion and conclusions

Overall, we found NOE suite assignment to be a useful tool for identifying suite conformers in NMR structures. By using the union of multiple NOEs, we were able to create conformer subsets that contained the Suitename-assigned suite 80% of the time (Table 11); these subsets on average contained only 20 of the 54 possible conformers. In

other words, we only need to search a third of the conformations before we have 80% certainty we have the correct one. When we incorporate a standard error of 10% for NOE measurements (Kuo 1980), this goes up to 92% certainty. When used in conjunction with chemical shifts that identify the pucker, NOE suite assignment becomes even more powerful, giving a 92% chance to find the correct suite in ~15% of the 54 possibilities. This could result in a substantial reduction in time needed to build initial NMR models, leaving more time for model validation and improvement.

Table 11: Results of NOE suite assignments

	Intersection Conformer Subsets	Union Conformer Subsets	Intersection Conformer Subsets +10% NOE error	Union Conformer Subsets +10% NOE error
Total suites (non!!)	118	118	118	118
Total matching suites	72	93	90	108
Total suites nonAform	42	42	42	42
Total matching suites (non-Aform)	17	34	24	39
Overall % matching suites	61%	79%	76%	92%
Overall % matching suites (non-Aform)	40%	81%	57%	93%
Average % matching suites per structure	51%	73%	67%	88%
Average % matching suites per struc (non-Aform)	46%	81%	61%	90%
Average reduction in conf space per structure	84%	63%	80%	61%
Average reduction in conf space:non-A-form	80%	56%	76%	48%
Conformer space saved if puckers are known			91%	86%
Conf. space saved if puckers are known (nonAform)			88%	82%

Of our 10 structures, there were 118 non-outlier suites according to Suitename. Our NOE assignment method was considered a match if the union of conformer subsets contained the suite assigned by Suitename. The average number of suites in a NOE assignment subset are reported as conformation space saved, so a 63% reduction in conformation space means we are searching through only 37% of the 54 conformers (20). Each value is given for all suites and all non-A-form suites.

Overall, there are generally at least a few NOE distance restraints observed that are useful for this analysis in a given structure. Those NOEs that are seen between the two residues within a suite are of high value, allowing us to restrain the model to a subset of conformers and pare down the possible backbone dihedral space considerably. When the intersection of 3D backbone conformer subsets determined by multiple

observed NOE's does not match with the suite-string analysis of the model, there is either a problem with the structure model or with the data, or with the variability in the conformer not reported in the tables. In one sense, conformation may be incorrectly modeled if we believe that the data are correct and the structure model is problematic. In reverse, this could be an indication of an error in an NOE assignment, or a demonstration of the limitations of accuracy in the data. Understanding what factors impact the range of values of the observed NOE is therefore a critical factor for this analysis.

There are two primary effects that impact the range of values around the observed NOE value that should be included when defining a subset based on an observed NOE distance restraint. First, the ranges of the 54 backbone rotamers around each dihedral will impact the NOE distances in our tables (how much 'give' on the value is acceptable—especially how short a given distance can get for each conformer); this is the rotamer contribution. Second, effects such as NOE scaling and how constraint bounds are set impact NOE distance restraint values used to define subsets of conformations for each observed NOE, plus the possibility of an incorrect assignment; this is the NMR-side contribution. A systematic evaluation of these effects will be important for spectroscopists intending to use this system.

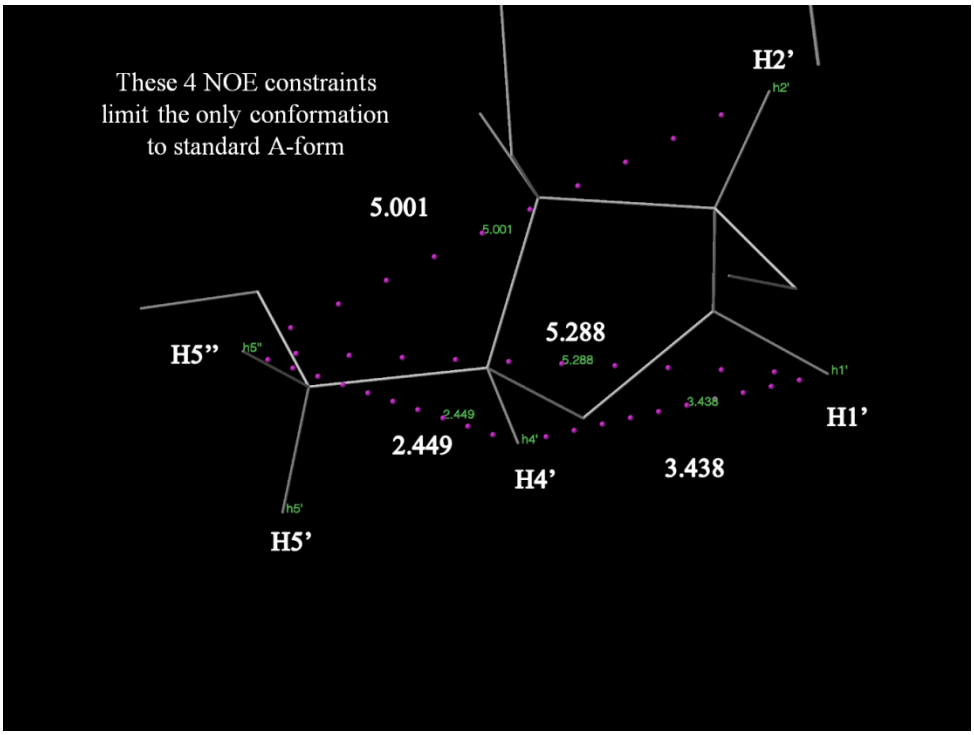


Figure 43: RNA Backbone NOE restraints that can conformationally restrict the suite to A-form

While some sets of restraints may conformationally restrict a suite to A-form (Figure 43), the structure determination packages make this observation difficult for non A-form RNA. This is to be expected, since no current structure determination package uses RNA backbone rotamers at all. Additionally, NOEs are seldom observed for some of the atom pairs in the lookup tables, for very technical experimental reasons, and are at the limits of the experimental methods currently available. In short, it is a challenge to determine an RNA structure by NMR and create a realistic and high quality model.

Despite the problem of errors from rotamer variability and from scaling, the backbone in-suite NOE distance restraints can usually narrow the backbone rotamer options for a given suite to a subset of the 54 rotamers. Our initial tests show that RNA structures done by NMR are not particularly well characterized or standardly created (not a surprise for those who struggle with such research). Future development of this work will include the addition of the more common base-base and base-backbone NOE's which, while not as directly conformationally restraining to the backbone, are useful in coarse measures such as distance vs. stacking of successive bases. This information would then be correlated with the rest of the observed NOEs to constrain the structure further.

It is known that many suite clusters contain few examples of the "ideal" suite, and in fact hover around the cluster center from which ideal dihedral values are derived. To improve the accuracy of suite assignments, it will be necessary to calculate the ranges of NOE constraints for each suite conformation. Additionally, bringing this information together into a single data quality metric for ease of use should be evaluated. One proposal would be to base it on the percentage of conformations fitting the particular residue's NOE constraints.

4.5 RNA structure correction methods

Now that we can diagnose errors in RNA structures, be they derived from crystals or from NMR data, we must now address how to correct them. NOE suite assignment and pucker-specific model building will go a long way towards getting the initial model correct, and pucker-specific refinement will fix much of the low-hanging fruit. Yet these methods are incapable of making large changes, and they are also inadequate for fixing regions with large amounts of error or buried by other parts of the structure. These difficult regions are often also the most interesting, and so we have helped develop several methods for correcting RNA structure.

4.5.1 Hand refits

At first, the only way to correct recalcitrant RNA backbone structure was by hand. To this end, David Richardson designed a kinemage to dock on a model and be set to each individual suite in MAGE. This "suitefit" kinemage contained ideal suite position information for each ideal suite conformer (updated as new conformers were identified), and allowed adjustments to be made via sliders corresponding to every dihedral angle. This ability to tweak the structure has proven to be particularly valuable, as most suites do not match their ideal definitions exactly. The suitefit kinemage building tool allows us to examine areas with poor geometry or impossible steric overlaps, and build new models which avoid such pitfalls. It also allows

experimentation with multiple suites in areas of the structure with poor density; if different conformers can fit into the same region, then we can choose the best one to fit the data, as well as learn more about which suites can substitute for each other. By using PROBE dots in conjunction with MolProbity's geometry analysis, we were able to correct a part of the 5S-rRNA S-motif that had previously been intractable due to its large steric clashes (Figure 44).

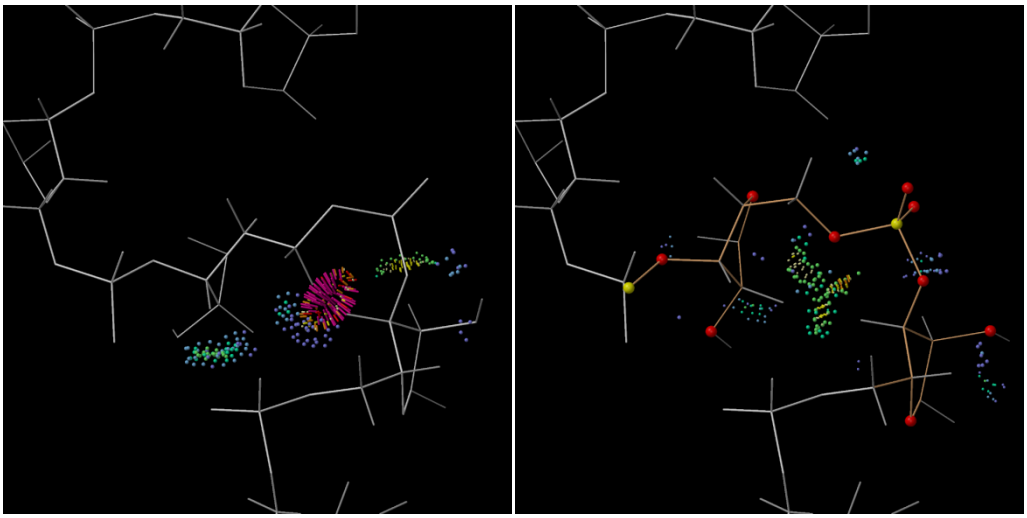


Figure 44: S-motif Suitefit hand refit.

Hand refit of the 5S S-motif, residues 76-77 (PDB: 1S72). The initial structure (left) has a large steric clash in pink; the refit structure removes the clash and fixes the suite from a !! to 5z (right).

In light of Suitefit's success in MAGE, Vincent Chen developed a version for KiNG, called the RNA Rotator. This system retained the ability to model in an ideal suite and then tweak it by changing the backbone dihedrals, using a dial system instead of

sliders. To aid in building correct backbone structure, each dial is shaded according to its ideal ranges; for example, δ has two shaded areas corresponding to the 55°-110° and 120°-175° ranges, while ζ has five shaded areas, representing the 5 separate peaks for C2'-endo and C3'-endo pucker. RNA Rotator also has the ability to generate PROBE contact dots on the fly, producing immediate feedback on a given change (Figure 45).

One of the largest problems with Suitefit and earlier attempts at hand rebuilding is the tendency of newly remodeled suites to be too far out of line to covalently bond with the rest of the structure. RNA rotator directly addresses this issue, and revolutionizes hand-rebuilt models in the process, by including the ability to continually superimpose the new model directly onto a series of user-selected atoms. This is accomplished via a large selection menu containing each atom from the chosen dinucleotide. By default, the new model superimposes on the central phosphate; if atoms from the menus are selected, the new model is superimposed upon them instead. This allows the user to keep the new model mostly static, while still being able to make adjustments; for example, if connectivity to the 5' Phosphorus becomes an issue, the user just specifies to superimpose on that phosphate and any adjustments from then on will attempt to minimize the RMSD to that atom. Furthermore, this allows the user to create on demand custom "anchors" for regions of high quality data; for example, if the C3', O3', and 3' P are known, then these can be chosen for superposition while the rest of the

new model is rotated freely. Conversely, a poorly fit sugar or phosphate may be deliberately deselected to allow it to move more freely in the hopes of getting a better model. We used this tool extensively in our modeling of the *E. coli* 70S structure discussed in Chapter 5.

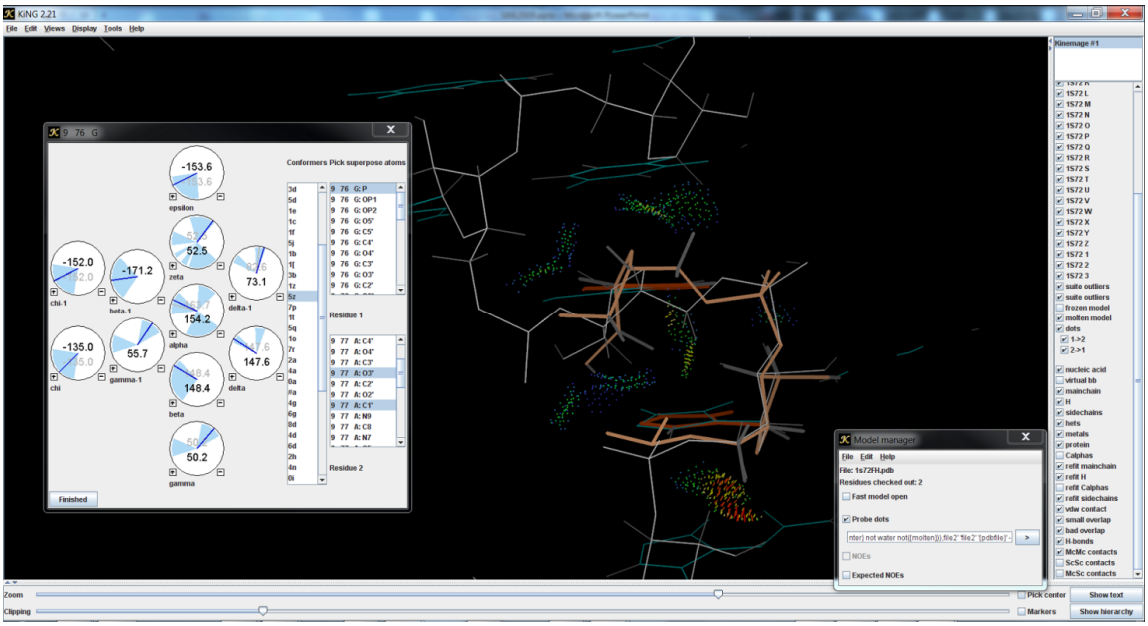


Figure 45: S-motif RNA rotator hand refit.

Hand refit of the 5S S-motif, residues 76-77 (PDB: 1S72) displayed in KiNG. RNA rotator dialog box showcases the dials and the selections for superimposed atoms. PROBE dots show the contacts for the new model (orange), which is superimposed on the original model in white.

4.5.2 RNABC

Even with the RNA rotator tool, hand correction of RNA backbone was still a tedious process in which a sterically impossible structure was replaced by an idealized conformation, whose dihedrals were then tweaked until the clashes were eliminated and the new structure fit the electron density. This process was time-consuming and required a high degree of skill and familiarity with RNA backbone structure to identify a new usable conformation. Often no acceptable conformation was found, but the difficulties of searching in 7D meant one could rarely be certain no such conformation existed. To allow faster, more complete corrections, the program RNABC, or RNA Backbone Correction, was developed by Xueyi Wang and myself in a collaboration between Snoeyink lab at UNC-CH and the Richardsons' lab (Wang 2008); Xueyi was responsible for the code, while I did much of the testing and analysis. RNABC uses input coordinates from a PDB coordinate file to rebuild a specified suite by anchoring phosphorus and base positions (Figure 46), which occupy the clearest electron density, and reconstructing the other atoms via forward kinematics. Geometric parameters are constrained within user-specified tolerance of canonical or original values, and torsion angles are constrained to ranges defined through empirical database analyses. Several optimizations reduce the time required to search the many possible conformations. The

output results are clustered and presented to the user, who can choose whether to accept one of the alternative conformations. Two test evaluations were conducted to show the effectiveness of RNABC, first on a set of S-motifs from 42 RNA structures, and second on the worst problem suites (clusters of bad clashes, or serious sugar pucker outliers) in 25 unrelated RNA structures.

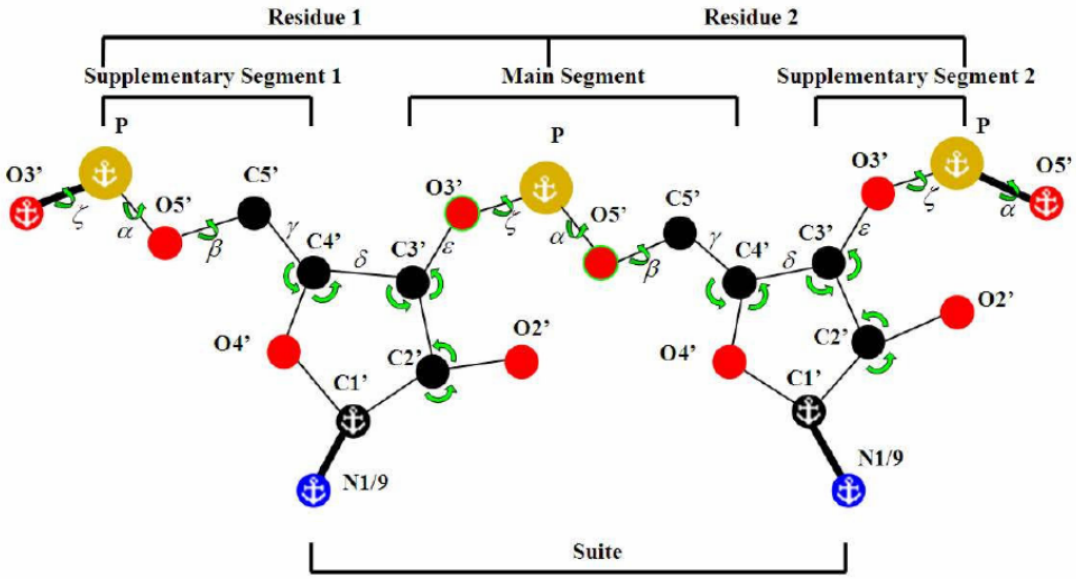


Figure 46: RNABC method.

Fixed points are indicated with anchors, and the main segment (the suite) and supplementary segments (rest of the dinucleotide) are indicated, along with affected angles and dihedrals.

102 S-motifs in 42 crystal structures are listed by the SCOR database of RNA motifs (Klosterman 2002). One S-motif, found in residues 8-12 of the sarcin/ricin loop of rat 28S rRNA (PDB:430D; Correll 1998) has a steric clash between the residue 12 C1’,

whose position is held fixed by RNABC, and an out-of-suite N6 on residue 20, and was removed from the test set, since RNABC only makes changes within a suite. We studied the three distinctive non-A-form suites on the primary strand. The sugar puckers are typically C3'-C2' for the first suite (**5z**), C2'-C2' for the second (**4s**), and C2'-C3' for the third (#**a**). The backbone conformations differ in each suite; they are not easy to fit accurately, so they often show serious steric clashes and sometimes deviant geometry—out of 101 S-motifs, all but 13 contain either steric clashes or bad geometry—making this dataset suitable for testing RNABC.

For the 88 S-motifs with clashes and poor geometry, we ran RNABC on the suites containing these errors, specifying clash-free output within 4 standard deviations of canonical parameters. For example, in the 5S rRNA S-motif with primary strand residues 76-79 (chain 9 of PDB: 1S72), shown in Figure 47, residues 76 and 77 contain steric clashes so we ran RNABC on suites 76-77 and 77-78, but not on suite 78-79.

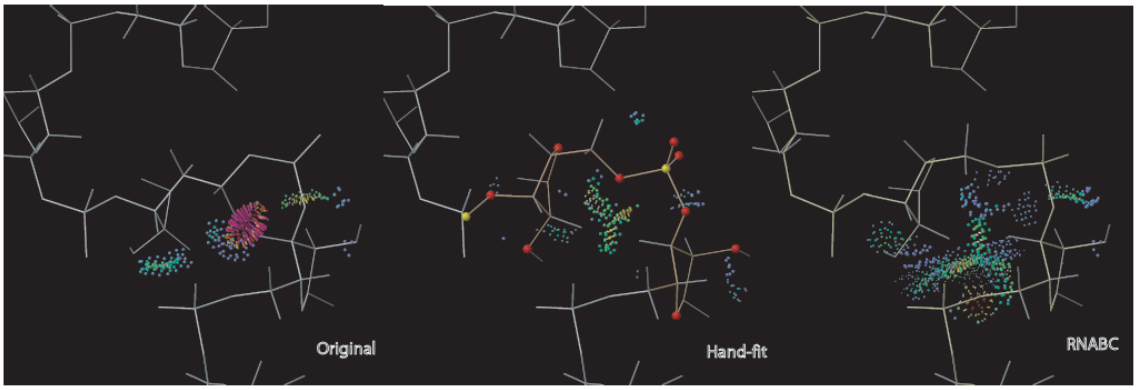


Figure 47: RNABC correction of S-motif.

