



universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

Seeing Secondary, Sampling Tertiary: A parallel journey through the prediction and visualization of RNA tertiary and secondary structure

verfasst von / submitted by

cand. scient. Peter Kerpedjiev

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2016 / Vienna, 2016

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on the student
record sheet:

A 794 685 437

Dissertationsgebiet lt. Studienblatt /
field of study as it appears on the student record sheet:

Biologie, DK: RNA Biology

Betreut von / Supervisor:

Univ.-Prof. Dipl.-Phys. Dr. Ivo Hofacker

Peter Kerpedjiev: *Seeing Secondary, Sampling Tertiary*: A parallel journey through the prediction and visualization of RNA tertiary and secondary structure

A thesis submitted in partial fulfillment of the requirements for the degree of:

Doctor of Philosophy (PhD)

at the

Institute for Theoretical Chemistry
Fakultät für Chemie
University of Vienna

© January 2016

ABSTRACT

This thesis presents two parallel tracks of research into the prediction of RNA tertiary structure and the display of secondary structure.

The first, a coarse-grain model of RNA tertiary structure, provides an ensemble based prediction method for modeling RNA's tertiary structure given its primary and secondary structure. In contrast with existing methods, it excels at quickly providing a multitude of predictions for use in downstream analysis. We aim to efficiently sample the conformational space that can be explored by a given RNA secondary structure. The results indicate that our work is a cautious step in the right direction. The quality of the predictions are comparable with those from more sophisticated models. We can generate numerous predictions in a relatively short amount of time and our predictions are consistent with the knowledge-based descriptors of global RNA structure derived from native structures. We hope that the conformations we generate can serve as valuable inputs for further downstream analysis corroborating evidence from experiments such as Förster resonance energy transfer (FRET) and atomic force microscopy (AFM).

The second, a visualization tool, aims to simplify the display of RNA secondary structures. While a number of tools exist for this purpose, ours is dependency free, accessible from any browser-equipped computer and offers a number of features not found in any other software. We have used our implementation to provide innovative, attractive, and revealing depictions of cotranscriptional folding which allow researchers to better understand the variety of structures which may potentially form as an RNA is being transcribed. This same layout was used to re-imagine the traditional dot-plot layout for displaying base pair probabilities and supplement it with a relevant and easily understandable depiction of the RNA structures that these base pairs are found in. We tie this second section back to the tertiary structure prediction work by showing how a flexible method for displaying secondary structure can be used to diagnose correctly and incorrectly predicted long range interactions in tertiary structures.

ZUSAMMENFASSUNG

Die vorliegende Dissertation umfasst zwei übergeordnete Schwerpunkte, die Vorhersage von RNA-Tertiärstruktur und die Visualisierung von Sekundärstruktur.

Der erste Schwerpunkt, ein coarse-grained Modell der RNA-Tertiärstruktur, ermöglicht die Ensemble-basierte Modellierung der Tertiärstruktur einer RNA, aufbauend auf deren Primär- und Sekundärstruktur. Dieser neue Ansatz bietet, im Gegensatz zu bestehenden Methoden, die Möglichkeit rasch eine Vielzahl von Strukturen für spätere Analysen zu generieren. Ziel ist es, ausgehend von einer RNA-Sekundärstruktur, den sich ergebenden Konformationsraum größtmöglich zu erkunden. Die Ergebnisse zeigen, dass die entwickelte Methode einen bedachtem Schritt in die richtige Richtung darstellt. Die Strukturvorhersagen sind qualitativ vergleichbar mit Ergebnissen wesentlich komplexerer Ansätze. Die Methode generiert in einem Bruchteil der Zeit eine große Anzahl an Strukturen. Diese Strukturen sind konsistent mit Deskriptoren globaler RNA-Struktur, abgeleitet aus nativen RNA-Strukturen. Die vorhergesagten Strukturen eignen sich prinzipiell hervorragend als Ausgangspunkt für weitere Analysen und sollen Experimenten wie Förster Resonance Energy Transfer (FRET) oder Atomic Force Microscopy (AFM) als hilfreiche Untermauerung ihrer Ergebnisse dienen.

Der zweite Schwerpunkt, ein Visualisierungstool, zielt auf die einfache Darstellung von RNA-Sekundärstrukturen ab. Obwohl für diesen Zweck bereits Tools entwickelt wurden, sticht die hier vorgestellte Neuentwicklung durch die Freiheit von jeglichen Software-Dependencies, Browser-Integration und eine Reihe von neuen Features hervor. So ist es möglich mit der Applikation in innovativer, attraktiver und einleuchtender Weise cotranskriptionelles Falten darzustellen, was zukünftigen Forschern die Verschiedenheit der dabei potentiell involvierten Strukturen elegant verständlich macht. Aufbauend auf jener Darstellungsweise, wurde das klassische Layout zur Visualisierung von Basenpaar-Wahrscheinlichkeiten, der Dot-Plot, auf eine Darstellung übertragen welche sich im Rahmen einer leicht verständlichen Sekundärstruktur mit allen relevanten Basenpaaren bewegt. Dieser Teil ist direkt mit der Tertiärstrukturvorhersage verbunden, da gezeigt werden konnte dass die flexible Art der Sekundärstruktur-Visualisierung in der Lage ist korrekte und inkorrekte long-range Interactions zu identifizieren.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Ivo Hofacker for promoting this project and providing help, ideas, and most importantly, explanations for complicated and difficult to understand topics. I would also like to sincerely thank the other two members of my PhD committee, Thomas Hamelryck for fruitful discussions about the reference-ratio method, for thorough explanations about how to formulate a knowledge-based energy and ideas about what to incorporate as energy terms and Michael Jantsch for pointing me in the direction of biological work relevant to the chimera of theory and application that is this thesis.

I would like to thank Christoph Flamm, Christian Höner zu Siederdisen and Sven Findeiß, for their involvement and advice regarding the project. Christoph was never at a loss for enthusiastic ideas and advice about which topics to pursue. His organization of 3D printing workshops and general enthusiasm for summer lunch beverages provided numerous distractions from the creeping tedium of working on a particular topic for four years. Christian was the impetus behind the publication of the first paper and was extremely helpful in propelling my experiments and explorations towards a publication. Sven initially suggested that I work on 3D structure prediction as a potential PhD topic. Without their involvement, this thesis would likely be about a completely different topic.

No work is done in a vacuum, and the days of solitary science are long gone. The visualization section of this thesis would not have been possible without the help of Stefan Hammer and Stefan Badelt. I have to sincerely thank Stefan H. for his diligent work in turning the FORNA concept into a working web application. Stefan B. was a constant source of constructive criticism and a fountainhead of practical ideas for improving the tools we created. His desire to display his results in a practical and relevant manner lead to the idea of creating a specialized method for visualizing cotranscriptional folding.

Writing a thesis is only one of the many tasks I performed at the TBI. The rest consisted of research, travel, classes and paperwork. All of these activities required the unwavering help and patience of Judith Ivansits, Nicola Wiskocil and Gerlinde Aschauer. I would like to thank Juli and Nicola for helping me in every organizational matter and making sure all my paperwork was submitted in a timely manner. A big thanks to Juli as well as Gerlinde for processing the payments for Open Access and color figure charges on the papers associated with this work.

Published work is invaluable as a reference. Just as valueable is the input of knowledgeable colleagues who pointed me in the direction of relevant

information and helped me to use existing tools. I would like to thank Craig Zirbel for sharing a wealth of knowledge about RNA 3D structure and helping me with using JAR3D and FR3D for 3D motif detection in RNA crystal structures. Ronny Lorenz deserves a large commendation for the work he has done on the Vienna RNA Package and his willingness to assist me in making it work with FORNA. I would also like to thank Andrea Tanzer for explanations involving anything biology or evolution related, as well recommending books and papers about the origin of life. The list of valuable expertise would not be complete without Torsten Möller who contributed to fruitful discussions and brainstorming about DR. FORNA.

There is an indescribable feeling of purpose when one's work is used by others. I have to thank Jing Qin and Nikolai Hecker for actually applying ERNWIN to real world problems and showing that it can be used to explain experimental data. Nikolai deserves an extra show of gratitude for his help in setting up and using RNAFDL for RNA secondary structure display. I also owe Sarah Berkemer a big thanks for being one of the first users of FORNA and for sharing the structure and conservation of the C/D box RNA with me.

While PhD work should be focused on one or two core topics, one of its privileges is the ability to take part in different projects and learn about complementary areas of research. To this end, I have to thank Jan Gorodkin and Anne Wenzel for including me in the RISEARCH2 project and thus giving me the opportunity to learn about and contribute to the search for RNA-RNA interactions.

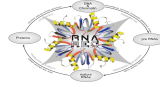
As no thesis written in Austria is complete without an abstract (Zusammenfassung) in German, I have to thank Roman Ochsenreiter and Ronny for translating my English version to German.

Being a PhD student includes many moments of **not** writing a PhD thesis. A fantastic work environment helped make that task a breeze. I have to thank Marcel Kucharik, Roman, Bernhard Thiel, Stefan Hammer and Dominik Steininger for helping solve offbeat yet mildly relevant puzzles related to our work. Florian Eggenhofer and Jörg Fallman deserve a special thanks for their contagious laughter, never-ending stream of jokes and general maintenance of an exceptionally entertaining working environment. Fabian Amman and Michael T. Wolfinger are owed a sincere thanks for their culinary companionship in times of serious Speck consumption as well as their input on next generation sequencing mapping techniques.

Last but not least, I have to thank my family and friends for their support and encouragement through the ups and downs of the past four years.

Funding

This work was funded, in part, by the Austrian DK RNA program FG748004, and by the Austrian FWF, project “SFB F43 RNA regulation of the transcriptome”.



universität
wien



FWF

Der Wissenschaftsfonds.

CONTENTS

i	INTRODUCTION	1
1	INFORMAL INTRODUCTION	3
2	INTRODUCTION TO RNA SECONDARY STRUCTURE	5
2.1	What is RNA secondary structure?	5
2.2	The importance of secondary structure	6
2.3	Determining secondary structure	9
2.4	Secondary structure prediction	11
3	RNA TERTIARY STRUCTURE	17
4	PREVIOUS METHODS FOR RNA TERTIARY STRUCTURE PREDICTION	23
4.1	Birds-eye view of existing methods	23
4.2	FARNA and fragment assembly	24
4.3	FARFAR adds high resolution energy	25
4.4	The MC-Fold MC-Sym pipeline	25
4.5	ModERNA for homology modeling	26
4.6	Discrete molecular dynamics for fast force fields	27
4.7	NAST coarse-grains and uses a knowledge-based potential	28
4.8	BARNACLE creates models in continuous torsion angle space	29
4.9	3DRNA assembles fragments from SCOR and RNAJunction	30
4.10	RNACOMPOSER assembles fragments using machine translation	31
4.11	RNAJAG predicts helical geometries used to sample coarse-grain structures	31
4.12	GARN uses game theory to sample coarse-grain structures	32
4.13	Summary	32
5	VISUALIZATION	35
5.1	The standard polygonal layout is aesthetically pleasing but prone to overlaps	36
5.2	Radial drawing draws spindly diagrams devoid of overlaps	36
5.3	The NAView algorithm trades potential overlaps for aesthetic appeal	38
5.4	Force-directed layouts apply a general graph layout algorithm to RNA structures	38
5.5	RnaViz can store of layouts for future reuse	39

5.6	PseudoViewer draws pseudoknotted structures	40
5.7	VARNA is a cross-platform application with extensive functionality	41
ii	PUBLISHED WORK	43
6	ERNWIN PUBLICATION	45
7	FORNA PUBLICATION	85
iii	TERTIARY STRUCTURE PREDICTION	101
8	A HELIX-CENTERED COARSE-GRAIN MODEL	103
8.1	Helices are defined by 10 parameters	104
8.2	Inter-helical orientations can be defined by six parameters	105
8.3	Terminal loops can be defined by three parameters	105
8.4	Virtual residues are interpolated nucleotide positions	106
8.5	Virtual atoms are interpolated atom positions	107
8.6	Avoiding excluded volume	107
8.7	Maintaining junction integrity	108
9	MEASURING 3D MODEL DIFFERENCES	113
9.1	root mean square deviation (RMSD) and distance root mean square deviation (dRMSD)	113
9.2	Interaction network fidelity	114
9.3	The recall, precision, F-measure (RPF) measure	115
9.4	The adjacency correlation coefficient (ACC)	116
10	METHODS OF SAMPLING AND ENERGY EVALUATION	119
10.1	How to sample from a probability distribution	119
10.2	Distribution approximation using kernel density estimates	121
10.3	Target distribution describing global RNA structure	123
10.4	Sampling local RNA structure	125
10.5	The proposal distribution consists of fragments of known structure	126
10.6	Sampling of loop parameters is conditioned on the loop's size	127
10.7	Motif-based loop parameter sampling uses sequences to find matching motifs	127
10.8	Long-range interaction energy format	128
10.9	A-minor motif energy	128
10.10	Conclusion	131
11	RESULTS	133
11.1	Local structure model quality evaluation	133
11.2	Comparison to the ROSETTA inter-helical statistics	135
11.3	Local structure constraints from JAR3D	139
11.4	Hypothetically perfect interior loop information	142

11.5	Imperfect secondary structure	144
11.6	Interpreting FRET data	147
12	SUMMARY AND FUTURE WORK	153
12.1	Covariation analysis for long-range constraints	153
12.2	Using selective 2'hydroxyl acylation analyzed by primer extension (SHAPE) data to improve predictions	154
12.3	ERNWIN for multimeric structures	155
12.4	Guiding energy for multiloop construction	155
12.5	Kink turn evolution	155
12.6	Tetraloop (TR) - TR receptor (TRR) evolution	156
12.7	Using known long-range interactions as constraints	157
12.8	Corroborating atomic force microscopy (AFM) images	157
12.9	Summary	158
iv	SECONDARY STRUCTURE VISUALIZATION	161
13	FORNA AND FORNAC: SUPER SIMPLE SHAREABLE SECONDARY STRUCTURES	163
13.1	See, simplify and share with FORNA	163
13.2	Potential improvements	164
14	THE DOT-STRUCT PLOT: AN IMPROVED DOT PLOT	167
14.1	Augmenting the dot plot	167
14.2	Design contest submission	170
15	FOLLOWING COTRANSCRIPTIONAL FOLDING WITH DR. FORNA	173
16	TERTIARY STRUCTURE PREDICTION DIAGNOSIS	181
16.1	Long-range interaction overlay	182
16.2	Diagnosing the accuracy of predicted adjacencies	183
16.3	Showing multiple predictions using small multiples	184
16.4	Summary and further work	186
17	CONCLUSION	187
v	APPENDIX	189
	BIBLIOGRAPHY	205
	CV	

LIST OF FIGURES

Figure 1	RNA folding introduction	12	
Figure 2	Suboptimal RNA secondary structures	13	
Figure 3	Standard polygonal secondary structure diagram		36
Figure 4	Radial secondary structure diagram	37	
Figure 5	NAView secondary structure diagram	38	
Figure 6	jViz.RNA secondary structure diagram	39	
Figure 7	PseudoViewer secondary structure diagram		40
Figure 8	RNA torsion angles	103	
Figure 9	Coarse-grain stem parameters	104	
Figure 10	Coarse-grain hairpin loop parameters		106
Figure 11	Virtual residues	107	
Figure 12	Virtual atoms	108	
Figure 13	Junction closure distances 1	109	
Figure 14	Junction closure distances 2	109	
Figure 15	Junction closure distances 3	110	
Figure 16	Junction closure calculation illustration		110
Figure 17	Interacting element distances	114	
Figure 18	All-atom vs coarse grain (CG) RMSD		114
Figure 19	Element interaction probability	117	
Figure 20	Element distance histogram	118	
Figure 21	Stem base pair rise	118	
Figure 22	Single number distribution	120	
Figure 23	Reference ratio sampling	121	
Figure 24	Adaptive rejection sampling	122	
Figure 25	RNA substructures for statistics	124	
Figure 26	A-minor interaction parameters	130	
Figure 27	Four types of A-minor interactions		131
Figure 28	CG helix definition	134	
Figure 29	Inter-helical angle distributions	137	
Figure 30	Inter-helical angle distributions	138	
Figure 31	Best predicted group I intron	139	
Figure 32	Well predicted motif	140	
Figure 33	Poorly predicted motif	142	
Figure 34	Interior loop prediction examples		143
Figure 35	Tertiary structure from predicted secondary structure	146	
Figure 36	Inter-nucleotide distance approximation	148	
Figure 37	ERNWIN inter-nucleotide distance calculation		149
Figure 38	FRETTRANSLATOR structure assignments	150	

Figure 39	AFM image with ERNWIN annotations	157
Figure 40	Ribosome secondary structure illustration	164
Figure 41	Original dot plot	167
Figure 42	A dot-struct plot	168
Figure 43	Riboswitch dot-struct plot	169
Figure 44	Expanded riboswitch dot-struct plot	169
Figure 45	Cotranscriptional folding illustration example	174
Figure 46	Cotranscriptional folding illustration example	174
Figure 47	Cotranscriptional folding illustration example	175
Figure 48	Cotranscriptional folding illustration example	175
Figure 49	Cotranscriptional folding illustration example	176
Figure 50	Original DR. TRANSFOMER output	177
Figure 51	DR. FORNA Screenshot	178
Figure 52	DR. FORNA Montage	179
Figure 53	Connections for distant nodes	182
Figure 54	Color for correct and incorrect predictions	184
Figure 55	Long-range interaction small multiples	185
Figure 56	Protein Data Bank (PDB) atom names	192
Figure 57	ERNWIN statistics generation	193

LIST OF TABLES

Table 1	A-minor template nucleotides	129
Table 2	A-minor interaction prevalence	129
Table 3	A-minor interaction prevalence	130
Table 4	Cheating energy	136
Table 5	Adding motif predictions	140
Table 6	Fixed loop parameters	141
Table 7	Inter-nucleotide distances from ERNWIN	148

Table 8	A list of all the structures used for benchmarking ERNWIN. The length of each structure is given in nucleotides (nt).	196
---------	---	-----

LISTINGS

Listing 1	Metropolis-Hastings algorithm	120
-----------	-------------------------------	-----

ACRONYMS

NMR	nuclear magnetic resonance
DMS	dimethyl sulfate
STMV	satellite tobacco mosaic virus
PDB	Protein Data Bank
RMSD	root mean square deviation
dRMSD	distance root mean square deviation
INF	interaction network fidelity
RPF	recall, precision, F-measure
MFE	minimum free energy
ROG	Radius of Gyration
KDE	kernel density estimate
DBN	dynamic bayesian network
RNP	ribonucleoprotein
SRP	signal recognition particle
snRNA	small nuclear RNA
tRNA	transfer RNA

SSE	secondary structure element
cryo-EM	cryo-electron microscopy
MCC	Matthews correlation coefficient
ACC	adjacency correlation coefficient
siRNA	small interfering RNA
miRNA	micro RNA
smRNA	small RNA
SAXS	small angle X-Ray scattering
HMM	hidden Markov model
AFM	atomic force microscopy
FCCD	full cyclic coordinate descent
RNAJAG	RNA Junctions as Graphs
FRET	Förster resonance energy transfer
CG	coarse grain
NCM	nuclear cyclic motif
SHAPE	selective 2'hydroxyl acylation analyzed by primer extension

Part I

INTRODUCTION

This thesis will be long. The audience will be learned and the material will be confusing. So where should it begin? What is an appropriate place to start introducing the topic which has taken up countless hours of time over the past four years? There is no good answer to this question but to make sure nothing is omitted, it's probably safe to start at the beginning. Which beginning? Why, the beginning of life, of course.

In the beginning, there was RNA. Today, there's still RNA and we are still trying to figure out what it does and how it does it. To this end, researchers have created computational tools to predict what shape it takes within the cell. This is called its structure. The chapters in this section will discuss why it's important to determine RNA's structure, what we gain from this knowledge, and what methods have been developed to predict and display it.

INFORMAL INTRODUCTION

Where do we come from? How do we work? How do we begin to answer such questions? Is there an answer? The answers to such broad question only open up more questions. We end up having to learn more and more about less and less to fill the gaps in our understanding. So how did we get here? Where are we going? How does this work aid our journey? These are the questions the readers should keep in mind while reading this thesis or any other scientific publication. This brief informal introduction will summarize my understanding of our scientific situation and the context within which the work for this thesis was done.

At first there were humans. We could see each other and communicate with each other. We knew we existed. We knew we needed food and water to survive and that we needed to have sex to reproduce. Over time we learned about our internal organs and that the heart was responsible for pumping blood through our body and that our brain was responsible for thinking. We learned that our body was composed of cells and that these cells, in turn, had nuclei and mitochondria. We learned that cells communicate via receptors and that hormones regulate essential bodily processes. We learned that DNA mediates heredity and that RNA is involved in translation, regulation and translocation. We learned the structure of molecules. We learned that changes in the conformations of enzymes were responsible for their catalytic activity and we learned that changes in the conformation of RNA are responsible for regulation and splicing. We learned more and more about less and less.

How did we learn these things? We saw that when people didn't eat or drink, they died. When they didn't have sex, they didn't have babies. We found out that when the heart stopped, blood stopped flowing through our veins. When our brain was damaged, our thought processes changed. When we looked through a microscope, we saw tiny compartments called cells. Extracting fluid from organs and re-injecting it into the bloodstream led to the discovery of hormones. Transforming bacteria with the DNA of other bacteria confirmed its role as the mediator of heredity. Analysis of bacterial and phage mutants revealed the importance of RNA as a messenger between DNA and protein synthesis. Further mutational analysis expanded its role to regulation and catalysis. X-Ray crystallographs of DNA, RNA and proteins revealed their structure and how it influences their function. Advances in experimental techniques expanded the tools available for exploration (what happens if ...?), explanation (how does ...?), and confirmation (if I change x does y change too?).

So how does this relate to this thesis? What are we trying to achieve?

The purpose of the work described in this thesis is to provide tools for *exploration* and *explanation*. X-Ray crystallography, nuclear magnetic resonance (NMR) spectroscopy as well as a host of other experimental techniques allow us to see the structure of macromolecules as it exists in crystals or solution, respectively. They provide means for exploration, explanation and confirmation because they record a property of a physical molecule. In our tertiary structure prediction work, we try to mimic their function by way of simulation. We try to provide researchers with the same information that they may obtain from crystallography or from NMR, much more quickly and cheaply. This comes at the cost of accuracy. The information we provide is not a snapshot of reality, but rather a prediction. It is an attempt to use our current knowledge and understanding of RNA to provide an informed estimate of what tertiary structure(s) a given sequence can form.

Even in the case where our predictions aren't accurate, we hope that merely providing a range of conformations will be useful for *exploring* how an RNA molecule may fold. Which conformations are possible? Which aren't? Which are likely? Which distal sections can potentially form interactions? Which can't? Such exploratory questions can theoretically also serve as *explanations*. Why can't a particular pseudoknot be formed? If a particular tertiary structure is necessary for degradation, knowing that it can't be formed can provide a potential explanation for the lack of degradation. While these are theoretical examples, they're mentioned to provide a glimpse into what we are trying to enable with our tertiary structure prediction tool.

Our secondary structure visualization work focuses on *exploration*, *interaction* and *dissemination*. We strove to create tools that make it easy to explore the secondary structure present in 3D crystal structure models of RNA, to modify the presented diagrams and to easily share the diagrams with others. The framework and software components we created ended up being useful for more than just that. They allowed us to place RNA structure in more meaningful contexts and to highlight and clarify the output of other predictive models. Most of all, they enable us to use our uniquely human intuition to diagnose the efficacy of our methods and to seek out new topics for exploration.

It is our hope that our work will be a step forward in the progress of discovery by providing researchers with more powerful, easier to use and more accessible tools to make hypotheses about the mechanisms of RNA biology, as well as to document and disseminate their work online. With a more capable toolbox, we hope that researchers will be better equipped to uncover greater details about the myriad of processes occurring within the cells of all living things.

2

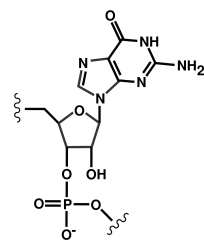
INTRODUCTION TO RNA SECONDARY STRUCTURE

2.1 What is RNA secondary structure?

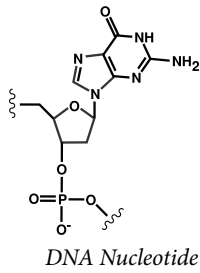
The secondary structure of RNA is an abstraction of its tertiary structure which describes the intra-molecular base pairing of its nucleotides. While physiological effects are a consequence of tertiary structure, there is enough information contained in the secondary structure to describe molecular processes such as protein binding and RNA interference, among others. This can be used to deduce mechanisms of action for particular mutations (e.g. disrupting a stem), or to place sequence variants into context for further biological interpretation (e.g. necessity of an unpaired region). Perhaps the strongest evidence for the functional role of RNA's secondary structure is its degree of conservation in a variety of diverse transcripts from bicoid localization signals [109], to CRISPR repeats [96], to rRNA [189], to tRNA [159]. The rest of this chapter will show that RNA's secondary structure plays an important role in the cell. It will also enumerate ways to experimentally measure base pairing in RNA and end with an introduction to how we can computationally predict secondary structure from sequence.

RNA is composed of a string of covalently linked nucleotides which are able to pair with each other by forming hydrogen bonds between their aromatic base rings. This pairing ability, however, is constrained to couples of compatible nucleotides: A (adenine) is only able to form a base pair with U (uracil) using two hydrogen bonds whereas G (guanine) can form a strong pair with C (cytosine) using three hydrogen bonds or a weaker one with uracil using two hydrogen bonds. Base pairing can only occur exclusively between nucleotides (i.e. a nucleotide can only be involved in one base pair at a time) and requires the two nucleotides to be in a specific physical orientation relative to each other. Consecutive base pairs imply a double helical tertiary structure. In addition to the canonical base pairs described here, nucleotides can also be found in noncanonical pairs where they form weaker bonds and take on geometries which deviate from the double helical structure [170]. These interactions will be briefly mentioned in Section 2.4.

The same pairing principle is behind the double-stranded nature of DNA. A double helical DNA molecule is actually two single stranded molecules that are held together by the base pairing of their respective nucleotides (A, C, G and T). The hydrogen bonds which form between the bases of the paired nucleotides hold the two molecules together. The



RNA Nucleotide



chemical structure of the nucleotides allows a variety of geometrical conformations when unpaired. When paired, however, the structure becomes locked into the familiar double helical geometry seen in textbooks, logos and infographics around the world.

In the case of DNA, the tertiary structure is simple because all nucleotides are paired with their complement. Each DNA dimer is composed of two complementary molecules (each called a *strand*). Due to the perfect pairing of the nucleotides in the two strands, the resulting complex is a simple double helix. RNA, unlike DNA, is usually found as a monomer. Due to the presence of the extra alcohol group on its sugar ring, RNA is more flexible and is able to fold back onto itself in such a way that nucleotides of the same molecule can pair with each other. This is called its *secondary structure* and is responsible for the large variety of functional roles that RNA plays in the cell, most of which are mediated through some form of protein binding [58].

The rest of this chapter will describe why knowledge of the secondary structure of RNA is important, how we can predict it, measure it and verify our predictions.

2.2 The importance of secondary structure

Why is secondary structure important? The simple answer is that RNA's secondary structure is a proxy for its tertiary structure. In other words, the secondary structure of an RNA molecule implies a lot about its tertiary structure. All of the paired regions in the secondary structure will adopt a roughly uniform double helical structure. Even without knowledge of the exact orientations of the helices (the work in this thesis), simply knowing they exist and how they are connected can provide valuable information about the identity and function of the molecule.

Why do we care? The following few sections will provide examples of why secondary structure is functionally important in cells. The first will demonstrate that its capacity to bind proteins depends on its secondary structure. This will be followed by an overview of where secondary structure is found within the transcriptome and what types of functions it is associated with. To cap off the list of examples, I'll mention its role in RNA interference. For a more comprehensive overview of RNA's many roles in the cell, the reader is encouraged to consult a more thorough review, as can be found in Wan et al. [180].

RNA BINDS PROTEINS IN A STRUCTURE / SEQUENCE DEPENDENT MANNER Because RNA helices typically are typically A-form, their base pairs are sequestered deep within the helix making sequence-specific binding of proteins to double stranded RNA difficult. This is in contrast to DNA double helices, which are typically B-form with a shallow major

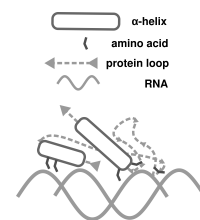
groove, allowing binding proteins to interact with and recognize paired nucleotides [118]. The simple consequence of this fact is that specific sequence-dependent recognition is much more difficult in RNA than in DNA helices. Due to its secondary structure, however, RNA can expose single-stranded regions that are held in a fixed position relative to the double stranded regions that flank them.

Such regions are selectively bound by RNA-binding domains found in a variety of different proteins [58, 118]. These domains tend to be sequence-specific and bind preferentially to single-stranded regions in a stem-loop context [7, 147]. This is commonly demonstrated by creating a pool of different RNAs, pulling them down via a protein intermediate and analyzing the sequences [177, 147] of those that were bound. The presence of strongly conserved motifs indicates that protein binding is sequence dependent, while the fact that association constants are higher when the motif is present in a stem-loop structure demonstrate the importance of the structural context [64, 147]. More detailed atomic level examination of the protein-RNA complexes reveals that binding and recognition do indeed rely on unpaired RNA nucleotides interacting with the side chains of proteins [7].

Like single-stranded binding, double-stranded binding is also mediated by a protein domain found across many different families of RNA-binding proteins [46]. Ryter et al [152] show that the dsRNA-binding protein Xlrbpa interacts with two minor grooves and one major groove spanning 16bp of a double-stranded RNA molecule. The interactions within the minor groove are mediated by contacts with the OH groups of the sugars, which are independent of the sequence. What little protein base interactions are present are mediated by water molecules which are believed to be able to adjust their position depending on the sequence present. What is important here is that the binding is structure-, rather than sequence- dependent. The main requirement for association between a dsRNA-binding protein and an RNA is the presence of a long double-stranded region.

These two examples of single and double-stranded binding seek to illustrate that RNA recognition and binding by proteins can depend on the structure in three ways:

1. Sequence-specific protein binding requires an unpaired region containing a particular sequence motif.
2. Sequence-specific protein binding can also require a structural context wherein the unpaired sequence is flanked by paired regions.
3. Largely sequence-independent binding occurs on double stranded RNA.



A protein binding two minor and one major groove of a double stranded RNA molecule. Figure inspired by [152].

This is not entirely true, there is evidence that dsRNA binding can be sequence dependent [70]. The mechanism for recognizing nucleotide identity is poorly understood.

GENOME-WIDE STUDIES CORRELATE SECONDARY STRUCTURE WITH TRANSCRIPT ABUNDANCE AS WELL AS FUNCTIONAL REGIONS OF THE INDIVIDUAL TRANSCRIPTS The importance of RNA secondary structure can be indirectly ascertained by noting where it occurs and how it correlates with other qualities, and/or properties of the transcripts. By exclusively sequencing single-stranded or double-stranded RNAs, Li et al. created a map of where transcripts contain structure (double stranded regions) as well as how structured they are [103]. They found consistent biases in the degree of structure in different regions of transcripts. The area around the translation start site at the end of the 5' UTR as well as the start of the 3' UTR are significantly less structured than other regions. Transposons and miRNA binding sites are statistically more structured. Studies on yeast and mouse [88, 78] have shown similar patterns of secondary structure prevalence using different probing methods, indicating that there is some consistent cause for the varying levels of pairing in RNA transcripts.

The excellent review by Mortimer et al. [124] provides copious examples of where RNA secondary structure is found, and what genome wide probing studies have revealed about the transcripts it is found in. Other regions and functions associated with secondary structure include regulation in the 5' and 3' UTRs. A high degree of structure in and around the Shine-Dalgarno sequence can impair the ability of ribosomes to bind and initiate translation. Structured sections of the coding regions can slow down the ribosomes and lead to alternate co-translational folding pathways for the proteins they encode. Secondary structure, as it pertains to its association with certain cell transport proteins, can confer it an important role in localization. The propensity for secondary structure to depend on temperature can allow it to function as an 'RNA thermometer', wherein the melting of a secondary structure leads to altered stability and/or binding of regulatory proteins. Finally, the presence of secondary structure can have an enormous impact on RNA regulation by small RNAs such as miRNAs. Indeed the very biogenesis of miRNAs depends on the presence of a stem loop structure which is cleaved by the Dicer protein.

While such aggregate measures don't directly implicate the secondary structure in any cellular function, their consistent findings across different studies provide strong evidence that it plays some role.

DOUBLE STRANDED RNA IS NECESSARY FOR RNA INTERFERENCE (RNAI) Secondary structure is crucial to the action of small RNA (smRNA) such as micro RNA (miRNA) and small interfering RNA (siRNA). Their biogenesis depends on the formation of a double stranded hairpin structure which is cleaved by Dicer [47, 3, 116]. Their downstream action requires nearly perfect base pair complementarity with a target region.

While this is likely due to some nuance of the tertiary structure formed, it is sufficient to be able to recognize that a particular secondary structure forms between an miRNA strand and a target molecule in order to predict where the silencing effect will take place. These two silencing pathways, siRNA and miRNA have revolutionized the field of molecular biology and are slowly making their way into pharmaceuticals targeting aberrant transcripts via the duplex it makes with an siRNA [186].

2.3 *Determining secondary structure*

The previous section outlined a few examples demonstrating why RNA's secondary structure is functionally important. In order to make these inferences, we need to know what the secondary structure of an RNA molecule actually is. Full knowledge would imply that we know which pairs of nucleotides are paired. Partial knowledge implies that we know whether a particular nucleotide is paired at all. But how do we measure this? Is there a way to interrogate the secondary structure of RNA? The rest of this section will present three methods of measuring the secondary structure of RNA. Phylogenetic analysis seeks to extract base pairing information through patterns of covariation in evolutionarily related RNA sequences. Chemical probing exploits differences in the reactivity of paired and unpaired nucleotides to determine their pairing status. Finally, X-ray crystallography and NMR spectroscopy yield snapshots of the full three dimensional structure of RNA molecules from which the secondary structure can be extracted.

COMPARATIVE ANALYSIS USES COVARIATION AMONG PAIRED NUCLEOTIDES IN SEQUENCE ALIGNMENTS TO REVEAL PAIRING PATTERNS Comparative analysis of RNA sequences yields a rich source of information about secondary structure. While it's not necessarily a method for confirming a putative structure, the evidence it provides can substantiate other methods of structure determination, or vice versa. In fact, some of the most significant secondary structures, such as those of the 5S [51], the 16S [189] and the 23S [128] RNA were initially determined by using a combination of comparative phylogenetic analysis and chemical probing.

Comparative analysis works on the simple assumption that nucleotides which are paired will evolve compensatory mutations to maintain their pairing [188]. If a nucleotide at position x is paired with a nucleotide at position y and a mutation changes the identity of nucleotide x, then we would expect another mutation to change the identity of nucleotide y to maintain the pairing. Searching for pairs of nucleotides that change identity in tandem is known as covariation analysis. Covariation between

pairs of nucleotides is strong evidence that these nucleotides are paired or functionally dependent on each other.

DIFFERENCES IN REACTIVITY HINT AT SECONDARY STRUCTURE
Chemical probing experiments use a reagent that either cleaves or modifies nucleotides which are either paired, unpaired, or have some reactive group exposed to chemical reaction. When RNA is chemically modified, the identity of the modified nucleotides is usually determined by reverse transcribing the modified RNA and sequencing the resulting DNA. As the reverse transcriptase moves along the RNA molecule, the modifications cause it to dissociate from the template RNA strand, leading to truncated DNA molecules. By analyzing where reverse transcription was terminated, researchers can infer the sites of modification and assign a pairing status to them (i.e. paired, unpaired, protected, etc...).

Dimethyl sulfate (DMS) footprinting, for example, is a method for ascertaining the secondary structure and/or protein binding status of an RNA molecule by treating it with the eponymous chemical reagent [174, 135]. Upon treatment, dimethyl sulfate (DMS) methylates either the N1 atom of adenosine or the N3 atom of cytidine, but only when they are *accessible* (i.e. not occluded by other nearby bases). The reagent will only react with its target, when that target (i.e. the N1 or N3 atom) is not sequestered in a base pair or a protein interaction. Knowing which nucleotides have reacted with DMS, a researcher can infer which nucleotides are not paired or interacting with a protein. Because the methylation performed by DMS stops the progress of reverse transcriptase, the procedure described previously can be used to determine where the reactions took place.

In addition to DMS footprinting, a variety of other methods exist for probing RNA secondary structure [183]. All work on the principle that some detectable reaction takes place at a certain location in the RNA and reveals some property of the secondary (or tertiary) structure at that position. While these methods used to be performed on single sequences, recent advances in high-throughput methods have enabled nucleotide-resolution characterization of multiple sequences and secondary structures at once [107, 91]. This, in turn, has enabled the genome-wide scans of RNA secondary structure presented in Section 2.2.

X-RAY CRYSTALLOGRAPHY AND NMR! (NMR!) PROVIDE PHYSICAL SNAPSHOTS OF THE TERTIARY STRUCTURE OF RNA MOLECULES.
Perhaps the most definitive methods for determining the secondary structure of RNA involve solving its tertiary structure first. Methods such as X-Ray crystallography and NMR spectroscopy allow us to build precise 3D models of the tertiary structure of an RNA. Because the secondary structure is implicit in the tertiary structure, extracting it is only a matter of analyzing the data and using a program such as MC-ANNOTATE [56],

RNAView [194] or 3DNA [106]. These methods paint the most detailed pictures of the conformation of RNA, but they come with significant drawbacks.

X-Ray crystallography requires the presence of crystals of the RNA molecule which due to RNA's flexibility are significantly harder to obtain than for proteins of a similar size. Even in the presence of crystals, the images produced have to be of adequate quality and resolution to be able to infer regions of electron density and atom positions. The images produced by X-Ray crystallography are necessarily static snapshots of the molecule frozen in time. There is little room for extracting information about dynamics or conformational changes under varying ion concentrations. Nevertheless, with effort and some luck, crystallographic images can yield significant insight into the structure and function of important RNAs such as the ribosome [10], the group I intron [28], ribonuclease P [149] as well as various others.

Nuclear magnetic resonance spectroscopy relies on exposing a sample to a large magnetic field, irradiating it with a radio frequency pulse and measuring the resonance frequencies of the atoms in the sample. Analysis of the resulting spectra can reveal which atoms are connected via covalent bonds as well as which are physically near each other. Because the measurements are performed over a period of time, there is some variation in the structure of the molecule as it is being recorded. This, however, is rarely enough to record catalytic events or major structural changes. NMR can work well for smaller molecules but becomes intractable for larger ones due to the crowding of the frequency signals. Nevertheless, the fact that the molecules are suspended in solution makes preparation significantly easier than for X-ray crystallography.

Tertiary structures resolved by X-ray crystallography and NMR spectroscopy are traditionally deposited in the Protein Data Bank (PDB) with a unique 4 character identifier. X-ray structures in the PDB are used extensively in our work to benchmark our prediction method as well as to extract fragment data.

2.4 Secondary structure prediction

Having discussed the importance of secondary structure as well as how to measure it, this section will introduce computational approaches to predicting it. By accurately predicting secondary structure, we can bypass the costly and time-consuming experimental methods for its measurement and go straight to creating hypotheses about its functional roles. It should be emphasized that while secondary structure prediction is immensely useful, it remains conjectural. It is a prediction and not a measurement. It can be used by researchers to explain their findings and to formulate new

Venkatraman Ramakrishnan, Tom Steitz and Ada Yonath won the 2009 Nobel prize in chemistry for solving the structure of the ribosome, which can be found in the PDB under the accession 1FFK.

ideas about mechanisms of action, but it is not an experimental technique which can be used to confirm hypotheses.

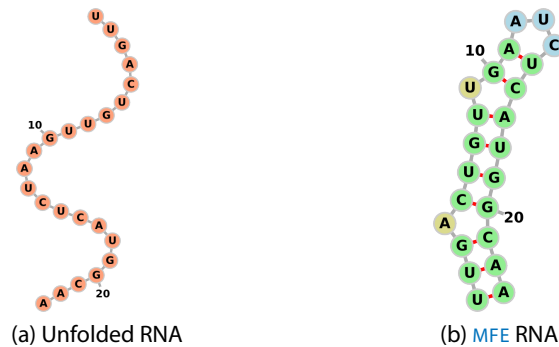


Figure 1: RNA folding requires intra-molecular base pairing.

Figure 1 illustrates secondary structure by showing an unfolded (unpaired) structure on the left and its folded equivalent on the right. The lines in grey indicate the backbone, or the covalent phosphodiester connections between adjacent nucleotides whereas the red lines indicate the base pairs between complementary nucleotides. The figure on the right has secondary structure because it has intra-molecular base pairs. One may wonder why we see the secondary structure on the right of Figure 1 as opposed to any other possible secondary structure? Surely the U at position 2 could also pair with the A at position 1, and the C at position 5 could pair with the G at position 20, etc ...

Figure 2 shows two potential structures for a given sequence. So why would we expect to see the structure on the right in nature rather than the structure on the left? The answer is that the structure on the right is more energetically favorable than the structure on the left. It has more base pairs. But how do we know this? How can we calculate which secondary structure is optimal?

To calculate the optimal secondary structure for a given sequence, we need to find the set of base pairs present in the lowest energy structure, where the energy of a structure is calculated according to an *energy model*. The simplest such model might count the number of base pairs that the structure contains [130]. A slightly more complex version may count the number of hydrogen bonds that are formed. The most widely used energy model takes into account the stacking of adjacent base pairs, assigns energies to unpaired loops and considers a variety of other parameters whose energy contributions are determined using thermodynamic melting experiments [115]. This energy model is combined with an algorithm for finding the set of pairings which define the structure with the minimum free energy (the MFE structure).

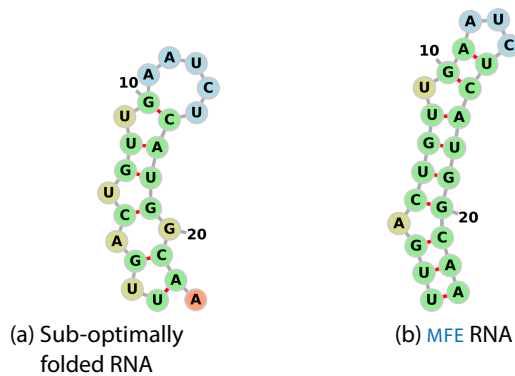


Figure 2: RNA can fold into different secondary structure conformations.

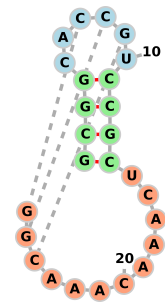
How does such an algorithm work? One option is to enumerate every possible structure, evaluate its energy and return the one with the lowest. Given the exorbitant amount of possible structures, such an approach is computationally infeasible. A more rational method uses dynamic programming, wherein the solution to the problem of finding the MFE structure for the entire sequence is built up by finding MFE structures for subsequences. The score for any subsequence from position i to position j can thus be expressed as a recurrence relation [42, 130, 181]:

$$S(i, j) = \begin{cases} S(i+1, j-1) + 1, & \text{if } i \text{ and } j \text{ are paired} \\ S(i+1, j), & \text{if } i \text{ is unpaired} \\ S(i, j-1), & \text{if } j \text{ is unpaired} \\ \max_{i < k < j} S(i, k) + S(k+1, j) & \end{cases} \quad (1)$$

By storing intermediate values for $S(i, j)$ we build up a matrix of solutions to sub-problems that are eventually extended to the entire sequence. Backtracking through the matrix yields the structure that corresponds to the optimal solution. This algorithm runs quickly, is deterministic, and produces the minimum free energy structure (according to the energy model). While the recurrence relations in Equation 1 provide a simple example of how a dynamic program algorithm builds up its table of scores, modern implementations use a more complicated energy model which distinguishes different types of loops (e.g. hairpin, interior, multi) and assigns them corresponding energies [200, 178].

It is important to note that one of the reasons for inaccurate predictions will never be that the MFE energy is not found. The algorithm described above is guaranteed to always return the optimal structure in terms of its

Pseudoknot-containing RNA structure



The dashed lines indicate nested base pairs.

energy. Any discrepancy between the [MFE](#) prediction and the structure observed in the cell can be attributed to one of three causes:

1. Inaccurate energy model: the energy calculated does not reflect the real energy of the structure.
2. The structure found in the cell is not the [MFE](#) structure: this can occur because of kinetic traps, protein binding, varying ion concentrations, etc. . .
3. The structure contains long-range and tertiary interactions which can not be predicted by the dynamic programming algorithm described above.

This is in sharp contrast to the algorithms employed for tertiary structure prediction which often never find the minimum possible free energy dictated by their energy model. This discrepancy can be attributed to two facts: that RNA secondary structure is discrete with a finite number of possible conformations and that the energy for subsequences is additive. In other words, the secondary structure can be decomposed into independent sections whose energies can be combined, allowing algorithms to quickly find the minimum free energy structure. Tertiary structure, in contrast, has many more potential conformations as well as non-nested interactions which make a systematic search for the [MFE](#) much more complicated.

SHORTCOMINGS OF SECONDARY STRUCTURE PREDICTION While immensely useful, secondary structure prediction does have its shortcomings. Most notably, it occasionally fails to predict results consistent with experimental data. Evidence for this can easily be found in benchmarks where no single program can predict every target accurately [145]. A more concrete, applicable and challenging example can be found in the prediction of a secondary structure for the satellite tobacco mosaic virus ([STMV](#)) which does not contain the helices seen in the cryo-electron microscopy ([cryo-EM](#)) photos of the particle [155]. Such examples are no surprise due to the fact that RNA secondary structure alone explains neither long-range interactions, nor inter-molecular interactions, nor (or extremely rarey) does it take into account varying ion concentrations. This is yet another impetus for the development of tools which strive to take long-range tertiary interactions into account. While our tertiary structure prediction method does not seek to modify the input secondary structure, other prediction methods can operate on the sequence alone and return an implicit prediction of the secondary structure along with the tertiary [35, 82, 131].

One can imagine opening up a base pair to allow a stable long-range interaction to form.

In between the prediction of secondary structure and tertiary structure (Chapter 4) is the emerging field of extended secondary structure prediction, which seeks to identify the optimal canonical as well as noncanonical pairings between nucleotides given a sequence. Progress on this front, however, has only recently begun and has much room for improvement.

EXTENDED SECONDARY STRUCTURE PREDICTION SEEKS TO PROVIDE A STEPPING STONE BETWEEN SECONDARY AND TERTIARY STRUCTURE As mentioned in the previous section, noncanonical base pairing plays a major role in the full tertiary structure of RNA molecules [171, 120]. While progress on the prediction of canonical secondary structure has been swift and comprehensive, the prediction of noncanonical interactions has only recently begun. Programs such as RNAWOLF [199] and JAR3D [198] can take a sequence and predict the non-Watson Crick interactions present among the available unpaired nucleotides. Unfortunately, the energy parameters required to describe noncanonical interactions are not as well defined and much more numerous leading to difficulty in achieving accuracy comparable to that of canonical structure prediction. Nevertheless, such interactions bring us one step closer to the prediction of the tertiary structure of an RNA molecule and can be used as an input to full tertiary structure prediction methods, an approach that we have begun to develop as described in Section 11.3.

While RNA secondary structure is defined by the base pairing between intra-molecular nucleotides, tertiary structure describes the structure of the molecule in 3D space. The tertiary structure describes where helices, nucleotides and atoms are located with respect to each other. A full tertiary model of an RNA molecule is defined by the positions of all its atoms and can be obtained using X-Ray crystallography or NMR spectroscopy (Chapter 2). Lower resolution methods such as cryo-EM provide a coarser view of electron density wherein only the positions of the larger regular components such as helices can be accurately resolved. . At an even lower resolution, small angle X-Ray scattering (SAXS) experiments yield information about the general shape and radius of gyration of a molecule.

Much as experimental techniques yield tertiary structure data at varying levels of resolution, tertiary structure prediction tools also create models at varying levels of resolution. As described in detail in Chapter 4, the available tools yield predictions with details from the level of individual atoms to the level of entire helices. Before these methods are introduced in Chapter 4, this chapter will provide examples of the importance of tertiary structure in explaining molecular mechanisms. These examples were chosen to demonstrate how RNA's tertiary structure plays a role in key cellular processes. They seek to highlight the importance of features that can not be described by the sequence or secondary structure alone and thus require some knowledge of the geometric arrangement of the atoms of the RNA to explain their mechanism of action. This is not meant to be an exhaustive list, and it intentionally omits the most striking example of the ribosome in favor of simpler, perhaps less well known anecdotes involving shorter stretches of RNA.

THE TELOMERASE PSEUDOKNOT IS CRUCIAL TO THE ELONGATION OF TELOMERES Telomeres are sections of repeated nucleotides at the ends of chromosomes which are essential for DNA replication and chromosome protection [104]. After DNA replication, they are extended by an enzyme known as *telomerase* in a process with far-reaching implications for the health and longevity of the organism. Key to the function of telomerase is a bound pseudoknotted RNA strand which serves as a template for DNA extension. This pseudoknot is stabilized by a triple-helix structure, the necessity of which has been demonstrated not only by mutational studies but also by comparative analysis. A lack of conservation at

Recent advances can actually yield near-atomic resolution from cryo-EM data [8].

The 2009 Nobel Prize in Physiology was awarded to Jack Szostak, Carol Greider, and Elizabeth Blackburn for the discovery of "how chromosomes are protected by telomeres and the enzyme telomerase".

the sequence level is balanced by a strong conservation on the structural level which is highly suggestive of a structural basis for its function [166].

PKR-MEDIATED TRANSLATION INHIBITION DEPENDS ON NUCLEOSIDE MODIFICATIONS One of the innate cell immunity mechanisms to defend from invading viruses is PKR-mediated translation inhibition. By recognizing foreign RNA particles, the signaling cascade triggered by PKR leads to the inhibition of translation in the cytoplasm and thus interferes with the production of new viral particles. The basis for the recognition of foreign RNA molecules is not completely clear, but it is thought to be dependent on the tertiary structure of the target molecules. More specifically, transfer RNA (tRNA) isolated from the mitochondria, which lack nucleoside modifications are able to activate PKR whereas cytoplasmic tRNA molecules, which tend to be heavily modified do not trigger PKR. Even though the two molecules have the same sequence, secondary structure, and likely tertiary structure, the nucleoside modifications affect its ability to activate PKR. This could be due to the modifications' tendency to prevent the dimerization of tRNA molecules, a state that also leads to the activation of the innate immune response. A detailed view of the tertiary structure could shed light onto the mechanisms and factors responsible for PKR recognition and activation [126].

PROTEIN TRANSLOCATION TO PLASMA MEMBRANE DEPENDS ON THE TERTIARY STRUCTURE OF THE SIGNAL RECOGNITION PARTICLE Many proteins need to be translocated from the ribosome to the plasma membrane upon translation termination or during transcription itself (along with the ribosome). Nascent membrane-bound peptides have a hydrophobic *signal* amino-acid sequence which indicates that they should be inserted into the membrane. The exact sequence of events leading to this translocation, while not entirely clear at the moment, is known to rely on the presence of a highly conserved 4.5S signal recognition particle (SRP) RNA. Two proteins, FtsY and Fhf interact with the SRP RNA and the nascent signal sequence to attach it to the translocon and move the entire complex to either the endoplasmic reticulum or the plasma membrane [6]. The structure of the SRP RNA is essential to this process as it provides a scaffold for the FtsY and Fhf proteins to attach and mediate the translocation process. Furthermore, the presence of a single bulged nucleotide is essential for the function of the entire complex.

COOPERATIVE RNA STRUCTURE DEPENDENT PROTEIN BINDING HELPS REGULATE TRANSLATION The protein sponge RsmZ acts to sequester the translation repressor RsmE. A single RsmZ RNA can bind up to five RsmE dimers which serves to sequester it and inhibit its ability to repress translation. The binding of this protein is cooperative insofar

as the binding of the first and second dimers increases the affinity for the third dimer by changing the tertiary structure of the RNA molecule and bringing the two binding sites for the third dimer into an optimal configuration for protein binding[41]. This mechanism of action requires no changes to the secondary structure of RsmE and is mediated purely by alteration of its tertiary structure induced by the binding of the RsmZ protein.

RNA SPLICING DEPENDS ON SMALL NUCLEAR RNAS RNA splicing, the mechanism for excising introns from a nascent RNA transcript, relies on a major assembly of proteins and RNA molecules (dubbed the spliceosome) to convene, cut out the intron and ligate the two ends of the exons. While the exact mechanisms of the well orchestrated cleavage and ligation reactions are still not completely characterized, it is known that non-messenger RNA molecules play an important role in the process. In particular, the U6 small nuclear RNA (*snRNA*) is responsible for the catalysis of both splicing reactions [44]. Its major role lies in positioning metal ions such that they catalyze the phospho-transesterification of splicing. This is achieved by forming a triple-helix structure similar to that found in the group II intron [45]. Tertiary interactions between an AGC triad (in a helix) and a conserved bulge 5bp away provide a scaffold for coordinating the two metal ions. Thus, the *snRNA* plays an important role in the splicing reaction in large part because of its tertiary structure.

The Group II intron, a putative predecessor of the spliceosome, provides a striking example of the importance of RNA tertiary structure in catalyzing splicing reactions. The Group II intron is a self-splicing ribozyme which is capable of inserting itself back into a genome [19]. The self-splicing functionality can be detected *in vitro* in the absence of auxiliary proteins, making it particularly curious in light of the RNA world hypothesis. It is thought to share functional and evolutionary similarities with the spliceosome responsible for pre-mRNA processing in eukaryotes. In this light, insights into the process of splicing on an atomic level are revealing on both a mechanistic as well as an evolutionary level. Such studies delve deep into the role that tertiary structure motifs such as triple helices, kink turns, and ion binding active sites play in positioning and catalyzing the excision of the intron from the containing transcript [111].

TRNA BINDING REQUIRES A PRECISE FIT BETWEEN THE TRNA AND THE T-BOX RNA *tRNAs* are ubiquitous within all living cells and serve to transport amino acids to the peptidyl transferase center of the ribosome so that they can be incorporated into the nascent polypeptide chain. They are often subject to modification [63, 168, 2] and cleavage and are the targets of multiple regulatory pathways [90]. While the importance of the tertiary structure in modification and cleavage is self-evident,

its role in regulation is less trivial. To maintain a constant concentration of tRNA, over-abundance should lead to decreased production and scarcity should lead to over-expression. In many scenarios, these regulatory constraints are mediated by the pairing of a ncRNA to an mRNA transcript and thus causing its (the mRNA's) degradation. While this process relies on base pair complementarity between the ncRNA and the mRNA, tRNA is recognized by a regulatory element by way of its tertiary structure [59].

The T-box riboswitch is a regulatory element found in the 5' end of gram-positive bacteria that regulates molecular machinery tasked with maintaining the pool of charged amino-acyl tRNAs. It recognizes uncharged tRNAs and changes its conformation leading to the expression of genes associated with the charging of tRNAs [59, 110]. When there is no bound uncharged tRNA, a downstream terminator stem is formed which leads to a downregulation of gene expression.

The tertiary structure formed by the T-box RNA recognizes the tertiary fold of tRNA molecules. This recognition, in turn, effects a change of conformation within the T-box RNA bringing the 3' terminus of stem I closer to the acceptor end of the tRNA and thus allowing the proper regulatory function of the T box RNA. The recognition of the tRNA requires the presence of an internal "E-loop" which recognizes the anticodon sequence of the tRNA molecule. The stem loop at the apex of stem I then stacks onto the t-loop of the tRNA molecule allowing for a sort of docking of the tRNA along the stem I structure of the T-box RNA. This binding leads to the folding of the proximal residues of stem I into a K-turn. With little to no change in the secondary structure of either molecule, a regulatory effect is achieved by the recognition of the tertiary fold of the tRNA by the T-box RNA and the alteration of its tertiary structure upon binding.

GROUP I INTRON The Group I intron is a class of self-splicing ribozymes found in protozoa (*Tetrahymena*) and some bacteriophage T4 genes [104]. While they can function alone, they are also found in ribonucleoprotein (RNP) complexes [134]. In the example cited above, a mitochondrial tyrosyl-tRNA synthetase (TyrRS) "*provides an extended scaffold for the phosphodiester backbone of the conserved catalytic core of the intron RNA, allowing the protein to promote the splicing of a wide variety of group I introns*" [134].

SUMMARY These examples should provide an introduction to how tertiary structure influences RNA function. It should also serve as justification for the work done predicting tertiary structure. If we could accurately measure or predict tertiary structure, we would be better positioned to understand the underlying mechanisms of the processes mentioned above. The road to *accurate* prediction, however, is long and winding. To reach our goal, we have to explore many different paths. This, both proverbially

and literally, is the goal of this work. We wish to sample and evaluate many different structures in the hope of finding some which are similar to the native (real) structure.

PREVIOUS METHODS FOR RNA TERTIARY STRUCTURE PREDICTION

4.1 *Birds-eye view of existing methods*

The field of tertiary RNA structure prediction started in earnest in the mid-2000s with the publications of FARNA [35] and the MC-Fold | MC-Sym pipeline [131]. Since then, the field has expanded to encompass a number of different ways of building tertiary RNA structures. With the exception of BARNACLE, the discrete molecular dynamics-based approaches [39, 165] and GARN [22], most methods employ some form of fragment assembly to stitch together pieces of known structures to form a prediction or set of predictions of the tertiary structure of the input sequence. The following two paragraphs will briefly describe how each method generates its structures, while the sections thereafter will provide a more comprehensive overview of how they come to their final predictions.

Fragment assembly can come in many different forms, depending on the size of the fragments and how they are selected. FARNA and FAR-FAR [36] use fragments of consecutive nucleotides up to 3 nucleotides long which are taken from the ribosome crystal structure. NAST [82] appears to use torsion angles obtained from consecutive nucleotides found in the ribosome. The MC-Fold | MC-Sym pipeline uses nuclear cyclic motifs, which represent fragments that contain both covalently linked as well as base-paired nucleotides. 3DRNA [197], RNACOMPOSER [142], and ERNWIN [87] all use fragments from entire secondary structure elements (i.e. stems or loops). While RNACOMPOSER and ERNWIN use fragments that are either automatically generated, present in the PDB or found in previously predicted structures, 3DRNA extracts fragments from the SCOR Database [172], the RNA Junction database, or from the PDB (in the case of pseudoknots).

The graph-based sampling approach of Kim et al. uses junctions predicted using the RNA Junctions as Graphs module (RNAJAG) [98] and fixed angle and distance increments for changing the orientations of the non-junction bound helices. GARN [22, 99], in a similar fashion, uses fixed torsion angle increments of 60 degrees in order to adhere to the triangular lattice on which its simulations take place. The discrete molecular dynamics-based approaches use a discretized force field to push the structures toward energetically favorable conformations and thus do not need a special method for assembling structures. Finally, BARNACLE uses a

generative probabilistic model which generates realistic local structures in continuous torsion angle space.

The remainder of this chapter will introduce each of the methods previously mentioned in more detail. This is intended to serve as a description of the context within which we developed ERNWIN [87] as well as to underscore which aspects of other prediction methods we tried to improve. Our approach is detailed in a publication [87] in the journal RNA and is reproduced in Chapter 6 here. More information about our coarse-grain model can be found in the Methods (Chapter 10) while its uses and potential improvements can be found in the Results (Chapter 11).

4.2 FARNA and fragment assembly

Building on the success of the ROSETTA package [169] for protein structure prediction FARNA (fragment assembly of RNA) [35] employs fragment assembly coupled to a knowledge based potential to create models of tertiary RNA structure. The fragments used for building the structure are selected based on a reduced two letter, purine - pyrimidine based alphabet. The energy function contains general terms for compactness ("*proportional to the radius of gyration*"), steric clashes for avoiding excluded volume as well as a number of RNA specific terms such as the relative orientations of the nucleotides involved in base pairs, the stagger (how misaligned the base centroids are along the axis normal to the plane of one base) and the stacking (how aligned the base normals are) between successive paired nucleotides, as well as a number of other minor energy terms relating the orientations of interacting nucleotides.

The results claim that FARNA can and does recapitulate noncanonical structures found in small fragments, but has a harder time with larger, more complex structures. The authors state that deficiencies in both the sampling procedure and the energy function are to blame for the poor predictions. Evidence of the former lies in the fact that for some sequences no structures are sampled with energy lower than the native, while the latter is diagnosed by the presence of non-native like structures with energy lower than that of the native. Both of these inconsistencies can be observed when generating structures for different inputs indicating that their approach suffers from both insufficient sampling as well as from an inaccurate energy function. The authors hint that the problem of insufficient sampling can be addressed by using more computational resources whereas the problem of inaccurate energy can be tackled by incorporating fine-grained terms such as hydrogen bonds as well as explicit inclusion of water locations into the energy function. These suggestions are prescient of the publication of the next paper on using ROSETTA as an RNA structure prediction tool (FARFAR).

Diagnosing sampling vs. energy functions helps distinguish which aspect of a structure prediction method needs improvement.

As the FARNA method is implemented within the ROSETTA structure prediction suite, all future references to the use of ROSETTA for RNA structure prediction will refer to the method described here as FARNA as well as the FARFAR extension described below.

4.3 *FARFAR adds high resolution energy*

As an extension of the low-resolution FARNA method, FARFAR [36] adds an additional high resolution refinement step using the ROSETTA all-atom energy model. It achieves exemplary results for half of the 32 models (with lengths of 6-23 nt) of short noncanonical segments that it was tested on. In the cases where poor models were created, the blame fell on the sampling procedure due to its failure to produce models with energy as low as or lower than the native structure. The successful predictions were limited to motifs with 18 or fewer nucleotides. In addition to computing the structure for a given sequence, the reverse problem of computing sequences to fit a particular structure was attempted and verified using structure mapping experiments.

If we could accurately predict noncanonical motifs, then they could be spliced together to generate larger structures. We could create libraries of accurate predicted conformations for any number of RNA motifs and use those as building blocks to assemble larger structures. Assembly methods for larger structures already exist and include the structure prediction topic of this thesis. Having a library of **accurate** structural fragments should serve to significantly constrain the conformational space that needs to be explored while sampling RNA tertiary structures. Such an approach has been tried within the ROSETTA framework [91] and could also be combined with our implementation in order to improve the efficiency with which the conformational space is explored.

4.4 *The MC-Fold | MC-Sym pipeline*

In one of the first and most basic attempts at piecing together secondary structure elements, Parisien et al. introduced the use of nuclear cyclic motif (NCM) for the automated generation of both secondary and tertiary structures using the MC-Fold | MC-Sym pipeline [131]. NCMs are small indivisible cycles that are present in any RNA secondary structure when it is represented as a graph with nucleotides as nodes and backbones and base-pairs (both canonical and noncanonical) as edges [100]. These cycles can be extracted using Horton's algorithm [74] and the smallest such cycle would contain 3 nucleotides.

The MC-Fold step of the MC-Fold | MC-Sym pipeline generates secondary structures by combining NCM fragments. For hairpin loops, MC-

Fold enumerates a set of [NCMs](#) which can be used to construct the loops. For multiloops a recurrence relation is defined and the generation procedure descends into subsections of the molecule while filling in a dynamic programming table of energies. The energies contain terms derived from the probability of seeing an [NCM](#) given a sequence as well as the probability of seeing combinations of [NCMs](#). The resulting output is a list of secondary structures sorted by energy values.

This list of secondary structures can then be passed to the MC-Sym portion of the simulation, which generates tertiary structures from the assigned [NCMs](#). Because each [NCM](#) has a number of 3D fragments associated with it, and each structure is composed of multiple [NCMs](#), it is the job of MC-Sym to align adjacent fragments and to weed out structures with steric clashes. For this it uses a Las Vegas algorithm to "try and explore as many structures as possible within a 12 hour period". These structures (all in [PDB](#) format) can then be piped to MC-Cons to extract a consensus structure which maximizes the similarity score between it and all other predicted structures.

The approach proves well suited to predicting both secondary as well as tertiary structure of RNAs given their primary nucleotide sequence. Parisien et al. argue that it yields secondary structures with a higher Matthews correlation coefficient ([MCC](#)) than RNAsubopt for a corpus of hairpin loop structure derived from the [PDB](#). It furthermore yields numerous tertiary structure predictions with an [RMSD](#) of less than 4Å from the native for 13 tertiary structures described as hairpin loops, multi-branch (Y-shape) and pseudoknot structures. The quality of these results, however, is balanced by the running time which is listed as taking 12 hours (for structures up to 47nt long). The results are promising, having low [RMSD](#) and high [MCC](#) values, with the caveat that the presented results are chosen as the best based on knowledge of the real structure. This does little to demonstrate the predictive power of the model in the case when the real structure is not known and [MCC](#) and [RMSD](#) values can not be calculated.

We improve upon this by reporting not only the best [RMSD](#) values, but also the entire distribution of sampled [RMSD](#) values of our simulation (in the case of benchmarks when the real structure is known). We also sample faster and show results for structures up to 298 nucleotides long (Ribonuclease P, [PDB](#) ID: 2A64 [84]).

4.5 *ModeRNA for homology modeling*

Adopting a technique commonly used for proteins, ModeRNA [151] uses template based modeling to generate a tertiary structure starting from a template (already known structure) and an alignment which indicates how the query relates to the template. ModeRNA then performs a number

of actions including "copying coordinates or residues that are invariant between the target and the template, introducing substitutions for aligned residues that differ, adding or removing post-transcriptional modifications, processing insertions / deletions (indels) and adding structural fragments for short regions without a template", in order to generate a putative model of the query structure.

While changing bases and adding nucleotide modifications is relatively trivial, dealing with insertions and deletions is slightly more involved. Deletions require mending the backbone at the point of deletion while insertions require finding a fitting fragment to insert. Its flanking 5' and 3' ends must approximately overlap the 3' and 5' ends of the nucleotides flanking the insertion point in the template. It must have a sequence similar to the query sequence and it must not introduce steric clashes when inserted into the template structure. A search for such a fragment is run on a database of resolved 3D structures. Insertion of the fragment is followed by a round of full cyclic coordinate descent (FCCD) [20] to connect its ends to the backbone of the template.