Suite 76-77 of the 5S rRNA S-motif is on the left, followed by hand refitting and RNABC refitting (right). The RNABC refit took seconds, while the hand refit took hours to find, even though the conformations are almost exactly the same.

Table 12: Performance on removing steric clashes and correcting bad geometry for the 101 S-motifs studied.

Smotif Corrections	Total for each clash category	Good starting geometry	Geometry corrected by RNABC	Bad starting geometry
Total for each geometry category	**101**	**68**	**30**	**3**
No starting clashes	**17**	13	4	0
Clashes corrected by RNABC	**71**	47	23	1
Clashes remain after RNABC	**13**	8	3	2

Table 12 summarizes the results by cross-referencing geometry and clashes; for example, there are 68 structures with good starting geometry, 13 of which start with no clashes, 47 have clashes that RNABC can fix, and 8 have clashes that RNABC does not fix. For the 101 original S-motifs, 84 have at least one steric clash; RNABC proposes at least one clash-free conformation for 71 of those (85%). In the 33 Smotifs with bad geometry, RNABC found conformations with good geometry for 30 of them (91%). Electron density was only available for 30 of the 42 structures (71 of the 101 Smotifs), but the output conformations were checked for acceptable fit to the electron density where available; only two S-motif outputs were rejected at this stage. Combining both criteria, the overall success rate on this first test was 72 proposed corrected conformations out of the 88 S-motifs originally having problems (82%). As shown in Figure 47, the RNABC refit shown is very similar to the hand refit, but took significantly less time and expertise.

Having shown the consistent usefulness of RNABC in correcting a specific backbone motif, a second test was conducted to determine the program's ability to handle severe local problems in a variety of contexts. A set of 25 diverse structures were chosen from the RNA03 database (Murray 2003), with representatives ranging from simple duplex RNA to the ribosomal subunits and tRNAs. For each of these structures, we used MolProbity and KiNG to identify suites with especially bad clashes and sugar-

pucker outliers. RNABC was run on those suites, as well as suites immediately before and after. If an RNABC run with default parameters failed to yield results, parameters were relaxed in a sequential manner, ensuring that new conformations were found whenever possible.

RNABC suggested new conformations for 72 of the 154 suites tested. However, 8 of these new suites were later rejected, 3 due to remaining steric overlaps and/or sugar pucker outliers, 2 because of poor fit to the electron density, and 3 for both of those reasons. Thus, RNABC produced new clash-free conformations and/or better sugar puckers, with satisfactory geometry and density fit, for 64 of the 154 suites tested (42%); 19 of those successes were obtained with default parameters. Table 13 shows the most common problems identified among the original 72 suites, along with how well RNABC improved them. A given suite may have multiple problems, which are categorized into steric clashes (separated by specific pairs of clashing atoms), pucker outliers, and unfavorable ϵ dihedral values. Pucker and ϵ dihedral problems often occur together since distortion of ϵ is often the result of fitting a ribose into the wrong pucker state. RNABC does best at correcting steric clashes, as these were its central design emphasis. It can usually improve and sometimes correct sugar puckers that are misfit as 3′ or 4′ when they should be 2′, as in the example of Figure 48. The "other" puckers are extreme distortions, which the program finds difficult to improve or correct. Each of the bad ϵ

values was related to a bad sugar pucker; RNABC corrects 5 of them; the 14 ϵ values that remain unfavorable correspond to 14 sugar puckers that are improved but are not corrected completely. For all but three suites, when RNABC aggravated a problem in one category, it greatly improved the other two categories.

Table 13: RNABC results for worst case model outliers.

Common problems	# of instances	# fixed completely	# improved	# unchanged	# worse	% fixed	% fixed or improved
Steric Clashes							
1H5′—O2′	29	17	6	3	3	59	79
2HO′—P	23	13	5	4	1	57	78
C5′or H5′— C2′or H2′	19	11	4	3	1	58	79
1H2′—O4′	16	10	2	2	2	63	75
Others	80	45	17	7	11	56	78
Pucker outliers							
C4′	12	2	8	2	0	17	83
C3′ → C2′	11	2	7	1	1	18	82
Others	11	1	0	4	6	9	9
Unfavorable ε dihedrals (-45 to +155)							
Bad ε	19	5	0	14	2	26	26

A suite was considered unchanged unless the new one differed by 5 clash spikes, 10° dihedral, 0.5Å Pperp-to-line length, or 40° ϵ dihedral. The total number of clashes is greater than 72 because many suites contained several clashes.

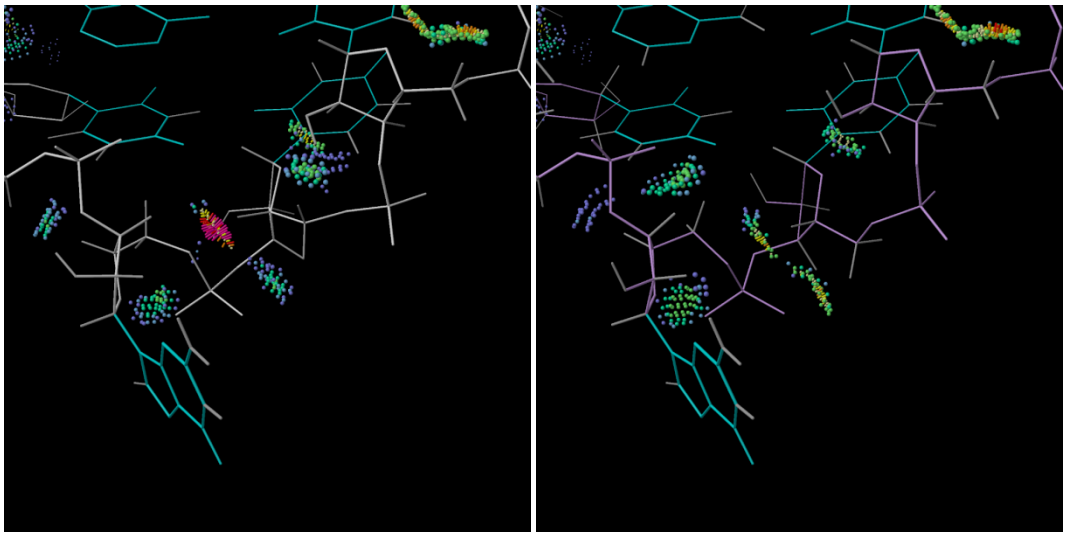


Figure 48: Pucker correction in tRNA^{ILE} .

The original tRNA^{ILE} structure (PDB: 1FFY) has a large clash due to a pucker outlier (left). Correction via RNABC fixes the pucker outlier and relieves the clash.

The final filter was to determine, for the 10 structures (42 of the 72 suites) that had structure factors available, how well RNABC's proposed new conformations fit into the electron density. Although RNABC currently incorporates no constraints for electron density, the fit improved in almost every case — dramatically for some suites, as depicted in Figure 49. Five suites were exceptions; three conformations already targeted for elimination by other geometric offenses and two new cases were found that lay significantly outside the density compared to the initial structure. Thus, 8 of the 72 outputs were rejected by these post-filtering steps, with 89% of the suggested suite

conformations deemed acceptable for future refinement. Overall, this test of RNABC on extreme structural deviations had a 42% success rate, with a fairly low rate of false positives.

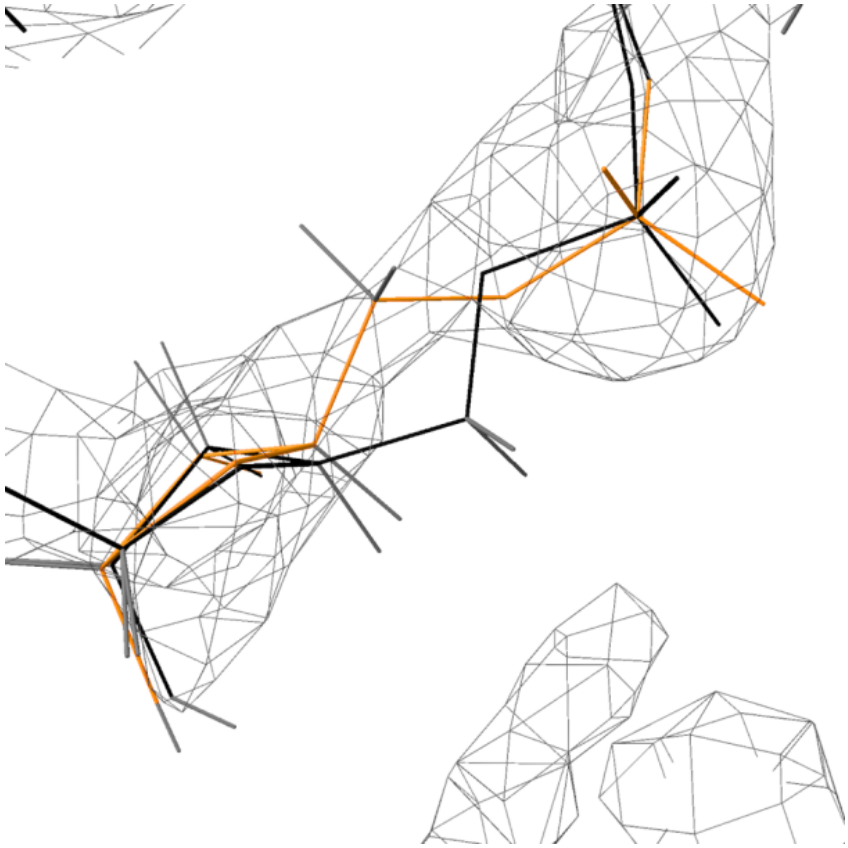


Figure 49: RNABC and electron density.

The original version of the 50S rRNA residue 1942 (black; PDB: 1S72) does not fit the density well. The RNABC refit (gold) moves the structure into the density without causing steric clashes.

We close this discussion on tests with a look at how many different sets of conformations are output by RNABC, and how different these are from the original structure. In the 235 tested suites for which RNABC produced output conformations, the output dihedral angles differed from the original by 20° ± 3° RMSD across the 6-dihedral sets, with the extremes ranging from 2° (tiny wiggles) to 100° (large backbone shifts). Often a single dihedral undergoes a relatively large change while the other dihedrals adjust slightly to accommodate; sometimes two dihedrals change 30°-50°(usually α and γ in the long-recognized "crankshaft" motion). Cases in which 3 or more dihedrals change more than 35° were rare. Moreover, 30% of the time RNABC yields two conformations that are different from each other as well (dihedral RMSD > 20°) ; a further 5% yield 3 or more different conformations. Thus, RNABC is capable of giving the user significantly new and sometimes varied options with which to replace the original local conformation.

4.5.3 ERRASER

RNABC was a fantastic program, but it was still limited by its assumption that the phosphate and base were fit correctly and therefore did not need to move. Our search to bypass this limitation led us to Rhiju Das' lab, where he developed RNA applications for Rosetta (Simons 1999). A collaboration between the Das lab and the

Richardson lab resulted in the development of a new program for improving RNA structure, called ERRASER, which stands for Enumerative Real-space Refinement ASisted by Electron density under Rosetta (Chou 2013). Coded by Fang-Chieh Chou, a grad student in the Das lab, ERRASER aims to improve problem residues in RNA structures, either by searching the entire structure for errors, or by rebuilding a single user-specified residue. In a typical ERRASER rebuild, the model is refined for several cycles with PHENIX, and then ERRASER uses the refined model and the electron density to rebuild areas of the RNA that need correction. This rebuilding has several stages: to begin with, ERRASER minimizes all torsion angles and all backbone bond lengths and bond angles in the model using the Rosetta high-resolution energy function, and uses an electron density correlation score to ensure the new model is consistent with the experimental data. After this step, remaining bond length, bond angle, pucker, and suite outliers are identified by PHENIX's RNA validation tools. Such residues, as well as residues with large RMSD between their original position and the minimized position, are then rebuilt one at a time through single-nucleotide stepwise assembly, an *ab initio* method of building each residue by enumerating many conformations covering all build-up paths. Lower-energy rebuilds are accepted; afterward the entire new model is minimized again. The model generated by ERRASER should then be subjected to a final set of rounds of PHENIX refinement against the experimental data.

To test this new program, I ran ERRASER on a G-riboswitch structure (PDB: 1U8D; Batey 2004) that we had previously used to make corrections by hand and with RNABC. Not only did ERRASER correct the two pucker outliers in residues 63 and 48, but also fixed the geometry outliers in eight other residues. Encouraged by this, we ran additional tests on the Twort Group I intron (PDB: 1Y0Q; Golden 2005), a 229 residue structure solved at 3.6Å resolution. RNABC had previously failed to find corrections to more than a few residues in this structure, mostly due to problems with anchored atoms at low resolution and the sheer number of clashes. ERRASER gets around these issues by not paying attention to clashes, and by rebuilding each offending residue, rather than anchoring some of its existing atoms.

The first step in ERRASER is to refine the structure in PHENIX. The result of this refinement was a reduction in clashscore by 43 clashes per thousand atoms, boosting the structure from 24th percentile MolProbity rank to 86th. Refinement also corrected 2 sugar puckers, but it made the backbone bond angles worse by 3.5%. The results from the ERRASER rebuild and subsequent PHENIX refinement are even better. Clashscore drops from the original 69.5 to 5.21, making the new model in the 100th percentile (Table 14) . Of the original 20 pucker outliers, 10 are fixed, and of the 52 suite outliers, only 28 are left. Furthermore, this new model had only 3% bad bond angles, vs. the 3.86% from the original or the 7.3% from PHENIX refinement without ERRASER. Finally, the most

important evaluation of the new model, the R/R_{free} values, shows a slightly higher R, and a slightly lower (by .1%) R_{free}, indicating that with this new model, overfitting has been reduced, while at the same time many known errors have been erased. Similar results were obtained for a second Group I intron at 3.1Å(PDB: 1U6B,), thus showing that ERRASER can consistently improve RNA backbone structures at low resolution.

Table 14: MolProbity statistics for 1Y0Q, original and after ERRASER and refinement.

All-Atom Contacts	Clashscore, all atoms:	69.53	24th percentile* (N=37, 3Å - 9999Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.		
Nucleic Acid Geometry	Probably wrong sugar puckers:	20	Goal: 0
	Bad backbone conformations#:	52	Goal: 0
	Residues with bad bonds:	0.00%	Goal: 0%
	Residues with bad angles:	3.86%	Goal: <0.1%

All-Atom Contacts	Clashscore, all atoms:	5.21	100th percentile* (N=37, 3Å - 9999Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.		
Nucleic Acid Geometry	Probably wrong sugar puckers:	10	Goal: 0
	Bad backbone conformations#:	28	Goal: 0
	Residues with bad bonds:	0.00%	Goal: 0%
	Residues with bad angles:	3.00%	Goal: <0.1%

4.6 Discussion

This chapter has given a detailed overview of our journey from almost no RNA-specific parameters to a variety of tools available for all RNA-oriented structural biologists to use, be they crystallographers or NMR spectroscopists. We designed the datasets to have the best of the best data available, and are currently developing an update to RNA11 that will provide even more accurate, empirically derived RNA parameters. Our error diagnosis tools for sugar pucker and backbone geometry outliers have been implemented in PHENIX and in MolProbity, the latter of which is available as a free webservice for all users. These diagnostics and our newly derived pucker-specific parameters are currently available as part of the model building and refinement software in PHENIX, ensuring that crystallographers have the most accurate tools available when solving RNA-containing structures. For NMR, we have tools that will pare down the possible suite conformations based upon NOEs, reducing the overall time it takes to solve the structure and ensuring that RNA backbone motifs are quickly identified so they can be modeled accurately despite the often ambiguous NMR data. Finally, for the most difficult problems in modeling and rebuilding RNA structure, we have collaboratively developed RNABC and, subsequently, ERRASER, to provide all scientists with tools allowing them to acquire and publish the best RNA models possible.

5. Diagnosis, correction, and refinement of RNA-protein complexes

RNA-protein interactions allow ribosome assembly, alternative splicing, and post-transcriptional and post-translational processing. It is these processes that control the final product of the genetic code, as well as when and how much of the product is produced. Thus, RNA-protein interactions are at the heart of biological life. With our new tools for analyzing and correcting RNA backbone structure, we launched the first investigation of RNA-protein interactions done primarily from the point of view of the RNA structure.

We found four archetypes of RNA-protein interactions. Non-specific interactions involve electrostatically driven attachment to the RNA backbone with no regard to the sequence, similar to non-specific DNA binding. Such interactions are common among RNA-specific deaminases and Dicer structures (Kim 1994; Bernstein 2001); the proteins recognize any dsRNA regardless of the sequence. Specific static interactions involve protein interacting with a static, often unusual, RNA backbone structure whose presence is unperturbed by said cognate protein—tetraloops, kink-turns, and S-motifs fall into this category. An induced interaction is formed when both the protein and the RNA backbone are pulled out of position to interact, such as the U1 hairpin II interaction with protein U1A to make the U1 snRNP. And, finally, the purely sequence-specific

interaction category involves the protein binding to specific base sequences, found either with aptamers and other small, single-stranded RNAs, or else with the protein unfolding the RNA backbone structure to locally free the bases for "reading" the sequence, such as for the restrictocin/sarcin loop complex (Yang 2001).

By combining our knowledge about RNA-protein interactions and RNA backbone motifs with our newly developed tools for building and validating RNA backbone, we set out to improve some of the most important RNA-protein complexes. We started with the U1 snRNP and moved on to several other important complexes, most notably improving the highest-resolution (at the time) structure of the *E. coli* ribosome.

5.1 Rerefinement of U1 snRNP and HDV ribozyme

As mentioned in chapter 3, the U1 snRNP recognition site is often spliced into an RNA sequence to aid crystallization. This method was first used in the high resolution crystal structure of the HDV ribozyme (Ferre-D'Amare 1998; PDB: 1CX0). This ribozyme catalyzes self-cleavage by a transesterification reaction, resulting in a 2', 3'-cyclic phosphate. The 1CX0 crystal structure is of the post-cleaved ribozyme; in order to crystallize it, the solvent exposed P4 stem was engineered to contain the U1 hairpin II,

which binds tightly to the U1A spliceosomal protein. This section covers the errors in the original structure and our work to correct them.

5.1.1 Initial structure

Using the methods described in chapters 3 and 4 for diagnosing errors in the RNA backbone, I discovered anomalies in the U1 snRNP interface portion of the high-resolution HDV ribozyme structure 1CX0, where part of the RNA backbone was misfit along the RNA-protein interface. Rather than the U1 hairpin II suitestring of **1a:1[:6g:1[:0a:1a**, as found in the native U1 snRNP structure 1URN (Oubridge 1994), the RNA backbone suitestring in this region (residues 148-154) was **1a:1[:6g:1m:!!:1a.** The **1m!!** conformation was caused by a pucker outlier in residue 152, which had mistakenly been fit as C3′-endo rather than C2′-endo; this caused several bond angle geometry outliers. Residue C150 (**[C6** in the modular nomenclature) contained another bond angle outlier due to a clash with R83 from the CTD of U1A. Finally, selenomethionine (MSE) 51 from the U1A RNP1 had several clashes with both the protein and RNA backbone. Several rotamer outliers and clashes also affected the U1A protein.

The HDV ribozyme portion of the structure had many problems, particularly near the active site. There were 7 more sugar pucker outliers and 15 more bad backbone conformations. There were also a number of bond angle outliers, bringing the total

amount in the structure up to 14. The overall clashscore was 13.79 for the entire structure, which isn't bad—it is still in the 83rd percentile for a 2.3Å structure.