ModeRNA works well for modeling homologous sequences but suffers when two sequences diverge evolutionarily. It will never find a completely novel arrangement of helices, nor does it give any consideration to the long-range interactions which are so important in stabilizing a tertiary structure. Active sites or functionally important loop regions which differ between the target and query are modeled with little regard to the context beyond the immediate backbone vicinity and may thus miss important non-local interactions. Nevertheless, ModeRNA can be a worthwhile tool for obtaining a glance at what the tertiary structure of a new sequence may look like as long as it is related to a known sequence. By building up structures from much smaller fragments, we avoid the pitfall of needing a close evolutionary relative and allow users to get an idea of novel helical arrangements that may be possible for a given secondary structure.

4.6 *Discrete molecular dynamics for fast force fields*

Shortly after the work of Das et al. and Parisien et al. in predicting RNA structure using fragments and nuclear cyclic motifs came one of the first uses of coarse-graining for RNA structure prediction. Ding et al., in attempting to speed up molecular dynamics-based folding simulations, introduce the use of a three-bead per nucleotide model of RNA structure [39]. They run simulations using discrete molecular dynamics (DMD), which replaces the traditional continuous energy potential functions used in molecular dynamics situations with discrete analogs [143] in order to reduce their computational demands. By discretizing the potential function, the forces between two nucleotides can often be omitted when they are far enough from each other.

The simulations outlined in Ding et al. extol a fast method capable of elucidating not only the predicted structure, but also the path it followed to to the minimum free energy structure. They boldly claim that the "majority of DMD-predicted 3D structures have less than 4Å deviations from experimental structures" (for their test set of structures between 10 and 100 nt), which would be an extraordinary achievement even now seven years after the publication. Unfortunately, a cursory examination of the structures predicted reveals that a large majority are simple helices with a short hairpin loops which are likely also modeled well using FARNAs. Nevertheless, there is no doubt that their approach is a novel way to tackle the RNA folding problem, albeit one with limited applicability to larger RNAs.

iFoldRNA is a web-based implementation of the DMD method outlined above [165]. The three bead per nucleotide representation obtained is decorated with the positions of the individual atoms and returned to the user. While useful as an easy-to-use web-based prediction tool, it only works for queries containing less than 50 nucleotides. The time required to run a simulation can take up to a half hour.

As no binary is available for the original DMD application and structures longer than 50 nucleotides can not be submitted to iFoldRNA, it is difficult to assess the performance of this approach on longer RNAs with more convoluted secondary structure. The stated running time of 5h for 2×10^6 DMD time units (presumably long enough to closely approach the global minimum energy) on 8 cpus, furthers dims the prospects of its utility for longer structures.

4.7 *NAST coarse-grains and uses a knowledge-based potential*

Extending the concept of coarse-graining, Jonikas et al. introduced a one-point per nucleotide model of RNA tertiary structure [82] as part of the Nucleic Acid Simulation Tool (NAST). They use an energy function derived from observed dihedral angles and distances between two, three and four sequential nucleotides. This training data is obtained from the crystal structures of three different ribosomes. They include a repulsive term between unbonded nucleotides to compensate for excluded volume and constrain paired nucleotides to geometries favoring an ideal right-handed A'-form helix. Finally, NAST allows the input of tertiary structure constraints for long-range interactions. These, however, must be manually added from known sources of data such as probing experiments, FRET studies and/or already known crystal structures. The article makes scant mention of how local structure is sampled other than that it "samples local geometries observed in ribosomal RNA".

The authors report an RMSD of 16.3Å for the 158-residue P4-P6 fragment of the Group I intron (PDB ID: 1GID [28]), which is achieved with the

help of long-range contact information from covariance analysis [120]. They claim that their method is feasible for structures up to and above 377 nucleotides in length. While this may be true, enforcing secondary structure via terms added to the energy function increases the ruggedness of the energy landscape and increases the difficulty of finding the global minimum. The simulation will have to find both the ideal secondary structure as well as the tertiary structure to yield a correct prediction. We overcome the former challenge by starting with already formed secondary structure features and focusing on arranging them in 3D space. This eliminates the need to bring pairing partners together and allows us to focus on finding potential long-range interactions.

4.8 *BARNACLE creates models in continuous torsion angle space*

So far, every method listed has either used fragments from known structures [35, 131, 151] or a physics-based simulation to gradually push atoms towards their minimum free energy state [36, 39]. In the former, the proposal distributions have relied on stitching together discrete fragments from the corpus of known structures. While effective, these methods run the inherent risk of missing any fragment which hasn't currently been cataloged in the (rather meager) collection of known RNA structures. Local conformations in the vicinity of a known fragment can be completely invisible to the sampling procedure due to their absence from the structure database. This situation is antithetical to what actually occurs in nature, where conformations exist in continuous space and can occupy states infinitesimally close to others. The sheer number of potential conformations available to even the smallest RNA render the idea of explicitly storing every possible fragment lamentably unfeasible.

That is not to say it's not possible sample such a variety of conformations. It just requires an adequate representation in the form of a continuous probability distribution. The work of Frellsen et al. [52] introduces the use of dynamic bayesian networks (DBNs) [57] to create a generative probabilistic model of RNA tertiary structure. They build on a prior approach [21] to sampling realistic local protein structure to create a framework for not only generating local RNA structure, but also evaluating its probability. The authors show that the distribution of dihedral angles sampled from BARNACLE closely matches that of known 3D structures. By modeling torsion angles using the circular von Mises probability distribution [112], they avoid the problem of discretizing the conformational space and create a proposal distribution that yields infinitely variable tertiary structures. Combining this model with an energy function is demonstrated in a didactic manner and largely left to future endeavors.

The software is available as a stand-alone program or as a set of python bindings. Due to the lack of energy terms for secondary structure or long-range interactions, its utility is limited to sampling short single-stranded fragments. With minor tweaking it could be retooled to generating all-atom models of loop regions. If the helical regions are positioned using a coarse-grain model such as ours, then the loop regions can be accurately and efficiently built using a fine-grain tool such as BARNACLE. Such a fusion would require implementing an energy term to coerce BARNACLE fragments to start and end at certain points and is left as a potential avenue for further exploration.

4.9 *3DRNA assembles fragments from SCOR and RNAJunction*

3DRNA [196] is a method of assembling secondary structure elements (SSEs) to form complete 3D models. It takes loop fragments from the Structural Classification of RNA (SCOR) database, junction fragments from the RNAJunction database, and pseudoknots from structures in the RCSB PDB database. Fragments are selected from the databases based on sequence similarity and then stitched together by aligning the overlapping regions (i.e. the 3' end of a forward fragment to the 5' end of the rear fragment). Structures are then subject to molecular dynamics energy minimization using the AMBER 98 force field to correct any inconsistencies before the results are returned to the user. Its reliance on databases of structural fragments makes this method difficult to extend to structures with previously unseen geometries. Furthermore, the work is not available as a binary or source code while the online tool produces only empty PDB files (at the time of writing of this document).

The Structural Classification of RNA (SCOR) database [172] catalogs 3D RNA structures according to the motifs they contain. These are organized into a directed acyclic graph, wherein more general classes (e.g. internal loops or hairpin loops) have more specialized children (e.g. loops with base triples or loops with dinucleotide platforms).

The RNAJunction database [16] contains records of 3D junctions, interior loops, bulges and loop-loop interactions. It claimed, at the time of writing, to be the only database to store information about kissing hairpins as well as the only database that can be searched by inter-helical angle values. This makes it a potentially useful tool in creating de novo designed 3D RNA folds. Helix orientations are calculated by superimposing an idealized helix onto the query helix using the C4' atoms. These two databases and their role in assembling RNA tertiary structure are mentioned to inform the reader of the presence of catalogs of motifs and the efforts to assemble them into accurate models of RNA tertiary structure. They pave the way for the next two methods, both of which

use a related technique as well as for our work on creating ensembles of assembled structures.

4.10 *RNACOMPOSER assembles fragments using machine translation*

RNACOMPOSER [142] is a tool for quickly generating de novo 3D models of RNA structures from a sequence and secondary structure. It uses machine translation to convert secondary structure representations to 3D models. The translation uses a dictionary which maps individual secondary structure elements such as stems and loops to tertiary structure fragments. These structural fragments are then glued together via superposition of their overlapping ends and then subject to energy minimization using XPLOM-NIH [157].

Suitable fragments for building the tertiary structure are first chosen based on the topology of the secondary structure element they are in (e.g. interior loop with one unpaired nucleotide on one strand and two on the other), "*sequence similarity, pyrimidines/purines compatibility, source structure resolution and the energy*", (where the energy is presumably obtained from the molecular dynamics energy minimization of the fragment). In cases when no appropriate fragments are present, RNACOMPOSER generates a fragment using the CYANA package [61]. When used with perfectly accurate secondary structure, RNACOMPOSER generates exemplary models for many (not all) test structures. Its accuracy, however, drops precipitously when used with imperfect secondary structure, as obtained from prediction tools such as RNAFOLD, for example (See Section 11.5).

While useful as a tool for obtaining a quick glance at what an RNA structure may look like, the lack of source code or a binary tool precludes RNACOMPOSER's use locally or on a large scale. The online application provides up to ten models per sequence/secondary structure combination. This is a step toward providing ensembles of structures, but inadequate for any application that may require a thorough exploration of the conformational space of an RNA molecule.

4.11 *RNAJAG predicts helical geometries used to sample coarse-grain structures*

More recently Kim et al. [89], created an ultra-coarse-grain model comprised solely of line segments. The RNA secondary structure is represented as a graph with nodes at the ends of each stem and in the middle of each junction. Key to their approach is the use of the RNA Junctions as Graphs (RNAJAG) module [98] for predicting the topologies of RNA multiloop junctions. This method uses a random forest trained on the lengths of the

unpaired segments in junctions to predict the orientations of the adjacent helices. This allows them to fix one of the greatest sources of variation in an RNA structure, namely the junction, and to focus on sampling values for the other elements.

By varying the angles between adjacent stems and evaluating an energy based on the bending and torsion angles between helical segments and including a term for the radius of gyration Kim et al. are able to generate a variety of structures containing reasonable local and global structure. The sampling proceeds according to the standard Metropolis Hastings criterion and the lowest energy structure is compared to those obtained by other methods. The results are underwhelming but do show a slight preference for better [RMSD](#) (as compared to MC-SYM, FARNA and NAST) for structures composed of less than 70 nt). The publication makes no mention of the availability of this approach as a usable program.

4.12 *GARN uses game theory to sample coarse-grain structures*

With a stated goal very similar to ours, GARN [22] seeks to explore the conformational space of coarse-grained RNA molecules. To do this, it represents a tertiary structure as a set of game theoretical *players*, each of which corresponds to some [SSE](#) (i.e. helices and loop segments). The players are placed on a triangular lattice and allowed a set of moves which determines where they will be placed in the next step of the simulation (subject to certain constraints). Each move is associated with a knowledge-based score derived from previously solved tertiary structures. Which moves the players choose to take depends on the strategy they adopt in playing the game.

The results show that this approach does indeed sample a wide range of conformations, many of which are in the low [RMSD](#) range. While the reported structures are coarse-grain, the authors recommend a number of tools for reconstructing all-atom models. The program is provided as a Java executable with examples and documentation. While its stated aim of conformational space exploration is similar to ours, the sampling approach of GARN differs greatly from our rejection sampling algorithm. We furthermore explore structures in real space rather than on a lattice, hopefully providing a slightly more realistic representation and an easier route to generating all-atom models.

4.13 *Summary*

The examples of RNA structure prediction programs presented here are intended to inform the reader of the different strategies used for creating RNA structure models (short fragments, cyclic motifs, force-fields, [SSE](#)

fragments, DBNs, and simple geometric constraints) as well as some of the sampling techniques (Las Vegas, Metropolis Hastings, game theory, discrete molecular dynamics) used for exploring conformational space. It should provide the reader with an overview of the state of the art in the field as well as to give points of comparison and references to techniques both related to and orthogonal to those that we employ.

From the crystal of the first RNA molecule (a tRNA) [150] researchers have resorted to a simplified representation to display intra-molecular base pairs. Because these base pairs are important not only to the function of the molecule, but also to the distinction of different classes of RNA from each other, their display is essential for forming a visual model of the molecule. Questions such as "*which base pairs are near each other?*", "*how many consecutive base pairs are in a region?*", "*is this paired region near the 5' end of the molecule?*", or "*how many helices branch out from this unpaired region?*" can be answered at a mere glance.

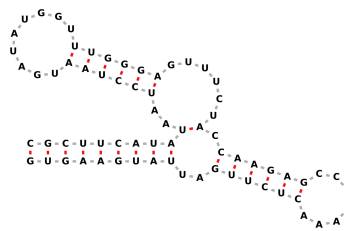
While secondary structure diagrams were already being drawn when tRNA was sequenced, it wasn't until the emergence of the initial wave of secondary structure prediction programs [141, 40, 181, 129] that the need for automated methods became pressing. As soon as a structure could be predicted without the hardships of crystallization and X-ray diffraction, many more structures could be generated and would need to be analyzed. Given that RNA's function is structure dependent, researchers wanted a visual representation of the secondary structures being predicted by the new algorithms. This provided a strong impetus for the creation of automatic secondary structure drawing tools.

Prior to the development of the first automatic RNA drawing algorithms, RNA secondary structure was either sketched by hand or crudely depicted by tediously arranging letters on a typewritten sheet of a paper in a manner roughly reminiscent of ASCII art. Automated drawing allowed researchers to easily create larger, more sophisticated diagrams with a consistent look and feel. Secondary prediction programs could be coupled to a visualization tool to immediately display a structure diagram without the need to manually position each individual nucleotide.

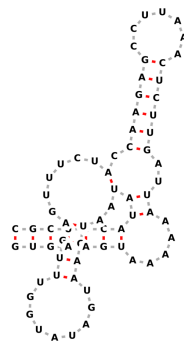
The tradition of automated RNA secondary structure diagram generation has continued unabated for the past three decades, and half of this thesis is dedicated to furthering the state of the art by implementing existing layouts using modern tools, making them easily accessible and applying them to new contexts. The rest of this introduction will report on some of the key advancements in RNA secondary structure display.

5.1 *The standard polygonal layout is aesthetically pleasing but prone to overlaps*

The first recorded paper (as far as I can tell) devoted to the creation of RNA secondary structure diagrams introduced the *standard polygonal* layout [163]. This method relies on drawing every loop as a regular polygon, where a loop was either two adjacent paired nucleotides, an interior loop, or a junction. This produces aesthetically pleasing drawings for smaller structures, but can lead to significant overlaps when structures become larger or simply have an inconvenient geometry. Figure 3 provides an example of this layout for both a well drawn structure and one with overlaps.



(a) Well-drawn Structure



(b) Overlap Structure

Figure 3: An illustration of the standard polygonal method of drawing secondary structure where all loops are drawn as regular polygons and all base pair and backbone distances are equal. Drawn using the *fornac* container described in [86].

5.2 *Radial drawing draws spindly diagrams devoid of overlaps*

Shortly after publishing the standard polygonal layout for drawing RNA secondary structure Bruce et al., set out to address the issue of overlapping nucleotides by introducing the *radial drawing* layout [164]. It guarantees no overlaps which results in a spindly spread out figure. For larger

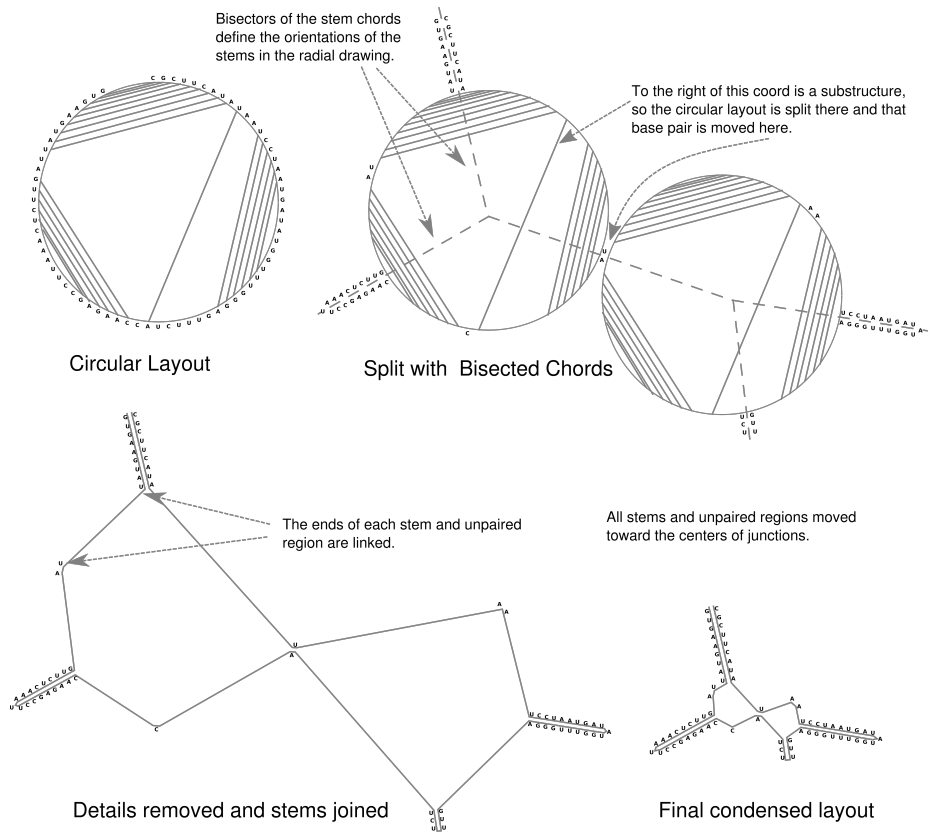


Figure 4: Creating a radial drawing starting from a circular layout. First chords are drawn through the base pairs in the circular layout. Then the stems and unpaired region are drawn outside the circle. The edges of the stems and unpaired regions are linked. Finally the linker regions are shortened to compress the drawing.

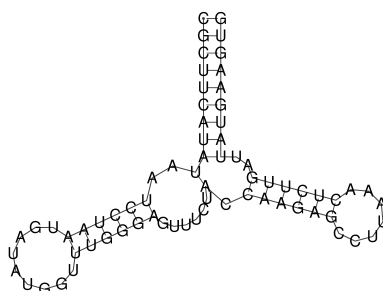


Figure 5: The output of the NAView algorithm as implemented in RNAplot [71, 105].

molecules, the diagrams can become exceptionally large and illegible. Bruce et al., expanded the concept one step further and proposed using the radial layout as a guide to untangle the standard polygonal drawing. Doing this in an automated fashion still leaves some overlapping regions, for which the authors provide manual manipulation facilities to help the user adjust and fine tune the drawing to their preferred taste. This is the first use of manual manipulation in an automatically generated RNA layout. This feature is prominent in our online structure display program, FORNA [86], and essential for arranging the layout of larger structures to correct overlaps.

5.3 *The NAView algorithm trades potential overlaps for aesthetic appeal*

Taking inspiration from the radial drawing layout, Brucoleri et al. set out to give it a more aesthetically pleasing appearance [24] (Figure 5). To do this, they modified the algorithm to display the loop regions as collections of polygons. By *extruding* or placing a long single stranded sections outside of the loop they comprise, they are able to shrink the radius of the loops and thus give the entire structure a more proportional look while simultaneously avoiding (but not completely escaping) the large overlaps that are introduced by the standard polygonal algorithm. Furthermore, because the directions of the axes of the stems match the directions of the bisectors of their chords in the radial drawing layout, they argue that similar structures will have similar layouts, thus facilitating comparison.

5.4 *Force-directed layouts apply a general graph layout algorithm to RNA structures*

With the increase in the availability of computational power, more sophisticated layouts became possible. One such method, the force-directed layout, can be applied to any graph by treating links between nodes as

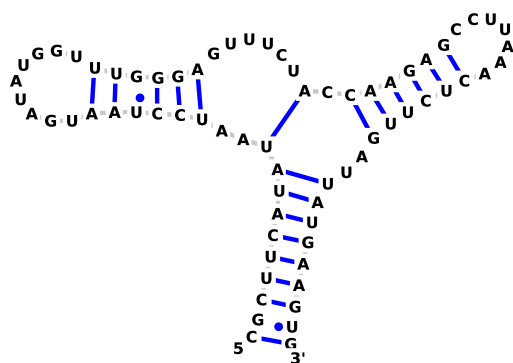


Figure 6: An adenine riboswitch as laid out by the jViz.RNA force-directed graph layout [187].

springs and applying a repulsive force between nodes to avoid overlaps. In the case of RNA, nucleotides become nodes and backbone and base pair bonds become links. A force is calculated at each iteration and the particles (nodes) are moved to their next position until the structure converges to a certain layout. Nakaya et al. [125] first demonstrated this method on RNA. To improve the running time from $O(n^2)$ (where n is the number of nodes), they used the Barnes-Hut [11] approximation for n -body simulations, where groups of distant points are treated as single bodies. Furthermore, the algorithm itself is executed in parallel. This leads to improvements in running time for up to 20 processors, beyond which the overhead of the extra processor negates the extra computational power provided. The lack of a publicly available program precludes the display of a sample structure, but on qualitative level, they resemble those produced by jViz.RNA (Figure 6).

The force-directed graph layout makes another appearance in the suite of RNA visualization tools called jViz.RNA [187]. This program offers an array of different options for display an RNA secondary structure. Its force-directed graph implementation lets the user drag and rearrange the molecules. As soon as the user drags on an element, the force simulation is triggered pulling the remaining nodes into an RNA-like conformation.

The force-directed layout also plays a prominent role in our visualization tool, FORNA [86] largely due to its ability to arrange nodes to resemble RNA structure on a local scale, allowing for easy interaction and manipulation of structures.

5.5 *RnaViz can store of layouts for future reuse*

User interaction was seen to be an important element of correcting imprecise or aesthetically displeasing layouts from the very beginning of auto-

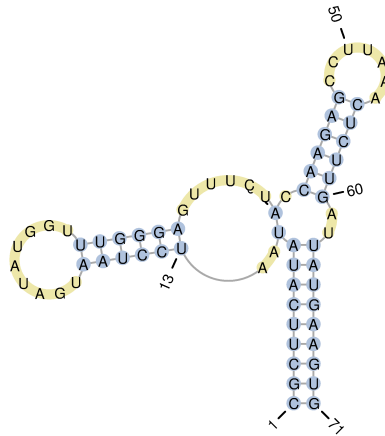


Figure 7: An adenine riboswitch as laid out by PseudoViewer [26].

mated RNA secondary structure drawing [164]. The concept of manually creating a layout, and then applying to it new structures was introduced by De Rijk et al. [37]. RnaViz creates an initial layout and lets user modify it (e.g. to remove overlaps or to place certain elements near each other). They can then store this layout as a template and re-use it for other similar molecules. A subsequent version of their tool, RnaViz2 [38], improves annotation, zooming and scaling.

5.6 *PseudoViewer draws pseudoknotted structures*

On top of the inherent difficulty of creating aesthetically pleasing layouts of planar non-pseudoknotted RNA structures lies the added challenge of trying to depict pseudoknots. Pseudoknots, which consist of non-nested base pairs ruin the planar property of RNA secondary structures. A pseudoknotted structure may require intersecting lines in order to be drawn on a flat piece of paper. While that is not always the case, pseudoknotted structures do deviate from the standard branching pattern of regular nested RNA secondary structures. Pseudoviewer [66, 26] is both a Java as well as a web application which can draw layouts including pseudoknotted base pairs.

For non-pseudoknotted structures, Pseudoviewer’s layout is vaguely reminiscent of a cross between the standard polygonal and the NAView layouts. It arranges the nucleotides in loops in a circular fashion. Rather than varying the distance between the nucleotide labels, however, it leaves gaps at the start or end of the loop when there aren’t enough nucleotides to fill the space (Figure 7). This gives it flexibility in arranging the helices so as to avoid overlaps (which it does admirably). The online application

is sleek and easy to use, but offers users little opportunity to rearrange a layout they don't like.

5.7 *VARNA is a cross-platform application with extensive functionality*

Until the publication of VARNA [34], all of the tools available for drawing RNA secondary structure were available only as desktop programs tied to a particular architecture, or requiring re-compilation to work on different types of computers. VARNA introduced a Java application that allowed users to draw RNA secondary structures on any operating system. While it did not pioneer any new layout, it provided the facilities for creating plots using one of four (standard polygonal, circular, linear, and NView) previously created layouts. By providing a web application as well as a command line tool, users could either interactively create a quick sketch or build a pipeline to automate the process for numerous different inputs.

Part II

PUBLISHED WORK

The two major topics of this thesis, tertiary structure prediction and secondary structure visualization are each represented by one publication each. These publications describe much (though not all) of the work done on these topics over the course of my time as a PhD student. Additional descriptions of the methods as well as unpublished results are presented in the subsequent two parts of this thesis:

ERNWIN Publication (Chapter 6)

- Detailed description of the model: Chapter 8
- Detailed description of the sampling procedure: Chapter 10
- Additional diagnostics and results: Chapter 11
- Suggestions for future exploration: Chapter 12

FORNA Publication (Chapter 7)

- Additional details and suggestions for improvement: Chapter 13
- Used to improve dot-plots: Chapter 14
- Used to show cotranscriptional folding: Chapter 15
- Used to diagnose tertiary structure predictions: Chapter 16

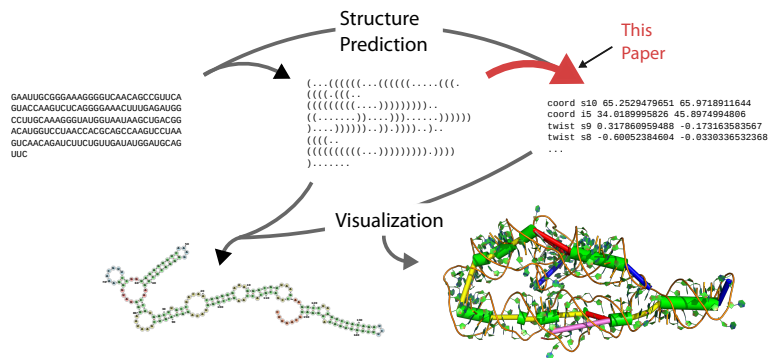
6

ERNWIN PUBLICATION

Peter Kerpedjiev, Christian Höner zu Siederdissen, and Ivo L. Hofacker.
Predicting RNA 3D structure using a coarse-grain helix-centered model.
in *RNA* 2015, 5:6.
doi: [10.1261/rna.047522.114](https://doi.org/10.1261/rna.047522.114)

PK, CH and IL designed the study. PK implemented the software and wrote substantial portions of the article. IL and CH wrote portions of the article.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Predicting RNA 3D structure using a coarse-grain helix-centered model

PETER KERPEJIEV,¹ CHRISTIAN HÖNER ZU SIEDERDISSEN,^{1,2,3} and IVO L. HOFACKER^{1,4,5}

¹Institute for Theoretical Chemistry, A-1090 Vienna, Austria

²Bioinformatics Group, Department of Computer Science, Universität Leipzig, D-04107 Leipzig, Germany

³Interdisciplinary Center for Bioinformatics, Universität Leipzig, D-04107 Leipzig, Germany

⁴Research Group Bioinformatics and Computational Biology, University of Vienna, A-1090 Vienna, Austria

⁵Center for non-coding RNA in Technology and Health, Department of Veterinary Clinical and Animal Science, University of Copenhagen, DK-1870 Frederiksberg, Denmark

ABSTRACT

A 3D model of RNA structure can provide information about its function and regulation that is not possible with just the sequence or secondary structure. Current models suffer from low accuracy and long running times and either neglect or presume knowledge of the long-range interactions which stabilize the tertiary structure. Our coarse-grained, helix-based, tertiary structure model operates with only a few degrees of freedom compared with all-atom models while preserving the ability to sample tertiary structures given a secondary structure. It strikes a balance between the precision of an all-atom tertiary structure model and the simplicity and effectiveness of a secondary structure representation. It provides a simplified tool for exploring global arrangements of helices and loops within RNA structures. We provide an example of a novel energy function relying only on the positions of stems and loops. We show that coupling our model to this energy function produces predictions as good as or better than the current state of the art tools. We propose that given the wide range of conformational space that needs to be explored, a coarse-grain approach can explore more conformations in less iterations than an all-atom model coupled to a fine-grain energy function. Finally, we emphasize the overarching theme of providing an ensemble of predicted structures, something which our tool excels at, rather than providing a handful of the lowest energy structures.

Keywords: RNA tertiary structure; coarse-grain model; knowledge-based energy function; structure prediction

INTRODUCTION

Structured noncoding RNAs (ncRNAs) are an integral part of every cell. In contrast to mRNAs, whose main duty is being the messenger in the construction of proteins from DNA genes, noncoding RNAs are involved in many regulatory and functional processes. In these roles, the three-dimensional structure of an ncRNA is of more importance than the sequence of nucleotides making up the molecule. The structure, however, is largely determined by the self-folding of the sequence.

This structural importance has led to many approaches to predict either the two-dimensional secondary structure (Zuker 2003; Do et al. 2006; Lorenz et al. 2011) or the three-dimensional tertiary structure (Das and Baker 2007; Ding et al. 2008; Parisien and Major 2008; Frellsen et al. 2009; Jonikas et al. 2009; Popena et al. 2012; Zhao et al. 2012). Compared with the former, predicting the tertiary structure is both costly in terms of computational resources

and less accurate than secondary structure prediction. These downsides are, however, balanced by the additional information encoded in the tertiary structure.

In this work, we propose an approach that bridges the gap between abstract secondary structure prediction and concrete all-atomic prediction with a coarse-grained tertiary structure prediction and sampling approach for RNAs. This approach is centered on the helix as the main immutable structural feature.

We provide three interlinked contributions toward predicting RNA 3D structures.

- I. We first introduce a coarse-grained graph that captures the main structural elements of an RNA structure. It is derived from RNA secondary structures and defines the structural relations of individual helices. Similar graph representations and their use in structure prediction have been mentioned by Zhao et al. (2012), Lamiable et al. (2013), and Kim et al. (2014) but we aim to

Corresponding author: pkerp@tbi.univie.ac.at

Article published online ahead of print. Article and publication date are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.047522.114>. Freely available online through the RNA Open Access option.

© 2015 Kerpedjiev et al. This article, published in *RNA*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

formalize their definition and illustrate its use as a guide for building a coarse-grain 3D structure.

- II. Each helix, or consecutive stack of Watson–Crick base pairs in the form of a cylinder, is one coarse-grained building block of our 3D model. Compared with all-atom models, this greatly reduces the number of parameters that need to be considered, while the property of helices of forming regular and consistent structures makes this model feasible. We also give statistics on the actual fit of this cylinder abstraction to observed helices.
- III. Finally, we provide a sampling algorithm that suggests candidate folded 3D structures, which allows us to explore the ensemble of structures matching a particular knowledge-based distribution of descriptors of the coarse-grain tertiary structure. This leads to a sampling of structures containing not only a realistic local structure but also a plausible global arrangement of the secondary structure elements.

Together these contributions yield a fast algorithm that produces structural predictions competitive with more advanced methods as we will show.

Other methods and what we contribute

Initial approaches to the prediction of RNA 3D structure simply adapted the methods developed for predicting tertiary protein structure (Das and Baker 2007). This yielded modest accuracy for smaller molecules but suffered from extremely low accuracy for any structure beyond ~30 nt in length. A subsequent approach broke the structure down into nuclear cyclic motifs which could be assigned energy values and assembled to form full structures (Parisien and Major 2008). The work of Jonikas et al. (2009) introduced a coarse-grained model which focused on the individual nucleotide as the salient building block of the RNA structure and used an energy function based solely on dinucleotide statistics obtained from the corpus of known structures. Such models have been successfully used, e.g., for modeling the folding dynamics of noncoding RNAs (Chen et al. 2013), or characterizing RNA protein interactions (Vincent et al. 2012).

Since the turn of the decade newer approaches have focused on the statistically sound and efficient prediction of local tertiary structure (Frelsen et al. 2009), on the assembly of larger structures based on the knowledge of the structure of existing secondary structure elements (SSEs) (Popenda et al. 2012; Zhao et al. 2012) and motifs (Reinharz et al. 2012). An underlying theme of modern RNA structure prediction approaches is the abstraction of the secondary structure of RNA into distinct elements with distinct properties. With few exceptions, the structure of helices is relatively uniform. Similarly, interior loops, hairpins, junctions and 5' and 3' unpaired regions all share certain structural constraints, respectively. In this article, we formalize the definition of each

element and introduce a framework for sampling different instances of each element in order to sample the space of coarse-grained 3D structures consistent with the given secondary structure. Whereas previous attempts at reducing the degrees of freedom in an RNA molecule have ranged from using three points to represent a nucleotide (Ding et al. 2008), to using one point to represent a nucleotide (Jonikas et al. 2009), we represent the helix using one line segment and two vectors and consider elements linking helices as the degrees of freedom. It should be noted that a recent approach (Kim et al. 2014) has presented a very similar model using a helix-as-a-stick representation of RNA 3D structure and combining it with predictions of local junction topology to provide accurate predictions of RNA structures. While our approaches overlap in the abstraction of the structure, our method for sampling local structure as well as our energy function formulations differ significantly. Moreover, we emphasize our ability to generate ensembles of structures competitive with the predictions of more sophisticated all-atom models.

The remainder of the article first describes the conversion of a secondary structure to a graph representing the connectivity between the different secondary structure elements. This is followed by a description of the coarse-grain representation of a helix and the methods used to fit a helix to a known all-atom structure. We then shift the focus to the parameters used to assemble tertiary structures and the energy function used to direct the sampling toward realistic structures. We demonstrate the efficacy of this approach in generating structural ensembles that conform to the target distributions and finish with a short comparison to other structure prediction methods. The software implementing this approach is titled Erwin, is licensed under the GPL-V3 license, and is freely available on Github (<http://github.com/pkerpedjiev/ernwin>).

MATERIALS AND METHODS

Secondary structure elements and graph definition

The secondary structure of an RNA molecule can be represented as a collection of elements that share similar characteristics in terms of how they link the canonical helices within the structure. The individual structural elements and their connectivity are depicted in Figure 1. The graph representation (Fig. 1B), which is used to direct the construction of the 3D model, is almost identical to the skeleton graph described by Lamiable et al. (2013), and will be referred to as such in the rest of this article. The following definitions assume the lack of pseudoknots in the secondary structure.

“Stems” are canonical double-stranded helical regions. They are identified by the nucleotides at each “corner,” that is, the nucleotides at the 5' and 3' ends of each of the strands (see Fig. 1). The corners are numbered in increasing order from 5' to 3' such that $c_1(s) < c_2(s) < c_3(s) < c_4(s)$ where $c_n(s)$ is the index of the nucleotide at corner n of stem s . Stems may be connected to each other via interior loops or multiloop segments.

The “5' unpaired region” is the set of unpaired nucleotides at the 5' end of the molecule. It is defined by the first and last unpaired

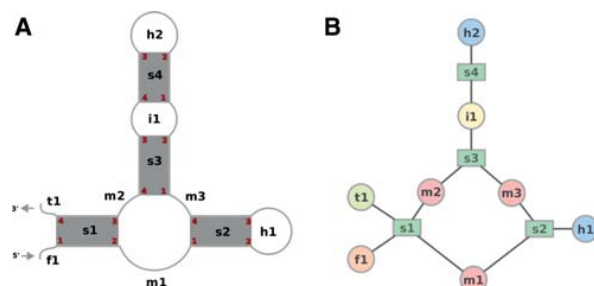


FIGURE 1. The coarse-grain representation of the 2D structure of an RNA molecule. (A) The paired regions are shown as gray rectangles. The arcs show the path of the strand in connecting the paired regions. The labels in black are names given to distinguish the different secondary structure elements in the graph. The elements $f1$ and $t1$ are the 5' and 3' unpaired regions, respectively. Elements starting with “s” correspond to base-paired canonical helices. Elements starting with an “h” are hairpins. Interior loops and multiloops are denoted by names starting with “i” and “m,” respectively. The numbers in red indicate the corners of the stem. (B) The skeleton graph representation of the structure.

nucleotides before the first stem. This section is always connected to the first stem. If there are no paired regions, then the entire molecule will be a single 5' unpaired region.

“Interior loops” are double-stranded regions which link exactly two stems and contain no canonical base pairs, although they may be rich in noncanonical base pairs (Leontis et al. 2006). These regions always connect corners 2 and 3 of one stem (s_j) to corners 1 and 4 of the next stem (s_k), where the term “next stem” implies that $c_1(s_j) < c_1(s_k)$. Since no pseudoknots are allowed in our representation we have $c_2(s_j) < c_1(s_k) < c_4(s_k) < c_3(s_j)$. Interior loops are defined by the nucleotides $c_2(s_j) + 1$, $c_1(s_k) - 1$, $c_4(s_k) + 1$, and $c_3(s_j) - 1$ for s_j and s_k next to each other. If one strand of an interior loop has no unpaired bases, then the interior loop is defined only by the unpaired nucleotides on the other strand. The interior loop i_1 in Figure 1 connects the two stems s_3 and s_4 .

“Multiloop segments” are single-stranded unpaired regions which connect two stems that are not separated by an interior loop. They can connect two stems s_j and s_k where $s_j < s_k$ in three different ways: $c_2(s_j) \rightarrow c_1(s_k)$, $c_4(s_j) \rightarrow c_1(s_k)$, and $c_3(s_j) \rightarrow c_4(s_k)$. In Figure 1 there are three multiloop segments: m_1 , m_2 , and m_3 .

The “3' unpaired region” denotes the unpaired nucleotides at the 3' end of the molecule. This region only connects with the last stem in the structure (s_j) and is defined by the nucleotide $c_4(s_j) + 1$ up to the final 3' nucleotide.

Creation of the secondary structure graph

The secondary structure graph is created from RNA secondary structure predictions. Currently, we use RNAfold from the ViennaRNA v2 package (Lorenz et al. 2011). The coarse-grained graph can be trivially created from any secondary structure representation or prediction algorithm (i.e., minimum-free energy folding, centroid structures, nonphysics based methods) which does not contain pseudoknots. Threading a coarse-grain model onto a known 3D structure requires the extraction of the secondary structure, for which we use the annotation produced by MC-Annotate (Gendron et al. 2001), removing the pseudoknots (conflict elimination method) (Smit et al. 2008), creating the secondary structure

graph and then fitting helices onto the all-atom model to get the 3D coordinates of the coarse-grain representation (see next section).

The helix and the 3D model

At the core of the Ernwin tertiary structure prediction package is the reduced cylinder-like model of an RNA helix. The representation of the helix is defined by a line segment indicating the start and end points of the axis of the helix (a_s , a_e) as well as two vectors pointing from the ends of the axis to the middle of the first and last base pairs, respectively (t_s , t_e) as depicted in the schematic (Fig. 2; Supplemental Fig. A.10). The calculation of these parameters cannot exactly represent a helix insofar as RNA helices deviate from an ideal double helix. While such a representation has previously been alluded to (Laederach et al. 2007; Popena et al. 2012), the calculation of the axis and twist vectors has never been explicitly defined. We tested four different methods for fitting idealized helices to real RNA double helices, the details of which are documented in Supplemental Section A.5. The position of the twist values is illustrated in Supplemental Section A.5.5 and Supplemental Fig. A.10.

Proposal distribution, model building, and sampling

The proposal distribution for new structures is based on a set of statistics relating the orientation of two adjacent helices, the orientations of hairpin loops, and the 5' and 3' unpaired regions relative to helices. Just as the position of 1 nt relative to the previous can be expressed as a function of the torsion angles and sugar pucker, the position of one coarse-grain helix relative to the previous can be expressed using a set of six different parameters (subsequently referred to as interhelical parameters) (Bailor et al. 2011; Sim and

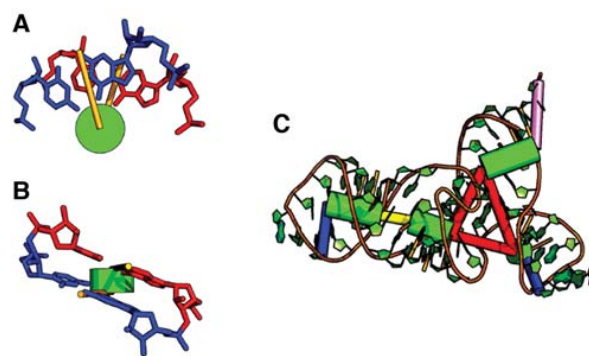


FIGURE 2. An illustration of the helix model for the 53 nucleotide SMK box (SAM-III) Riboswitch RNA structure (C, PDB: 3e5c). The helices are shown as green cylinders, interior loops as thinner yellow cylinders, multiloops as collections of red cylinders, hairpin loops as thin blue cylinders, and the 3' unpaired region is shown in magenta. A 5' unpaired region is missing from this structure since the first nucleotide is already paired. We denote the twist parameters as orange lines protruding from the axis of a helix as viewed along (A) and perpendicular (B) to the axis of a helix. Each one is perpendicular to the cylinder axis and points in the direction of the midpoint between the C1' atoms of the first and last base pair in the helix. In C, twist vectors are interpolated for the base pairs in the middle of the stem with values stored only for the vectors at the end of each stem.

Levitt 2011). Likewise, much as the distribution of potential torsion angles can be inferred by looking at solved structures, so can the range of potential interhelical parameters. This distribution, which should exclude parameters which lead to impossible configurations (due to steric hindrances, for example), is partitioned according to the size of the secondary structure element (i.e., interior loop or multiloop), which separates the two stems (see “Secondary structure elements and graph definition”).

The current implementation uses parameters mined from a large corpus of predicted 3D structures to ensure there is no overlap between the tested structures and the statistics used to predict them. This approach can be used to supplement statistics mined from known structures in cases where no instances of a particular secondary structure element are known. Each predicted 3D structure was created from a random sequence whose secondary structure was predicted using RNAfold (Lorenz et al. 2011) and whose 3D structure was predicted using FARNAs (Das and Baker 2007).

The 3D model is initially built by sampling orientation parameters for every interior loop, multiloop, hairpin loop, 5' and 3' unpaired region. Using the parameters for a coarse-grain element is analogous to inserting a fragment for that element into the overall structure. The length and twist parameters for each stem are also sampled from the list of known parameters to account for the slight variability seen in the structure of canonical helices. The initial model is built by traversing the skeleton graph (Fig. 1B) and placing each element in relation to the one preceding it. Due to the cyclical nature of junctions, one segment is necessarily determined by the orientations of the other segments. A break is introduced in the multiloop segment with the largest number of nucleotides of all the segments in a particular junction.

Direct sampling from the proposal distribution produces structures that have native-like local structure but lack long-range tertiary interactions and global structural properties found in real structures. We therefore need to add an energy function that enforces global features, such as compactness of structures, and favorable long-range interactions, which we will describe below. In order to sample from the corresponding distribution we implement a Markov chain Monte Carlo (MCMC) simulation. At each sampling step, one loop or stem is picked at random, its parameters are resampled, and the resulting structure's energy is evaluated. The new structure is accepted or rejected according to the Metropolis Hastings rule using the energy function. The secondary structure is kept fixed during the entire simulation.

Energy function

Before explaining the energy function, we will state the definitions of a few commonly used terms:

Measure: some quantifiable property of a 3D structure (e.g., its radius of gyration).

Proposal distribution: the distribution of structures obtained using only the statistics on the orientation of adjacent helices.

Target distribution: the desired distribution of a measure, i.e., the distribution observed in native structures of the appropriate size (i.e., smaller structures will have a greater chance of having a lower radius of gyration).

Sampled distribution: the distribution of a measure among all of the structures sampled over the course of a simulation.

Background distribution: the sampled distribution of a measure for a simulation run using only the constraint energies, equivalent to the distribution of the measure induced by the proposal distribution.

Reference distribution: the distribution used to calculate the energy values by comparison to the target distribution. Initially derived from a set of decoy structures (see below), the reference distribution approaches the sampled distribution as more samples are added from the MC simulation.

Our energy function is composed of five separate terms each of which is described in one of the next subsections. Two are based on physical forces to exclude impossible structures (called constraint energies, and described in the subsections “Clash detection” and “Junction closure detection”), the remaining three are knowledge-based potentials derived from known structures (called nonconstraint energies, and described in Radius of gyration, A-minor energy, and Loop–loop interaction energy). For comparison we also use an energy function which returns a value of zero for every structure (leading to constant acceptance of new structures and a direct sampling from the proposal distribution) and is intended to mimic the effect of using no energy.

The knowledge-based potentials are based on coarse-grained measures whose distributions differ between native structures (target distribution) and structures sampled from the proposal distribution (reference distribution). For each of these coarse-grained measures, we will present examples of the target distribution and the reference distribution (as calculated from a decoy) as well as the associated energy calculated by the reference ratio method (Hamelryck et al. 2010; Valentin et al. 2014). The energy associated with a value x of the measure is calculated as the log of the ratio of the target distribution $[p_t(x)]$ divided by the reference distribution $[p_r(x)]$ and multiplied by a factor c which serves as a parameter for tuning how closely the target distribution should match the sampled values (see Supplemental Section A.6.2):

$$E = -c * \log \frac{p_t(x)}{p_r(x)} \quad (1)$$

The target distribution is defined by subgraphs of the ribosome structure (PDB: 1JJ2). For a given structure we calculate the measure of interest on all subgraphs whose sequence length lies within a certain range of the target structure. The range is initially very narrow (within 1% of the length of the target structure) but is expanded until there are at least 500 measures that can be used to define a probability distribution for the target measure. For example, if trying to model a structure with a length of 100, we would consider the radii of gyration of all ribosomal subgraphs with a length between $100 - x$, and $100 + x$, such that the number of available subgraphs within that range is >500 .

The background, or reference distribution, is initially approximated from random subgraphs of an artificial ribosome structure (decoy) built using only the proposal distribution and the constraint energy terms. In a typical knowledge-based energy function, this corresponds to the reference state (Sippl 1995) and remains unchanged throughout the simulation. As pointed out in Hamelryck et al. (2010) and Valentin et al. (2014), however, the reference state depends on the structure being sampled. The reference state for the molecule being simulated is initially unknown, but can be approximated over the course of the simulation. This leads to a reference distribution which changes to reflect the ensemble of sampled

structures. The energy function is therefore variable at least during the burn-in phase of the simulation.

As more samples are produced by the MCMC, they are added to the reference distribution and used in the calculation of subsequent energies. Gaussian kernel density estimates are used to convert discrete frequencies into continuous distributions for both the target and sample distributions using a bandwidth selected using Scott's rule (Scott 2009). The bandwidth selection for the kernel density estimates smooths the distributions obtained from the training data relieving the threat of trying to match a distribution specific to the substructures of the ribosome which were used to estimate the parameters of the energy function. ERNWIN recalculates the reference distribution after every tenth MCMC step. This leads to a convergence of the distribution of sampled coarse-grain measures to their target distribution. It should be noted that while the reference ratio method (Hamelryck et al. 2010; Valentin et al. 2014) uses multiple complete sampling runs (iterations) to adequately describe the reference distribution such that samples are drawn from the target, we recreate it multiple times over a single simulation, thus enabling a close approximation of the target distribution over some predictable burn-in period (see "Energy function quality and simulation length").

An illustration of the calculation of each of the nonconstraint energy functions is shown in Figure 3. Immediately visible is the tendency for the energy to decrease in the regions where the probability density of the target distribution is greater than the probability density of the reference distribution. In practice, the reference distribution and the concomitant energy function change according to the values of the structures sampled over the course of the simulation. This process is described in more detail in Supplemental Section A.6.1.

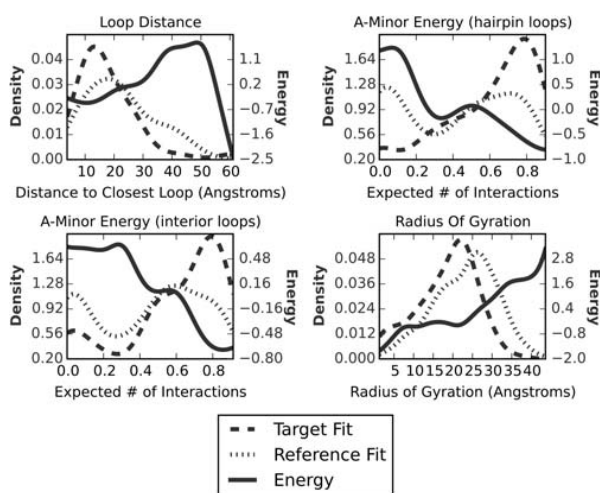


FIGURE 3. Frequency distribution and corresponding initial potential for the four different energy terms (Loop-loop distance [cf. subsection "Loop-loop interaction energy"], radius of gyration [cf. subsection "Radius of gyration"], and A-minor energy for interior and hairpin loops [cf. subsection "A-minor energy"]). The target (dashed) and reference (dotted) distributions are obtained from subgraphs of the native ribosome and a decoy ribosome structure obtained by simulating using only constraint energies (i.e., clash detection and junction closure), respectively. Here we used subgraphs of length 83, the length of the *Escherichia coli* thi-box riboswitch 2HOJ.

Clash detection

To prevent two or more atoms from occupying the same space, a heavy energetic penalty is imposed in such situations. As our model does not track individual atoms, such an energy function has to be somewhat indirect and imprecise. We have little to no hope of detecting clashes between nucleotides which are not part of a helix. There is simply too much variation in their spatial position, given the parameters that define our model. The position of the remaining nucleotides, which are in helices, can reasonably be approximated and accounted for (see Supplemental Section A.1.1). These estimated positions will be referred to later as the virtual base pair and virtual atom positions. Any clashes within the atoms of these nucleotides are given a heavy energetic penalty to ensure the rejection of that conformation.

Junction closure detection

The construction of multiloops by placing subsequent helices independently one after another leads to the problem that the parameters of the final segment of a multiloop will necessarily be determined by the previously sampled segments. Since this set of parameters is calculated, rather than chosen from the known values, it is possible that it corresponds to a sterically impossible structure, e.g., when the distance between the ends of the two adjacent stems is too large to be bridged by the nucleotides in between. To counter this occurrence, we penalize such situations by imposing a large energetic penalty. The allowed distances are determined as a function of the distance between the positions of the virtual P and O3' atoms of the capping nucleotides of the two adjacent stems.

Radius of gyration

Like proteins, albeit in a less pronounced manner, RNA molecules tend to form compact structures. To measure the compactness of the structure, we use the common radius of gyration (ROG) measure as calculated over the virtual residues of the stems of the structure (see "Clash Energy," Supplemental Section A.1.1). Instead of simply giving a bonus for a more tightly packed structure, we aim to sample structures whose distribution of ROG values matches the distribution we would expect from typical structures of that size.

A-minor energy

The A-minor motif is the most common long-range interaction found in RNA structures and contributes greatly to the overall tertiary fold of the molecule (Nissen et al. 2001). It involves an interaction between an unpaired adenine with the minor groove of a helix. The unpaired adenine (the donor) may be found in hairpins, interior loops, or junctions, but only instances where it occurs in a hairpin or interior loop are considered in this paper. Predicting the positions in the secondary structure where such an interaction might occur is difficult. We therefore assign a probability of forming an A-minor interaction to each helix-loop pair and score each loop by the weighted number of its A-minor interactions.

If we imagine that the interaction between a helix and a loop occurs over a vector connecting the closest points of the two elements, then we can parameterize it using its distance d , the angle it makes with the minor groove of the stem (ψ) and the angle (φ) between the axes of the two elements, as depicted in Figure 4.

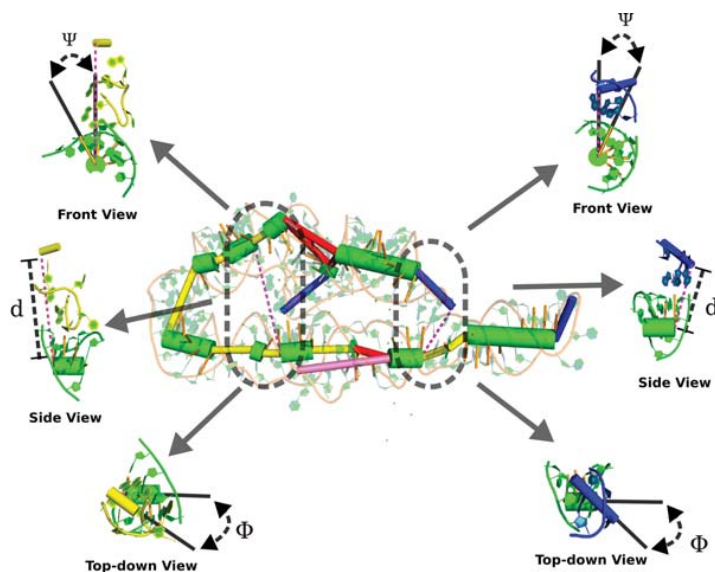


FIGURE 4. The parameterization of A-minor interactions in the Group I intron (PDB ID: 1GID). On the *left* is an interaction between an interior loop and a stem, while on the *right* is an interaction between a hairpin and a stem. Both are parameterized in the same manner, wherein the distance (d) along the interaction vector (the vector between the two closest points on the two interacting elements) is shown in the *side* view, the angle between the interaction vector and the minor groove of the stem (ψ) is shown in the *front* view, and the angle between the two interacting elements (ϕ) is shown in the *top-down* view. The direction of the view is relative to the receptor stem.

We now estimate the probability distribution for true A-minor interactions $P(d, \phi, \psi|I)$ as well as all helix-loop pairs $P(d, \phi, \psi)$ from the native ribosome structure. We can then calculate the probability that two elements (i, j) interact given their relative positions d, ϕ , and ψ :

$$P_{i,j}(I|d, \phi, \psi) = \frac{P_{i,j}(d, \phi, \psi|I) \times P_{i,j}(I)}{P_{i,j}(d, \phi, \psi)}$$

Figure 5 shows the probability distributions introduced above for hairpins. As expected, elements engaged in hairpin A-minor interactions are closer to each other and their interaction vector is generally more anti-parallel to the receptor minor groove than in the general population of pairs of proximate elements. The angle between the donor and receptor elements (ϕ) varies less, but shows a split toward a bimodal distribution in the interacting population. The distributions of parameters for interior loop A-minor interactions are similar although the minor groove-interaction angle (ψ) varies slightly more; see Supplemental Figure A.1. This is likely explained by the tendency for the A-minor interactions to occur at locations that do not correspond to the closest point between the coarse-grain interior loop donor and its stem receptor (see Supplemental Fig. A.2).

To obtain an energy function we calculate the expected number of A-minor interactions A_i that a particular loop, i , is involved in, by summing over all possible interacting helices which are not directly connected to loop i . Elements further than 30 Å have an almost negligible probability of participating in A-minor interac-

tions and are therefore excluded.

$$A_i(I|d, \phi, \psi) = \sum_{j \in A, \text{dist}(i,j) \leq 30} P_{i,j}(I|d, \phi, \psi).$$

Like all other energy terms, we obtain a target distribution from the ribosome and use the log odds ratio (Equation 1) to assign an A-minor energy to each loop. The corresponding distributions and energy function can be seen in Figure 6. As expected, the distribution for the native ribosome structure (target distribution) is shifted toward higher number of A-minor interactions compared with the reference distribution obtained from a decoy.

Loop-loop interaction energy

Unlike proteins, RNAs are polar molecules and thus lack the innate tendency to form tightly clustered structures. Their packing is more reliant on the presence of interacting motifs which tend to attract each other (Butcher and Pyle 2011). Among the variety of interactions which stabilize the global tertiary fold of an RNA molecule is the hairpin-hairpin interaction. This often occurs when two proximate hairpins are linked via hydrogen bonds and/or base stacking interactions. While there are attempts to predict such interactions (Theis et al. 2010; Sperschneider et al. 2011), we do not presume to have this ability and instead try to sample structures

which have native-like distances between the hairpins. The ribosome provides a training set from which to observe a distribution of distances from one hairpin to its nearest neighbor. This distribution, along with its analog from the background distribution of the thi-box RNA are shown in the upper left plot of Figure 6. In this structure, the loops happen to interact, but in cases where they do not it is expected that this energy will be balanced by potential A-minor interactions elsewhere or by the constraints of the local tertiary structure. An instance of this energy is created for each hairpin in the structure.

RESULTS

Structure sampling

Coarse-graining RNA structure to the level of secondary structure elements provides a fast, logical way of sampling only the regions whose 3D structure varies the most. We sampled using a Markov Chain Monte Carlo simulation for 10,000 iterations. Every nonclash structure was stored and used to calculate summary statistics about the distributions of the coarse-grain variables. In the Supplemental Material we show that as the simulation progresses the deviation between the target and sampled distributions decreases, indicating the efficacy of our sampling approach and hinting toward a potential criterion for when to terminate the simulation (see Supplemental Section A.6.2). The results indicate that

Kerpedjiev et al.

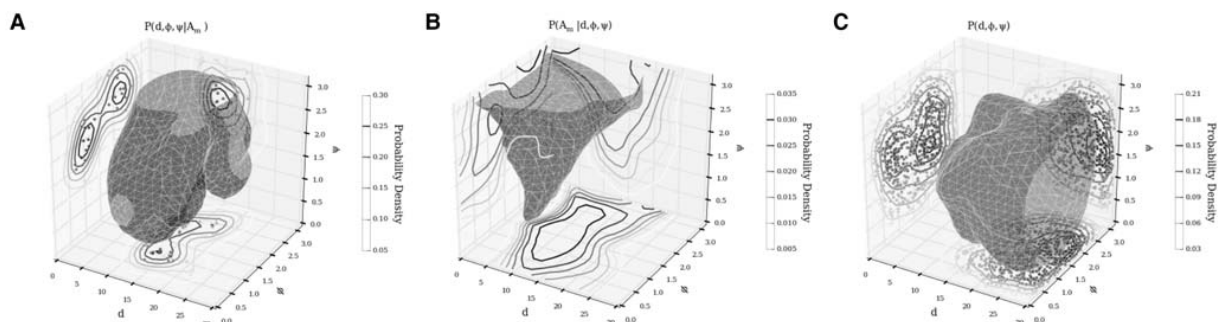


FIGURE 5. Cross sections and iso-surface showing the probability density of the parameters describing hairpin to stem A-minor interactions. (A) The probability density of seeing interaction parameters (d , ϕ , ψ) given an A-minor interaction. (B) The probability density of an interaction as calculated using Bayes' law. (C) The probability density of seeing a particular set of parameters among all adenine-containing hairpin loops within 30 Å of each other. The iso-surface in each plot corresponds to the mean probability density of all the points on the 3D grid describing the parameter space. The same plots for interior loops are presented in Supplemental Figure A.1.

applying energy functions which take only these coarse-grain elements into account can shift the distribution of sampled structures toward the native. To compute the similarity between two structures, we use the commonly used root mean square deviation (RMSD) between the superimposed positions of the virtual residues (see “Clash Energy,” Supplemental Section A.1.1) of their coarse-grain helix model representations (Kabsch 1976). The resulting structures are comparable in RMSD to the ones created by other tools such as FARNAs (Das and Baker 2007) and RNAcomposer (Popenda et al. 2012).

By applying the described energy functions, the sampling can be directed toward regions of the conformation space that share similar characteristics with native structures. In the case of the radius of gyration, constant-energy sampling yields larger, more spread out structures due to the preference for coaxial arrangements of helices (see Supplemental Fig. A.3).

Figure 6, which uses the *E. coli* thi-box riboswitch as an example, shows that structures sampled with no energy function tend to have a radius of gyration of slightly >20 Å. Structures sampled using an energy function including a term for the radius of gyration have a radius of gyration distribution peaking at ~ 18 Å. The application of the energy term has slightly broadened the distribution of sampled structures, which fortuitously happens to peak at the true value of ~ 18 Å. Clearly visible in this example is the limitation in trying to sample from the target distribution. As it includes structures smaller and greater than the native, it is more spread out and cannot be adequately approximated by the topology of the thi-box riboswitch structure. Fortunately, for larger structures, such effects become less noticeable due to the greater variety of conformations that can be adopted by larger structures.

The other two energy terms exhibit a pattern more in line with our expectations than that of the ROG energy. The A-minor energy for interior and hairpin loops (Fig. 6, upper

right and lower left, respectively, for the thi-box riboswitch [PDB ID: 2HOJ] and Supplemental Figures A.7, A.8 for all other structures) is broadened to resemble the target distribution. The peaks of the distribution, while slightly displaced from the native values are shifted toward them as compared with the background distribution.

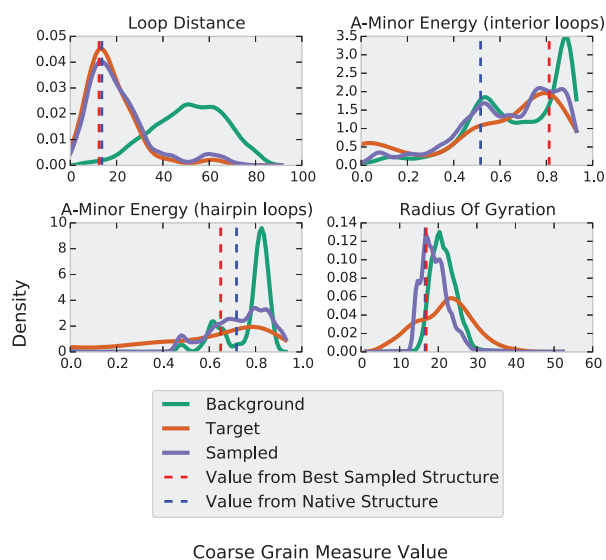


FIGURE 6. The four different coarse-grain measures and their distributions as applied to the *E. coli* thi-box riboswitch (PDB ID: 2HOJ, length: 83 nt). After including an energy value for the coarse-grain measures, the structures sampled begin to adopt values (“sampled” distribution above) similar to those expected from native structures (“target” distribution above). The radius of gyration is only slightly affected due to the constraints imposed by the topology of the RNA molecule. The blue and red dashed lines show the measures as calculated for the native and best sampled structure. The graphs for the loop distance, A-minor (interior loops), and A-minor (hairpin loops) are presented for the first hairpin, the first interior loop, and the first hairpin, respectively. A separate energy term is created for each element that the energy applies to.

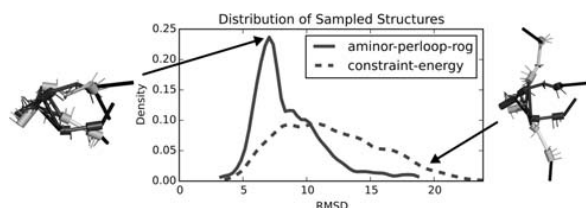


FIGURE 7. A visualization of the best (lowest energy) structure for the *E. coli* thi-box riboswitch (2HOJ, see “Prediction quality”) sampled using the full energy (left) versus the constraint energy (right). The darkened structure is native whereas the lighter is sampled. The plot shows the shift toward sampling more native-like structures using the full energy as opposed to the constraint energy. Superpositions of the lowest energy models and the native structures are provided for the whole benchmark set in Supplemental Table A.1.

The loop distance values for structures sampled with the constraint energy function are centered around a distance of 50 Å, while the real value is closer to 10 Å (Fig. 6, upper left). This is due to the presence of a kissing hairpin interaction in the structure and is reflected in both the target and sampled distributions. It should be noted that the target distribution includes structures which have hairpins that do not interact and thus peaks at a value beyond that expected for a structure with interacting hairpins. Nevertheless, the distribution of sampled structures is shifted in the direction of closer hairpins.

A stark example of applying the described energy functions to the sampling of a particular structure is illustrated in Figure 7. The distribution of RMSD values of the structures sampled with a constant-energy function (“constant-energy”) has a weighted mean at a value of ~13 Å. Applying our energy function shifts the weighted mean of the RMSD distri-

bution to a value slightly >7 Å. By visualizing the lowest energy sampled structure, we see a marked qualitative improvement in the model from the energy-based sampling as compared with the constraint energy sampling (Fig. 7, top left and top right, respectively). The shift toward lower RMSD does not always occur, as for the example of 3R4F (see Supplemental Section A.4), but in the RNAs tested, the general trend was toward an improvement. The results for all of the tested structures are shown in Supplemental Figure A.4.

Supplemental Figure A.6 shows the target, background and sampled distributions for a number of solved structures.

Comparison with other structure prediction methods

Prediction quality

The overall quality of sampled structures is comparable to some of the best structure prediction programs available. By calculating a coarse-grain model from the structures predicted using FARNa and RNACOMPOSER (where we provide the true secondary structure) we provide a comparison of the alignment between the predicted and native structures using the RMSD metric (Fig. 8). The structures used for the benchmarks were collected from the BGSU RNA 3D Hub nonredundant RNA structures list (Leontis and Zirbel 2012) and filtered to exclude structures with <70 or >500 nt as well as multimers and RNAs with bound proteins.

An example of a relatively successful simulation using Erwin is shown in Figure 7. The conformation of the lowest energy structure has two helical arms arranged in a roughly parallel fashion with the two hairpin loops near to each other, whereas a random structure sampled using a constant energy shows a worse configuration where each of the arms of the

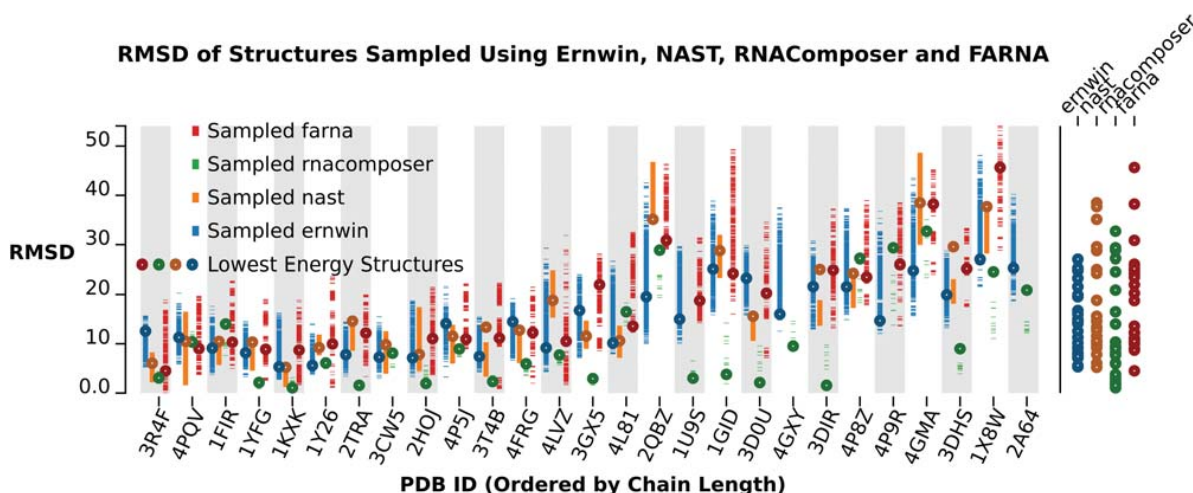


FIGURE 8. A comparison of the RMSD value between the structures sampled by each program and the native structure. Each dash in the chart represents one sampled structure. The circles represent the lowest energy structures. On the right is a tabulation of the lowest energy structures predicted by each program. The RMSD values were calculated by threading a coarse-grain representation onto the all-atom models generated by the other programs. Missing values indicate that the corresponding program failed to give a prediction for that structure.