5.1.2 RNA-protein interface refits

The initial corrections to the RNA-protein interface were done with KiNG and RNABC. I started by using RNABC to find an alternate suite to replace the **!!** conformation in the U1hpII suitestring. One of the possible conformations RNABC found was the standard **0a** conformation, which, when inserted into the U1hpII, recreated the **1[:0a** conformation that is found along other instances of the U1 snRNP interface.

For the protein corrections, I used the sidechain rotator and backrub tools in KiNG (Davis 2006). R83 in chain A has a large steric clash between its NH2 and O2 of U150. Even with other improvements to the local structure, this stubborn clash between the protein and RNA persisted. I used the sidechain rotator in KiNG to tweak the χ_2 and χ_3 angles to move the R83 away from U150, since the base was in very clear density and had little room for improvement. The new position held through refinement and alleviated the clash without causing new problems.

The correction for MSE51 was a little more difficult. It clashed slightly with the C5′ of Adenine 153, while at the same time causing a huge steric overlap with the

carbonyl oxygen of its immediate neighbor K50. The electron density around it showed that there was some room for interpretation, so I centered the selenium in the strongest density by rotating all three χ angles. This new fit had only one small clash, which was removed completely during refinement (Figure 50) .

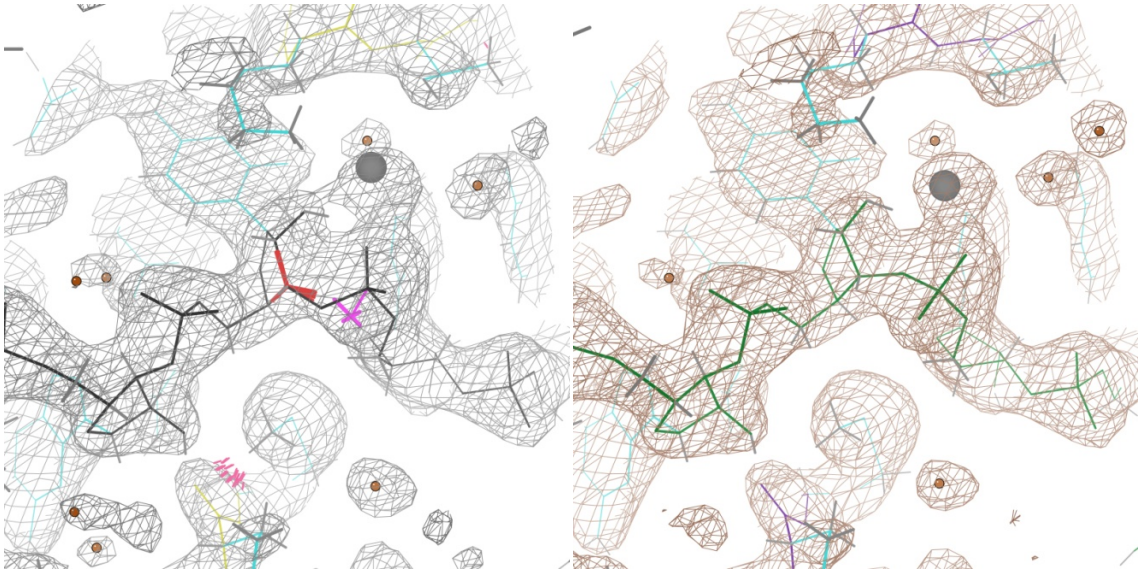


Figure 50: Fixes along the U1 snRNP interface.

The RNA residue 152 shows red bond angle and magenta sugar pucker outliers (left) which are corrected after RNABC replacement and refinement (right). Meanwhile, the MSE51 clash with its backbone in the bottom center of the left image is fixed in the right image. Note the slight changes in electron density (grey vs. brown) that accompany these corrections.

Outside of the RNA-protein interface, the U1A protein still had some errors. The MSE72 HE1 has a very large clash (>.75Å overlap!) with R36, despite its good rotamer position. However, its difference density shows a large peak that is readily filled by rotating from the mmt rotamer to the mmm rotamer, but when I changed it, I got positive density for its original position. Modeling both alternates together shows that both fit the density well and that the new mmm rotamer has an occupancy of 65% and the mmt rotamer has an occupancy of 35%. On its own, the mmm rotamer clashed with the old I33 position, so I was compelled to refit that sidechain as detailed below.

I33 contained a clash with the neighboring F34. When I looked for a way to alleviate the clash, the difference density map showed clear positive and negative density around the isoleucine, indicating a different rotamer should be fit into that position. The pt rotamer was chosen over the original mp rotamer, and a combination of χ rotation and backrub motions eased it into place. After refinement, positive density is gone altogether, the negative density has mostly disappeared, and Ile33 no longer clashes with MSe72 or Phe34.

5.1.3 ERRASER correction of RNA and refinement

After making as many corrections as possible to the protein and the protein-RNA interface, we set our sights on the rest of the RNA structure. RNABC could not find any

allowable conformations for the residues in the active site, and it became clear that we would need extensive movement of the phosphate and base positions to build any new conformations. ERRASER's step-wise assembly can accomplish this rebuilding, as detailed in chapter 4, so we applied it to our new model for the HDV ribozyme/U1snRNP structure. After PHENIX refinement with our pucker-specific parameters, even before applying ERRASER, we managed to fix 5 of the 8 pucker outliers, reduce the clashscore from 13.79 (83rd percentile) to 8.21 (97th percentile), and decrease the suite outliers by 5, as well as reducing the number of sidechain rotamer outliers and increasing the Ramachandran favored residues in protein.

ERRASER was run on this refined model and the results were spectacular. Despite ERRASER's limitations (no clash evaluation, can't see non-RNA), the clashscore dropped to 1, and the three remaining pucker outliers were fixed (Figures 51-52). Two more backbone suites were put into recognized conformations and the number of bad bond angles was reduced to 0. Furthermore, this new model dropped the R and R_{free} value from 23.51% and 26.96% to 21.37% and 25.06%, an improvement of 2.14% and 1.9% over the original R and R_{free}, respectively. Tables 15 and 16 show the MolProbity statistics for the original structure and the ERRASER refined one.

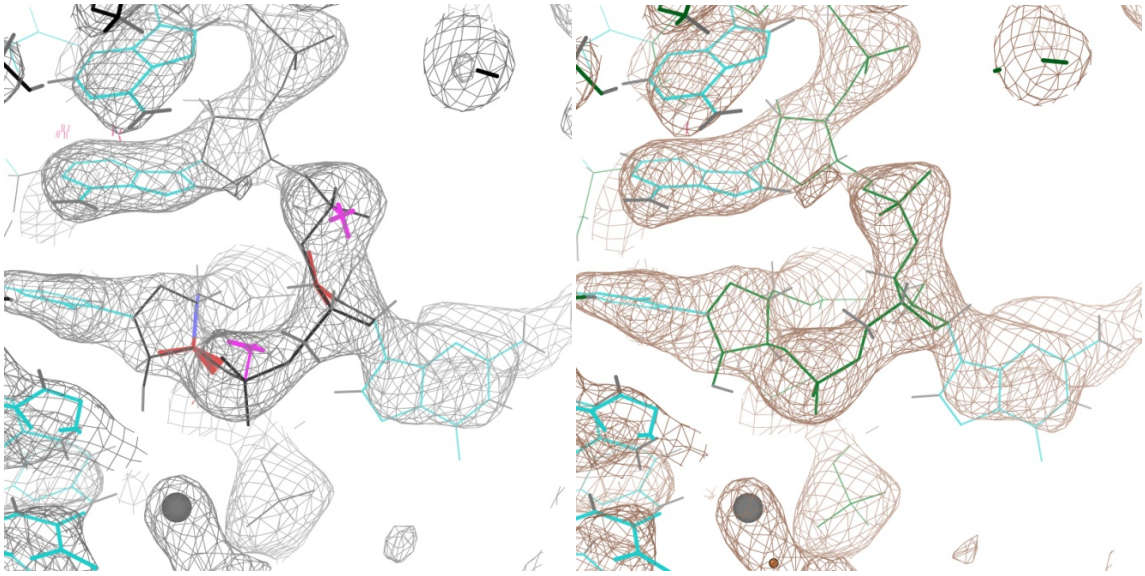


Figure 51: Correction of C163-G164 in 1CX0.

C163 and G164 had difficult to fix sugars with pucker outliers and bond angle outliers; C163 alone has four angle outliers (left). After ERRASER refinement, these outliers have been corrected, and the new models fit the electron density just as well (right).

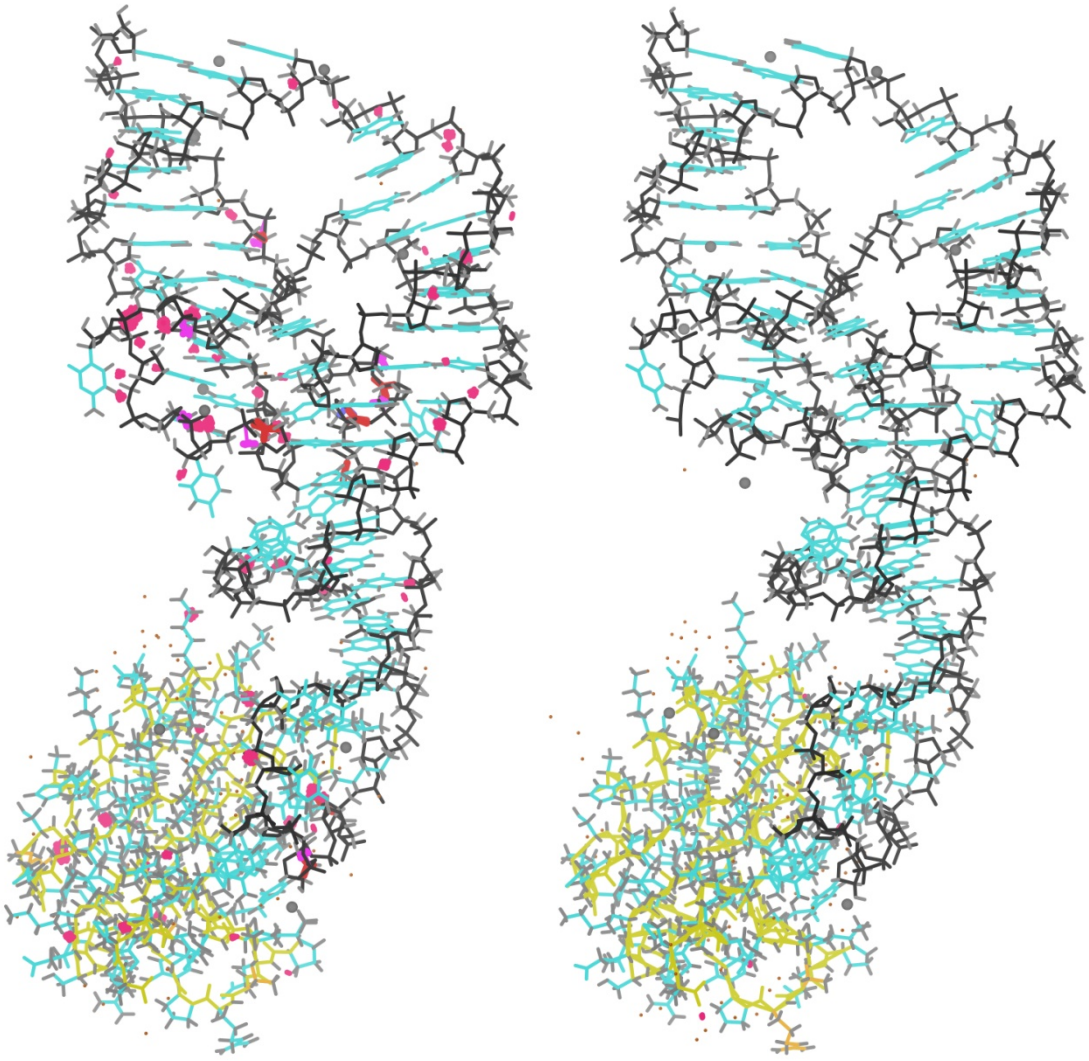


Figure 52: Correction overview in 1CX0.

Hot pink regions are impossible steric overlaps, magenta and purple crosses are sugar pucker outliers, and red and blue traces are angle outliers (left). Almost every error has been corrected in the new model (right).

Table 15: MolProbity statistics for 1CX0, original.

All-Atom Contacts	Clashscore, all atoms:	13.79		83rd percentile* (N=355, 2.30Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	3	3.53%	Goal: <1%
	Ramachandran outliers	0	0.00%	Goal: <0.05%
	Ramachandran favored	89	95.70%	Goal: >98%
	MolProbity score^	2.36		71st percentile* (N=8909, 2.30Å ± 0.25Å)
	Cβ deviations >0.25Å	0	0.00%	Goal: 0
	Bad backbone bonds:	0 / 379	0.00%	Goal: 0%
	Bad backbone angles:	0 / 472	0.00%	Goal: <0.1%
Nucleic Acid Geometry	Probably wrong sugar puckers:	8	11.11%	Goal: 0
	Bad backbone conformations#:	16	22.22%	Goal: <= 5%
	Bad bonds:	0 / 932	0.00%	Goal: 0%
	Bad angles:	14 / 1576	0.89%	Goal: <0.1%

Table 16: MolProbity statistics for 1CX0, after ERRASER refinement.

All-Atom Contacts	Clashscore, all atoms:	1		100th percentile* (N=355, 2.300Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	2	2.35%	Goal: <1%
	Ramachandran outliers	0	0.00%	Goal: <0.05%
	Ramachandran favored	91	97.85%	Goal: >98%
	MolProbity score^	1.11		100th percentile* (N=8909, 2.300Å ± 0.25Å)
	Cβ deviations >0.25Å	0	0.00%	Goal: 0
	Bad backbone bonds:	0 / 379	0.00%	Goal: 0%
	Bad backbone angles:	0 / 472	0.00%	Goal: <0.1%
Nucleic Acid Geometry	Probably wrong sugar puckers:	0	0.00%	Goal: 0
	Bad backbone conformations#:	9	12.50%	Goal: <= 5%
	Bad bonds:	0 / 932	0.00%	Goal: 0%
	Bad angles:	0 / 1576	0.00%	Goal: <0.1%

5.2 Ratcheted ribosome and refinement

As spectroscopic techniques improve, we have been able to learn much more about how the ribosome operates during translation. The ribosome controls movement of tRNA and mRNA by means of large-scale rearrangements. During elongation, tRNA translocation is facilitated by a ratchet-like motion of the small subunit relative to the large ribosomal subunit (Frank 2000). The ratcheting motion forces the tRNA into a temporary hybrid state, in which the CCA-stem is advanced by one site on the large subunit while the anticodon remains in the same place on the small subunit (Moazed 1989). Jamie Cate's lab collected data on a crystal with two 70S ribosomes in the same asymmetric unit, one in the unratcheted position and one in the fully ratcheted position. As part of our collaboration with them, Vincent Chen and I helped build the model for the protein and RNA portions, respectively (Dunkle 2011).

5.2.1 Initial structure

The first structure we worked on, Crystal A (3I1M,N,O,P; Zhang 2009), contained two entire, apo 70S ribosomes in the asymmetric unit, one of which was highly disordered. That made refinement very difficult, even though the resolution of 3.2Å was better than the initial *E. coli* apo 70S model at 3.5Å (2AVY,2AW4,7,B; Schuwirth 2005).

Several alternative refinement methods were tried and numerous corrections were implemented, but unfortunately, none of these were able to account for more than a handful of changes in the RNA.

A second crystal containing two 70S structures was solved to 3.0Å; this crystal, Crystal B, contained a partially ratcheted state and the inhibitor gentamicin. This structure became an initial testing ground for many of the crystallographic techniques for identifying and improving RNA backbone. Vincent Chen made many extensive corrections to the proteins of the better-ordered molecule (Chen 2010). Using a combination of Vince's RNA rotator tool (Chen 2010) and RNABC, I found corrections for 79 suites that reduced clashes or improved geometry. Many of the clash-busting corrections remained after refinement, but the corrections that improved geometry were undone, due to a bug in that version of PHENIX. We corrected the bug, but moved on to a more interesting crystal structure, as detailed below.

Only a few months after solving the 70S structure with gentamicin, the Cate lab produced a second crystal form at 3.0Å resolution, Crystal C, but this time with both copies of the ribosome well ordered. Crystal C contained two full 70S ribosomes, one with a P-site tRNA in the native, unratcheted state and one bound to Ribosome Release Factor (RRF) and a hybrid P/E-site tRNA, and in the most fully ratcheted state ever seen. Due to their similarity, the ordered ribosome from Crystal B was used as the starting

point for the new refinements of Crystal C. We first did 16 steps of iterated homology modeling and refinement, resulting in 70S structures with interesting, new features that had heretofore not been observed at atomic detail. For example, these models included the first hybrid P/E-site tRNA ever solved in a crystal structure, and also the first direct comparison between a P/P site-tRNA and a hybrid P/E site in the same asymmetric unit. The tRNA in the P/E hybrid site is bent at the D-stem by ~37°, and contains some unusual features, such as an interruption in the base stacking in the D-stem that accommodates the twist (Figure 53) .

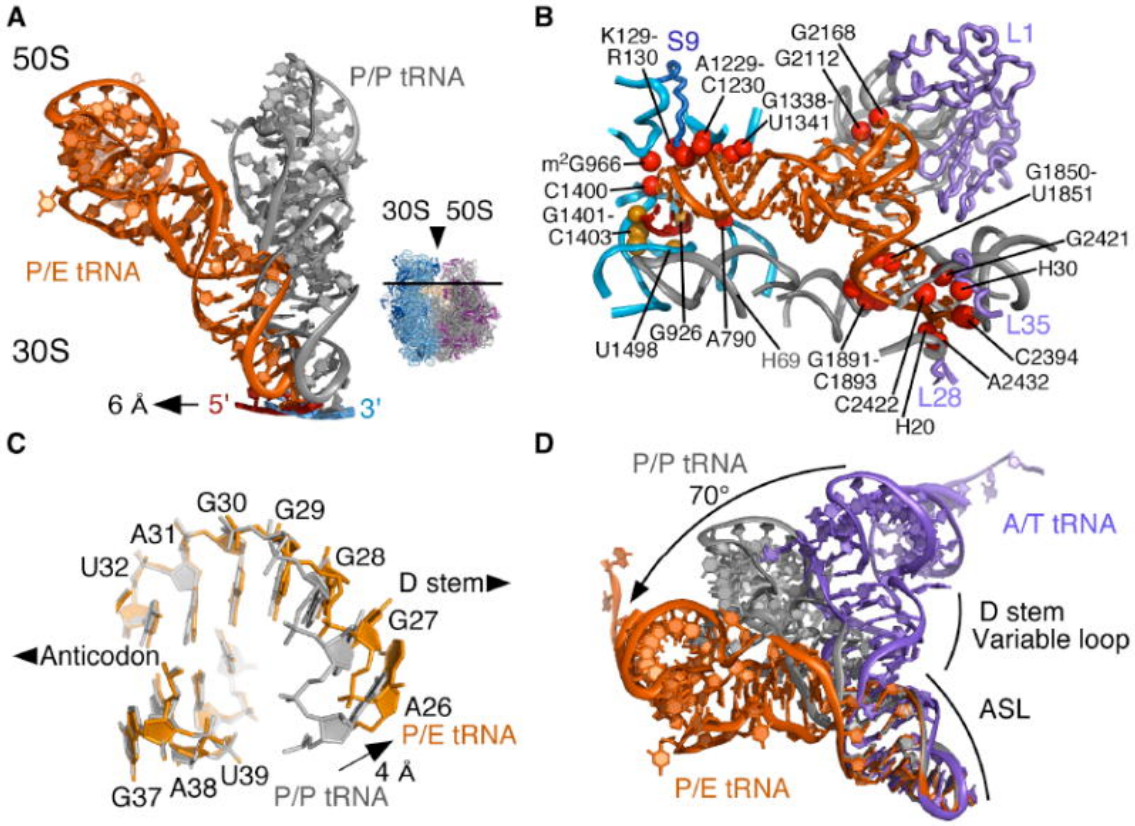


Figure 53: Conformation of tRNA in P/P state vs. P/E hybrid state.