Kerpedjiev et al.

structure points in a separate direction and the loops are on opposite ends of the molecule. The quality of the better prediction is largely due to the accurate sampling of most coarse-grain measures shown in Figure 6. The loop distance, expected number of hairpin loop A-minor interactions, and the radius of gyration are sampled at values extremely close to the native. The sampled values for the entire ensemble match the target values very well, except for the case of the ROG which is, in this case, constrained by the topology of the secondary structure. An examination of the energy landscape (Fig. 9) indicates that our energy function describes this structure particularly well, showing the desired negative correlation between RMSD and energy. While this is the case for most structures (Supplemental Table A.2), there are some notable exceptions that lead to poor predictions. One of these is presented and discussed in more detail in Supplemental Section A.4.

The circled RMSD values in Figure 8 correspond to the lowest energy structure for FARNa, Ernwin, and NAST. For RNAComposer they correspond to the structure returned when asking for one structure. As is expected, smaller structures are predicted with greater accuracy than larger ones. RNAComposer performs exceptionally well on a handful and significantly worse on others. This is likely explained by its use of known fragments for the interior and multi-loop sections, leading to near exact matches for structures with unique junction topology and sequences (i.e., tRNA,

2TRA). FARNa exhibits more tempered performance over the smaller structures which degrades over the larger structures and Ernwin exhibits measured performance over the whole data set. Over the entire sampling run, Ernwin consistently and thoroughly samples a wide range of available conformations, often yielding structures in the more native range of the samples. FARNa can sample a wide range of values, but does so more sparsely which is likely due to the fact that its simulated annealing approach falls into an energy basin that is difficult to escape as the temperature decreases. This performs well in the context of smaller structures, but leads to poor sampling of larger structures. In such cases, Ernwin can sample more structures closer to the native than both FARNa and RNAComposer.

NAST samples many structures but in very narrow ranges of the conformational landscape whereas RNAComposer only returns a maximum of 10 structures. The seemingly exemplary performance of RNAComposer in sampling low-RMSD structures should be looked upon with slight suspicion due to its use of loop topology and sequence to pick out large fragments for constructing sampled structures. Given the presence of the benchmark structures in the PDB database, these fragments are likely in RNAComposer's database of building blocks and thus accurately assembled into the known structures. While this works well with structures containing seen-before and unambiguous motifs, it can quickly backfire when a motif is absent from the database,

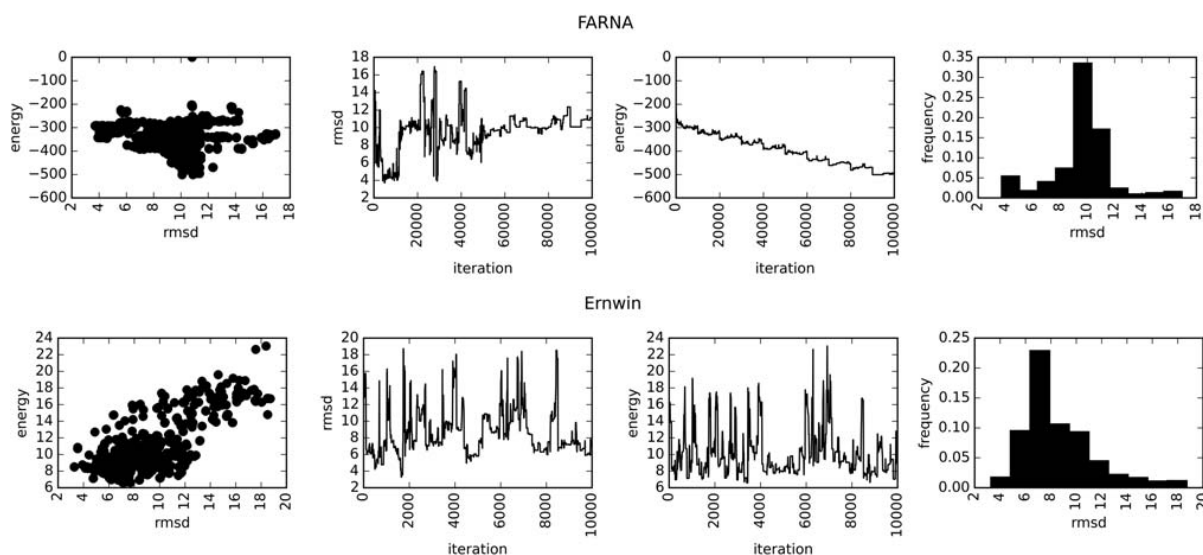


FIGURE 9. Statistics for the structure prediction procedure of FARNa (*top*) and Ernwin (*bottom*). The energy of the best structure constantly decreases up to 1,000,000 iterations with FARNa, whereas it plateaus very rapidly with Ernwin. This indicates that the broad conformational space has been mostly explored shortly within the start of the Ernwin simulation and subsequent MC steps only explore around the low-energy basin. The *left* plots display the RMSD of the structures as a function of their energy, where Ernwin displays the desired correlation. The next plot shows the RMSD of the structure as a function of the MC iteration showing no clear downward trend as the simulation progresses. The third plot shows the energy as a function of the iteration number showing the clear downward trend throughout the entire FARNa simulation and the quick arrival at a steady for Ernwin. Finally the histogram shows the RMSD of every structure sampled by both methods. The RMSDs are not directly comparable as FARNa's are for an all-atom model while Ernwin's are calculated over the coarse-grain representation.

or has multiple geometries (as may be the case for the structures 4GMA or 4P9R).

A thorough and wide sampling of the potential solution structures, as provided by Erwin, can help explore the enormous conformational space accessible to larger RNA molecules and provide many potential structures for further examination. It quickly samples unique structures (see Supplemental Table A.4 for timing information), and can be readily expanded with more accurate and more numerous fragments to expand the range of accessible conformations.

Energy function quality and simulation length

One of the challenges in structure prediction is determining the burn-in period before sampling structures from the distribution, as well as determining the thinning factor. This, of course, depends on the sampling procedure as well as the energy function. In an attempt to quantify this, we recorded the energy value and RMSD of the current structure at each iteration of the simulation (see Fig. 9). In contrast to FARNA, Erwin quickly reaches a locally minimal energy and samples structures around it. The histogram of RMSD values indicates the propensity for sampling low-energy, low-RMSD structures. FARNA, in contrast, continuously finds lower energy structures throughout the entire simulation, but due to the lack of correlation between the energy and RMSD (left panel) ends up sampling many suboptimal structures.

While FARNA's energy function works well for smaller structures, it seems to fail for larger structures and leads to sampling of high-RMSD conformations. Erwin's energy function, based on global helical arrangements provides a more robust measure of the general quality of a structure. While the example provided in Figure 9 is particularly fortuitous, most structures tested show a characteristic linear correlation between the energy and RMSD of the sampled conformations (see Supplemental Table A.2 for Erwin and Supplemental Table A.3 for FARNA). By examining the trajectory of the sampled energy values, we propose that Erwin achieves an adequate sampling within <2000 iterations, whereas FARNA requires many more iterations to reach the lowest energy values. Given Erwin's method of sampling coarse-grain measures from target distributions, one can also assess how well it has sampled from each distribution by examining the Jensen-Shannon divergence (Endres and Schindelin 2003) of the sampled values from the target values (see Supplemental Section A.6.2 and Supplemental Fig. A.13). When the divergence levels off, we have adequately sampled from our target distribution indicating that additional MC steps provide no new conformations. While there is no consistent number of iterations that is applicable to all structures, examining the progress of the distribution can provide an empirical method for determining when to end a simulation. A thorough treatment of this topic, however, is out of the scope of this paper and is merely mentioned to highlight the utility of having a probability-distribution based energy function.

DISCUSSION AND CONCLUSION

In this paper, we have presented a coarse-grained model of RNA structure parameterized by the angles and shifts between helices. We have shown that coupling a simple proposal distribution with a probability-based energy function can yield predictions that match those of programs with much more sophisticated models and energy functions. We propose that our model can be used for quick exploration of the macroscale conformational space of an RNA molecule.

We suggest that such a model can also be useful for the elucidation and identification of different RNA species in atomic-force microscopy images where the positions of the individual atoms are largely indistinguishable (Petkovic et al. 2015). Given fluorescence resonance energy transfer (FRET) data, the structures generated by our model can provide the experimentalist with an overview of the global structure of the RNA molecule without the overwhelming precision (and uncertainty) of an all-atom model. Simple diagnostics such as determining whether two loops have the potential to localize within a certain distance of each other, while maintaining steric integrity, can also easily be performed.

A particularly compelling future application is the combination of our sampling method with data from a SAXS experiment. As RNA in solution can adopt a multitude of conformations, its true structure in a solution may not be accurately represented by the crystal structures used as benchmarks (Ali et al. 2010; Brenner et al. 2010). Spectra obtained from SAXS experiments, however, reflect the true distribution of conformations present in a solution. Furthermore, coarse-grained models as presented here, are sufficient to generate theoretical SAXS profiles. Thus, SAXS data could be incorporated directly in the simulation as an additional potential based on the difference between the theoretical and measured SAXS profile. A similar approach can be envisioned for FRET data which can be directly interpreted as a probability distribution on the distance between some donor and acceptor groups, which can be turned into an energy function in the same way as our coarse-grain measures. Other low-resolution methods such as hydroxyl radical footprinting offer information about how accessible a particular nucleotide is to solvent (Tullius and Greenbaum 2005), while multiplexed hydroxyl radical cleavage analysis (MOHCA) yields potential interactions between nucleotides within 25 Å of each other (Das et al. 2008). Each of these can be encoded as a potential and sampled from, yielding an ensemble of structures which conform to the constraints imposed by the experimental method. Given the probabilistic nature of the potentials, uncertainty about the constraints (due to difficulty in resolving gel bands, for example) can be encoded in the target distribution imposed by the experimental data.

Beyond the potential applications, this work aims to provide a platform for further exploration into the determinants of global tertiary RNA structure. The inclusion of

predicted local structural motifs (Lescoute et al. 2005; Sarver et al. 2008; Petrov et al. 2013; Theis et al. 2013) provides an immediate avenue for the improvement of the prediction quality. Information about the extended secondary structure (Höner zu Siederdisen et al. 2011) of a sequence could provide a more fine-grain partitioning of the statistics used in generating the proposal distribution. The framework makes it straightforward to add additional energy terms for long-range interactions and thus provides an orthogonal path for determining what information is necessary for the accurate prediction of global RNA structure.

In summary, coarse-grained 3D RNA structures provide a fast, efficient way toward tertiary structure prediction. They also point toward an information mismatch that we aim to fill with future research. In particular, sequence information is only taken into account during the initial graph construction phase, when the skeleton graph is created from predicted secondary structures. Even using this simplified representation, the lowest energy structure are comparable and often better than some of the more fine-grained prediction methods. In addition, Ernwin provides a more thorough and wider sampling of the conformational space than existing methods. Such an accomplishment without sequence information calls into question the efficacy of the sampling approaches of other more fine-grained methods and provides a simplified model for exploring new methods of building and sampling *de novo* structures.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank Christoph Flamm and Thomas Hamelryck for inspiration and fruitful discussions and Craig Zirbel for helpful advice and comments about the manuscript. This work was funded, in part, by the Austrian DK RNA program FG748004, and by the Austrian FWF, project “SFB F43 RNA regulation of the transcriptome.”

Received August 20, 2014; accepted February 13, 2015.

REFERENCES

- Ali M, Lipfert J, Seifert S, Herschlag D, Doniach S. 2010. The ligand-free state of the TPP riboswitch: a partially folded RNA structure. *J Mol Biol* **396**: 153–165.
- Bailor MH, Mustoe AM, Brooks CL III, Al-Hashimi HM. 2011. 3D maps of RNA interhelical junctions. *Nat Protoc* **6**: 1536–1545.
- Brenner MD, Scanlan MS, Nahas MK, Ha T, Silverman SK. 2010. Multivector fluorescence analysis of the *xpt* guanine riboswitch aptamer domain and the conformational role of guanine. *Biochemistry* **49**: 1596–1605.
- Butcher SE, Pyle AM. 2011. The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Acc Chem Res* **44**: 1302–1311.
- Chen C, Mitra S, Jonikas M, Martin J, Brenowitz M, Laederach A. 2013. Understanding the role of three-dimensional topology in determining the folding intermediates of group I introns. *Biophys J* **104**: 1326–1337.
- Das R, Baker D. 2007. Automated *de novo* prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci* **104**: 14664–14669.
- Das R, Kudaravalli M, Jonikas M, Laederach A, Fong R, Schwans JP, Baker D, Piccirilli JA, Altman RB, Herschlag D. 2008. Structural inference of native and partially folded RNA by high-throughput contact mapping. *Proc Natl Acad Sci* **105**: 4144–4149.
- Ding F, Sharma S, Chalasani P, Demidov VV, Broude NE, Dokholyan NV. 2008. Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA* **14**: 1164–1173.
- Do CB, Woods DA, Batzoglou S. 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**: e90–e98.
- Endres DM, Schindelin JE. 2003. A new metric for probability distributions. *IEEE Trans Inform Theory* **49**: 1858–1860.
- Frelsen J, Moltke I, Thiim M, Mardia KV, Ferkinghoff-Borg J, Hamelryck T. 2009. A probabilistic model of RNA conformational space. *PLoS Comput Biol* **5**: e1000406.
- Gendron P, Lemieux S, Major F. 2001. Quantitative analysis of nucleic acid three-dimensional structures. *J Mol Biol* **308**: 919–936.
- Hamelryck T, Borg M, Paluszewski M, Paulsen J, Frelsen J, Andreatta C, Boomsma W, Bottaro S, Ferkinghoff-Borg J. 2010. Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PLoS One* **5**: e13714.
- Höner zu Siederdisen C, Bernhart SH, Stadler PF, Hofacker IL. 2011. A folding algorithm for extended RNA secondary structures. *Bioinformatics* **27**: i129–i136.
- Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D, Altman RB. 2009. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA* **15**: 189–199.
- Kabsch W. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr A* **32**: 922–923.
- Kim N, Laing C, Elmetwaly S, Jung S, Curuksu J, Schlick T. 2014. Graph-based sampling for approximating global helical topologies of RNA. *Proc Natl Acad Sci* **111**: 4079–4084.
- Laederach A, Chan JM, Schwartzman A, Willgohe E, Altman RB. 2007. Coplanar and coaxial orientations of RNA bases and helices. *RNA* **13**: 643–650.
- Lamiable A, Quesette F, Vial S, Barth D, Denise A. 2013. An algorithmic game-theory approach for coarse-grain prediction of RNA 3D structure. *IEEE/ACM Trans Comput Biol Bioinform* **10**: 193–199.
- Leontis NB, Zirbel CL. 2012. Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In *RNA 3D structure analysis and prediction*, pp. 281–298. Springer, Berlin, Heidelberg.
- Leontis NB, Lescoute A, Westhof E. 2006. The building blocks and motifs of RNA architecture. *Curr Opin Struct Biol* **16**: 279–287.
- Lescoute A, Leontis NB, Massire C, Westhof E. 2005. Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic Acids Res* **33**: 2395–2409.
- Lorenz R, Bernhart SH, Höner zu Siederdisen C, Tafer H, Flamm C, Stadler PF, Hofacker L. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**: 26.
- Nissen P, Ippolito JA, Ban N, Moore PB, Steitz TA. 2001. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc Natl Acad Sci* **98**: 4899–4903.
- Parisien M, Major F. 2008. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**: 51–55.
- Petkovic S, Badelt S, Block S, Flamm C, Delcea M, Hofacker I, Müller S. 2015. Sequence-controlled RNA self-processing: computational design, biochemical analysis, and visualization by AFM. *RNA* (in press).
- Petrov AI, Zirbel CL, Leontis NB. 2013. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA* **19**: 1327–1340.

- Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, Blazewicz J, Adamiak RW. 2012. Automated 3D structure composition for large RNAs. *Nucleic Acids Res* **40**: e112.
- Reinharz V, Major F, Waldispühl J. 2012. Towards 3D structure prediction of large RNA molecules: an integer programming framework to insert local 3D motifs in RNA secondary structure. *Bioinformatics* **28**: i207–i214.
- Sarver M, Zirbel CL, Stombaugh J, Mokdad A, Leontis NB. 2008. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J Math Biol* **56**: 215–252.
- Scott DW. 2009. *Multivariate density estimation: theory, practice, and visualization*, Vol. 383. Wiley, New York, NY.
- Sim AY, Levitt M. 2011. Clustering to identify RNA conformations constrained by secondary structure. *Proc Natl Acad Sci* **108**: 3590–3595.
- Sippl MJ. 1995. Knowledge-based potentials for proteins. *Curr Opin Struct Biol* **5**: 229–235.
- Smit S, Rother K, Heringa J, Knight R. 2008. From knotted to nested RNA structures: a variety of computational methods for pseudoknot removal. *RNA* **14**: 410–416.
- Sperschneider J, Datta A, Wise MJ. 2011. Heuristic RNA pseudoknot prediction including intramolecular kissing hairpins. *RNA* **17**: 27–38.
- Theis C, Janssen S, Giegerich R. 2010. Prediction of RNA secondary structure including kissing hairpin motifs. In *Algorithms in bioinformatics*, pp. 52–64. Springer, Berlin, Heidelberg.
- Theis C, Höner zu Siederdisen C, Hofacker IL, Gorodkin J. 2013. Automated identification of 3D modules with discriminative power in RNA structural alignments. *Nucleic Acids Res* **41**: 9999–10009.
- Tullius TD, Greenbaum JA. 2005. Mapping nucleic acid structure by hydroxyl radical cleavage. *Curr Opin Chem Biol* **9**: 127–134.
- Valentin JB, Andreetta C, Boomsma W, Bottaro S, Ferkinghoff-Borg J, Frelsen J, Mardia KV, Tian P, Hamelryck T. 2014. Formulation of probabilistic models of protein structure in atomic detail using the reference ratio method. *Proteins* **82**: 288–299.
- Vincent HA, Henderson CA, Stone CM, Cary PD, Gowers DM, Sobott F, Taylor JE, Callaghan AJ. 2012. The low-resolution solution structure of *Vibrio cholerae* Hfq in complex with Qrr1 sRNA. *Nucleic Acids Res* **40**: 8698–8710.
- Zhao Y, Huang Y, Gong Z, Wang Y, Man J, Xiao Y. 2012. Automated and fast building of three-dimensional RNA structures. *Sci Rep* **2**: 734.
- Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406–3415.

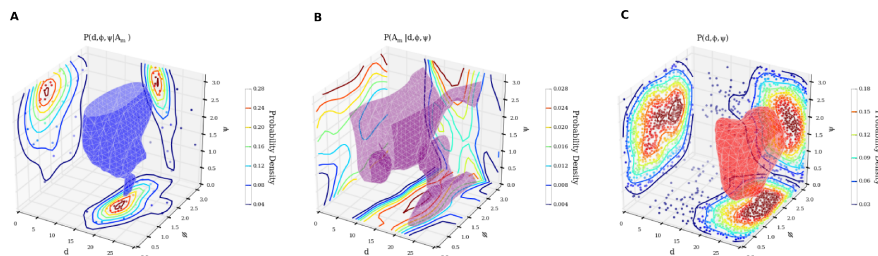


Figure A.1: Cross sections and iso-surface of the A-Minor energy as calculated for interior loop to stem A-minor interactions. (A) The probability density of seeing interaction parameters given an A-Minor interaction. (B) The probability density of an interaction as calculated using Bayes' law. (C) The probability density of seeing a particular set of parameters among all adenine-containing hairpin loops within 30 Å of each other. The iso-surface in each plot corresponds to the mean probability density of all the points on the 3D grid describing the parameter space.

A. Supplementary Material

A.1. Energies

A.1.1. Clash Energy

Having established that we can re-position the atoms of a helix reasonably well given its coarse grain parameters (see Sec. 2.2), we can use this technique to detect clashes within stem regions. This is done by 'virtual' base pairs, or interpolations of where a base pair and its constituent nucleotides and atoms would lie given its position within the stem and the stem's coarse-grain parameters.

Thus, if a stem has a length of n base pairs, a starting position vector s , an ending position vector e and two twist vectors t_1 and t_2 , the position of the i 'th virtual base pair along the axis of the helix will be $s + \frac{i-1}{n-1}(e-s)$. The direction of the angular position can be calculated in a similar fashion by taking the corresponding proportion of the angle that one would need to rotate the first twist around the stem axis to align it with the second twist. This total angle, naturally, needs to be adjusted for helices in which the base pairs twist fully around the stem axis one or more times (usually around the 11th basepair).

Using the stem axis vector $a = (e-s)$, the starting position of the virtual base pair v_s , and the virtual twist vector v_t , we can define a coordinate system for each base pair. Within this coordinate system, the position of all the atoms of the two nucleotides in the base pair can be represented in a manner that only depends on the strand which contains the nucleotide and the identity of the atom. These positions were calculated for all stem-contained atoms in the ribosome structure 1jj2 and averaged to create an average base-pair representation.

The clash energy function determines if the positions of the atoms calculated from the virtual base-pairs of each helix intersect (i.e. are positioned within 1.8 Å of each other). The number of intersections is multiplied by a large energetic penalty to yield a certain rejection of the sampled structure.

A.1.2. A-Minor Energy

A.2. Best Sampled Structures

A.3. Enumeration

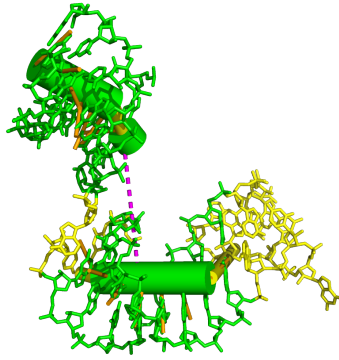
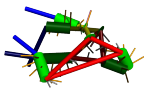
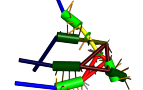
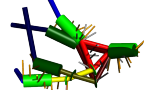
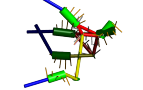
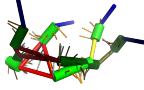
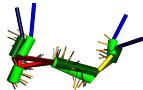
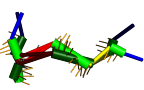
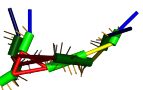
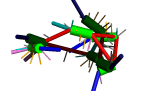
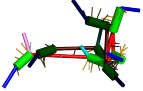
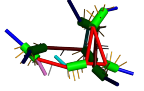

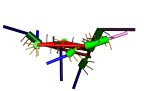
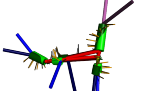

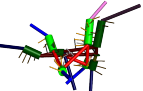


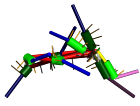
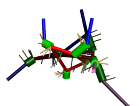

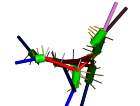
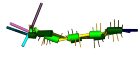
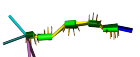
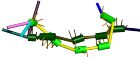
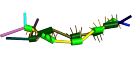
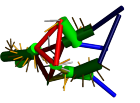
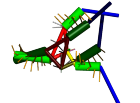
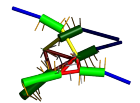
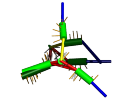
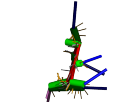
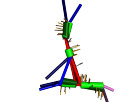
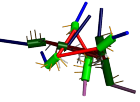





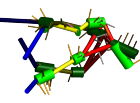
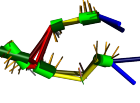
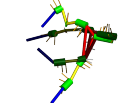
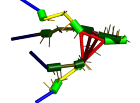
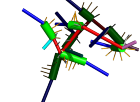

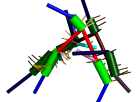
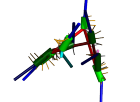
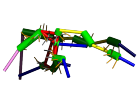
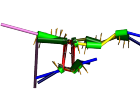
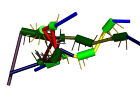
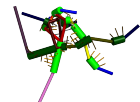
Figure A.2: An example of an A-Minor interaction where the line between the closest points on the two coarse-grain elements (dashed purple line) does not exactly correspond to the location of the interaction, which is on the far left of end of the bottom helix.

Table A.1: The RMSD values and best models produced by each prediction method. Missing structures indicate an error in the prediction pipeline or program.

PDB ID	Length	Ernwin	RNAComposer	FARNA	NAST
3FO4	63	 6.7	 7.7	 4.6	 10.2
3R4F	66	 12.6	 3.1	 4.5	 6.1
4PQV	68	 11.3	 10.4	 9.0	 10.5
1YFG	69	 8.2	 2.1	 8.9	 10.3

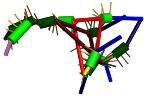
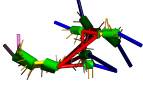
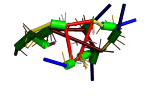
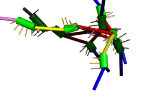
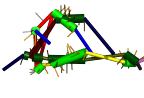
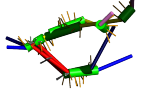
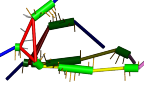
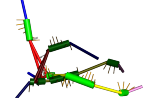
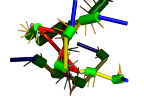
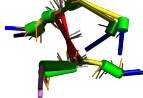

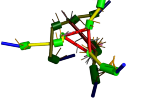
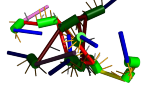
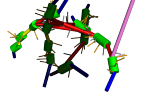
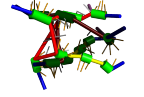
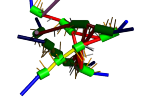
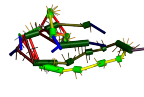
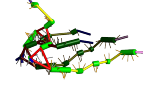
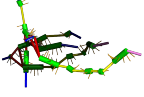
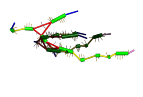
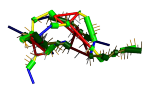
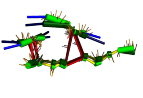
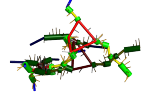
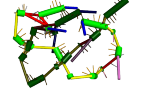
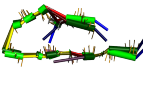
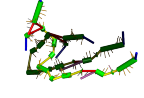
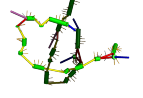
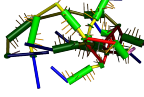
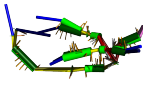
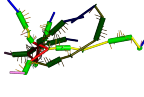
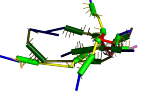
Continued on next page

Table A.1 – continued from previous page

PDB ID	Length	Ernwin	RNAComposer	FARNA	NAST
1FIR	69	 9.1	 14.0	 10.3	 10.5
1KXK	70	 5.3	 1.1	 8.7	 5.2
1Y26	71	 5.6	 6.1	 10.0	 9.2
2TRA	73	 7.8	 1.6	 12.2	 14.6
3CW5	75	 7.3	 8.1	 9.8	 9.8
2HOJ	78	 7.1	 1.9	 11.1	 7.8
4P5J	83	 14.1	 9.0	 10.9	 11.5
3T4B	83	 7.4	 2.3	 11.2	 13.4

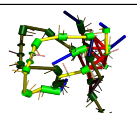

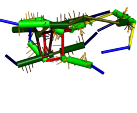
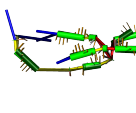
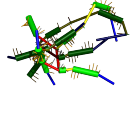
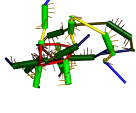
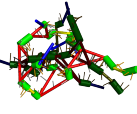
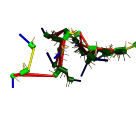
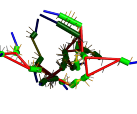
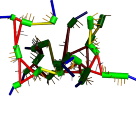
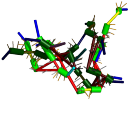
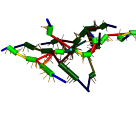
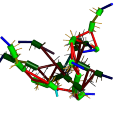
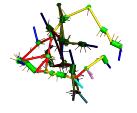
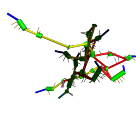
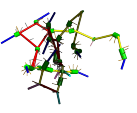
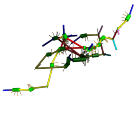
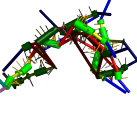
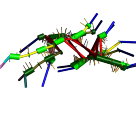
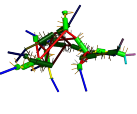
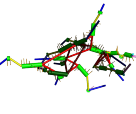
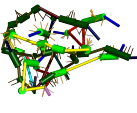
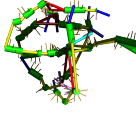
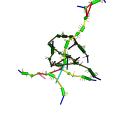
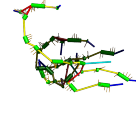
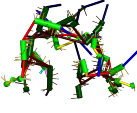
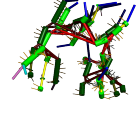
Continued on next page

Table A.1 – continued from previous page

PDB ID	Length	Ernwin	RNAComposer	FARNA	NAST
4FRG	84	 14.5	 5.9	 12.3	 12.8
4LVZ	89	 9.2	 7.7	 10.5	 18.8
3GX5	94	 16.8	 2.9	 22.0	 11.6
4L81	96	 10.1	 16.5	 13.6	 10.6
2QBZ	153	 19.5	 28.9	 30.9	 35.1
1U9S	155	 15.0	 3.0	 18.8	
1GID	158	 25.2	 3.8	 24.3	 28.8
3D0U	161	 23.3	 2.1	 20.3	 15.6

Continued on next page

Table A.1 – continued from previous page

PDB ID	Length	Ernwin	RNAComposer	FARNA	NAST
4GXY	161	 16.0	 9.5		
3DIR	172	 21.6	 1.6	 24.9	 25.1
4P8Z	188	 21.6	 27.2	 23.5	 24.3
4P9R	189	 14.7	 29.4	 26.0	
4GMA	192	 24.8	 32.7	 38.2	 38.5
3DHS	215	 19.9	 9.0	 25.2	 29.6
1X8W	242	 27.1	 24.6	 45.7	 37.7
2A64	298	 25.4	 20.9		

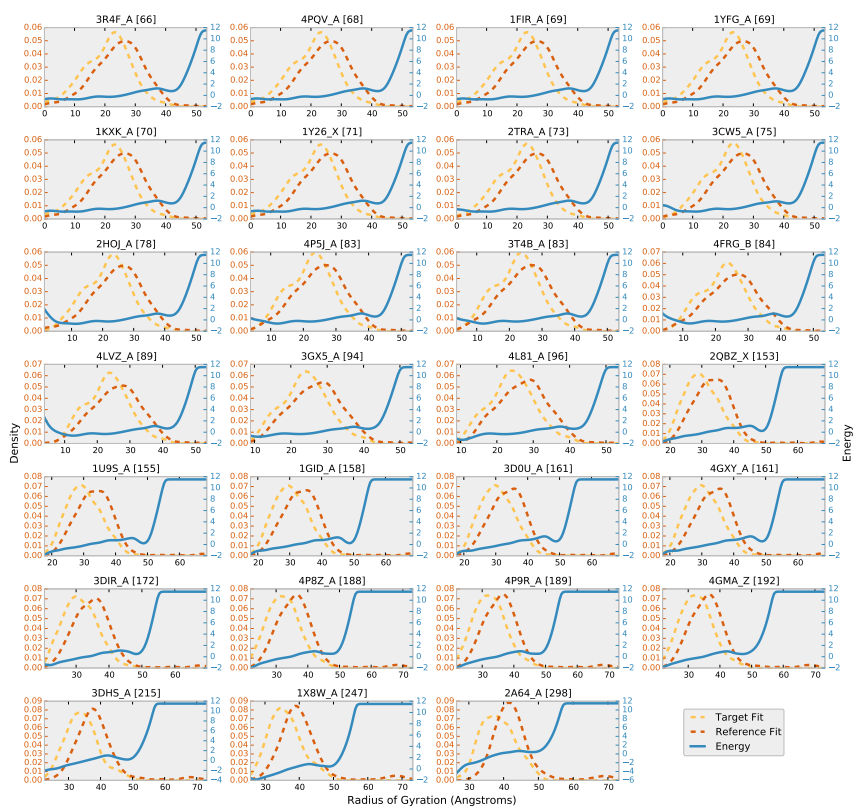


Figure A.3: The target and reference distributions for the radius of gyration. The target and reference measures are obtained from subgraphs of a native ribosome and an artificially constructed structure without clashes or broken junctions. Clearly evident is the tendency for native structures to have a smaller radius of gyration than larger structures. The energy values are thus lower for more compact structures and higher for more spread out structures. The effect becomes more pronounced for larger structures due to their inherent ability to spread out more.

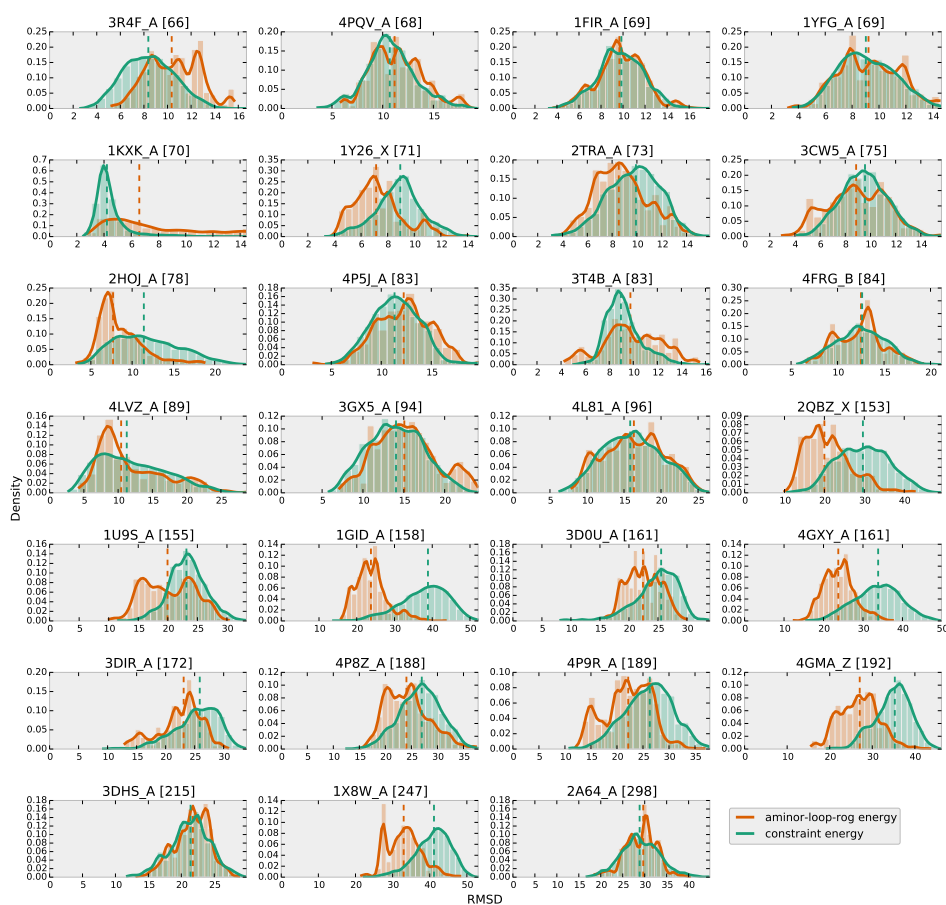


Figure A.4: The distribution of the RMSD between the native and the sampled structures using a naive (constant-energy) energy function and the combination of the three non-constraint coarse grain energies (aminor-perloop-rog).

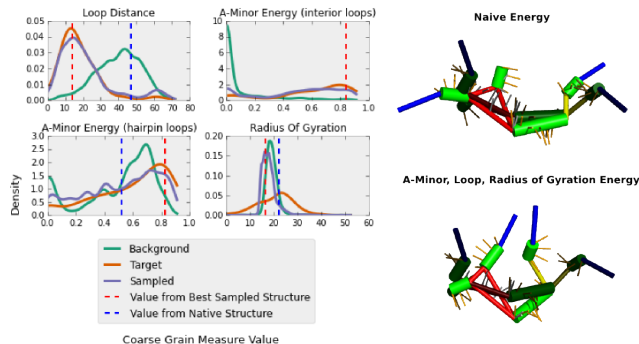


Figure A.5: The best model for the structure 3R4F (bright colors) aligned to the native structure (dark colors) from a simulation using the naive energy function (above right) and the energy described in Section 2.4 (below right). On the left are the graphs indicating the coarse grain value distributions for each of the measures. The value of the best sampled structure is shown as an orange dashed line, whereas the value from the native is shown as a dashed blue line. The two lines show a perfect overlap for the interior loop A-Minor Energy in the upper right hand corner and thus only the red line is visible.

A.4. Example of a Prediction Failure

As seen in figure A.4, sampling using the presented energy function does not always lead to an improvement over a naive energy. A good example of this is the structure 3R4F, show in Supplementary Figure A.5. The poor prediction performance is most likely attributed to the incorrect close positioning of the two hairpin loops in the predicted structure. The other coarse-grain values differ by smaller magnitudes and likely contribute less to the wrong configuration. Examining the energy of all of the sampled structures show that this structure has a very unfavorable energy-rmsd profile (see Supplementary Table A.2). This example illustrates one of the pitfalls of using a knowledge-based energy function wherein a strong pattern in the training data is not necessarily reflected in individual structures. We plan on addressing this issue by adding additional terms to our energy function such that a global maximum probability (minimum energy) occurs near the native structure. This, however, is an ongoing pursuit and is beyond the scope of this paper.

A.5. Helix Fitting

Four different methods for fitting a coarse grain helix (consisting of an axis segment and two twist vectors) to an all-atom helix were tested for their ability to accurately represent the positions of the atoms on the helix.

The **ad-hoc** method uses the directions of the base normals as well as the vector between the estimated centers of the outermost two base pairs of the helix as a way of calculating the axis vector.

The **fit** method assumes that the projection of the positions of backbone atoms onto the plane normal to the axis vector should form a circle. The axis is calculated by optimizing its vector so as to minimize the root-mean square deviation (RMSD) between the projected heavy atom positions and a circle fit onto their positions.

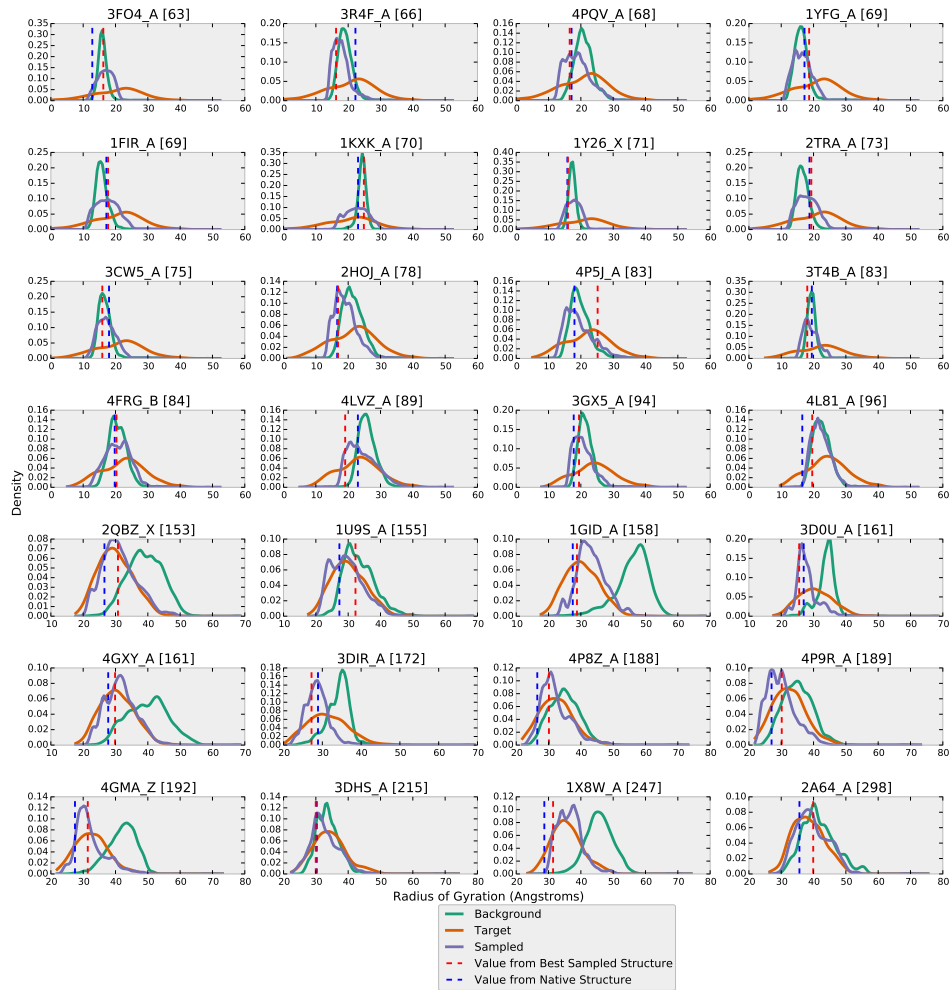


Figure A.6: The distribution of the radius of gyration values for structures derived from the native ribosome (target, orange), the naive sampling of the modeled structure (background, green) and the energy-directed sampling of the modeled structure (sampled, purple).

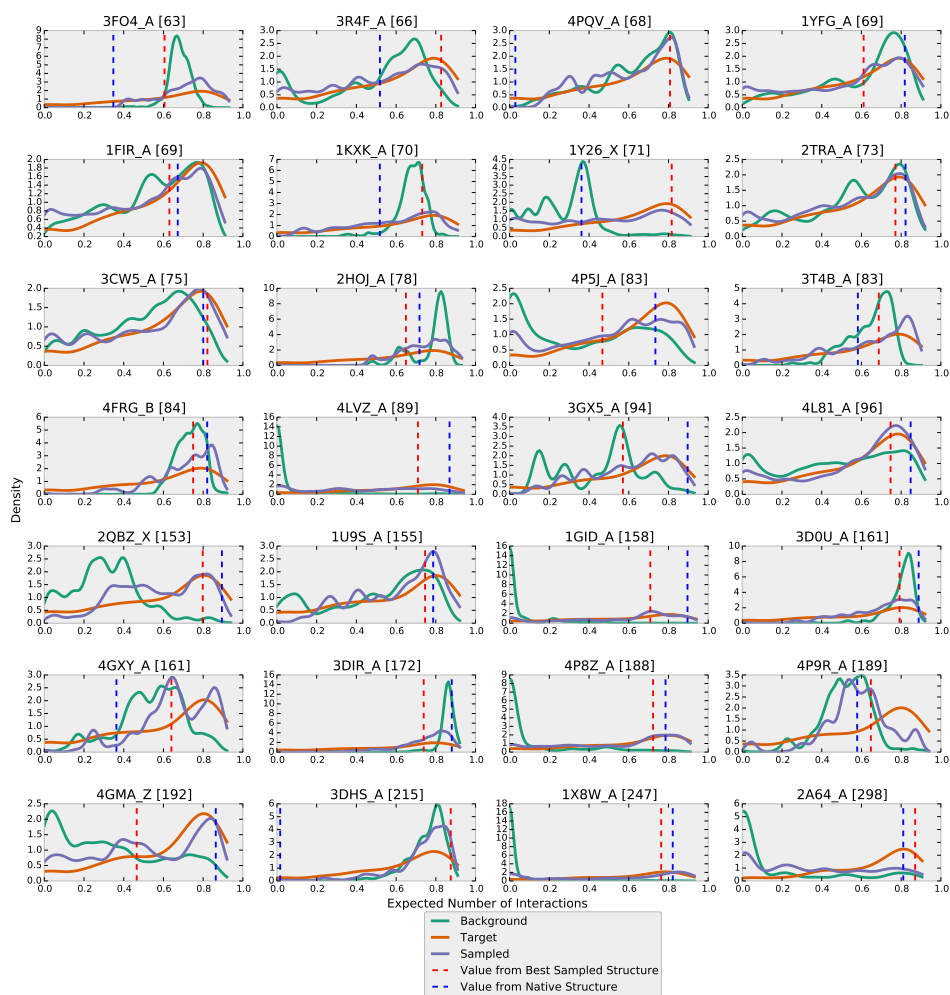


Figure A.7: The distribution of A-Minor interaction probabilities for hairpin loops for structures derived from the native ribosome (target, orange), the naive sampling of the modeled structure (background, green) and the energy- directed sampling of the modeled structure (sampled, purple).

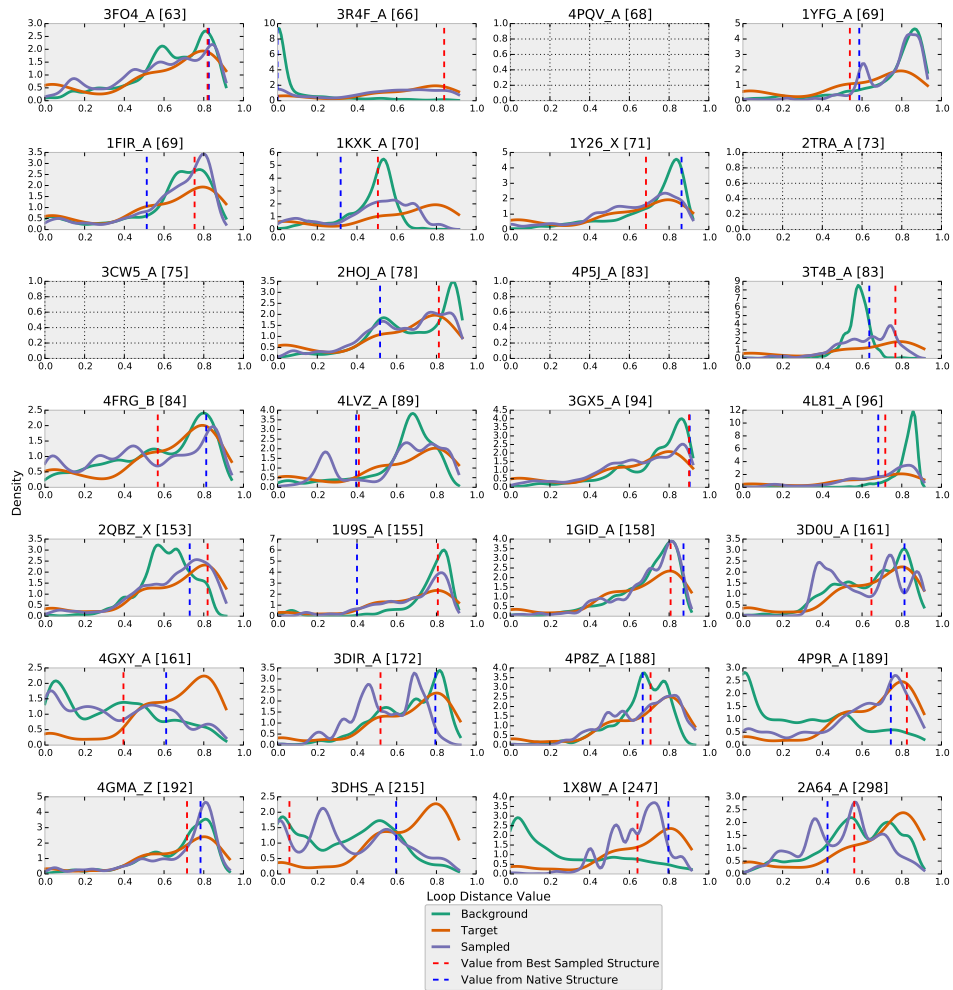


Figure A.8: The distribution of A-Minor interaction probabilities for interior loops for structures derived from the native ribosome (target, orange), the naive sampling of the modeled structure (background, green) and the energy- directed sampling of the modeled structure (sampled, purple).

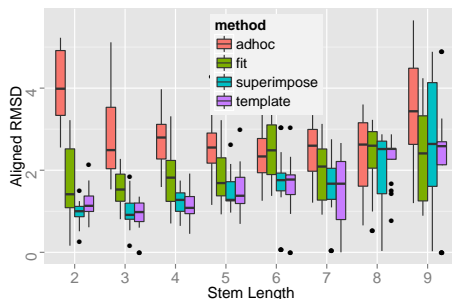


Figure A.9: A comparison of the ability of the helix-fitting methods to re-orient a different helix onto the original helix, given only the fitted helix parameters. The difference between the different methods is most pronounced in shorter stems which contain the least amount of data points available for helix-fitting. For the best method, the average aligned root mean square deviation increases for larger stems due to the increasing number of nucleotides present.

The **superimpose** method fits an axis as per the **fit** method onto an ideal helix (generated by using the fiber program of the 3DNA package [25]). The ideal helix is then superimposed onto the target helix using Kabsch’s algorithm [17] and the resulting transformation is applied to the fitted axis to yield an axis for the target helix.

The **template** method differs from the **superimpose** method insofar as it uses an extremely long (30 base pair) helix to fit the axis parameters and then superimposes the first n base pairs of the long ideal helix onto the target helix to generate the transformation for the axis.

A.5.1. Results

To measure the quality of each helix fitting method described in the previous section, we took pairs of helices containing the same number of base-pairs and calculated the parameters for both helices using each of the four described methods. For both helices, we calculated a twist parameter equal to a halfway rotation of one twist onto the other around the helix axis. The direction of this twist parameter as well as the line segment defining the axis of the helix provide enough parameters to direct the superposition of one helix onto another. That is, we can define a translation and a rotation that will superimpose one coarse grain representation onto another. By applying this transformation to each backbone atom of the second helix, we attempt to move and rotate it to its equivalent location on the first helix. We would expect a consistent helix fitting method to yield the lowest RMSD deviation between the atoms of the first helix and the atoms of the second (fitted) helix when superimposed on the first using the axes and twist parameters defining both helices.

The results (illustrated in Fig. A.9) confirm the expected dominance of the superposition-based methods (**superimpose** and **template**). The ad-hoc method’s performance on short helices was notably worse than any of the others due to the fact that the conformations of terminal base pairs tend to have a variable geometry which is not tempered by the length of the helix. In longer helices, the vector between the computed centers of the terminal base pairs is longer and thus has less room for variation, leading to better performance of the ad-hoc method. The **fit** method, which fits a helix axis by optimizing

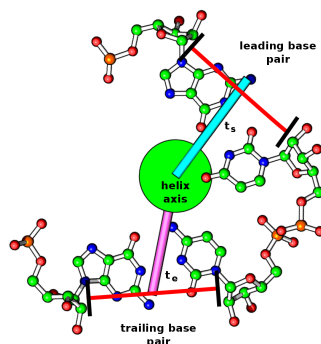


Figure A.10: An illustration of a fitted cylinder along with the 'twist' vector indicating the direction of the first and last base pair of the stem. The view is top-down looking down the axis of the stem. The red segments indicate the vectors between the C1' atoms of the first and last base pairs. The point bisecting these segments provides the direction of the 'twist' vectors.

the mean error of a circle fit to the projection of the backbone atoms onto the plane perpendicular to the axis, also lags behind the superposition based methods on shorter helices. The paucity of data points (backbone atoms) leads to semi-circular and noisy projections which in-turn lead to potentially inaccurate circle fits. The two superposition-based methods show almost identical performance due to the regular nature of the ideal helices to which the axis and circle are fit. The quality of all methods degrades slightly for larger helices due to the increased tendency of helices to bend slightly and deviate from the ideal geometry as they increase in length.

A.5.2. Ad-hoc

Let the residues on one strand of the helix be numbered $s_{a_1}, s_{a_2}, \dots, s_{a_n}$, and on the other strand $s_{b_1}, s_{b_2}, \dots, s_{b_n}$, where $s_{a_{n-1}} < s_{a_n}$ and $s_{b_{n-1}} < s_{b_n}$ then the two residues that cap one end of the helix are s_{a_1} and s_{b_n} while the residues that cap the other end of the helix are s_{a_n} and s_{b_1} . Let $c_\alpha(s)$ define the location of the C_α atom in the residue numbered s . The helix vector which is used as an initial estimate for the optimization function described above is calculated as follows:

$$\begin{aligned} V_{s1} &= c_\alpha(s_{a_1}) - c_\alpha(s_{b_n}) \\ V_{s2} &= c_\alpha(s_{a_1} + 1) - c_\alpha(s_{b_n} - 1) \\ V_{e1} &= c_\alpha(s_{a_n}) - c_\alpha(s_{b_1}) \\ V_{e2} &= c_\alpha(s_{a_n} - 1) - c_\alpha(s_{b_1} + 1) \end{aligned}$$

The vectors V_{s1}, V_{s2}, V_{e1} , and V_{e2} are between the C_α atoms of two base-paired nucleotides. They should be roughly orthogonal to the axis of the helix.

$$\begin{aligned}
S_n &= V_{s_1} \times V_{s_2} \\
E_n &= V_{e_1} \times V_{e_2} \\
S_y &= S_n \times V_{s_1} \\
E_y &= E_n \times V_{e_1}
\end{aligned}$$

The vectors S_n and E_n should be two estimates of the helix axis as calculated from the top and bottom two nucleotides. S_y and E_y should be orthogonal to the estimated helix-axis as well as the vector between the first and last C_α atom of the helix, respectively.

$$\begin{aligned}
S_c &= \left(\frac{V_{s1}}{|V_{s1}|} + \frac{S_y}{|S_y|} \right) / 2 \\
E_c &= \left(\frac{V_{e1}}{|V_{e1}|} + \frac{E_y}{|E_y|} \right) / 2 \\
S &= c_\alpha(s_{a1}) + \frac{8.4 \cdot S_c}{|S_c|} \\
E &= c_\alpha(s_{b1}) + \frac{8.4 \cdot E_c}{|E_c|} \\
V_{est} &= E - S
\end{aligned}$$

Taking the average of the normalized C_α vectors and the previous orthogonal vector should yield a vector roughly towards the center of the helix. Finally, adding a multiple of this vector to the C_α atoms of the start and end residues, yields rough estimates for the start and end of the helix. This is the crudest method of estimating the cylinder axis vector and performs consistently worse than the other methods (figure A.9).

A.5.3. Fit Method

An improvement over this method comes from the realization that a helix is a path along a cylindrical manifold. We proceed by fitting a cylinder to the backbone atoms of an RNA helix. Given a cylinder with an axis vector C_v , then a coordinate system can be created with the z-axis lying along C_v . Assuming the axis vector defines a cylinder with radius r which minimizes the root-mean-square distance of the backbone atoms from the surface of the cylinder, then transforming the locations of the backbone atoms into the coordinate system defined by the cylinder vector should yield a circle on the plane defined by the x and y axes which has a minimum root-mean-square deviation from the circle created by intersecting our ideal cylinder with the x-y plane of the our transformed coordinate system. These assumptions yield a straightforward method of fitting a cylinder to an RNA helix by fitting an axis such as to minimize the root-mean-square deviation of the best circle that can be fit onto the transformed backbone atoms of the RNA helix. The use of this method significantly improves the ability to align a second helix onto the first over the naive 'estimate' method (Fig. A.9).

Fitting the circle is done via the method described by [4], while fitting the axis is done using Python's leastsq optimization function. An initial estimate for the axis vector is calculated using the ad-hoc method.

A.5.4. Superimpose and Template Methods

RNA helices, especially of the shorter variety, often vary slightly in their structure such that their atoms do not form an ideal circle when viewed along the helix axis. To compensate for this, we first created an idealized stem using the fiber program [25], superimposed it onto the stem being parameterized and then calculated the cylinder axis for the ideal stem. The resulting method (called **superimpose**) showed a marked improvement over all of the previous attempts at approximating a helix.

Longer helices, as one may expect, should be easier to parameterize due to the larger number of data points available. The final method we tried was to create an idealized 30 base-pair helix and to approximate its axis using the 'fit' method described above. For every subsequent helix of length n ($n < 30$) for which we wanted to calculate an axis, we simply calculated the best rotation and translation to align its atoms onto those of the first n base pairs of long reference helix. The resulting transformation was applied to the reference axis which was further cropped to match the length of the query helix. The results indicate that while this method (called **template**) was more consistent, the parameters it created were on average no better than those created by the 'superimpose' method (Fig. A.9). Nevertheless, it is the method that is used throughout this paper for fitting coarse grain helices onto all-atom models.

A.5.5. Twist Parameters

In addition to the cylinder defining the helix, we need an approximation for how the base pairs are positioned along this cylinder. This is accomplished simply by storing vectors (henceforth called twist vectors) which indicate the direction from each end of the cylinder to the middle of the terminal base pairs. If the axis of the stem cylinder is defined by a start point, C_s , and an end point, C_e , then the twist vectors are calculated by the taking the vector rejection of a , (the average of the vectors from the end of the cylinder to the terminal nucleotides' C_α atoms), from the cylinder's direction vector.

$$\begin{aligned} a &= c_\alpha(s_{a_1}) - C_s + (c_\alpha(s_{b_n}) - C_s) \\ t &= a - \left(\frac{a \cdot V}{V \cdot V} \right) V \end{aligned}$$

A.6. Sampling Quality and Energy Factor

A.6.1. Energy Function Evolution

As the sampling proceeds, the distribution of the sampled values for each coarse grain measure change and the reference distribution for the energy function must also change. This is starkly illustrated in the case of the adenine riboswitch (PDB: 1Y26). The initial energy function, as calculated from substructures of the native and artificially constructed ribosome (PDB: 1jj2), favors more compact structures due to the relative paucity of such structures in the initial reference distribution (see Fig. A.3, upper left hand plot labeled 1Y26). If, however, we look at conformations sampled for this particular secondary structure (as opposed to the collection of ribosome substructures of roughly the same size), it becomes apparent that the background distribution actually favors more compact structures than the target distribution (see green and orange, respectively in Fig. A.6, upper left corner labeled 1Y26). Thus as the sampling proceeds, the energy values for more compact structures stay stable whereas the energy values for more spread out structures decrease (see Fig. A.11).

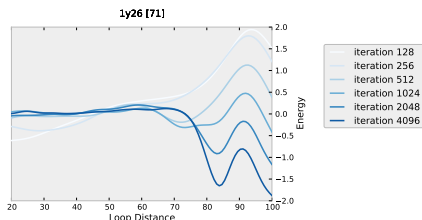


Figure A.11: The change in the values of the energy function as the sampling progresses. With few samples, more compact structures are favored. As more of those accumulate, more spread out structures assume the lower energy values facilitating the sampling of structures matching the target distribution.

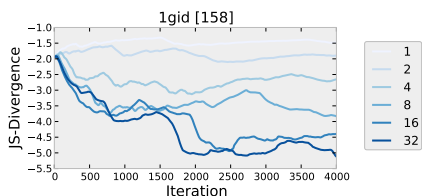


Figure A.12: The Jensen-Shannon divergence between the target and sampled distribution at various points during the simulation. The color of the lines corresponds to the value of the energy weight. Larger energy weights lead to closer matches between the target and sampled distributions and thus lower KL values.

A.6.2. Energy Factor

The energy for each coarse grained value (x) is calculated as the log of the ratio of the target distribution ($p_t(x)$) divided by the reference distribution ($p_r(x)$) and multiplied by a factor c which serves as a parameter for tuning how closely the target distribution should match the sampled values.

$$E = -c \cdot \log \frac{p_t(x)}{p_r(x)} \quad (\text{A.1})$$

Since the reference distribution is re-calculated every ten sampling steps, it tends to converge toward the target distribution. Under-sampled and over-sampled regions of the conformation space will contain structures whose energy values have large absolute values. For over-sampled regions, these values will be large and lead to a quick traversal into a different region of the conformational space whereas in under-sampled regions the energy values will be low, yielding more structures in this region. The factor c in the energy formulation above determines how strictly the sampled distribution should follow the target distribution.

The measure of the Jensen-Shannon (JS) divergence between the target and sampled distributions provides a natural measure of when to terminate the sampling. While other methods sample a fixed number of steps in the hope that a sufficiently low energy structure is found, we can simply identify when the JS divergence stabilizes and return the most frequently sampled conformations at that point. Fig. A.13, shows how the JS divergence between the sampled and target distributions varies as a function of how many sampling

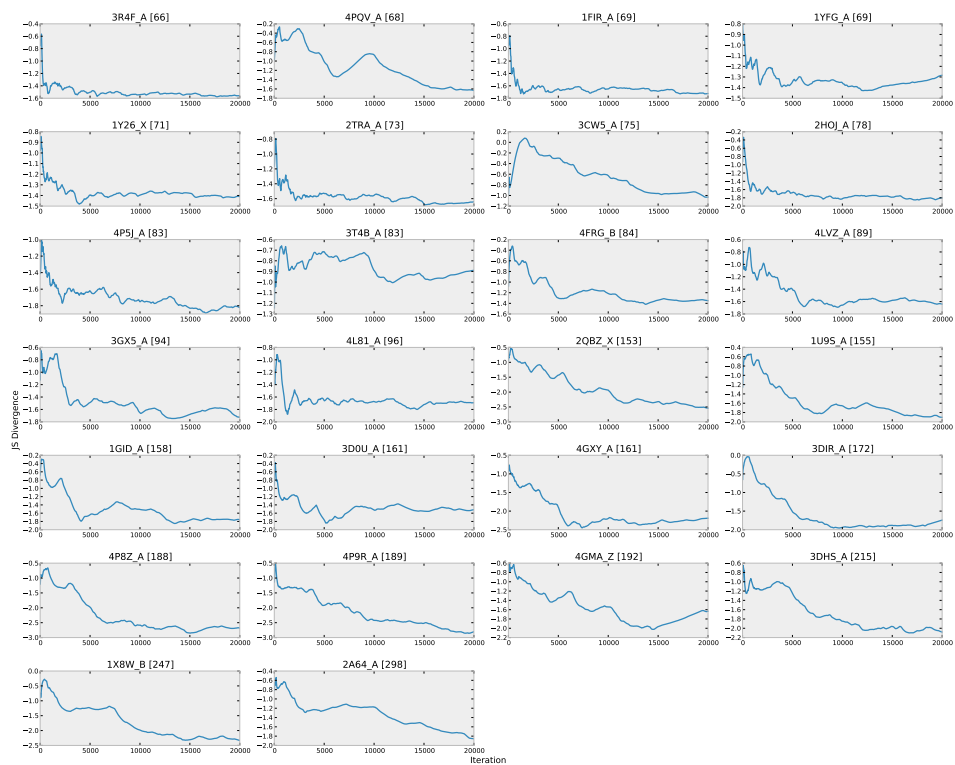
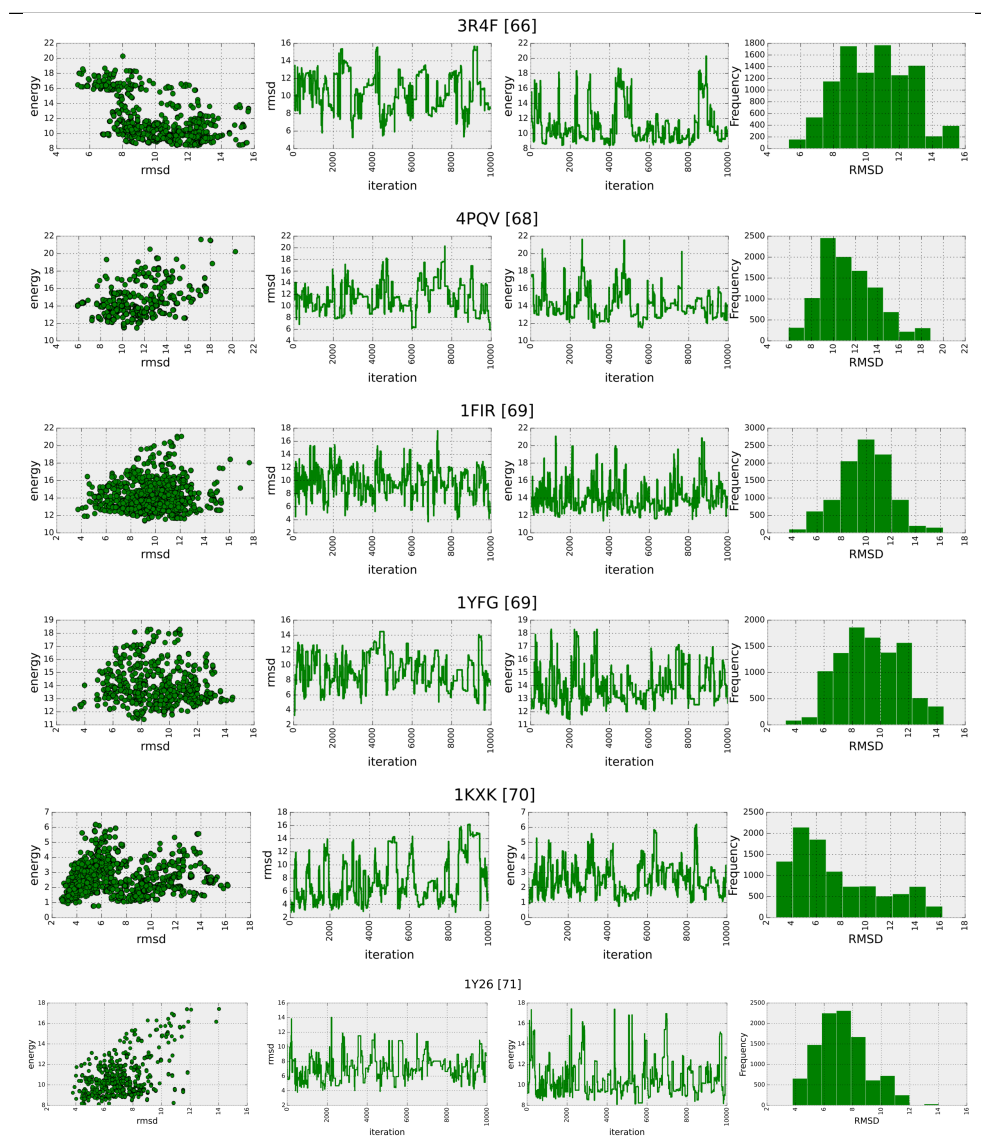


Figure A.13: The Jensen-Shannon divergence between the target and sampled distribution at various points during the simulation for all of the structures tested.

iterations have been performed. We present this merely as an avenue for further research. In actuality, all of our simulations were run to 20000 sampling steps.