(A) movement of the P/E tRNA (orange) compared to the P/P tRNA (grey) and mRNA. The location relative to the full 70S is shown in the inset. (B) All residue contacts to the P/E tRNA and mRNA (gold). (C) Detail of P/E tRNA vs. P/P tRNA. The P/E hybrid state has a wider turn around the helix between the anticodon and D loops. (D) Comparison of anticodon stem-loop (ASL) and D stem between P/E hybrid state, P/P state, and A/T hybrid state (purple) tRNAs. The latter represents tRNA normally bound to EF-Tu during its initial entry into the A-site (Voorhees 2010). Note the bend of 70° around the D-stem as tRNA goes from the A/T to P/E states.

The structure from Crystal C also provides the first high-resolution view of just how different the fully ratcheted state is from the native ribosome. Superimposing the 50S subunits, the 30S subunits are rotated 9° relative to each other (Figure 54). The head domain of the 30S swivels 4° in the direction of the E-site on the 50S subunit; combined, these shifts result in movements of 20Å on the ribosome periphery. The tRNA anticodon stem itself moves ~6Å relative to the 50S, breaking its interactions with the 23S helix H69, but maintaining contact with the 30S head and platform domains. Meanwhile, the top half of the tRNA (the CCA stem and T-loop) moves into the E-site, where it contacts the G2112 and G2168 of the 23S rRNA, approximately as if it were a normal E-site bound tRNA (Selmer 2006).

Overall, the ratcheting mechanism uses the secondary structural elements to control large-scale conformational rearrangements in the ribosome. RNA bridges, regions where the RNA from the 16S contacts that of the 23S, undergo great shifts, resulting in changes for bridges B2a, B4, and B7a. Helices 68, 76, and 42 in the large subunit also are disrupted by ratcheting, and the L1 and L11 arms move nearly 15Å (Figure 55). Helix h28 in the 16S rRNA appears to serve as a spring in the ratcheting process, helping position the 16S to intercalate with the mRNA, and thus act as a pawl for the ratchet to prevent reverse translocation from the 23S E to P sites.

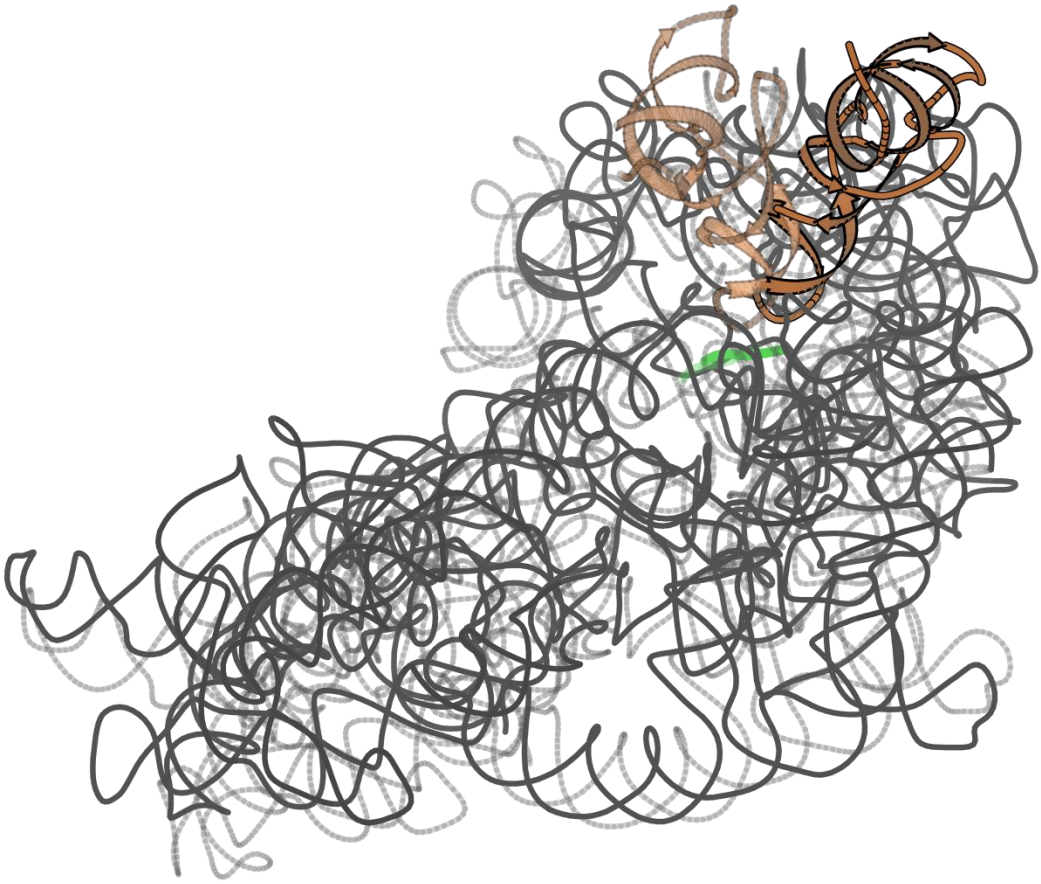


Figure 54: Rotation of 30S subunit during ratcheting.

The 30S ribosomal subunit rotates ~9° while moving from the unratcheted state (transparent grey) to the fully ratcheted state (dark grey). The P/P tRNA (transparent peach) bends along the D stem and shifts ~6Å down the mRNA to produce the P/E hybrid tRNA state (dark peach).

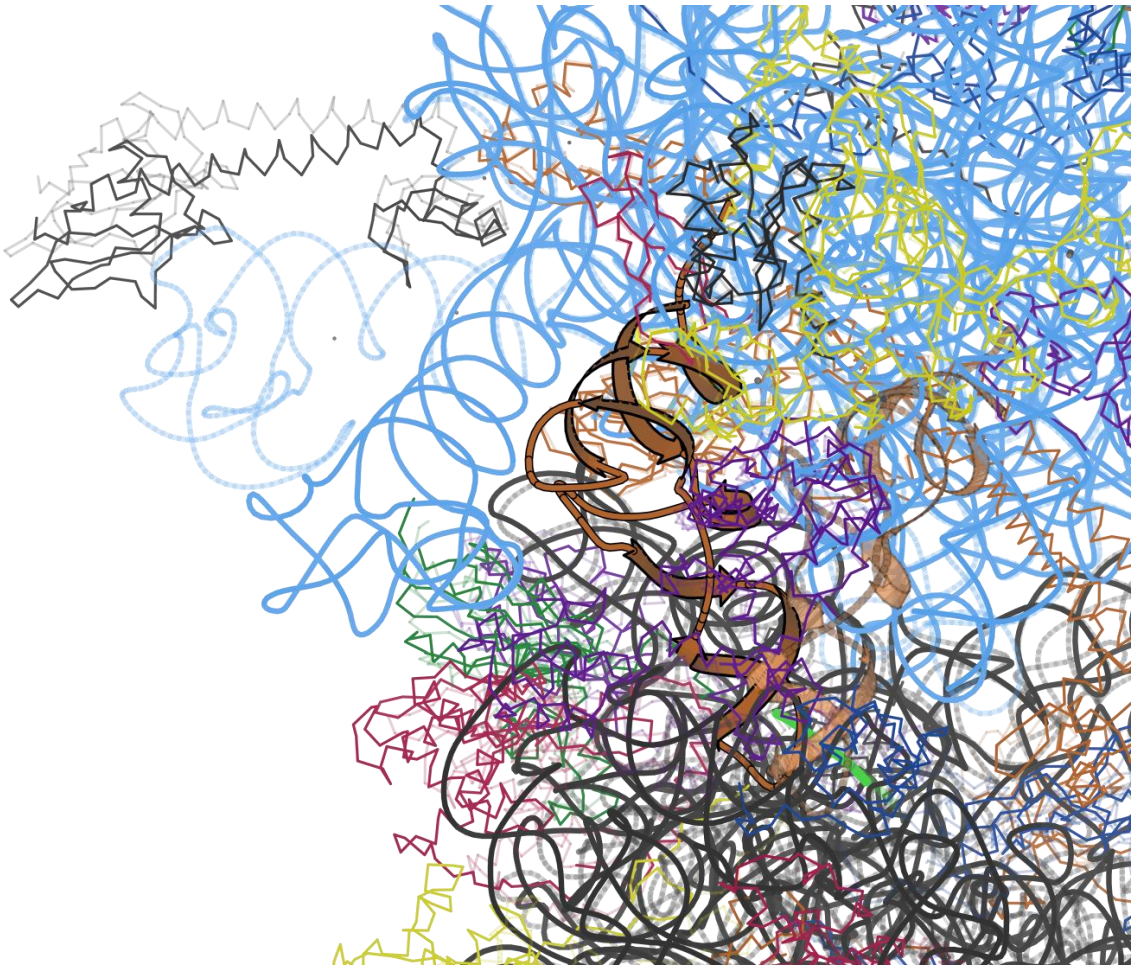


Figure 55: L1 stalk movement.

The L1 stalk in the unratcheted state (blue, transparent) shifts ~15Å to the ratcheted state (dark blue). This allows residues 2112 and 2168 (pink) on the L1 stalk to stack with the conserved G19-C56 tRNA tertiary basepair, stabilizing the hybrid P/E tRNA (peach).

As is common for large structures at low resolution, both the ratcheted and unratcheted structures had hundreds errors scattered throughout the structure, and after the sixteenth model (Model 16) of Crystal C was produced, Vincent Chen and I began a pattern of hand corrections to the protein and RNA, respectively, followed by refinement. After over a hundred corrections, I chose to focus on the most novel part of the this structure: the tRNA in the P/P and hybrid P/E sites. Starting with our Model 30, I began correcting the P/P-site tRNA since it looked like a normal $tRNA^{PHE}$ and there were prior high-resolution ribosomes with tRNA in the P/P site (PDB 2J00, Selmer 2006) and prior structures of $tRNA^{PHE}$ (PDB 3L0U, Byrne 2010). Overall, the P/P loop had several clashes and pucker outliers, and many geometry outliers; I corrected 16 geometry outliers between residues 40 and 47 by using the RNA rotator tool. Residue G44 had a sugar pucker outlier that was outside the density, and U45 contained a large clash between its H5'' and the OP2 of G46 . Once again making use of the RNA rotator tool, I was able to model a new version of the G44-U45 suite that corrected the poor sugar pucker and alleviated most of the clash (some overlap still remained, unfortunately); this had the added benefit of correcting the sugar pucker of U45 as well (Figure 56).

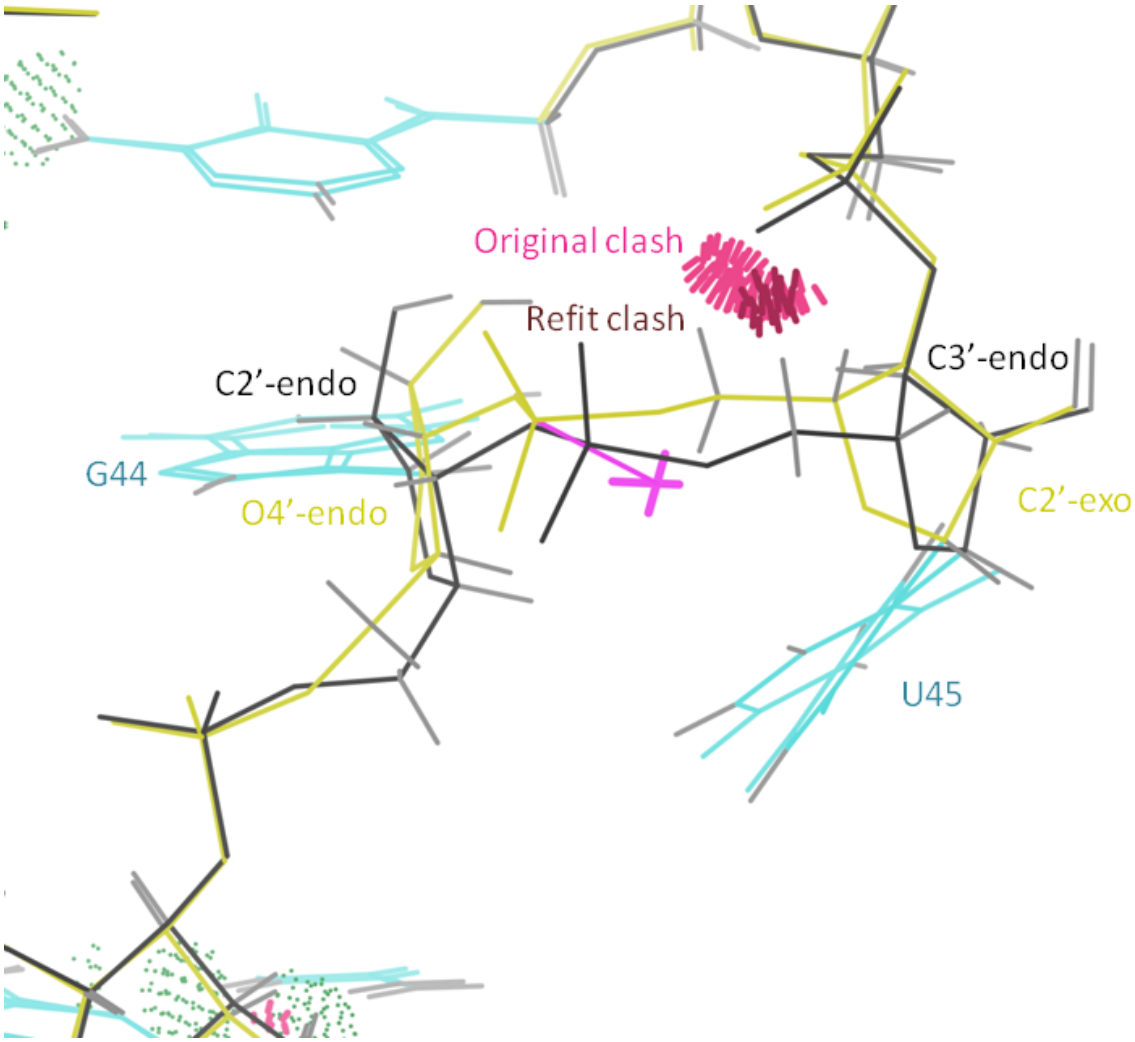


Figure 56: Correction of G44-U45 of tRNA^{PHE} in P/P site

The original model (yellow) has a pucker outlier indicator and a large pink clash; G44 is puckered at O4’-endo and U45 has C2’-exo pucker, both of which are very unusual. Manual correction via RNA rotator followed by PHENIX refinement yielded the black model, which corrects G44 to C2’-endo, and gives U45 a C3’-endo pucker while shifting and reducing the clash to the much smaller maroon region.

Crystal C had poor density for the P/E hybrid near the T-loop, around residues 48-64, so it was within this region that I did the bulk of my corrections, eventually extending them along the CCA stem to residue 72. The TΨC loop region was compressed in our early automated model builds, causing many steric clashes. In addition, residue 48 was too close to residue 59, causing a series of large clashes and distorting the TΨC loop even further. For Model 31, I superimposed the P/P-site T-loop and CCA stem on residues 48-72 to serve as a starting point, and suite by suite, I used the RNA rotator tool to rebuild the backbone to better fit the density and alleviate the clashes. In so doing, I created a structure with much better geometry and fewer clashes overall, but with two huge clashes between residues A58 and G18, and C48 and J21 (the T-loop clashing with the D-loop). Refinement shifted the D-loop and T-loop slightly to compensate for these clashes, and kept most of the other adjustments along the T-loop, resulting in better R-factors (Figure 57). By Model 38, we had a version of the hybrid P/E-site $tRNA^{PHE}$ that we deemed worthy of deposition.

Model 38 of Crystal C was deposited in the PDB, but because of the limitations on the number of chains in PDB files, it was deposited as 4 separate files: the 50S and 30S subunits were deposited as 3R8S and 4GD1 for the ratcheted ribosome, and 3R8T and 4GD2 for the unratcheted ribosome (the 30S subunits originally had PDB IDs 3R8N and

3R8O, but were superseded by 4GD1 and 4GD2, respectively, to correct an accidental omission in the coordinates).

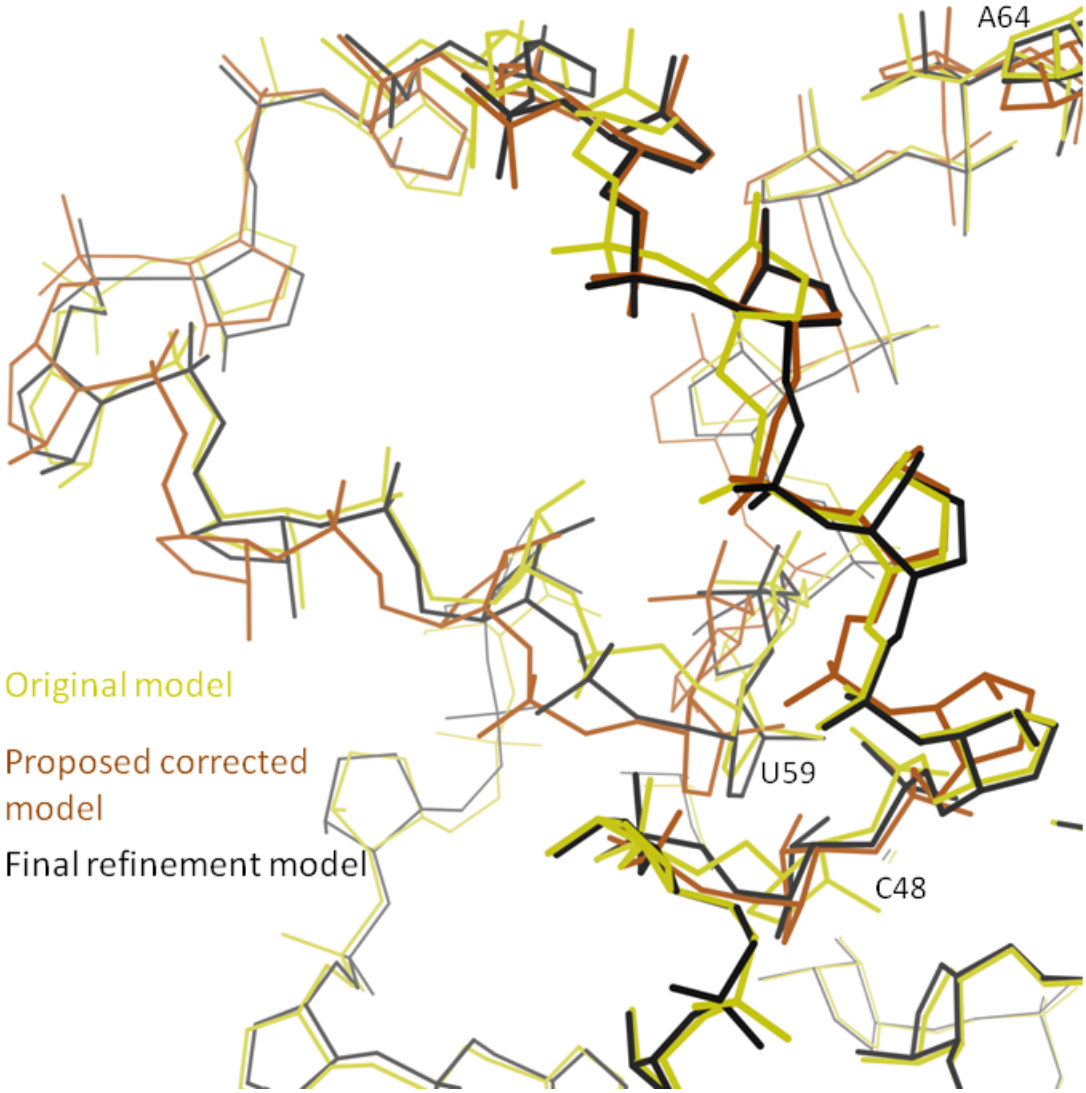


Figure 57: Remodeling of T-loop of tRNA^{PHE} in P/E site

The P/E-site tRNA from Model 30 (yellow) is compressed and has clashes between C48 and U59. The proposed Model 31 (orange) alleviates the clash and widens the entire loop by nearly 1.5Å all around. The final refined model (black) matches elements of the original structure and the new model, and, overall, results in fewer clashes (not pictured) than the original.

5.2.2 ERRASER refinement

Our deposited structure was as accurate as the best tools of the time could make it. Even so, there was still much room for improvement. Once we had successfully tested ERRASER on smaller structures, we decided to turn it to the biggest test of all—the 70S ribosome. Such a task had technical challenges—trying to do a full structure minimization on such a large structure would take a tremendous amount of time and memory. Furthermore, because ERRASER does not recognize the existence of protein in its current iteration, we would have to be very careful when modeling anything along the RNA-protein interface.