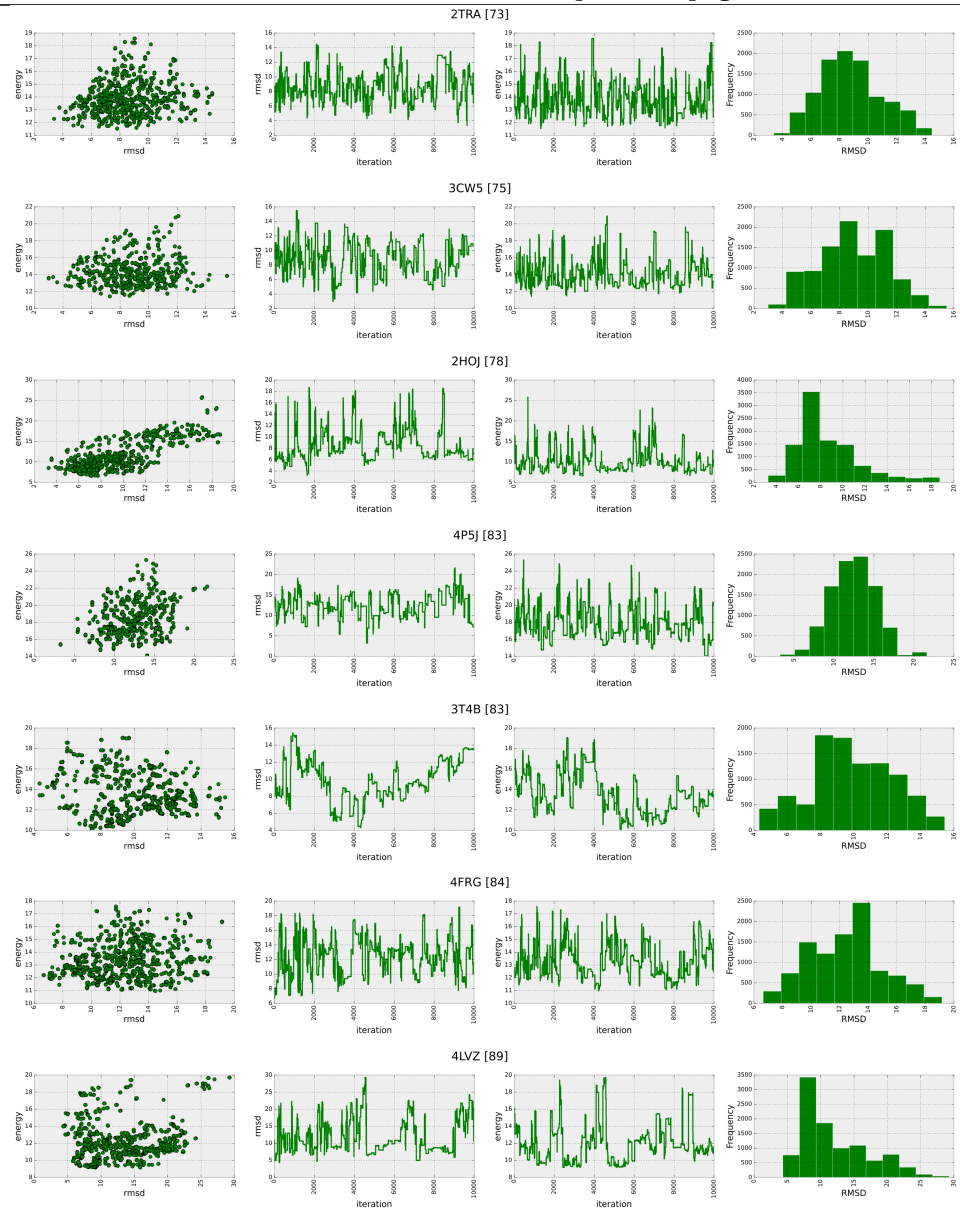
A.7. Termination Criterion and Sampling

Table A.2: Energy landscape and profile over the course of an Ernwin simulation. See Main Text Figure 9 for a more detailed description.



Continued on next page

Table A.2 – continued from previous page



Continued on next page

Table A.2 – continued from previous page

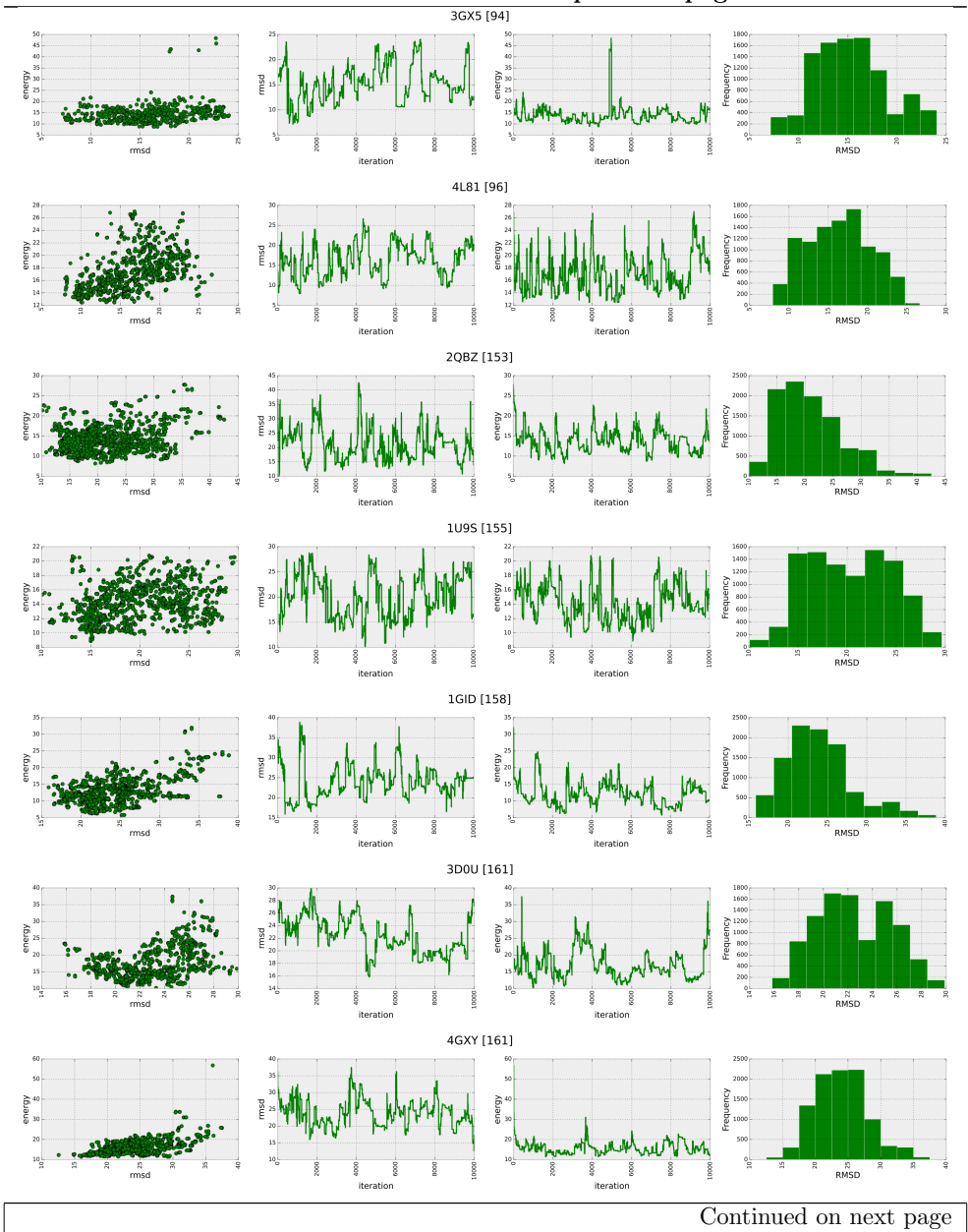


Table A.2 – continued from previous page

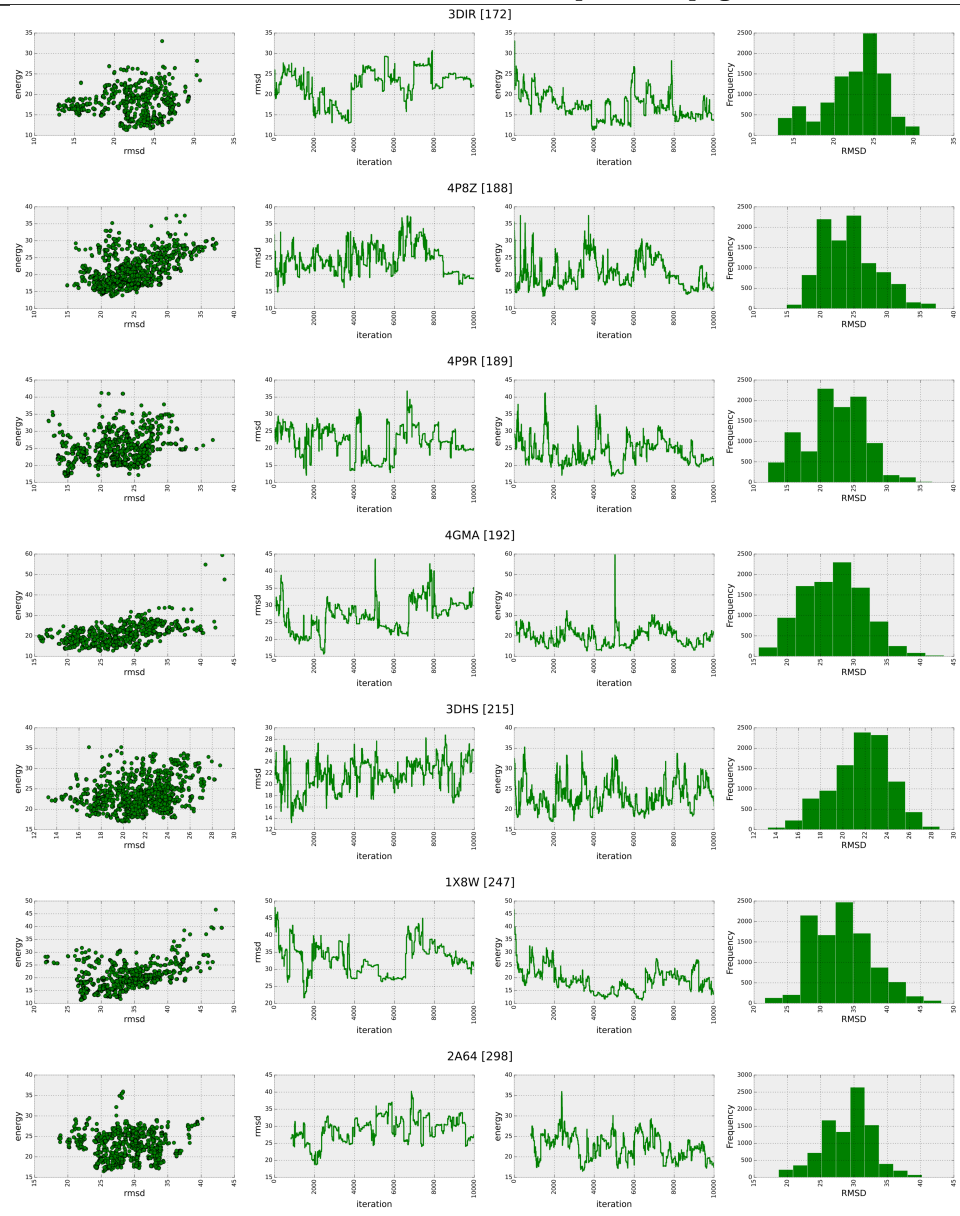
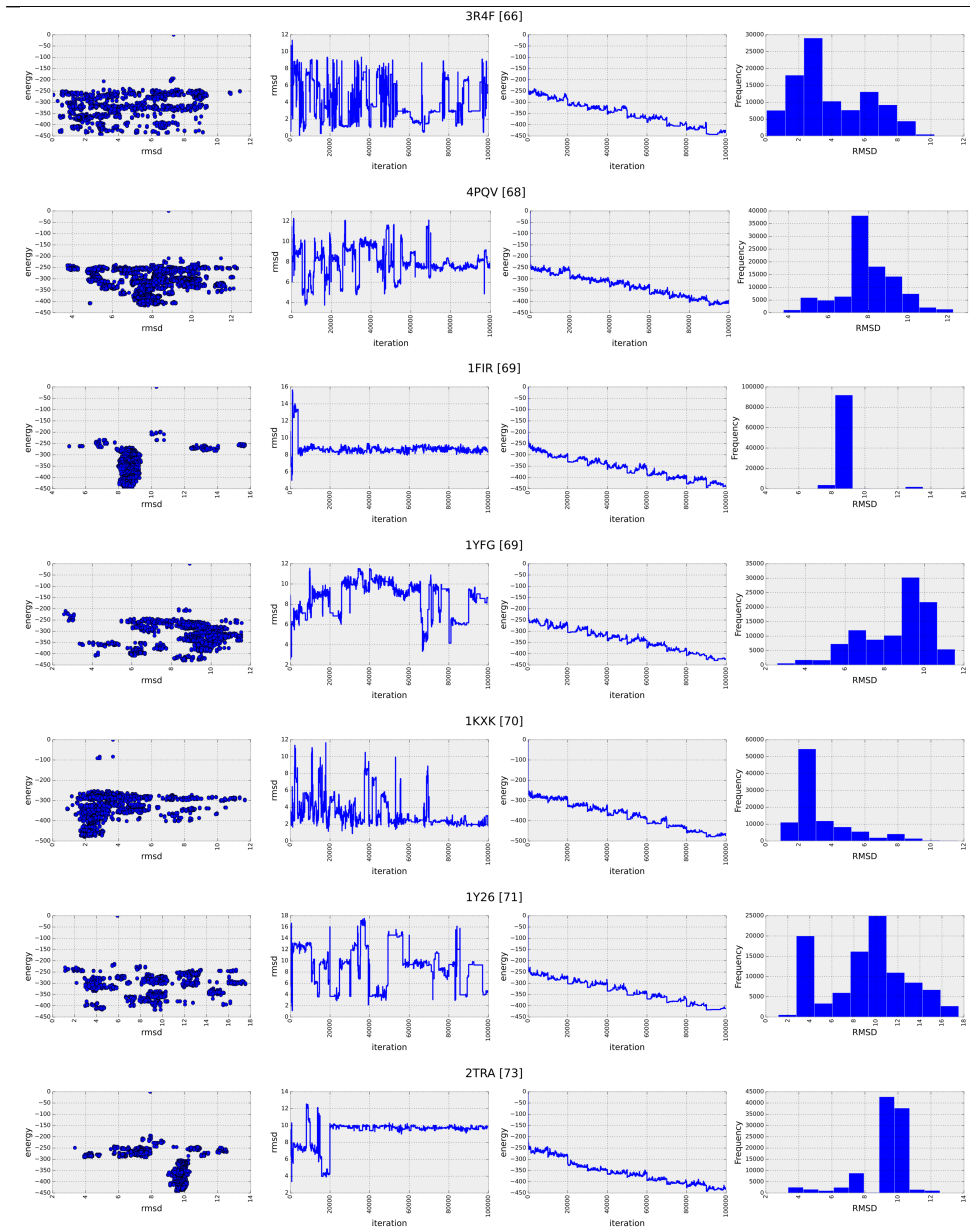
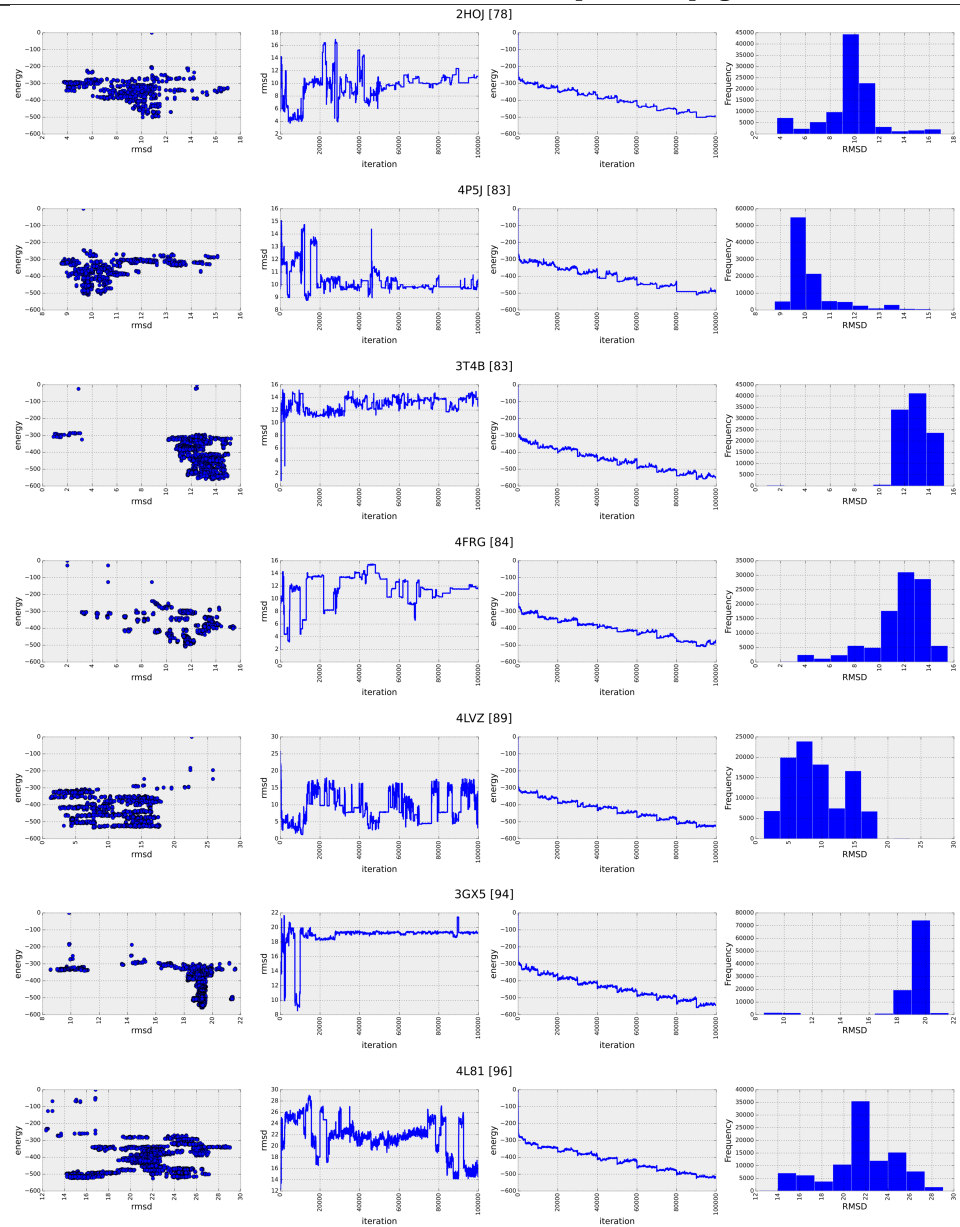


Table A.3: Energy landscape and profile over the course of an Erwin simulation. See Main Text Figure 9 for a more detailed description.



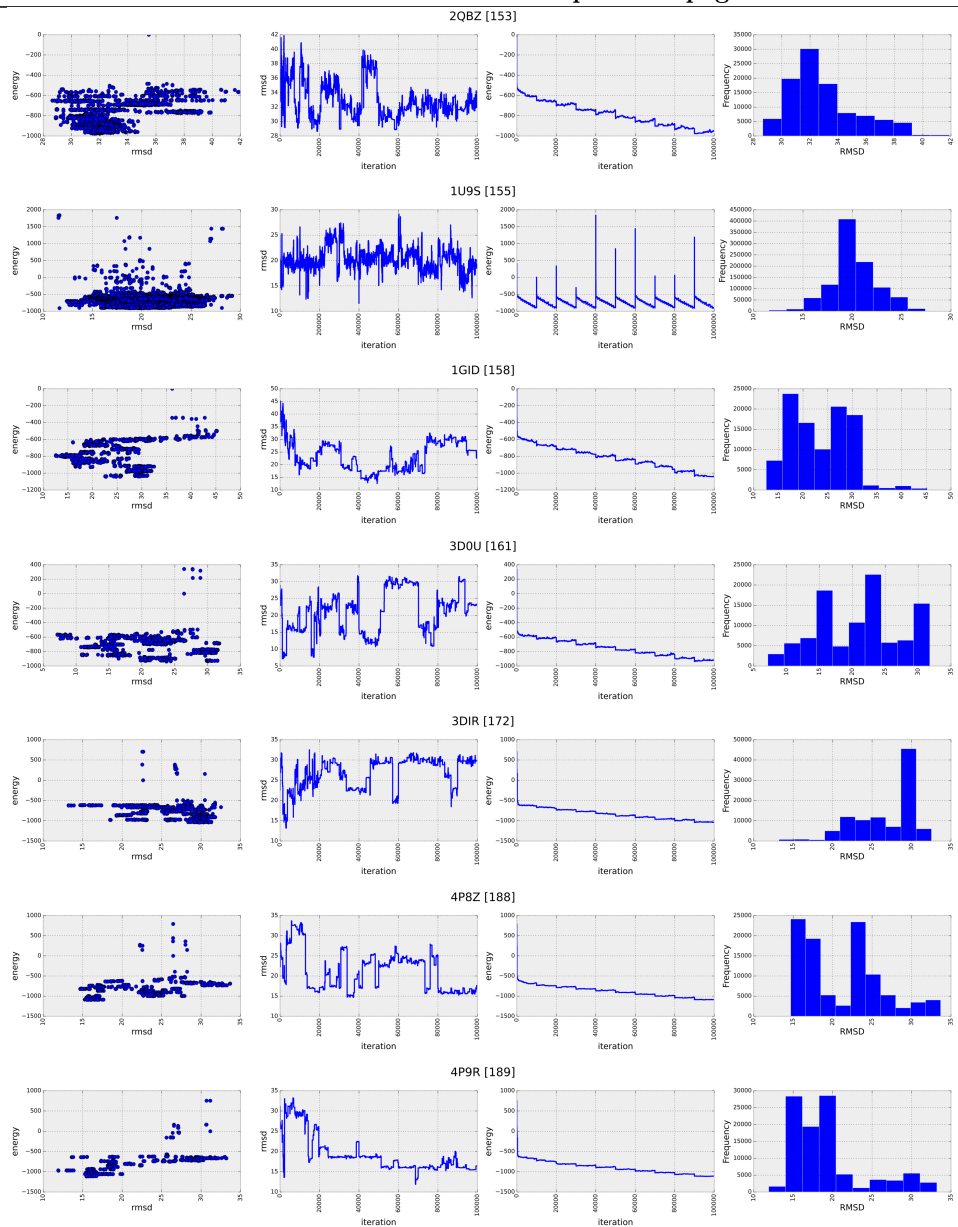
Continued on next page

Table A.3 – continued from previous page



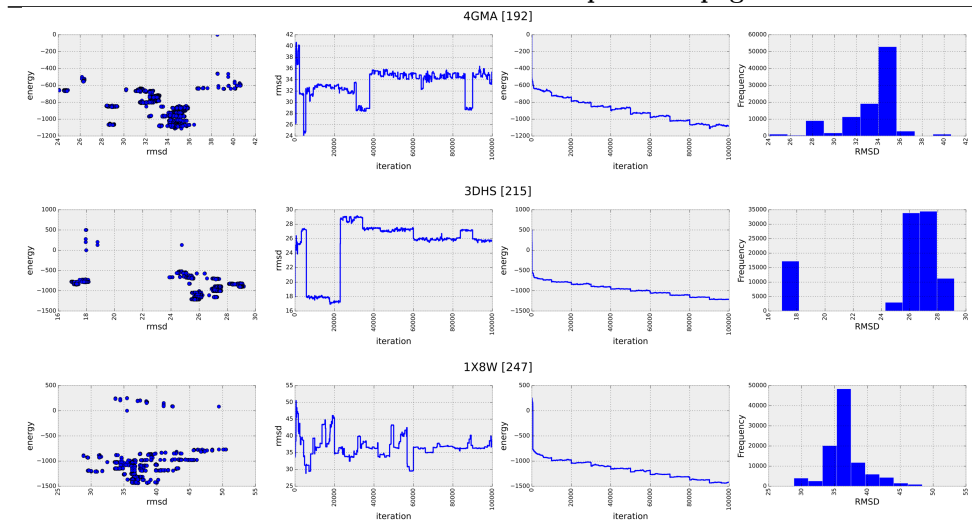
Continued on next page

Table A.3 – continued from previous page



Continued on next page

Table A.3 – continued from previous page



A.8. Running Time

The times listed in Table A.4 represent the entire time spent for each simulation. Due to the lack of sequence dependence, all fragment statistics used by Ernwin are pre-computed so normal usage does not require their recalculation. FARNa, however, requires the creation of sub-models for each loop region, leading to the large running time listed in Table A.4. Such models could, in principle, be computed ahead of time, stored externally and assembled to form a complete structure which would make simulations using FARNa significantly faster.

PDB ID	Chain Length	Ernwin	NAST	RNAComposer	FARNA
3R4F	66	1475	319	16	5482
4PQV	68	1812	235	15	6141
1FIR	69	952	48	14	6422
1YFG	69	1034	262	14	6319
1KXK	70	2284	277	15	5592
1Y26	71	1775	48	13	6747
2TRA	73	1282	150	13	6444
3CW5	75	1915	316	16	*
2HOJ	78	3332	341	16	7381
4P5J	83	2004	225	17	8669
4FRG	84	2993	63	17	9227
4LVZ	89	1976	302	20	8047
3GX5	94	4469	171	20	10751
4L81	96	2982	277	21	10539
2QBZ	153	3529	536	31	14549
1U9S	155	9204	*	35	21692
1GID	158	3852	328	33	15047
4GXY	161	6171	*	34	*
3D0U	161	5395	354	36	16796
3DIR	172	3167	485	38	16832
4P8Z	188	3398	1530	39	19630
4P9R	189	3607	*	39	20118
4GMA	192	4122	1354	47	22965
3DHS	215	4279	446	52	30740
1X8W	247	6334	1194	71	28799
2A64	298	2687	*	72	*

Table A.4: The total time (in seconds) spent assembling structures. It should be noted that the results are a run of 10000 MCMC steps using Ernwin, 10000 MCMC steps for 10 structures using FARNA, 1000000 MCMC steps using NAST and obtaining 10 structures using the RNAComposer web interface. Asterisks denote a failure to obtain models using the given program. The times correspond to the samples and predictions presented in Figure 8.

Peter Kerpedjiev, Stefan Hammer, and Ivo L. Hofacker.

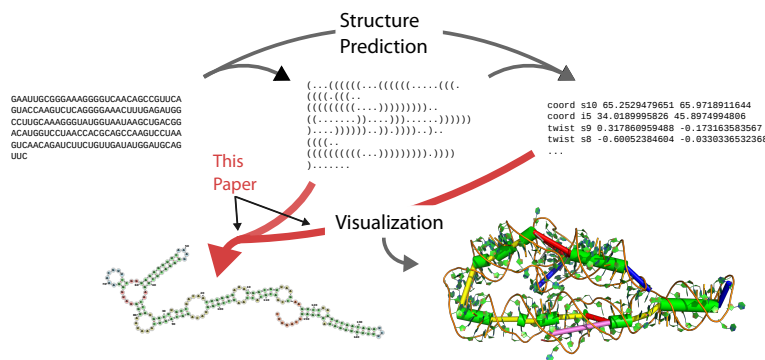
forna (force-directed RNA): Simple and Effective online RNA secondary structure diagrams.

in *Bioinformatics* 2015, 31:20.

doi: [10.1093/bioinformatics/btv372](https://doi.org/10.1093/bioinformatics/btv372)

PK and SH implemented the software. PK wrote most of the article. SH and IL contributed ideas and wrote portions of the article.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Bioinformatics Advance Access published July 15, 2015

Bioinformatics, 2015, 1–3

doi: 10.1093/bioinformatics/btv372

Advance Access Publication Date: 22 June 2015

Applications Note

OXFORD

Structural bioinformatics

Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams**Peter Kerpedjiev^{1,*}, Stefan Hammer^{1,2} and Ivo L. Hofacker^{1,2}**¹Institute for Theoretical Chemistry, University of Vienna, Währinger Straße 17/3, A-1090 Vienna, Austria and²Research Group Bioinformatics and Computational Biology, University of Vienna, Währinger Straße 29, A-1090 Vienna, Austria

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on January 20, 2015; revised on May 29, 2015; accepted on June 11, 2015

Abstract

Motivation: The secondary structure of RNA is integral to the variety of functions it carries out in the cell and its depiction allows researchers to develop hypotheses about which nucleotides and base pairs are functionally relevant. Current approaches to visualizing secondary structure provide an adequate platform for the conversion of static text-based representations to 2D images, but are limited in their offer of interactivity as well as their ability to display larger structures, multiple structures and pseudoknotted structures.

Results: In this article, we present *forna*, a web-based tool for displaying RNA secondary structure which allows users to easily convert sequences and secondary structures to clean, concise and customizable visualizations. It supports, among other features, the simultaneous visualization of multiple structures, the display of pseudoknotted structures, the interactive editing of the displayed structures, and the automatic generation of secondary structure diagrams from PDB files. It requires no software installation apart from a modern web browser.

Availability and implementation: The web interface of *forna* is available at <http://rna.tbi.univie.ac.at/forna> while the source code is available on github at www.github.com/pkerpedjiev/forna.

Contact: pkerp@tbi.univie.ac.at

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

The use of secondary structure diagrams is ubiquitous within the field of RNA biology. They convey not only which nucleotides are paired, but also and perhaps more importantly, which are unpaired. The contents and positions of sub-structures such as hairpin loops, interior loops, multiloop junctions and external loops are immediately evident. Such information is of great value to researchers seeking to identify putative mutations to perform when seeking to isolate the structural basis of a biological effect, to find protein binding, and to provide a context for observed behavior. It is used as both an exploratory as well as a communicative tool. Researchers examine secondary structure diagrams to gain insights about potential functions and mechanisms as well as to describe and disseminate them.

Although there are a number of available tools (Byun and Han, 2009; Darty *et al.*, 2009; Hecker *et al.*, 2013; Wiese *et al.*, 2005) for visualizing the secondary structure of RNA molecules, with the exception of PseudoViewer, none are available online without java and none offer the flexibility in exploring, arranging and manipulating the structure that *forna* does (Table 1 for an enumerated comparison of features).

2 Approach

Our tool, called *forna* (for *f*orce-directed *r*na), consists of a web interface and a server which allows users to input RNA secondary structures as dot-bracket strings, and displays it as a force-directed graph (Screenshot in Fig. 1). In a manner previously demonstrated by jViz.RNA (Wiese *et al.*, 2005) the user can then position each of

Table 1. Comparison of the features of existing RNA visualization tools (where PV = PseudoViewer)

	forna	VARNA	PV	jVizRNA	RNAfdl
Editing	✓	✓			✓
Pseudoknots	✓	✓	✓	✓	
PDB files	✓				
Struct. prediction	✓				
Probing data	✓	✓			
Custom coloring	✓	✓			✓
Color schemes	✓	✓			
RNA-RNA pairs	✓	✓	✓		
Circular RNA	✓				
Annotations		✓			
Circular layout		✓		✓	✓

forna provides at least three convenient features not found in other programs.

the nucleotides and stems by dragging them. Each of the nucleotides is represented as a node, whereas backbone and base-pair bonds are considered links. Connections are treated as springs and a force is calculated to keep them a fixed distance from each other. Hidden helper nodes and extra links help to maintain the familiar RNA secondary structure layout. The initial position of each node (nucleotide) is calculated using the NAView algorithm (Bruccoleri and Heinrich, 1988), but is subsequently optimized by the force-directed layout algorithm. This can (especially for larger molecules) lead to artifacts such as twisted helices and nested loops, but these are easily rectified by dragging the affected nodes to their correct positions.

2.1 Input/output

Users can enter structures in dot-bracket format (Supplementary Material Sections S1.3, S1.6 and S1.7). When done, the diagram can be saved as either a vector (SVG) or raster (PNG) graphic. If one wishes to edit the structure again in the future, it can be saved and reloaded in forna using the JSON format.