With this in mind, I split the ribosome structure into stem-loops of no more than 100 residues. In this way, I hoped to use ERRASER to correct small pieces of regular structure, and trust our pucker-specific PHENIX refinement parameters to take care of the rest. This approach paid off mightily; Figures 58 and 59 show the corrections to an OHO motif and S-motif structure, respectively, illustrating how well ERRASER rebuilding deals with clashes. Table 17 shows the MolProbity statistics for the deposited coordinates, while Table 18 shows them for my ERRASER/PHENIX refinement of the same data. The results show a clashscore reduction of 34.98, moving it from the 6th

percentile to the 97th. The number of poor rotamers decreases by 3%, while the Ramachandran outliers decrease by 4.5%. The protein geometry greatly improves, with 9 of the 14 Cβ deviations being fixed, as well as 59 bond angle outliers. For RNA, the number of incorrect sugar puckers goes down by 63 (.68%) and the number of suite outliers goes down by 519 (6.59%). The geometry parameters are interesting—the bond length outliers go down, from 69 to 0, but the bond angle outliers go up from 24 to 72. This may be likely due to the balance of geometry constraints battling the data constraints for dominance in the model; RNA and protein have different optimal balances between the data and parameter definitions of a given residue, and this results in two different weighting factors within PHENIX refinement. Using an intermediate weighting factor as a compromise will give the best overall results but has a side effect of causing angle outliers in the RNA. Even so, with such a large structure, anything but the strictest adherence to the geometry constraints will result in some new geometry outliers during refinement—for the total number of outliers to be only 72 in over 9000 residues is still amazing, particularly considering the number of pucker outliers. The most convincing evidence of the success of our new model is the R and R_{free} values: R and R_{free} of the deposited structure are 19.24% and 25.19%, respectively, while the new model has values of 18.80% and 24.21%. In a structure where a hundred residues can be

added without changing the R/R_{free} values, an improvement of .44% and .98% to R and R_{free}, respectively, is a significant improvement. At this time, neither of these structures has been deposited, pending improvements to the ERRASER code that will make it protein-aware.

Table 17: MolProbity Statistics for the full asymmetric unit of the deposited ratcheted and unratcheted 70S structures

All-Atom Contacts	Clashscore, all atoms:	42.8		6th percentile* (N=1784, all resolutions)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	2019	21.29%	Goal: <1%
	Ramachandran outliers	1364	11.95%	Goal: <0.05%
	Ramachandran favored	7804	68.36%	Goal: >98%
	MolProbity score^	3.97		3rd percentile* (N=27675, 0Å - 99Å)
	Cβ deviations >0.25Å	14	0.13%	Goal: 0
	Bad backbone bonds:	0 / 46371	0.00%	Goal: 0%
	Bad backbone angles:	96 / 57787	0.17%	Goal: <0.1%
Nucleic Acid Geometry	Probably wrong sugar puckers:	155	1.67%	Goal: 0
	Bad backbone conformations#:	2508	27.00%	Goal: <= 5%
	Bad bonds:	69 / 120754	0.06%	Goal: 0%
	Bad angles:	24 / 204324	0.01%	Goal: <0.1%

Table 18: MolProbity Statistics for the full asymmetric unit of the ERRASER/PHENIX refined ratcheted and unratcheted 70S structures

All-Atom Contacts	Clashscore, all atoms:	7.82		97th percentile* (N=75, 3.000Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	1767	18.52%	Goal: <1%
	Ramachandran outliers	859	7.41%	Goal: <0.05%
	Ramachandran favored	8875	76.53%	Goal: >98%
	MolProbity score^	3.17		64th percentile* (N=3130, 3.000Å ± 0.25Å)
	Cβ deviations >0.25Å	5	0.05%	Goal: 0
	Bad backbone bonds:	0 / 47120	0.00%	Goal: 0%
	Bad backbone angles:	37 / 58717	0.06%	Goal: <0.1%
Nucleic Acid Geometry	Probably wrong sugar puckers:	92	0.99%	Goal: 0
	Bad backbone conformations#:	1989	21.41%	Goal: <= 5%
	Bad bonds:	0 / 120771	0.00%	Goal: 0%
	Bad angles:	72 / 204354	0.04%	Goal: <0.1%

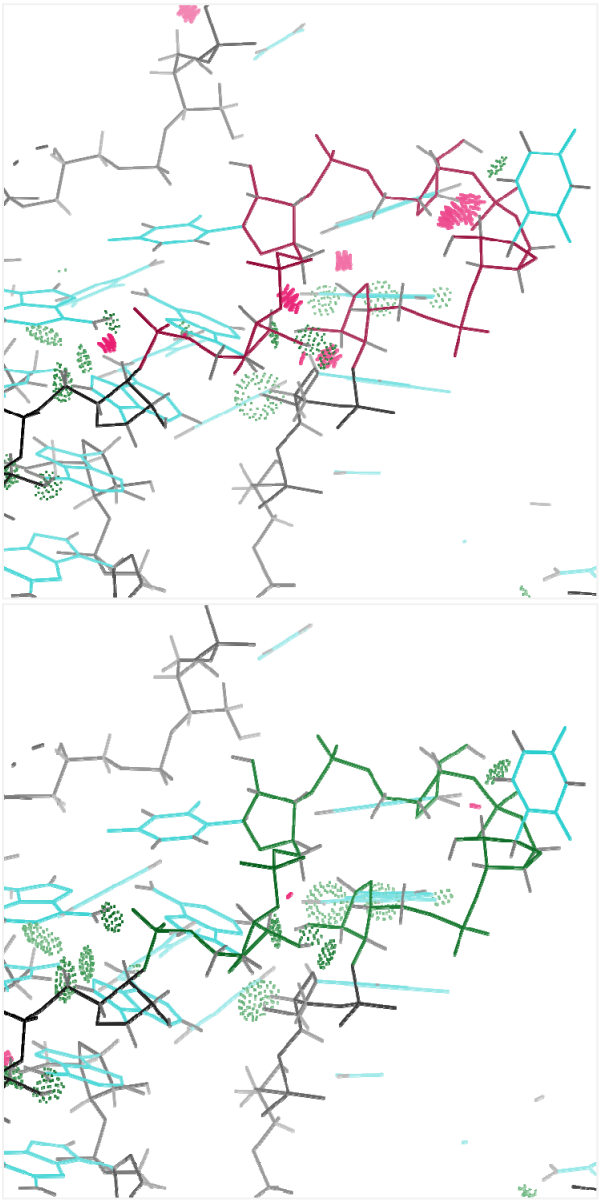


Figure 58: Sample corrections to ratcheted 23S rRNA.

The deposited structure contained a poorly modeled OHO loop in residues 402-406 (pink, top). The new model corrects nearly all the clashes, improves hydrogen bonding, and fixes the OHO loop to match its standard 1b4b6p2a suitestring (green, bottom).

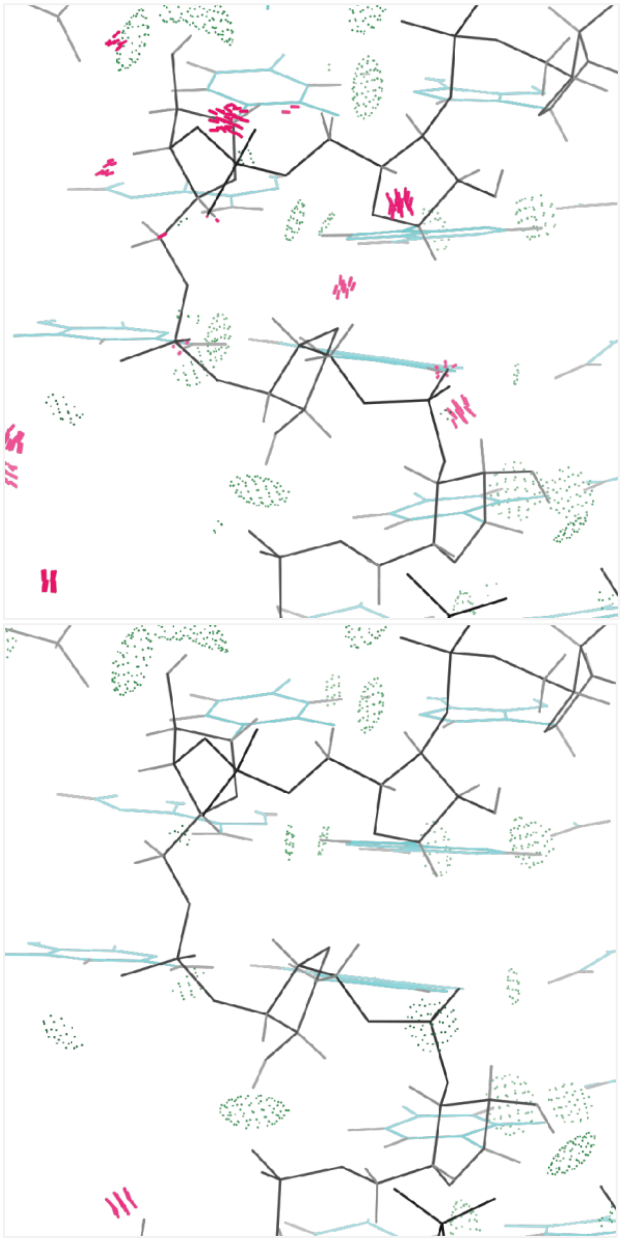


Figure 59: S-motif correction in ratcheted 16S rRNA.

The deposited structure fits the standard suitestring, but has many local clashes within the motif (top). The new model corrects nearly all the clashes and improves hydrogen bonding within the S-motif (bottom).

5.3 Histone mRNA stem-loop ternary complex

The histone mRNA contains a conserved stem-loop at the 3' end that is specifically recognized by human stem-loop binding protein (SLBP). Tan, *et al.* crystallized a ternary complex of human SLBP RNA binding domain, human 3' Exonuclease (3'hExo), and the histone mRNA stem-loop (SL) at 2.6Å resolution (PDB: 4HXH; Tan 2013) . Two versions of the complex were crystallized in the same asymmetric unit: the first contained the full ternary complex with SL bound to both 3'hExo and SLBP, while the second contained the SL bound only to 3'hExo. This fortuitously allowed a direct comparison between the SL both in the presence and absence of SLBP. The initial paper emphasized the role of the loop region of the SL in SLBP binding, but assumed that all four nucleotides in the loop had C2'-endo ribose puckers and were mostly static, leading to the assumption that the RNA backbone and a specific interaction with the base of G7 were responsible for recognition. Validation and correction of the structure using MolProbity, PHENIX, and ERRASER, has led to a new model which more accurately reflects the role of the loop region in SL recognition, and demonstrates the usefulness of these tools for practical modeling of RNA structure.

An initial run of the structure through MolProbity indicated three residues with incorrect puckers as indicated by their δ dihedrals; two of these were also ϵ outliers (see Table 19). Furthermore, the structure was found to have 18% bond angle outliers in the

RNA, most of which were in backbone of the loop region of the SL (Figure 60). Suitestring analysis revealed eight suites to have unknown conformations, seven of which were in the loop region. Out of these possible candidates for correction, the most compelling was residue 12 of the SL bound to 3'hExo only, as its perpendicular and δ values indicated that its ribose should be fit with a C3'-endo, in contrast to the initial assumption in the paper.

Table 19: Starting MolProbity Statistics for the full asymmetric unit of the SLBP-SL-3'hExo complex.

All-Atom Contacts	Clashscore, all atoms:	15.99	46th percentile* (N=1784, all resolutions)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.		
Protein Geometry	Poor rotamers	5.25%	Goal: <1%
	Ramachandran outliers	0.33%	Goal: <0.2%
	Ramachandran favored	96.67%	Goal: >98%
	Cβ deviations >0.25Å	0	Goal: 0
	MolProbity score^	2.46	49th percentile* (N=27675, 0Å - 99Å)
	Residues with bad bonds:	0.00%	Goal: 0%
	Residues with bad angles:	0.65%	Goal: <0.1%
Nucleic Acid Geometry	Probably wrong sugar puckers:	3	Goal: 0
	Bad backbone conformations#:	8	Goal: 0
	Residues with bad bonds:	0.00%	Goal: 0%
	Residues with bad angles:	18.18%	Goal: <0.1%

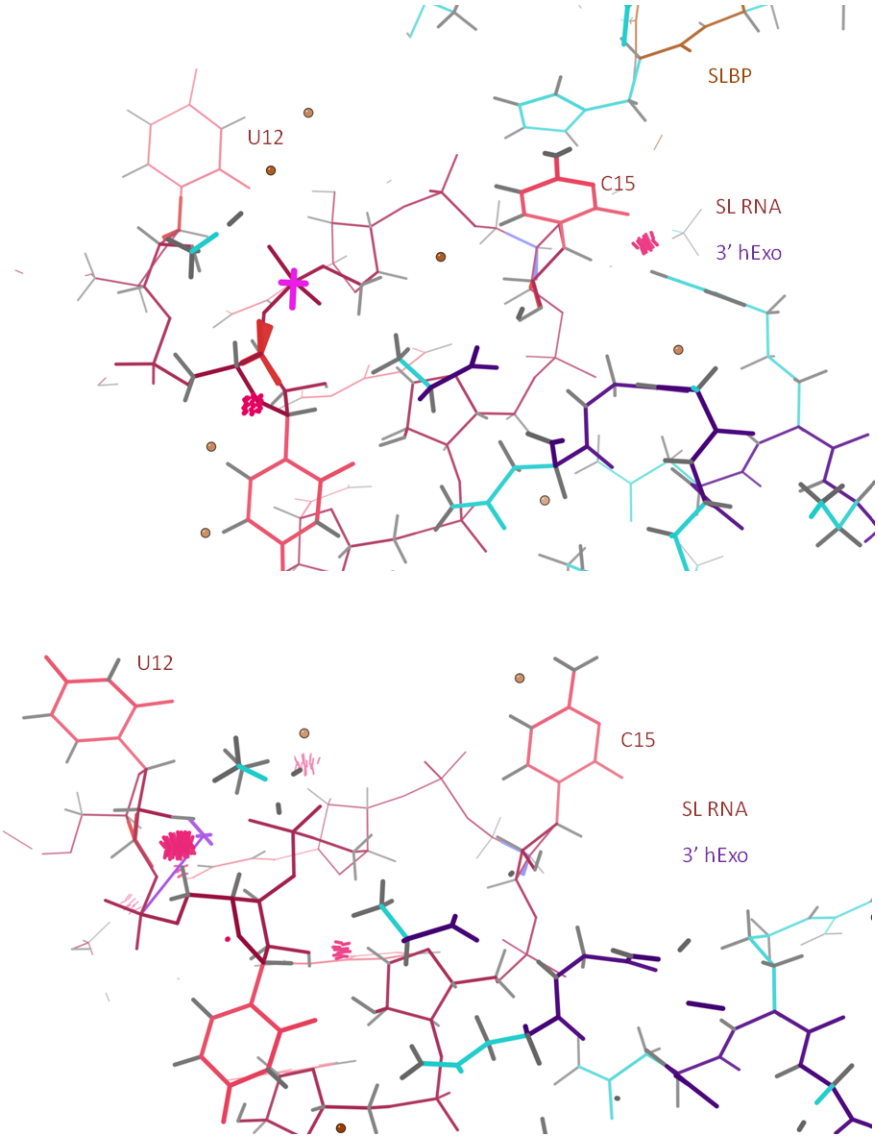


Figure 60: Original structure of SL RNA bound to SLBP and 3'hExo (top) and SL bound to 3'hExo only (bottom).

The stem-loop RNA (pink) interacts with 3'hExo (lilac); in the top panel, it also interacts with SLBP (peach). The magenta crosses indicate that the sugar pucker should be C2'-endo, while purple indicates a C3'-endo pucker; the pink spikes are clashes and the red and blue traces are bond angle geometry outliers.

I used ERRASER and PHENIX refinement to rebuild regions that showed up in MolProbity as having outlier suites or sugar puckers, resulting in a new model that improves the R and Rfree by ~0.5% (Figure 61), as well as improving the overall MolProbity statistics (Table 20). Angle geometry in several places was corrected in both loops. U13 of the SLBP-bound SL was initially modeled as a borderline C3'-endo/C2'-endo pucker and our validation criteria in MolProbity flagged it as a δ and ϵ outlier; the new model has a more pronounced C2'-endo sugar pucker and good δ and ϵ values. Similarly, the δ and ϵ dihedrals of C25 of SL bound only to 3'hExo were tweaked to make a more pronounced C2'-endo sugar pucker that satisfies our validation criteria and fits the density just as well. Both of these changes improve the model but don't change the reported pucker in the paper; rather, they emphasize it. Furthermore, they also move the local RNA backbone into recognized conformers; the U13 correction changes a **!!** to **4p**, and the C25 correction changes a **!!** to **2a**.

The most pronounced change in the new model is at U12 of the SLBP-free structure. Our validation criteria indicate a C3'-endo sugar pucker, rather than C2'-endo, and modeling it this way also gets rid of some bond angle outliers around the ribose; the new model also fits the density well. The implication, then is that this residue undergoes a pucker change from C3'-endo in the absence of SLBP to C2'-endo when interacting with SLBP by stacking on Y144. Superimposing the SL RNAs shows that this pucker

change is part of overall twisting in the loop region (Figure 62). When SLBP is present, U12 has a C2'-endo pucker and stacks on Y144, while C15 stacks with H195. When SLBP is absent, C15 rotates by about 30° towards where the H195 was (away from 3'hExo); U12 correspondingly rotates towards the 3'hExo and changes to C3'-endo pucker. U13 and U14 remain in roughly the same relative positions. The new model has been deposited into the PDB as 4L8R, superseding the original model 4HXH.

Table 20: MolProbity statistics for corrected 3'hExo-SL-SLBP ternary complex

All-Atom Contacts	Clashscore, all atoms:	5.93	99th percentile* (N=227, 2.604Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.		
Protein Geometry	Poor rotamers	4.38%	Goal: <1%
	Ramachandran outliers	0.17%	Goal: <0.2%
	Ramachandran favored	97.84%	Goal: >98%
	Cβ deviations >0.25Å	0	Goal: 0
	MolProbity score^	1.85	98th percentile* (N=6054, 2.604Å ± 0.25Å)
	Residues with bad bonds:	0.00%	Goal: 0%
	Residues with bad angles:	0.16%	Goal: <0.1%
Nucleic Acid Geometry	Probably wrong sugar puckers:	0	Goal: 0
	Bad backbone conformations#:	2	Goal: 0
	Residues with bad bonds:	0.00%	Goal: 0%
	Residues with bad angles:	0.00%	Goal: <0.1%

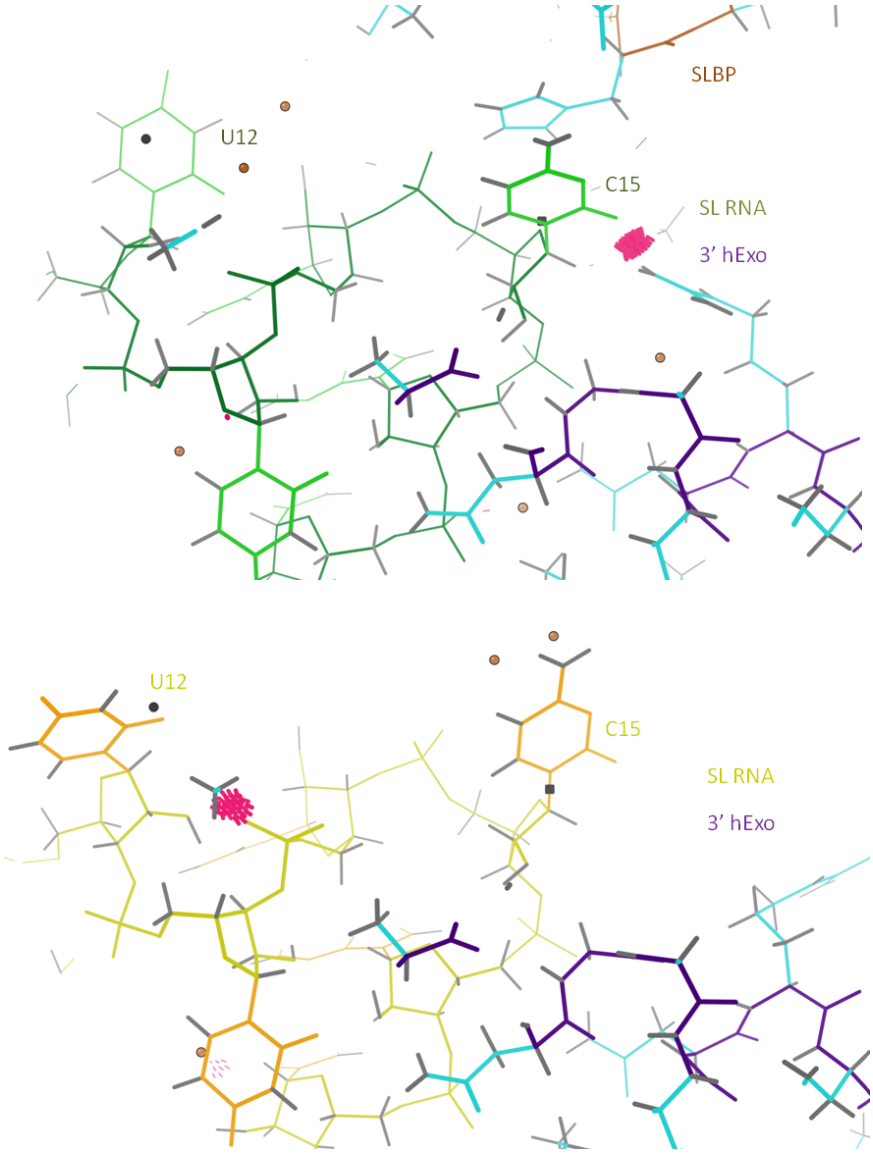


Figure 61: Corrected structures of SL RNA bound to SLBP and 3'hExo (top, green) and SL bound to only 3'hExo (bottom, gold).

The top and bottom panels correspond to the panels of the initial structure illustrated in Figure 60. The changes have removed the pucker and geometry outliers and greatly reduced the number of clashes. In addition, U12 now has a C2'-endo pucker in the top panel, and a C3'-endo pucker in the bottom one.

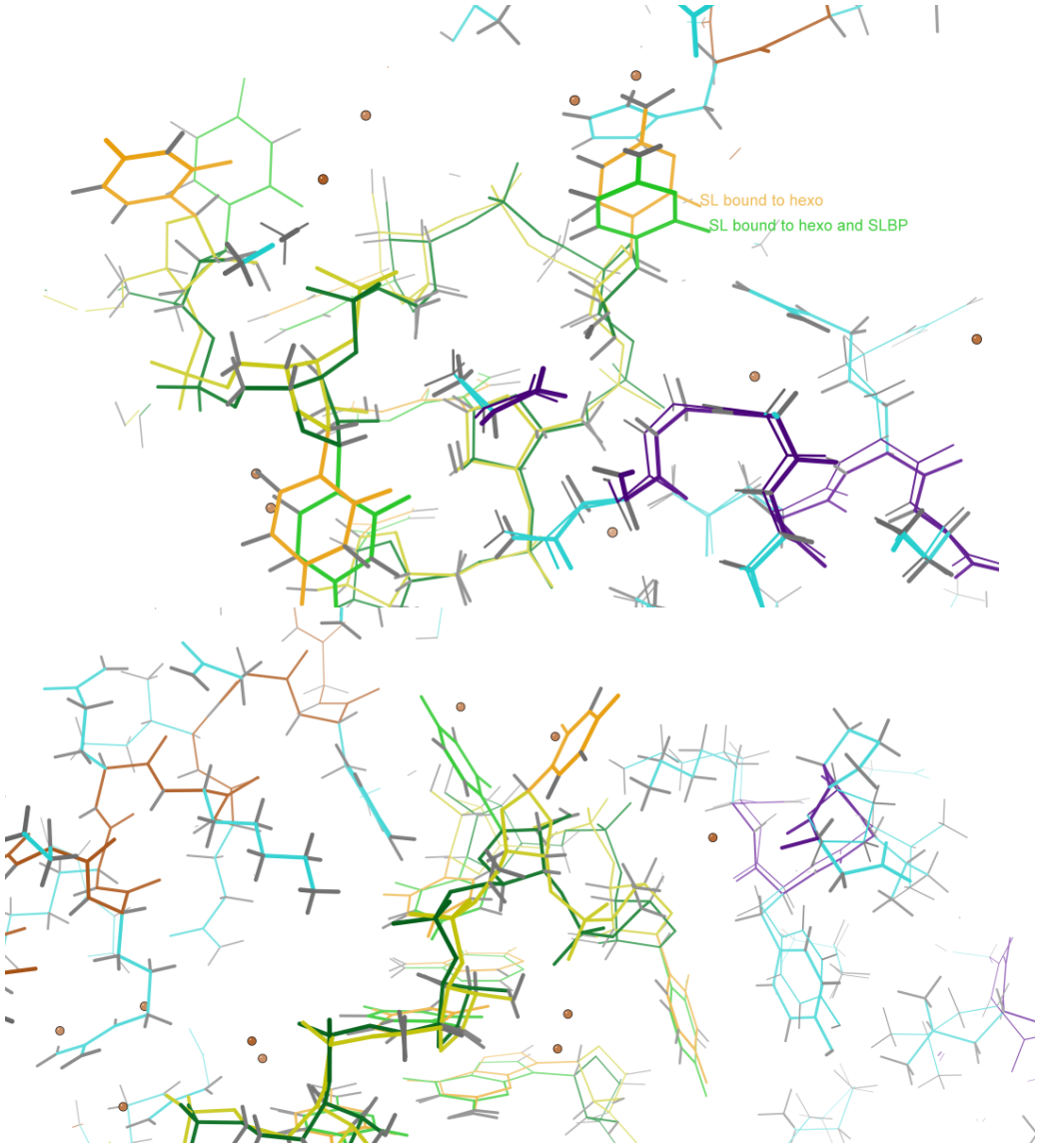


Figure 62: Two views comparing superimposed conformations of the loop with SLBP bound (green) and SLBP absent (peach)

Top shows the overall loop view with the superimposed RNA stem-loops. The bottom panel emphasizes the pucker change in U12, C2′-endo for the one that binds SLBP (green), and C3′-endo for the SLBP-free structure.

5.4 Discussion

The advancements made in pucker-specific PHENIX refinement and the model building procedures for ERRASER combine to form a powerful tool for building correct models of RNA structures. In the previous three examples, which represent very diverse interactions and cellular environments and over 21,000 residues, the new model was an improvement on the original model each time, correcting many suite and pucker outliers, alleviating steric clashes, and producing better fits to the data. Both PHENIX and ERRASER are freely available to academic researchers, and are intuitive to use. For even greater ease of use, we aim to have PHENIX automatically configurable to work with Rosetta and ERRASER from installation; users will still need to obtain both packages, but making them work together will no longer require difficult customization on the user's part. Our lab is also working on a major rewrite of the ERRASER/Rosetta code to allow simultaneous awareness of RNA and protein, essential for optimizing structures of the biologically crucial class of ribonucleoprotein complexes.

6. Conclusions and Future directions

The work presented in this dissertation represents the growth and development of the nascent studies on RNA backbone. When I embarked on this journey, my lofty goal was to elucidate RNA-protein interactions from the perspective of RNA backbone, but I faced two major challenges: a lack of unified language and a dearth of available high-quality structural data. The former was resolved by the lab setting up collaborations which brought together the three premier groups in RNA backbone study, spanning several universities across two continents; each group had developed their own way of dividing the backbone and their own language for describing it. Under the auspices of the RNA Ontology Consortium, we established a new consensus modular nomenclature for referring to RNA backbone structure, which could fully describe each group's initial divisions of the backbone as well as allow bases and backbone dihedrals to be incorporated into an easily parse-able string.

I set about tackling the second problem, lack of high-quality data, by helping create new tools (such as RNABC) to improve existing models, thereby improving the amount of high-quality data available. The amount of data was still too low, so our lab implemented new methods and parameters to aid RNA crystallographers and NMR spectroscopists to build better initial models and get more accurate models from refinements. These tools required extensive collaborations with other labs, some

working on RNA structures, some on computer science, but ultimately resulted in substantial improvements to the production and refinement of RNA structures in the PHENIX refinement package and the inclusion of better validation parameters in MolProbity.

Having helped invent and implement a new language that described the RNA backbone data, new tools to improve existing data, and new methods to acquire new high-quality data, I finally analyzed the way RNA backbone interacted with protein. The results show that sugar pucker plays a major role in interactions along an interface, particularly due to its influence on the accessibility of the 2'OH. Furthermore, most of the interactions between RNA and protein occur through hydrogen bonding and salt bridges along the backbone, rather than through stacking or H-bonding with bases. In addition, this work yielded new motifs that occur along protein-RNA interfaces and elsewhere in RNA structure, which will help crystallographers build new models and computational chemists design new interfaces.

In the future, I hope to expand the motifs and interfaces into Rosetta in such a way that RNA structure design could be as effective as protein design is now. Some steps have already been taken by ERRASER, but work needs to be done to integrate the RNA and protein sides of Rosetta before RNA-protein interface design can become a reality; Swati Jain, Steven Lewis and myself are working to bring this to fruition. Proper

integration of protein and RNA would also be useful for refinement, and I look forward to further improving the way PHENIX handles RNA-protein complexes by introducing separate weights for how RNA and protein models are handled with respect to the data and the starting model.

Concerning RNA biochemistry, future work will involve looking at some of the new motifs, particularly the OHO loop, and finding how resistant they are to local perturbations like SNPs or deletions, or global perturbations like truncation. It would be interesting to find which motifs are robust through such perturbations, indicating they are good, low-energy conformations that form due to their backbone geometry more than sequence. It would also be interesting to see which local perturbations resulted in variant motifs, such as using SNPs to successively mutate a UNCG tetraloop to bind to a GNRA tetraloop receptor and measuring how the backbone changes each step of the way.

The bioinformatics side of this work has one particularly obvious direction: adapt our work on the RNA backbone to DNA. This is harder than it seems, as while there is almost double the number of DNA structures vs. RNA, there is a severe lack of nonstandard DNA structures. Two attempts by our lab have been started and abandoned on applying the RNA database methods to DNA, because the DNA backbone dihedrals do not cluster clearly; it is uncertain whether this is due to poor

modeling or a more continuous conformer sampling region open to DNA since it lacks the 2'OH. Recent work between myself and my student assistant Kemberly McKinney has yielded strong clustering for DNA backbone dihedrals involved in G-quadruplexes, even to the point of accurately discriminating between antiparallel and parallel G-quadruplexes. Buoyed by this success, we are now looking for other backbone clusters in regions of known DNA motifs, while collaborating with Bradley Hintze and his student assistant Shouri Gottiparthi to find DNA backbone patterns associated with Hoogsteen basepairs. We are also working with Hashim Al-Hashimi to validate our results from this work via NMR.

Finally, the RNA databases, suite conformer, and motif libraries and their distribution must continue to evolve as new structures become available; we will continue to update the RNA databases as we have in the past. Swati Jain is already working on an update to RNA11 and has promoted the six of the eight previous wannabe suite conformations to full acceptance, as well as identifying two new wannabe suite conformations. I have begun implementing RNA backbone motif identification and searches to MolProbity that would allow crystallographers to have the motifs in their structure identified, as well as pointing out deviations from the normal definition (e.g., a **4b** conformation built as a **1b**) that may affect how they continue to refine the model. With this tool, and the other tools and parameters in PHENIX, MolProbity and Rosetta,

my work will help future crystallographers quickly obtain more accurate models of RNA structures and RNA-protein complexes.

References

Adams, Paul D., Pavel V. Afonine, Gábor Bunkóczi, Vincent B. Chen, Ian W. Davis, Nathaniel Echols, Jeffrey J. Headd, Li-Wei Hung, Gary J. Kapral, Ralf W. Grosse-Kunstleve, Airlie J. McCoy, Nigel W. Moriarty, Robert Oeffner, Randy J. Read, David C. Richardson, Jane S. Richardson, Thomas C. Terwilliger and Peter H. Zwart. "Phenix: A Comprehensive Python-Based System for Macromolecular Structure Solution." *Acta Crystallographica. Section D, Biological Crystallography* 66, no. Pt 2 (2010): 213-221.

Allen, F W. "The Biochemistry of the Nucleic Acids, Purines, and Pyrimidines." *Annual Review of Biochemistry* 10, no. 1 (1941): 221-244.

Allen, F. H., S. Bellard, M. D. Brice, B. A. Cartwright, A. Doubleday, H. Higgs, T. Hummelink, B. G. Hummelink-Peters, O. Kennard, W. D. S. Motherwell, J. R. Rodgers and D. G. Watson. "The Cambridge Crystallographic Data Centre: Computer-Based Search, Retrieval, Analysis and Display of Information." *Acta Crystallographica Section B* 35, no. 10 (1979): 2331-2339.

Altona, C. and M. Sundaralingam. "Conformational Analysis of the Sugar Ring in Nucleosides and Nucleotides. New Description Using the Concept of Pseudorotation." *Journal of the American Chemical Society* 94, no. 23 (1972): 8205-8212.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215, no. 3 (1990): 403-410.

Avery, Oswald T., Colin M. MacLeod and Maclyn McCarty. "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii." *The Journal of Experimental Medicine* 79, no. 2 (1944): 137-158.

Ban, Nenad, Poul Nissen, Jeffrey Hansen, Peter B. Moore and Thomas A. Steitz. "The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution." *Science* 289, no. 5481 (2000): 905-920.

Bandziulis, R J, M S Swanson and G Dreyfuss. "Rna-Binding Proteins as Developmental Regulators." *Genes & Development* 3, no. 4 (1989): 431-437.

Bashkirov, Vladimir I., Harry Scherthan, Jachen A. Solinger, Jean-Marie Buerstedde and Wolf-Dietrich Heyer. "A Mouse Cytoplasmic Exoribonuclease (Mxrn1p) with Preference for G4 Tetraplex Substrates." *The Journal of Cell Biology* 136, no. 4 (1997): 761-773.

Batey, Robert T., Sunny D. Gilbert and Rebecca K. Montange. "Structure of a Natural Guanine-Responsive Riboswitch Complexed with the Metabolite Hypoxanthine." *Nature* 432, no. 7015 (2004): 411-415.

Been, Michael D. and Thomas R. Cech. "One Binding Site Determines Sequence Specificity of Tetrahymena Pre-Rrna Self-Splicing, Trans-Splicing, and Rna Enzyme Activity." *Cell* 47, no. 2 (1986): 207-216.

Berman, H. M., W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S. H. Hsieh, A. R. Srinivasan and B. Schneider. "The Nucleic Acid Database. A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids." *Biophysical Journal* 63, no. 3 (1992): 751-759.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov and Philip E. Bourne. "The Protein Data Bank." *Nucleic Acids Research* 28, no. 1 (2000): 235-242.

Bernstein, Emily, Amy A. Caudy, Scott M. Hammond and Gregory J. Hannon. "Role for a Bidentate Ribonuclease in the Initiation Step of Rna Interference." *Nature* 409, no. 6818 (2001): 363-366.

Blaha, Gregor, Güliz Gürel, Susan J. Schroeder, Peter B. Moore and Thomas A. Steitz. "Mutations Outside the Anisomycin-Binding Site Can Make Ribosomes Drug-Resistant." *Journal of Molecular Biology* 379, no. 3 (2008): 505-519.

Blanchard, Scott C. and Joseph D. Puglisi. "Solution Structure of the a Loop of 23s Ribosomal Rna." *Proceedings of the National Academy of Sciences* 98, no. 7 (2001): 3720-3725.

Boelens, R., T. M. G. Koning, G. A. van der Marel, J. H. van Boom and R. Kaptein. "Iterative Procedure for Structure Determination from Proton-Proton Noes Using a Full Relaxation Matrix Approach. Application to a DNA Octamer." *Journal of Magnetic Resonance (1969)* 82, no. 2 (1989): 290-308.

Bolton, Philip H. and David R. Kearns. "Hydrogen Bonding Interactions of Polyamines with the 2′ Oh of Rna." *Nucleic Acids Research* 5, no. 4 (1978): 1315-1324.

Borgias, Brandan A., Miriam Gochin, Deborah J. Kerwood and Thomas L. James. "Relaxation Matrix Analysis of 2d Nmr Data." *Progress in Nuclear Magnetic Resonance Spectroscopy* 22, no. 1 (1990): 83-100.

Borgias, Brandan A. and Thomas L. James. "[9] Two-Dimensional Nuclear Overhauser Effect: Complete Relaxation Matrix Analysis." In *Methods in Enzymology*, edited by J. Oppenheimer Norman and L. James Thomas, Volume 176, 169-183: Academic Press, 1989.

Borgias, Brandan A. and Thomas L. James. "Mardigras-a Procedure for Matrix Analysis of Relaxation for Discerning Geometry of an Aqueous Structure." *Journal of Magnetic Resonance (1969)* 87, no. 3 (1990): 475-487.

Borovinskaya, Maria A., Raj D. Pai, Wen Zhang, Barbara S. Schuwirth, James M. Holton, Go Hirokawa, Hideko Kaji, Akira Kaji and Jamie H. Doudna Cate. "Structural Basis for Aminoglycoside Inhibition of Bacterial Ribosome Recycling." *Nat Struct Mol Biol* 14, no. 8 (2007): 727-732.

Branch, A D, B J Benenfeld and H D Robertson. "Ultraviolet Light-Induced Crosslinking Reveals a Unique Region of Local Tertiary Structure in Potato Spindle Tuber Viroid and Hela 5s Rna." *Proceedings of the National Academy of Sciences* 82, no. 19 (1985): 6590-6594.

Brunger, Axel T., Paul D. Adams, G. Marius Clore, Warren L. DeLano, Piet Gros, Ralf W. Grosse-Kunstleve, Jian-Sheng Jiang, John Kuszewski, Michael Nilges, Navraj S. Pannu, Randy J. Read, Luke M. Rice, Thomas Simonson and Gregory L. Warren. "Crystallography & Nmr System: A New Software Suite for Macromolecular Structure Determination." *Acta Crystallographica Section D* 54, no. 5 (1998): 905-921.

Bulkley, David, C. Axel Innis, Gregor Blaha and Thomas A. Steitz. "Revisiting the Structures of Several Antibiotics Bound to the Bacterial Ribosome." *Proceedings of the National Academy of Sciences* 107, no. 40 (2010): 17158-17163.

Byrne, Robert T., Andrey L. Konevega, Marina V. Rodnina and Alfred A. Antson. "The Crystal Structure of Unmodified Trnaphe from Escherichia Coli." *Nucleic Acids Research* 38, no. 12 (2010): 4154-4162.

Carter, Andrew P., William M. Clemons, Ditlev E. Brodersen, Robert J. Morgan-Warren, Thomas Hartsch, Brian T. Wimberly and V. Ramakrishnan. "Crystal Structure of an Initiation Factor Bound to the 30s Ribosomal Subunit." *Science* 291, no. 5503 (2001): 498-501.

Caspersson, T. and J. Schultz. "Pentose Nucleotides in the Cytoplasm of Growing Tissues." *Nature* 143, (1939): 602-603.

Cate, J. H., A. R. Gooding, E. Podell, K. Zhou, B. L. Golden, A. A. Szewczak, C. E. Kundrot, T. R. Cech and J. A. Doudna. "Rna Tertiary Structure Mediation by Adenosine Platforms." *Science (New York, N.Y.)* 273, no. 5282 (1996): 1696-1699.

Cavanagh, J., W.J. Fairbrother, A.G. Palmer, M. Rance and N.J. Skelton. *Protein Nmr Spectroscopy: Principles and Practice*. 2nd Ed. ed. San Diego: Academic Press, 2006.

Chargaff, E., R. Lipshitz and C. Green. "Composition of the Desoxypentose Nucleic Acids of Four Gera of Sea-Urchin." *J. Biol. Chem.* 195, (1952): 155-160.

Chen, Vincent B. "Building Better Backbones: Visualizations, Analyses, and Tools for Higher Quality Macromolecular Structure Models." Duke University, 2010.

Chen, Vincent B., W. Bryan Arendall, 3rd, Jeffrey J. Headd, Daniel A. Keedy, Robert M. Immormino, Gary J. Kapral, Laura W. Murray, Jane S. Richardson and David C. Richardson. "Molprobity: All-Atom Structure Validation for Macromolecular Crystallography." *Acta Crystallographica. Section D, Biological Crystallography* 66, no. Pt 1 (2010): 12-21.

Chen, Vincent B., Ian W. Davis and David C. Richardson. "King (Kinemage, Next Generation): A Versatile Interactive Molecular and Scientific Visualization Program." *Protein Science* 18, no. 11 (2009): 2403-2409.

Cheong, Chaejoon, Gabriele Varani and Ignacio Tinoco. "Solution Structure of an Unusually Stable Rna Hairpin, 5ggac(Uucg)Gucc." *Nature* 346, no. 6285 (1990): 680-682.

Chou, Fang-Chieh, Parin Sripakdeevong, Sergey M. Dibrov, Thomas Hermann and Rhiju Das. "Correcting Pervasive Errors in Rna Crystallography through Enumerative Structure Prediction." *Nat Meth* 10, no. 1 (2013): 74-76.

Correll, C. C., J. Beneken, M. J. Plantinga, M. Lubbers and Y. L. Chan. "The Common and the Distinctive Features of the Bulged-G Motif Based on a 1.04 a Resolution Rna Structure." *Nucleic Acids Research* 31, no. 23 (2003): 6806-6818.

Correll, Carl C., Alexander Munishkin, Yuen-Ling Chan, Zhong Ren, Ira G. Wool and Thomas A. Steitz. "Crystal Structure of the Ribosomal Rna Domain Essential for Binding Elongation Factors." *Proceedings of the National Academy of Sciences* 95, no. 23 (1998): 13436-13441.

Correll, Carl C. and Kerren Swinger. "Common and Distinctive Features of Gnra Tetraloops Based on a Guaa Tetraloop Structure at 1.4 Å Resolution." *RNA* 9, no. 3 (2003): 355-363.

Crick, F. H. C. "Ideas on Protein Synthesis." In *Symposia of the Society for Experimental Biology Symposium XII: The Biological Replication of Macromolecules*, 138-163. University College London: Cambridge University Press, 1958.

Crick, Francis. "Central Dogma of Molecular Biology." *Nature* 227, no. 5258 (1970): 561-563.

Darty, Kévin, Alain Denise and Yann Ponty. "Varna: Interactive Drawing and Editing of the Rna Secondary Structure." *Bioinformatics* 25, no. 15 (2009): 1974-1975.

Davis, Ian W., W. Bryan Arendall, David C. Richardson and Jane S. Richardson. "The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances." *Structure (London, England : 1993)* 14, no. 2 (2006): 265-274.

Davis, Ian W., Andrew Leaver-Fay, Vincent B. Chen, Jeremy N. Block, Gary J. Kapral, Xueyi Wang, Laura W. Murray, W. Bryan Arendall, 3rd, Jack Snoeyink, Jane S. Richardson and David C. Richardson. "Molprobity: All-Atom Contacts and Structure Validation for Proteins and Nucleic Acids." *Nucleic Acids Research* 35, no. Web Server issue (2007): W375-383.

Davis, Ian W., Laura Weston Murray, Jane S. Richardson and David C. Richardson. "Molprobity: Structure Validation and All-Atom Contact Analysis for Nucleic Acids and Their Complexes." *Nucleic Acids Research* 32, no. suppl 2 (2004): W615-W619.

Delagoutte, Benedicte, Dino Moras and Jean Cavarelli. "Trna Aminoacylation by Arginyl-Trna Synthetase: Induced Conformations During Substrates Binding." *EMBO J* 19, no. 21 (2000): 5599-5610.

DeLano, Warren L. "The Pymol Molecular Graphics System." (2002).

Duarte, C. M., L. M. Wadley and A. M. Pyle. "Rna Structure Comparison, Motif Search and Discovery Using a Reduced Representation of Rna Conformational Space." *Nucleic Acids Research* 31, no. 16 (2003): 4755-4761.

Duarte, Carlos M. and Anna Marie Pyle. "Stepping through an Rna Structure: A Novel Approach to Conformational Analysis." *Journal of Molecular Biology* 284, no. 5 (1998): 1465-1478.

Dunkle, Jack A., Leyi Wang, Michael B. Feldman, Arto Pulk, Vincent B. Chen, Gary J. Kapral, Jonas Noeske, Jane S. Richardson, Scott C. Blanchard and Jamie H. Doudna Cate. "Structures of the Bacterial Ribosome in Classical and Hybrid States of Trna Binding." *Science (New York, N.Y.)* 332, no. 6032 (2011): 981-984.

Egli, Martin and Wolfram Saenger. *Principles of Nucleic Acid Structure* Springer Advanced Texts in Chemistry: Springer, 1983.

Ellington, Andrew D. and Jack W. Szostak. "In Vitro Selection of Rna Molecules That Bind Specific Ligands." *Nature* 346, no. 6287 (1990): 818-822.

Emsley, Paul and Kevin Cowtan. "Coot: Model-Building Tools for Molecular Graphics." *Acta Crystallographica Section D* 60, no. 12 Part 1 (2004): 2126-2132.

Eulberg, Dirk and Sven Klussmann. "Spiegelmers: Biostable Aptamers." *ChemBioChem* 4, no. 10 (2003): 979-983.

Ferre-D'Amare, Adrian R., Kaihong Zhou and Jennifer A. Doudna. "Crystal Structure of a Hepatitis Delta Virus Ribozyme." *Nature* 395, no. 6702 (1998): 567-574.

Fire, Andrew, SiQun Xu, Mary K. Montgomery, Steven A. Kostas, Samuel E. Driver and Craig C. Mello. "Potent and Specific Genetic Interference by Double-Stranded Rna in Caenorhabditis Elegans." *Nature* 391, no. 6669 (1998): 806-811.

Frank, Joachim and Rajendra Kumar Agrawal. "A Ratchet-Like Inter-Subunit Reorganization of the Ribosome During Translocation." *Nature* 406, no. 6793 (2000): 318-322.

Gelbin, Anke, Bohdan Schneider, Lester Clowney, Shu-Hsin Hsieh, Wilma K. Olson and Helen M. Berman. "Geometric Parameters in Nucleic Acids: Sugar and Phosphate Constituents." *Journal of the American Chemical Society* 118, no. 3 (1996): 519-529.

Gielis, J. *Inventing the Circle: The Geometry of Nature*. Antwerp: Geniaal Press, 2003.

Gilbert, Sunny D., Francis E. Reyes, Andrea L. Edwards and Robert T. Batey. "Adaptive Ligand Binding by the Purine Riboswitch in the Recognition of Guanine and Adenine Analogs." *Structure* 17, no. 6 (2009): 857-868.

Gilbert, Walter. "Origin of Life: The Rna World." *Nature* 319, no. 6055 (1986): 618-618.

Golden, Barbara L., Hajeong Kim and Elaine Chase. "Crystal Structure of a Phage Twort Group I Ribozyme–Product Complex." *Nat Struct Mol Biol* 12, no. 1 (2005): 82-89.

Goldstein, L. and W. Plaut. "Direct Evidence for Nuclear Synthesis of Cytoplasmic Ribose Nucleic Acid." *PNAS* 41, no. 11 (1955): 874-880.

Grabow, Wade W., Paul Zakrevsky, Kirill A. Afonin, Arkadiusz Chworos, Bruce A. Shapiro and Luc Jaeger. "Self-Assembling Rna Nanorings Based on Rnai/Ii Inverse Kissing Complexes." *Nano Letters* 11, no. 2 (2011): 878-887.

Greenbaum, Nancy L., Claudius Mundoma and Dean R. Peterman. "Probing of Metal-Binding Domains of Rna Hairpin Loops by Laser-Induced Lanthanide(Iii) Luminescence†." *Biochemistry* 40, no. 4 (2001): 1124-1134.

Guerrier-Takada, Cecilia, Katheleen Gardiner, Terry Marsh, Norman Pace and Sidney Altman. "The Rna Moiety of Ribonuclease P Is the Catalytic Subunit of the Enzyme." *Cell* 35, no. 3, Part 2 (1983): 849-857.

Haurwitz, Rachel E., Martin Jinek, Blake Wiedenheft, Kaihong Zhou and Jennifer A. Doudna. "Sequence- and Structure-Specific Rna Processing by a Crispr Endonuclease." *Science* 329, no. 5997 (2010): 1355-1358.

Hermann, Thomas and Dinshaw J. Patel. "Rna Bulges as Architectural and Recognition Motifs." *Structure (London, England : 1993)* 8, no. 3 (2000): R47-R54.

Hershey, A. D. and Martha Chase. "Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage." *The Journal of General Physiology* 36, no. 1 (1952): 39-56.

Hershkovitz, E., E. Tannenbaum, S. B. Howerton, A. Seth, A. Tannenbaum and L. D. Williams. "Automated Identification of Rna Conformational Motifs: Theory and Application to the Hm Lsu 23s Rrna." *Nucleic Acids Research* 31, no. 21 (2003): 6249-6257.

Heus, H. and A. Pardi. "Structural Features That Give Rise to the Unusual Stability of Rna Hairpins Containing Gnra Loops." *Science* 253, no. 5016 (1991): 191-194.

Holbrook, Stephen R., Chaejoon Cheong, Ignacio Tinoco and Sung-Hou Kim. "Crystal Structure of an Rna Double Helix Incorporating a Track of Non-Watson-Crick Base Pairs." *Nature* 353, no. 6344 (1991): 579-581.

Holley, R. W., G. A. Everett, J. T. Madison and A. Zamir. "Nucleotide Sequences in the Yeast Alanine Transfer Ribonucleic Acid." *J Biol Chem* 240, (1965): 2122-8.

Hoogstraten, Charles G., Pascale Legault and Arthur Pardi. "Nmr Solution Structure of the Lead-Dependent Ribozyme: Evidence for Dynamics in Rna Catalysis." *Journal of Molecular Biology* 284, no. 2 (1998): 337-350.

Inselberg, A. and B. Dimsdale. "Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry." In *IEEE Visualization 1990*, 361-378, 1990.

Johnson, Philip E. and Logan W. Donaldson. "Rna Recognition by the Vts1p Sam Domain." *Nat Struct Mol Biol* 13, no. 2 (2006): 177-178.

Jurica, Melissa S. and Melissa J. Moore. "Pre-Mrna Splicing: Awash in a Sea of Proteins." *Molecular Cell* 12, no. 1 (2003): 5-14.

Kaluarachchi, Kumaralal, Robert P. Meadows and David G. Gorenstein. "How Accurately Can Oligonucleotide Structures Be Determined from the Hybrid Relaxation Rate Matrix/Noesy Distance Restrained Molecular Dynamics Approach?" *Biochemistry* 30, no. 36 (1991): 8785-8797.

Kim, S. H., G. J. Quigley, F. L. Suddath, A. McPherson, D. Sneden, J. J. Kim, J. Weinzierl and Alexander Rich. "Three-Dimensional Structure of Yeast Phenylalanine Transfer Rna: Folding of the Polynucleotide Chain." *Science* 179, no. 4070 (1973): 285-288.

Kim, U, Y Wang, T Sanford, Y Zeng and K Nishikura. "Molecular Cloning of Cdna for Double-Stranded Rna Adenosine Deaminase, a Candidate Enzyme for Nuclear Rna Editing." *Proceedings of the National Academy of Sciences* 91, no. 24 (1994): 11457-11461.

Kiss, Tamas. "Small Nucleolar Rna-Guided Post-Transcriptional Modification of Cellular Rnas." *EMBO J* 20, no. 14 (2001): 3617-3622.

Klein, D. J., P. B. Moore and T. A. Steitz. "The Roles of Ribosomal Proteins in the Structure Assembly, and Evolution of the Large Ribosomal Subunit☆." *Journal of Molecular Biology* 340, no. 1 (2004): 141-177.

Klein, D. J., T. M. Schmeing, P. B. Moore and T. A. Steitz. "The Kink-Turn: A New Rna Secondary Structure Motif." *The EMBO Journal* 20, no. 15 (2001): 4214-4221.

Klein, Daniel J., Sara R. Wilkinson, Michael D. Been and Adrian R. Ferré-D'Amaré. "Requirement of Helix P2.2 and Nucleotide G1 for Positioning the Cleavage Site and Cofactor of the Glms Ribozyme." *Journal of Molecular Biology* 373, no. 1 (2007): 178-189.

Klosterman, Peter S., Makio Tamura, Stephen R. Holbrook and Steven E. Brenner. "Scor: A Structural Classification of Rna Database." *Nucleic Acids Research* 30, no. 1 (2002): 392-394.

Krasilnikov, A. S. and A. Mondragon. "On the Occurrence of the T-Loop Rna Folding Motif in Large Rna Molecules." *RNA* 9, no. 6 (2003): 640-643.

Kulkarni, M., S. Ozgur and G. Stoecklin. "On Track with P-Bodies." *Biochem Soc Trans* 38, no. Pt 1 (2010): 242-51.

Ladner, J. E., A. Jack, J. D. Robertus, R. S. Brown, D. Rhodes, B. F. Clark and A. Klug. "Structure of Yeast Phenylalanine Transfer Rna at 2.5 a Resolution." *Proc Natl Acad Sci U S A* 72, no. 11 (1975): 4414-8.

Lakshminarayanan, A. and V. Sasisekharan. "Stereochemistry of Nucleic Acids and Polynucleotides. Ii. Allowed Conformations of the Monomer Unit for Different Ribose Puckerings." *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis* 204, no. 1 (1970): 49-59.

Lee, Brian M., Jing Xu, Bryan K. Clarkson, Maria A. Martinez-Yamout, H. Jane Dyson, David A. Case, Joel M. Gottesfeld and Peter E. Wright. "Induced Fit and "Lock and Key" Recognition of 5 S Rna by Zinc Fingers of Transcription Factor Iiia." *Journal of Molecular Biology* 357, no. 1 (2006): 275-291.

Leontis, N. B., R. B. Altman, H. M. Berman, S. E. Brenner, J. W. Brown, D. R. Engelke, S. C. Harvey, S. R. Holbrook, F. Jossinet, S. E. Lewis, F. Major, D. H. Mathews, J. S. Richardson, J. R. Williamson and E. Westhof. "The Rna Ontology Consortium: An Open Invitation to the Rna Community." *RNA* 12, no. 4 (2006): 533-541.

Leontis, N. and E. Westhof. "A Common Motif Organizes the Structure of Multi-Helix Loops in 16 S and 23 S Ribosomal Rnas." *Journal of Molecular Biology* 283, no. 3 (1998): 571-583.

Lerner, Michael R., John A. Boyle, Stephen M. Mount, Sandra L. Wolin and Joan A. Steitz. "Are Snrnps Involved in Splicing?" *Nature* 283, no. 5743 (1980): 220-224.

Lerner, Michael Rush and Joan Argetsinger Steitz. "Antibodies to Small Nuclear Rnas Complexed with Proteins Are Produced by Patients with Systemic Lupus Erythematosus." *Proceedings of the National Academy of Sciences* 76, no. 11 (1979): 5495-5499.

Lincoln, Tracey A. and Gerald F. Joyce. "Self-Sustained Replication of an Rna Enzyme." *Science* 323, no. 5918 (2009): 1229-1232.

Lovell, Simon C., Ian W. Davis, W. Bryan Arendall, Paul I. W. de Bakker, J. Michael Word, Michael G. Prisant, Jane S. Richardson and David C. Richardson. "Structure Validation by Cα Geometry: φ,Ψ and Cβ Deviation." *Proteins: Structure, Function, and Bioinformatics* 50, no. 3 (2003): 437-450.

Lu, X. J., W. K. Olson and H. J. Bussemaker. "The Rna Backbone Plays a Crucial Role in Mediating the Intrinsic Stability of the Gpu Dinucleotide Platform and the Gpupa/Gpa Miniduplex." *Nucleic Acids Research* 38, no. 14 (2010): 4868-4876.

Matt, Tanja, Chyan Leong Ng, Kathrin Lang, Su-Hua Sha, Rashid Akbergenov, Dmitri Shcherbakov, Martin Meyer, Stefan Duscha, Jing Xie, Srinivas R. Dubbaka, Déborah Perez-Fernandez, Andrea Vasella, V. Ramakrishnan, Jochen Schacht and Erik C. Böttger. "Dissociation of Antibacterial Activity and Aminoglycoside Ototoxicity in the 4-Monosubstituted 2-Deoxystreptamine Apramycin." *Proceedings of the National Academy of Sciences* 109, no. 27 (2012): 10984-10989.

McRee, Duncan E. "Xtalview/Xfit—a Versatile Program for Manipulating Atomic Coordinates and Electron Density." *Journal of Structural Biology* 125, no. 2–3 (1999): 156-165.

Merino, Edward J., Kevin A. Wilkinson, Jennifer L. Coughlan and Kevin M. Weeks. "Rna Structure Analysis at Single Nucleotide Resolution by Selective 2'-Hydroxyl Acylation and Primer Extension (Shape)." *Journal of the American Chemical Society* 127, no. 12 (2005): 4223-4231.

Moazed, Danesh and Harry F. Noller. "Intermediate States in the Movement of Transfer Rna in the Ribosome." *Nature* 342, no. 6246 (1989): 142-148.

Murphy, Frank V., Venki Ramakrishnan, Andrzej Malkiewicz and Paul F. Agris. "The Role of Modifications in Codon Discrimination by Trnalysuuu." *Nature Structural & Molecular Biology* 11, no. 12 (2004): 1186-1191.

Murray, L. J., J. S. Richardson, W. B. Arendall and D. C. Richardson. "Rna Backbone Rotamers--Finding Your Way in Seven Dimensions." *Biochem Soc Trans* 33, no. Pt 3 (2005): 485-7.

Murray, L. J. W., W. B. Arendall Iii, D. C. Richardson and J. S. Richardson. "Rna Backbone Is Rotameric." *Proceedings of the National Academy of Sciences* 100, no. 24 (2003): 13904-13909.

Murthy, Venkatesh L., Rajgopal Srinivasan, David E. Draper and George D. Rose. "A Complete Conformational Map for Rna☆." *Journal of Molecular Biology* 291, no. 2 (1999): 313-327.

Nagaswamy, Uma and George E. Fox. "Frequent Occurrence of the T-Loop Rna Folding Motif in Ribosomal Rnas." *RNA (New York, N.Y.)* 8, no. 9 (2002): 1112-1119.

Nahvi, Ali, Narasimhan Sudarsan, Margaret S. Ebert, Xiang Zou, Kenneth L. Brown and Ronald R. Breaker. "Genetic Control by a Metabolite Binding Mrna." *Chemistry & Biology* 9, no. 9 (2002): 1043-1049.

Nirenberg, M, P Leder, M Bernfield, R Brimacombe, J Trupin, F Rottman and C O'Neal. "Rna Codewords and Protein Synthesis, Vii. On the General Nature of the Rna Code." *Proceedings of the National Academy of Sciences* 53, no. 5 (1965): 1161-1168.

Nissen, P., J. A. Ippolito, N. Ban, P. B. Moore and T. A. Steitz. "Rna Tertiary Interactions in the Large Ribosomal Subunit: The a-Minor Motif." *Proceedings of the National Academy of Sciences* 98, no. 9 (2001): 4899-4903.

Nissen, Poul, Jeffrey Hansen, Nenad Ban, Peter B. Moore and Thomas A. Steitz. "The Structural Basis of Ribosome Activity in Peptide Bond Synthesis." *Science* 289, no. 5481 (2000): 920-930.

Noller, H. F. and J. B. Chaires. "Functional Modification of 16s Ribosomal Rna by Kethoxal." *Proc Natl Acad Sci U S A* 69, no. 11 (1972): 3115-8.

Numata, Tomoyuki, Yoshiho Ikeuchi, Shuya Fukai, Tsutomu Suzuki and Osamu Nureki. "Snapshots of Trna Sulphuration Via an Adenylated Intermediate." *Nature* 442, no. 7101 (2006): 419-424.

Olson, Wilma K. and Paul J. Flory. "Spatial Configurations of Polynucleotide Chains. I. Steric Interactions in Polyribonucleotides: A Virtual Bond Model." *Biopolymers* 11, no. 1 (1972): 1-23.

Oubridge, Chris, Nobutoshi Ito, Philip R. Evans, C. Hiang Teo and Kiyoshi Nagai. "Crystal Structure at 1.92 a Resolution of the Rna-Binding Domain of the U1a Spliceosomal Protein Complexed with an Rna Hairpin." *Nature* 372, no. 6505 (1994): 432-438.

Palade, G. E. and P. Siekevitz. "Liver Microsomes; an Integrated Morphological and Biochemical Study." *J Biophys Biochem Cytol* 2, no. 2 (1956): 171-200.

Pallan, Pradeep S., William S. Marshall, Joel Harp, Frederic C. Jewett, Zdzislaw Wawrzak, Bernard A. Brown, Alexander Rich and Martin Egli. "Crystal Structure of a Luteoviral Rna Pseudoknot and Model for a Minimal Ribosomal Frameshifting Motif†,‡." *Biochemistry* 44, no. 34 (2005): 11315-11322.

Parkinson, G., J. Vojtechovsky, L. Clowney, A. T. Brünger and H. M. Berman. "New Parameters for the Refinement of Nucleic Acid-Containing Structures." *Acta Crystallographica. Section D, Biological Crystallography* 52, no. Pt 1 (1996): 57-64.

Phannachet, Kulwadee, Youssef Elias and Raven H. Huang. "Dissecting the Roles of a Strictly Conserved Tyrosine in Substrate Recognition and Catalysis by Pseudouridine 55 Synthase†." *Biochemistry* 44, no. 47 (2005): 15488-15494.

Popenda, Mariusz, Jan Milecki and Ryszard W. Adamiak. "High Salt Solution Structure of a Left-Handed Rna Double Helix." *Nucleic Acids Research* 32, no. 13 (2004): 4044-4054.

Ramachandran, G. N., C. Ramakrishnan and V. Sasisekharan. "Stereochemistry of Polypeptide Chain Configurations." *Journal of Molecular Biology* 7, no. 1 (1963): 95-99.

Ramos, Andres and Gabriele Varani. "Structure of the Acceptor Stem of Escherichia Coli Trnaala: Role of the G3·U70 Base Pair in Synthetase Recognition." *Nucleic Acids Research* 25, no. 11 (1997): 2083-2090.

Richardson, D. C. and Jane S. Richardson. "Mage, Probe, and Kinemages." In *International Tables for Crystallography*, 727-730. Dordrecht: Kluwer Academic Publishers, 2001.

Richardson, David C. and Jane S. Richardson. "The Kinemage: A Tool for Scientific Communication." *Protein Science* 1, no. 1 (1992): 3-9.

Richardson, J. S., B. Schneider, L. W. Murray, G. J. Kapral, R. M. Immormino, J. J. Headd, D. C. Richardson, D. Ham, E. Hershkovits, L. D. Williams, K. S. Keating, A. M. Pyle, D. Micallef, J. Westbrook and H. M. Berman. "Rna Backbone: Consensus All-Angle Conformers and Modular String Nomenclature (an Rna Ontology Consortium Contribution)." *RNA* 14, no. 3 (2008): 465-481.

Richardson, Jane S., Bohdan Schneider, Laura W. Murray, Gary J. Kapral, Robert M. Immormino, Jeffrey J. Headd, David C. Richardson, Daniela Ham, Eli Hershkovits, Loren Dean Williams, Kevin S. Keating, Anna Marie Pyle, David Micallef, John Westbrook and Helen M. Berman. "Rna Backbone: Consensus All-

Angle Conformers and Modular String Nomenclature (an Rna Ontology Consortium Contribution)." *RNA (New York, N.Y.)* 14, no. 3 (2008): 465-481.

Rüdisser, Simon and Ignacio Tinoco Jr. "Solution Structure of Cobalt(Iii)Hexammine Complexed to the Gaaa Tetraloop, and Metal-Ion Binding to G·a Mismatches." *Journal of Molecular Biology* 295, no. 5 (2000): 1211-1223.

Rupert, Peter B., Archna P. Massey, Snorri Th. Sigurdsson and Adrian R. Ferré-D'Amaré. "Transition State Stabilization by a Catalytic Rna." *Science* 298, no. 5597 (2002): 1421-1424.

Sasisekharan, V. and A. V. Lakshminarayanan. "Stereochemistry of Nucleic Acids and Polynucleotides. Vi. Minimum Energy Conformations of Dimethyl Phosphate." *Biopolymers* 8, no. 4 (1969): 505-514.

Schluenzen, Frank, Ante Tocilj, Raz Zarivach, Joerg Harms, Marco Gluehmann, Daniela Janell, Anat Bashan, Heike Bartels, Ilana Agmon, François Franceschi and Ada Yonath. "Structure of Functionally Activated Small Ribosomal Subunit at 3.3 Å Resolution." *Cell* 102, no. 5 (2000): 615-623.

Schmitz, Uli and Thomas L. James. "[1] How to Generate Accurate Solution Structures of Double-Helical Nucleic Acid Fragments Using Nuclear Magnetic Resonance and Restrained Molecular Dynamics." In *Methods in Enzymology*, edited by L. James Thomas, Volume 261, 3-44: Academic Press, 1995.

Schneider, B., Z. Moravek and H. M. Berman. "Rna Conformational Classes." *Nucleic Acids Research* 32, no. 5 (2004): 1666-1677.

Schrodinger, LLC. "The Pymol Molecular Graphics System, Version 1.3r1." 2010.

Schuwirth, Barbara S., Maria A. Borovinskaya, Cathy W. Hau, Wen Zhang, Antón Vila-Sanjurjo, James M. Holton and Jamie H. Doudna Cate. "Structures of the Bacterial Ribosome at 3.5 Å Resolution." *Science* 310, no. 5749 (2005): 827-834.

Schwalbe, Martin, Oliver Ohlenschläger, Aliaksandr Marchanka, Ramadurai Ramachandran, Sabine Häfner, Tilman Heise and Matthias Görlach. "Solution

Structure of Stem-Loop A of the Hepatitis B Virus Post-Transcriptional Regulatory Element." *Nucleic Acids Research* 36, no. 5 (2008): 1681-1689.

Selmer, Maria, Christine M. Dunham, Frank V. Murphy, Albert Weixlbaumer, Sabine Petry, Ann C. Kelley, John R. Weir and V. Ramakrishnan. "Structure of the 70s Ribosome Complexed with Mrna and Trna." *Science* 313, no. 5795 (2006): 1935-1942.

Serganov, Alexander, Yu-Ren Yuan, Olga Pikovskaya, Anna Polonskaia, Lucy Malinina, Anh Tuân Phan, Claudia Hobartner, Ronald Micura, Ronald R. Breaker and Dinshaw J. Patel. "Structural Basis for Discriminative Regulation of Gene Expression by Adenine- and Guanine-Sensing Mrnas." *Chemistry & biology* 11, no. 12 (2004): 1729-1741.

Shi, H and P B Moore. "The Crystal Structure of Yeast Phenylalanine Trna at 1.93 a Resolution: A Classic Structure Revisited." *RNA* 6, no. 8 (2000): 1091-1105.

Sich, Christian, Oliver Ohlenschläger, Ramadurai Ramachandran, Matthias Görlach and Larry R. Brown. "Structure of an Rna Hairpin Loop with a 5‘-Cguuucg-3‘ Loop Motif by Heteronuclear Nmr Spectroscopy and Distance Geometry†." *Biochemistry* 36, no. 46 (1997): 13989-14002.

Siekevitz, Philip and George E. Palade. "A Cytochemical Study on the Pancreas of the Guinea Pig: Ii. Functional Variations in the Enzymatic Activity of Microsomes." *The Journal of Biophysical and Biochemical Cytology* 4, no. 3 (1958): 309-318.

Simons, K. T., R. Bonneau, I. Ruczinski and D. Baker. "Ab Initio Protein Structure Prediction of Casp Iii Targets Using Rosetta." *Proteins* Suppl 3, (1999): 171-6.

Smith, Kathryn D., Sarah V. Lipchock and Scott A. Strobel. "Structural and Biochemical Characterization of Linear Dinucleotide Analogues Bound to the C-Di-Gmp-I Aptamer." *Biochemistry* 51, no. 1 (2011): 425-432.

Stahley, Mary R. and Scott A. Strobel. "Structural Evidence for a Two-Metal-Ion Mechanism of Group I Intron Splicing." *Science* 309, no. 5740 (2005): 1587-1590.

Szep, S. "The Crystal Structure of a 26-Nucleotide Rna Containing a Hook-Turn." *RNA* 9, no. 1 (2003): 44-51.

Szewczak, A A, P B Moore, Y L Chang and I G Wool. "The Conformation of the Sarcin/Ricin Loop from 28s Ribosomal Rna." *Proceedings of the National Academy of Sciences* 90, no. 20 (1993): 9581-9585.

Tamura, M., D. K. Hendrix, P. S. Klosterman, N. R. B. Schimmelman, S. E. Brenner and S. R. Holbrook. "Scor: Structural Classification of Rna, Version 2.0." *Nucleic Acids Research* 32, no. 90001 (2004): 182D-184.

Tan, Dazhi, William F. Marzluff, Zbigniew Dominski and Liang Tong. "Structure of Histone Mrna Stem-Loop, Human Stem-Loop Binding Protein, and 3′Hexo Ternary Complex." *Science* 339, no. 6117 (2013): 318-321.

Theimer, Carla A., Craig A. Blois and Juli Feigon. "Structure of the Human Telomerase Rna Pseudoknot Reveals Conserved Tertiary Interactions Essential for Function." *Molecular Cell* 17, no. 5 (2005): 671-682.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins. "The Clustal_X Windows Interface: Flexible Strategies for Multiple Sequence Alignment Aided by Quality Analysis Tools." *Nucleic Acids Research* 25, no. 24 (1997): 4876-4882.

Tissieres, A. and J. D. Watson. "Ribonucleoprotein Particles from Escherichia Coli." *Nature* 182, no. 4638 (1958): 778-80.

Toor, Navtej, Kevin S. Keating, Sean D. Taylor and Anna Marie Pyle. "Crystal Structure of a Self-Spliced Group Ii Intron." *Science* 320, no. 5872 (2008): 77-82.

Tuerk, C. and L. Gold. "Systematic Evolution of Ligands by Exponential Enrichment: Rna Ligands to Bacteriophage T4 DNA Polymerase." *Science* 249, no. 4968 (1990): 505-10.

Tuschl, Thomas. "Expanding Small Rna Interference." *Nat Biotech* 20, no. 5 (2002): 446-448.

van de Ven, F. J. M., M. J. J. Blommers, R. E. Schouten and C. W. Hilbers. "Calculation of Interproton Distances from Noe Intensities. A Relaxation Matrix Approach without Requirement of a Molecular Model." *Journal of Magnetic Resonance (1969)* 94, no. 1 (1991): 140-151.

Varani, G., F. Aboulela and F. H. T. Allain. "Nmr Investigation of Rna Structure." *Progress in Nuclear Magnetic Resonance Spectroscopy* 29, no. 1-2 (1996): 51-127.

Wadley, Leven M., Kevin S. Keating, Carlos M. Duarte and Anna Marie Pyle. "Evaluating and Learning from Rna Pseudotorsional Space: Quantitative Validation of a Reduced Representation for Rna Structure." *Journal of Molecular Biology* 372, no. 4 (2007): 942-957.

Wadley, Leven M. and Anna Marie Pyle. "The Identification of Novel Rna Structural Motifs Using Compadres: An Automated Approach to Structural Discovery." *Nucleic Acids Research* 32, no. 22 (2004): 6650-6659.

Wang, Lincong and BruceRandall Donald. "Exact Solutions for Internuclear Vectors and Backbone Dihedral Angles from Nh Residual Dipolar Couplings in Two Media, and Their Application in a Systematic Search Algorithm for Determining Protein Backbone Structure." *Journal of Biomolecular NMR* 29, no. 3 (2004): 223-242.

Watson, J. D. and F. H. C. Crick. "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid." *Nature* 171, no. 4356 (1953): 737-738.

Westhof, E., B. Masquida and L. Jaeger. "Rna Tectonics: Towards Rna Design." *Folding & Design* 1, no. 4 (1996): R78-88.

Wijmenga, Sybren S., van Buuren, Bernd N. M. "The Use of Nmr Methods for Conformational Studies of Nucleic Acids." *Progress in NMR Spectroscopy* 32, (1998): 287-387.

Wilkinson, Kevin A., Edward J. Merino and Kevin M. Weeks. "Selective 2[Prime]-Hydroxyl Acylation Analyzed by Primer Extension (Shape): Quantitative Rna Structure Analysis at Single Nucleotide Resolution." *Nat. Protocols* 1, no. 3 (2006): 1610-1616.

Wimberly, Brian T., Ditlev E. Brodersen, William M. Clemons, Robert J. Morgan-Warren, Andrew P. Carter, Clemens Vonrhein, Thomas Hartsch and V. Ramakrishnan. "Structure of the 30s Ribosomal Subunit." *Nature* 407, no. 6802 (2000): 327-339.

Winkler, Wade C. and Ronald R. Breaker. "Regulation of Bacterial Gene Expression by Riboswitches." *Annual Review of Microbiology* 59, no. 1 (2005): 487-517.

Winkler, Wade, Ali Nahvi and Ronald R. Breaker. "Thiamine Derivatives Bind Messenger Rnas Directly to Regulate Bacterial Gene Expression." *Nature* 419, no. 6910 (2002): 952-956.

Woese, C. R. and G. E. Fox. "Phylogenetic Structure of the Prokaryotic Domain: The Primary Kingdoms." *Proc Natl Acad Sci U S A* 74, no. 11 (1977): 5088-90.

Woese, Carl R. *The Genetic Code : The Molecular Basis for Genetic Expression*. New York: Harper & Row, 1967.

Wool, Ira G., Anton Glück and Yaeta Endo. "Ribotoxin Recognition of Ribosomal Rna and a Proposal for the Mechanism of Translocation." *Trends in Biochemical Sciences* 17, no. 7 (1992): 266-269.

Word, J. Michael. "All-Atom Small-Probe Contact Surface Analysis: An Information-Rich Description of Molecular Goodness-of-Fit." Duke University, 2000.

Word, J. Michael, Simon C. Lovell, Thomas H. LaBean, Hope C. Taylor, Michael E. Zalis, Brent K. Presley, Jane S. Richardson and David C. Richardson. "Visualizing and Quantifying Molecular Goodness-of-Fit: Small-Probe Contact Dots with Explicit Hydrogen Atoms." *Journal of Molecular Biology* 285, no. 4 (1999a): 1711-1733.

Word, J. Michael, Simon C. Lovell, Jane S. Richardson and David C. Richardson. "Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-Chain Amide Orientation." *Journal of Molecular Biology* 285, no. 4 (1999b): 1735-1747.

Yang, Xiaojing, Timea Gerczei, LaTonya Glover and Carl C. Correll. "Crystal Structures of Restrictocin-Inhibitor Complexes with Implications for Rna Recognition and Base Flipping." *Nat Struct Mol Biol* 8, no. 11 (2001): 968-973.

Zaug, A. J. and T. R. Cech. "The Intervening Sequence Rna of Tetrahymena Is an Enzyme." *Science* 231, no. 4737 (1986): 470-5.

Zhang, Wen, Jack A. Dunkle and Jamie H. D. Cate. "Structures of the Ribosome in Intermediate States of Ratcheting." *Science* 325, no. 5943 (2009): 1014-1017.

Biography

Gary Joseph Kapral was born January 2, 1984 in Douglassville, Georgia, to Gary Mark Kapral and Rosemary Elizabeth Kapral. He has three younger brothers: Keith Jerome Kapral, Spencer Jesse Kapral, and Lance Jonathan Kapral.

He became a Christian at the age of six and the bulk of his primary education took place at Ebenezer Faith Christian School in Plymouth, Pennsylvania, where he quickly grew interested in science and history. The harbinger of Hurricane Fran greeted his move to North Carolina, where he later applied to and was accepted into the North Carolina School of Science and Mathematics (NCSSM) and where he promptly developed his first pneumothorax and was subsequently hit with Hurricane Floyd.

Undaunted, he began his research career in the summer of 2000 through a fellowship from NCSSM, conducting research at NC State under the auspices of Bruce Novak and Edward Tokas, developing a semi-batch emulsion polymerization method for Good-Year. While doing so, he discovered a new intermediate that acted as a mild adhesive that could be used for wallhanging (posters, etc.) without damaging paint, which became a hit in the NCSSM dorms. A concurrent (but unrelated) research initiative with SHODOR used computational methods to identify a transition state mechanism for the Ziegler-Natta catalyst. He also built a radon detector in a week-long special project in nuclear chemistry and discovered that on the whole, NCSSM is safe for

occupation (67% passing rate on radon tests). He majored in biochemistry at the Rochester Institute of Technology, but soon found his way into inorganic and polymer chemistry, synthesizing colorless polyimides with a non-linear optical pendant group with the goal of creating an organic film capable of second-harmonic generation, which would allow inexpensive laser frequency modulation.

In 2004, he made the decision to attend Duke University after seeing David Richardson's famous "Dancing Bears" presentation, and has been happily working on computational structural biophysics since then, with a side job of maintaining the Linux cluster for the lab. The led to a mix of RNA and protein focused research while maintaining and developing code to support the rest of the lab, and overall has resulted in an ability to collaborate, negotiate, and instruct under stressful conditions, skills which have proven invaluable in his large collaborative projects over the years. When he was not developing new collaborations, he began teaching students from NCSSM about structural biology and doing continuing development on KinImmerse to view the ribosome in the Duke Immersive Virtual Environment.

His time in the lab has resulted in many publications as the result of his collaborations with a variety of labs. They are listed as follows:

1. "The Phenix software for automated determination of macromolecular structures" Adams PD, Afonine PV, Bunkóczi G, Chen VB, Echols N, Headd JJ,

Hung L-W, Jain S, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner RD, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH (2011) Methods 55:94-106. doi: 10.1016/j.ymeth.2011.07.005

2. "Structures of the Bacterial Ribosome in Classical and Hybrid States of tRNA Binding" Dunkle JA, Wang L, Feldman MB, Pulk A, Chen VB, Kapral GJ, Noeske J, Richardson JS, Blanchard SC, & Doudna Cate JH (2011) Science 332:981-984. doi: 10.1126/science.1202692

3. "PHENIX: a comprehensive Python-based system for macromolecular structure solution" Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung L-W, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH (2010) Acta Cryst D66:213-221. doi: 10.1107/S0907444909052925 (open access)

4. "*MolProbity*: all-atom structure validation for macromolecular crystallography" Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS & Richardson DC (2010) Acta Cryst D66:12-21 doi: 10.1107/S0907444909042073 (open access)

5. "The other 90% of the protein: Assessment beyond the Cαs for CASP8 template-based and high-accuracy models" Keedy DA, Williams CJ, Headd JJ, Arendall WB, Chen VB, Kapral GJ, Gillespie RA, Block JN, Zemla A, Richardson DC & Richardson JS (2009) Proteins: Struc Func Bioinf 77(Suppl 9):29-49 doi: 10.1002/prot.22551

6. "RNA Backbone: Consensus All-angle Conformers and Modular String Nomenclature (an RNA Ontology Consortium contribution)" Richardson JS, Schneider B, Murray LW, Kapral GJ, Immormino RM, Headd JJ, Richardson DC, Ham D, Hershkovits E, Williams LD, Keating KS, Pyle AM, David Micallef d, Westbrook J & Berman HM (2008) RNA 14 :465-481. doi: 10.1261/rna.657708 (open access)

7. "RNABC: forward kinematics to reduce all-atom steric clashes in RNA backbone" Wang X, Kapral GJ, Murray LW, Richardson DC, Richardson JS & Snoeyink J (2008) J Math Biol 56:253-278. doi: 10.1007/s00285-007-0082

8. "MolProbity: all-atom contacts and structure validation for proteins and nucleic acids" Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, Snoeyink J, Richardson JS & Richardson DC (2007) Nucleic Acids Res 35 W375-W383. doi:10.1093/nar/gkm216 (open access)

While his graduate school career is at an end, there is much work to be done, and he plans to spend the rest of the semester working on yet more publications. Wherever his future career takes him, he hopes to continue working with the Richardson Lab to develop more features for KiNG so he can use it to educate budding biochemists in the importance of structural biology.

He married his high school sweetheart, Amy Lyndall Middleton on January 6, 2005. She gave birth to their first son, Gary James Kapral, on October 26, 2010, and their second son, Christopher Michael Yosef Kapral, on October 11, 2013. Together, Gary and Amy run the Duke/UNC branch of PhD Posters, a small business started by lab alumnus Ian Davis, through which they like to sponsor scientific meetings and provide funding for travel and poster awards to the local departments.