2.2 Dragging to position elements

The layout can be rearranged by selecting and dragging single or multiple nodes. The virtual forces then pull the structure toward an RNA-like layout with nearly uniform link distances. This behavior is similar to that available in jViz.Rna and valuable for arranging the nucleotides in a relevant, meaningful or simply aesthetically pleasing manner.

2.3 Pseudoknots and custom links (Supplementary Material Sections S1.1 and S1.6)

It is often necessary to display the interaction between two molecules or between different parts of the same molecule (i.e. pseudoknots). Although the user can enter pseudoknotted structures in dot-bracket notation (i.e. ((.[[...)]])), the pseudoknotted nucleotides in these cases are added as links with no strength. One may also add custom links by holding down shift and dragging from one nucleotide to another. This creates a spring-loaded link which can bring distal portions of a molecule together, or connect separate molecules. Such links are useful in depicting RNA-RNA interactions.

2.4 Coloring (Supplementary Material Section S1.4)

The coloring of nucleotides is essential for overlaying metadata on top of a structure. forna provides three default coloring modes: position, structure and sequence which color nucleotides according to their position in the molecule, the type of structural element they are in (i.e. stem, interior, hairpin, multi or exterior loop) or their identity (A,C,G

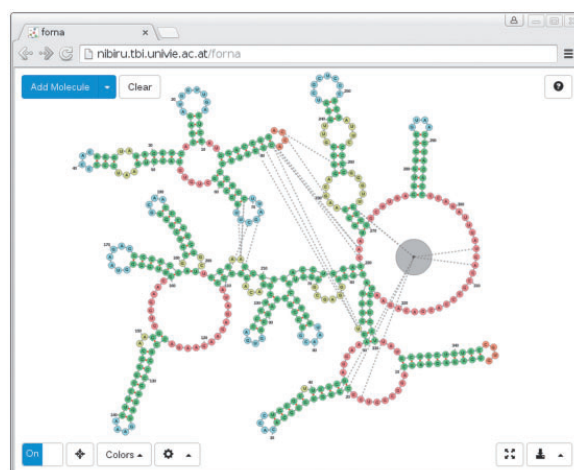


Fig. 1. Screenshot of forna web app displaying the 'Bacterial Ribonuclease P Holoenzyme in Complex with tRNA' (PDB ID: 3Q1Q). Immediately evident are the regions of the tRNA which are in contact with the ribonuclease, namely the 5' and 3' end nucleotides, as well as the TΨC loop. An RNA-binding protein is shown as a gray node in the lower-right hand region of the diagram

or U). A custom coloring mode is provided where bespoke values (as from probing data) can be entered in a text field.

2.5 Integrated structure prediction (Supplementary Material Section S1.5)

To simplify the process of going from sequence to secondary structure, forna provides a transparent interface to the Vienna RNA Package (Lorenz et al., 2011) which automatically calculates the minimum free energy for a particular sequence if no secondary structure is provided. This allows one to paste a sequence in the input field and immediately view its predicted secondary structure, without using additional tools.

2.6 Tertiary to secondary structure (Supplementary Material Sections S1.2 and S1.6)

One of the most important and unique features of forna is the automatic display of secondary structure information given a 3D structure as a PDB file. forna automatically extracts base-pair interaction information using MC-Annotate (Gendron et al., 2001) and displays the canonical secondary structure, which can be explored, manipulated or colored as described in the previous sections. Multiple chains are displayed as disconnected graphs. Proteins are displayed as larger gray nodes and interactions between different chains are represented as dashed lines (Fig. 1).

2.7 Reusable display container (Supplementary Material Section S2)

Researchers can effortlessly share RNA structures online by adding a few lines of javascript to their web page and showing a diagram of the secondary structure embedded as an SVG container attached to any specified element in the DOM tree. This rendering is purely client side and requires no calls to the server.

3 Conclusion

We provide an easy to use, accessible, free, open-source web tool for RNA secondary structure visualization that produces beautiful, highly customizable plots. Our tool requires no externally installed

software and is useful for both the exploration and dissemination of RNA secondary structure.

Acknowledgements

Thanks to developers of custom bootstrap elements and other handy tools written in javascript, e.g. saveSvgAsPng.js FileSaver.min.js, d3.js.

Funding

This work was funded by the Austrian DK RNA program FG748004, by the Austrian FWF, project 'SFB F43 RNA regulation of the transcriptome', and the European Commission under the Environment Theme of the 7th Framework Program for Research and Technological Development (Grant agreement number 323987).

Conflict of Interest: none declared.

References

- Bruccoleri,R.E. and Heinrich,G. (1988) An improved algorithm for nucleic acid secondary structure display. *Comput. Appl. Biosci.*, **4**, 167–173.
- Byun,Y. and Han,K. (2009) PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots. *Bioinformatics*, **25**, 1435–1437.
- Darty,K. *et al.* (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974.
- Gendron,P. *et al.* (2001) Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, **308**, 919–936.
- Hecker,N. *et al.* (2013) RNA secondary structure diagrams for very large molecules: RNAfdl. *Bioinformatics*, **29**, 2941–2942.
- Lorenz,R. *et al.* (2011) ViennaRNA package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
- Wiese,K.C. *et al.* (2005) jviz. rna-a java tool for RNA secondary structure visualization. *IEEE Trans. Nanobioscience*, **4**, 212–218.

Supplementary Material (Gallery)

July 2, 2015

The implementation and description of `forna` is divided into two parts: the interactive web application (Section 1) described in the main text as well as an additional section describing its use as reusable container for displaying RNA secondary structures on web pages (Section 2).

1 `forna` as a Web Application

1.1 Drawing a Secondary Structure Starting from an Open Configuration (Figure 1)

One of `forna`'s key features is the ability to intuitively edit an RNA structure. Links between unpaired nucleotides are added by holding the shift key and dragging from one unpaired nucleotide to another. The structure is immediately updated along with the coloring. The recalculation is performed on the client ensuring lag-free updates. Unwanted links can be removed by holding 'shift' and clicking on them. If the user introduces a pseudoknot with new link, it is detected and added as a force-less link.

```
>molecule_name
CGCUUCAUAUAAUCCUAAUGAUUGGUUUGGGAGUUUCUACCAAGAGCCUAAAACUCUUGAUUAUGAAGUG
.....
```

1.2 Visualizing a PDB File

PDB files store information about the 3D positions of each atom in a molecule as determined by structural biology methods such as X-Ray crystallography or NMR. While packed with information, they can be difficult to interpret without in-depth knowledge of the structure in question. Extracting the secondary structure requires the use of intermediate programs such as MC-Annotate 2. More recently, `rnapdbe` 1 has been developed as a web service to extract and display secondary structure from PDB files. The resulting images, however, are static and wedded to the layout provided by the visualization tool as well as the secondary structure present in the PDB file.

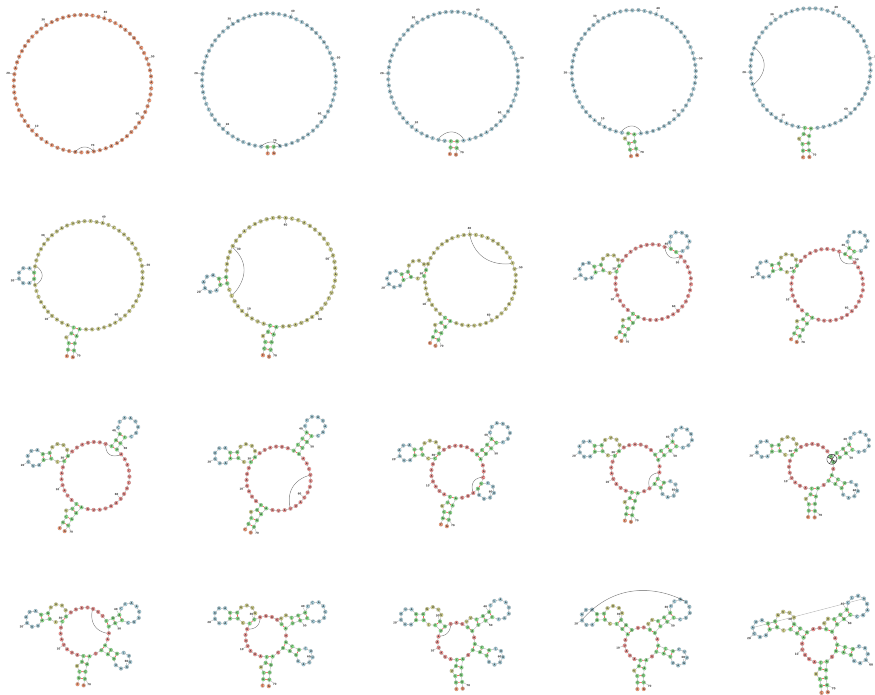


Figure 1: Drawing an RNA from an open configuration using **forna**. This sequence illustrates how one can draw a particular secondary structure and goes from left to right, top to bottom. Each arc represents the newly added base pair, which is added by shift-clicking on an unpaired nucleotide and dragging to a target nucleotide. In the last column of the third row, a link is removed by holding shift and clicking on the link. In the final step, a pseudoknotted interaction is detected and the resulting link is force-less.

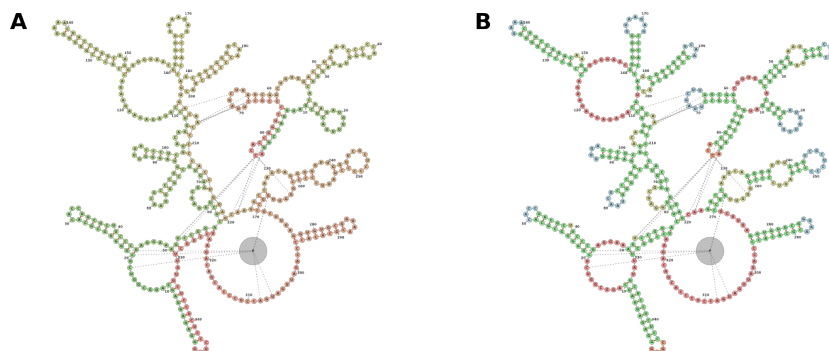


Figure 2: The secondary structure of a Bacterial Ribonuclease P Holoenzyme in Complex with a tRNA (PDB ID: 3Q1Q). The tRNA is shown in the upper right hand corner of each figure while a protein can be found in the middle of the large multi-loop junction. Two coloring schemes are shown, highlighting the positions of each nucleotide within the structure (A) and the type of structure at each nucleotide (B).

`forna` extends this functionality by allowing users to input a PDB file and displaying an interactive representation that can be explored and manipulated. Furthermore, `forna` includes information about protein interactions (an interaction in this case denoting the presence of a nucleotide and an amino acid within 2.8 Å of each other). Figure 2 displays the visualization of a Bacterial Ribonuclease P Holoenzyme in Complex with a tRNA. Immediately evident are the interactions between the ribozyme and the 5' and 3' ends of the tRNA as well as the $T\Psi C$ loop. A protein is seen interacting with the large junction and one of the interior loops of the ribozyme.

It should be noted that due to computational constraints, we set a limit of 2MB for the maximum size of a PDB file that can be uploaded. Users wishing to visualize larger molecules are encouraged to download and run the `forna` server locally.

1.3 Probing Data

Overlaying chemical probing data on a secondary structure gives researchers an informative perspective of where highly reactive regions lie. In the examples in Figure 3, it is clear that the probing data is consistent with the given secondary structure insofar as the highly reactive regions are unpaired whereas the paired regions exhibit lower reactivity. The example serves to showcase the ease with which probing data can be overlaid onto a given structure. The input sequence and structure for the molecule on the left of Figure 3 are as follows:

```
>GLYCFN_KNK_0002.rdat
ggaauuaaUCGGAUGAAGAUUAGAGGAGAGAUUUUCAUUUUAUGAAACACCGAAGAAGUAAAUCUUUCAGG
```



```

UAAAAAGGACUCAUAUUGGACGAACCUCUGGAGAGCUUAUCUAAGAGAUAAACACCGAAGGAGCAAAGCUAA
UUUUAGCCUAAACUCUCAGGUAAAAGGACGGAGaaaacaaaacaagaacaacaacaacaac
.....((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
).....))))))))))))))))))))))))))))))))))))))))))))))))))))))))
.....))))))))))))))))))))))))))))))))))))))))))))))))))))))))

```

The probing data was added by clicking on the 'Colors' drop-up and then on the 'Set' button. The following values (obtained from the RNA Mapping Database http://rmdb.stanford.edu/repository/detail/GLYCFN_KNK_0002) were pasted into the field.

```

247.6424 96.2278 54.8271 46.8534 64.6265 21.8767 39.5119 43.1716 14.4877
8.8179 2.8988 3.7053 23.1721 5.0993 3.4704 49.9487 37.8422 6.636 0.6161
0.3902 13.7014 2.9549 -0.376 1.0762 -0.5679 -0.5 2.0993 0.0671 2.4959
12.2261 13.3004 1.4647 -0.4664 -0.9908 0.0252 0.8216 0.7187 2.399 4.2495
4.4933 33.2972 11.1846 1.4183 -1.4358 1.7068 2.6488 0.4078 0.8826 4.6223
-0.0611 -0.8845 1.3948 15.4353 50.3329 6.4028 4.5445 6.3482 0.467 2.1263
39.3715 47.8274 50.979 9.8241 4.3092 0.4139 0.1997 -1.099 4.1683 32.5064
2.5253 -0.4245 10.2521 37.5462 25.9211 30.8512 21.1141 10.5151 -0.4247
1.2889 67.9884 6.7804 -0.2508 -0.1662 2.5511 -0.1782 1.4556 0.6963 -1.0435
-0.3248 4.0919 33.8002 3.647 -0.0516 33.2451 56.4639 8.8362 1.1629 -0.5062
1.1171 -0.2665 -0.5553 0.942 7.709 0.6988 17.9013 6.2786 3.2409 -1.07
0.1746 7.1433 0.1149 2.2959 9.1101 56.9181 56.5 30.1173 27.094 2.0416
6.6394 -0.1116 13.582 8.9534 3.2413 15.9977 0.9295 0.0206 0.7913 50.923
77.8383 -18.9486 7.2471 3.2602 10.3813 1.1856 61.4109 139.4171 131.8496
8.2972 2.6437 3.0427 74.7002 78.0617 5.4987 2.004 1.2521 -1.2755 12.413
0.4201 -0.4107 3.8533 2.0068 71.2049 108.6537 132.3693 6.8693 0.3981
0.0862 -0.773 11.2572 168.3761 -15.3035 14.3589 29.7187 59.0558 101.8314
110.0697 71.0021 1.9045 3.9161 156.4062 4.5086 0.7884 5.374 30.8158
15.1988 151.4786 126.5493 132.0433 150.3512 19.0894 130.44 174.8229
173.997 200.6308 228.1592

```

On the right of Figure 3 the input structure is:

```

GGAAAGCAAUUCGAGUAGAAUUGGAAAGGGAAAGAAACGCUUCAUAUAAUCCUAAUGAUUAGGUUU
GGGAGUUUCUACCAAGAGCCUUAACUCUUGAUUAUGAAGUGAAAACAAAGUUAAGGAGUACUUA
CACAAAGAAACAACAACAAC
.....((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))
.....))))))))))))))))))))))))))))))))))))))))))))))))))))))))
).....))))))))))))))))))))))))))))))))))))))))))))))))))))))))

```

The color information is also obtained from the RNA Mapping Database (http://rmdb.stanford.edu/site_media/rdat_files/ADDRSW_1M7_0006):

```

0.2731      0.5547      0.5066      0.2429      0.2110      0.2948
0.0237      0.0312      0.0589      0.0252      0.0250      0.0468
0.4216      0.4143      0.6900      0.4320      0.1618      0.0377

```

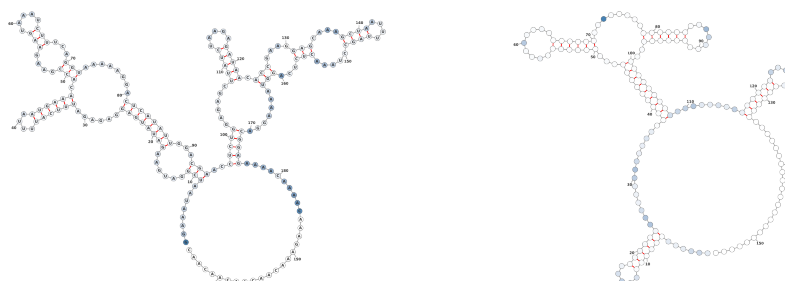


Figure 3: Probing data overlaid onto the secondary structure of an adenine riboswitch (left) and a glycine riboswitch (right). Darker colors indicate higher reactivity.

0.0555	0.0813	0.0714	0.0446	0.1151	0.8637
0.8640	0.4475	0.1755	0.1821	0.2529	0.7196
0.8994	0.5176	0.2022	0.2314	0.3745	0.2439
0.0354	0.0274	0.0218	0.0216	0.0222	0.0062
0.0148	0.0195	0.0207	0.0289	0.0232	0.0671
0.0316	0.0206	0.0094	0.0309	0.0296	0.0537
0.0360	0.0419	0.1660	0.0566	0.0546	0.7821
0.0637	0.0417	0.0963	0.0317	0.0280	0.0119
0.0262	0.0161	0.0189	0.0304	0.0490	0.0483
1.8979	0.0314	0.0071	0.0332	0.0305	0.0140
0.0191	0.0173	0.0070	0.0188	0.0159	0.0212
0.0211	0.0793	1.0460	0.6202	0.2630	0.0473
0.0220	0.0267	0.0072	0.0206	0.0175	0.0119
0.0504	0.0431	0.0422	0.0572	0.0379	0.0158
0.0500	0.0235	0.0293	0.0577	0.0413	0.4775
0.5773	0.6706	0.8049	0.2942	0.2483	0.3983
0.2593	0.6670	0.1990	0.0585	0.0596	0.0573
0.0267	0.1003	0.6188	0.2840	0.5888	0.9753
0.1816	0.0147	0.0138	0.0146	0.0152	0.0262
-0.0042	-0.0073	0.0000			

1.4 Arbitrary Coloring

It is often useful to color certain nucleotides a particular color to illustrate a region of interest. Figure 4 demonstrates how one can supply coloring information for specific ranges of nucleotides. The secondary structure in this example is extracted from the tertiary structure of the Ternary S-Domain Complex of Human Signal Recognition Particle (PDB ID). The coloring is entered by clicking on the 'Colors' drop-up, clicking 'Set' and then pasting the following text:

```
18-57:red 64-110:blue
```

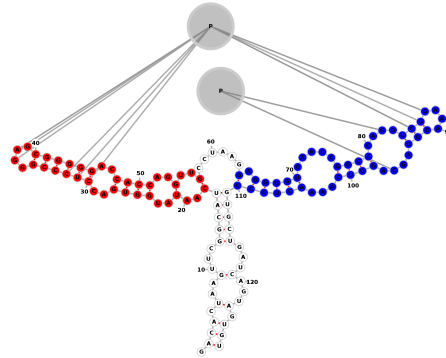


Figure 4: Specifying an arbitrary coloring scheme for an RNA. In this case, the secondary structure of the Ternary S-Domain Complex of the Human Signal Recognition Particle (PDB ID: 1MFQ) is colored to show the two branches which are involved in protein interactions.

1.5 Kissing Hairpins

One often needs to depict the interaction between two molecules. Figure 5 shows two small molecules interacting via a kissing hairpin interaction. It should be noted that this is difficult to display when the interactions are longer than a few nucleotides due to the layout constraints. Nevertheless, for shorter interactions, adding artificial links can provide an adequate view of where molecules interact. For this example, the following fasta sequences were entered in the 'Add Molecule' dialog:

```
>a
UCAAAUGAGCUACUCACGUGAGCUCAUCCUU
>b
CGAUAUGAGCUACGUGAGUAGCUCAUUGGU
```

The secondary structures are automatically predicted using `RNAfold`. The basepair nearest the hairpin is artificially broken (using 'shift'-click), and extra links are added between nucleotides 13,14,15,16 and 14,15,16,17.

1.6 Pseudoknots

Pseudoknots are detected in the input structure using a greedy algorithm which always marks the most nested base pairs as pseudoknots. These nested base

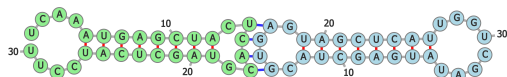


Figure 5: Two molecules interacting via a kissing-hairpin interactions. The inter-molecular base pairs are colored blue, whereas the the intramolecular base pairs are colored red.

pairs are then added as strength-less links and removed from the rest of the structure. Figure 6 shows two examples of structures with pseudoknots, one input as a pdb file (group II intron, PDB ID: 4FAW, left) and the other input from a dotbracket representation (corresponding to an adenine riboswitch). The dotbracket string for generating the structure on the right is shown below. Note the two different types of brackets used ('()' and '[]') in order to denote nested nucleotides. Other brackets such as '{}' and '<>' can also be used to denote multiply nested pairs.

```
>molecule_name
CGCUUCAUAUAAUCCUAAUGAUAUGGUUUUGGGAGUUUCUACCAAGAGCCUUAACUCUUGAUUAUGAAGUG
((((((((((..(((((((..[[[...)))))).....).((((((..]]]..))))))..))))))))))
```

1.7 Circular RNA

RNA usually exists as a single strand with distinct 5' and 3' ends, but it can also be found as a circular molecule. Such molecules have been ligated at their 5' and 3' ends and thus have no external loops. These can be displayed using forna (Figure 7) by appending an asterisk to the end of the dot-bracket string.

```
>circular_rna
CUGCUCACGCAAGGAGGUGGACUUAAGCGGCUCAUCCGGGUCUGCGGAUAUCCACUGCGCGG
UAUGCGCUCGCGAGUUCGAAUCUCGUCGCCAGUACACUGACUUCACUGGCGUGUCCGAGUGG
UUAGGCAA
..(((((((.....((((((((.....(((((((.....))))).))))).))))))((((.....
....))))..((((.....))))((((((((.....))))))..))))..))))
...))..*
```

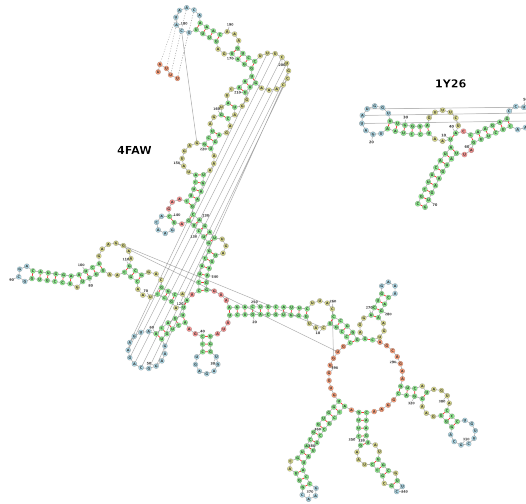


Figure 6: Pseudoknotted structures from an input dotbracket file (1Y26, upper right) and a PDB file (PDB ID: 4FAW, left).

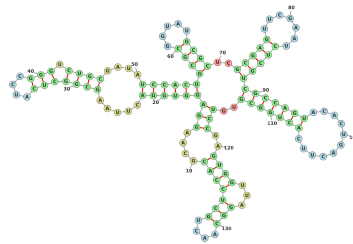


Figure 7: A circular RNA with no external loops and/or 5' and 3' ends.

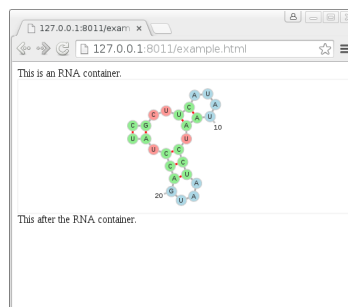


Figure 8: An example of using the `FornaContainer` to easily display an RNA structure within a web page.

2 forna as a Javascript Viewing Container

The web-application version of `forna` described in the main text relies on a server to calculate an initial layout which is then refined by the force-directed layout calculation. It provides an interface for adding and removing structures as well as for changing the coloring and the display parameters. There are, however, other applications where one may simply want to display a structure without allowing the user to display their own or to change coloring. This is often the case when one wants to share structures online, as for example, from a secondary structure prediction server. To accommodate this need, we provide an independent javascript container which is completely decoupled from the back-end server. The initial calculated layout is simpler, and features such as displaying a PDB file (which require server-side annotation) are disabled, but other features such as panning, zooming and dragging can be enabled using specific parameters.

The container is available as its own repository (called `fornac`: for `forna` container), and can be instantiated using only a few lines of javascript code. While the specifics of the API are detailed in the online documentation at <https://github.com/pkerpedjiev/fornac>, The general pattern for use is shown in the example web page below:

```
<!DOCTYPE html>
<meta charset="utf-8">

This is an RNA container.
<div id='rna_ss'> </div>
This after the RNA container.

<link rel='stylesheet' type='text/css' href='css/fornac.css' />
<script type='text/javascript' src='js/jquery.js'></script>
<script type='text/javascript' src='js/d3.js'></script>
<script type='text/javascript' src='js/fornac.js'></script>

<script type='text/javascript'>
  var container = new FornacContainer("#rna_ss",
    {'applyForce': false, 'allowPanningAndZooming': true});

  var options = {'structure': '(((..((...)).(((...))))))',
    'sequence': 'CGCUUCAUAUAAUCCUAAUGACCUAU'
  };

  container.addRNA(options.structure, options);
</script>
```

The two key features of the example are the `div` to contain the `forna` container and the javascript at the bottom which populates it with an RNA sequence, secondary structure and some optional parameters. The resulting web

page can be seen in Figure 8 where a visualization of the RNA secondary structure appears without the need to first create a static image or call a java library.

References

- [1] M. Antczak, T. Zok, M. Popena, P. Lukasiak, R. W. Adamiak, J. Blazewicz, and M. Szachniuk. RNAPdbee - a webservice to derive secondary structures from pdb files of knotted and unknotted RNAs. *Nucleic acids research*, page gku330, 2014.
- [2] P. Gendron, S. Lemieux, and F. Major. Quantitative analysis of nucleic acid three-dimensional structures. *Journal of molecular biology*, 308(5):919–936, 2001.

Part III

TERTIARY STRUCTURE PREDICTION

While tertiary structure information may be gleaned from experimental techniques such as X-ray crystallography or [NMR](#) spectroscopy, these methods are often expensive, cumbersome or simply impractical. An accurate computational method for determining the tertiary structure of an RNA molecule would be a boon to researchers looking to quickly and cheaply gain an insight into the three dimensional arrangement of the helices of an RNA molecule.

We have developed a new method for predicting tertiary RNA structures which uses statistical sampling to create an ensemble of structures conforming to a known distribution of coarse-grain measures such as the radius of gyration, the distance between loops, and the number of expected A-minor interactions. This section describes the coarse-grain model we use to represent RNA tertiary structure. It will explain the sampling procedure we use to generate structures conforming to a target distribution. It concludes with additional results and suggested avenues for further exploration.

A HELIX-CENTERED COARSE-GRAIN MODEL

In an all-atom molecular model, the global conformation of an RNA molecule can be defined by the values of the six backbone torsion angles, the glycosidic bond torsion angle (χ), and the sugar pucker (Figure 8). Given these values for each nucleotide, one can reconstruct an entire RNA molecule on an atomic level. When a coarse-graining is introduced, the dimensionality of the parameterization is reduced. This has a twofold effect of shrinking the conformational space and blurring the accuracy of the resulting model. A model that uses a single point to represent a nucleotide can be parameterized using just three parameters per nucleotide (e.g. x,y and z coordinates). At the same time, the exact position of the individual atoms will be ambiguous.

Our coarse-grain representation of an RNA molecule is dependent on the regularity of the double stranded helix. Given its uniform structure, we can represent its position using ten parameters (three coordinates for the start position, three coordinates for the end, and four for the location of its minor groove). The regions between helices are also defined by six parameters, but these parameters simply relate the positions of the helices which flank the loop region.

With respect to structure prediction, this coarse-graining has two main advantages:

- To shrink the conformational space: less potential conformations for the model.

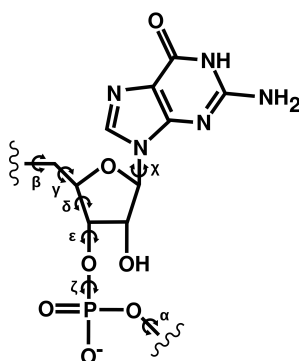


Figure 8: The torsion angles of an RNA nucleotide.

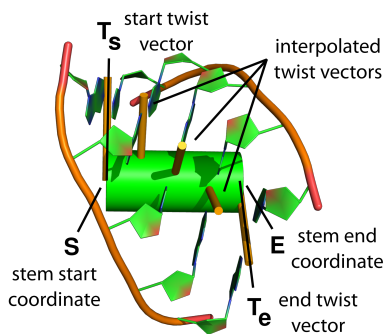


Figure 9: The parameters used to define a stem in our coarse-grain model. The stem start and end coordinates are points in 3-space, as are the twist start and end vectors. The interpolated twist vectors are not stored explicitly but calculated as an intermediate in determining virtual residue positions (Section 8.4).

- To restrict the computational workload required to generate and evaluate models.

To reap these benefits, it must sacrifice atomic details and the additional information they contribute to the energy/validity of the proposed structures. Our model takes coarse-graining one step beyond a one-point per nucleotide resolution and represents RNA helices as geometric helices having two end points and a *twist* parameter indicating the position of the minor groove and how much the helix turns along its length.

8.1 Helices are defined by 10 parameters

Each coarse-grain helix can be defined by a total of 10 parameters consisting of its start and end points, each in 3-space, and four parameters defining the location of its minor groove (Figure 9). In practice, however, each stem is described using 12 parameters. They can be found in each coarse-grain structure file under the `coord` section:

```
coord s2 24.10 -28.58 0.013 35.83 -33.22 -2.85
```

These values represent a start point (x,y, and z coordinates) and an end point (x,y, and z coordinates). The rotation of the helix is represented by another six parameters (T_s and T_e in Figure 9):

```
twist s2 0.34 0.32 0.88 -0.39 -0.53 -0.74
```

These six parameters define two vectors. The first starts at the beginning of the stem and the second at the end. These vectors point toward the middle of the minor groove for this helix. While six coordinates are used,

only four are actually necessary for this representation. Because the two *twist* parameters are coplanar and equal in length, the second twist can be represented as a rotation of the first around the axis of the helix.

In principle, each helix need only be defined by four parameters: a direction vector (3 parameters) and a helical rotation (1 parameter). If the start and minor groove direction are given, then the length and helical rotation will define its end point and final location of the minor groove. We use the 12 parameter representation simply for the sake of simplicity, ease of calculation of energy functions and statistics, as well as for visualization.

8.2 *Inter-helical orientations can be defined by six parameters*

The orientation of one helix (s_1) with respect to another (s_2), $O(s_1, s_2)$, can be represented by six parameters [9]:

- r, ϕ_d, ψ_d which describe the start of the axis of s_2 relative to the end of s_1
- ϕ_o, ψ_o , which describe the direction of the axis of s_2 , relative to the axis of s_1
- t , which describes how much s_2 is twisted relative to s_1

The values of the orientation parameters depend on the order of the two stems: $O(s_1, s_2) \neq O(s_2, s_1)$.

These parameters can be assigned to loops separating pairs of helices. Such loops function as degrees of freedom in our simulation. Changing their parameters alters the orientations of the helices and thus introduces regions of relative flexibility within the coarse-grain model.

8.3 *Terminal loops can be defined by three parameters*

Because terminal loops such as hairpin loops and exterior loops play such an important role in the tertiary structure of an RNA molecule, they must also be considered in our model. While the parameters of interior loops and sections of multiloops are implicitly defined by the stems that flank them, hairpin loops and exterior loops need to be explicitly parameterized. We have decided that each will be represented by two points specifying a start and an end. The start point is necessarily the end of the stem which that loop is connected to. The end point is defined as the C1' atom furthest away from the start point (Figure 10).

Because the end of the previous stem defines a coordinate system using the basis vectors of its axis and terminal twist parameter, the loop's end coordinates can be represented as a triple of spherical coordinates:

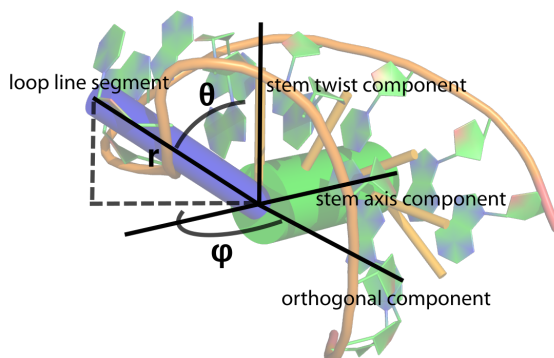


Figure 10: The parameterization of a terminal loop. r is the distance from the end of the stem to the furthest $C1'$ in the loop. θ is the polar coordinate and ϕ is the azimuth coordinate.

- r : The distance from the end of the stem.
- θ : The polar angle.
- ϕ : The azimuth angle.

In practice, loops are defined using six parameters indicating the start and end of the the line segment representing them. As implied above, three of these parameters are redundant as they are simply the end point of the previous stem (e.g. the coord definition in Section 8.1).

```
coord h0 35.83 -33.22 -2.85 46.66 -23.38 -20.99
```

8.4 Virtual residues are interpolated nucleotide positions

Because stems are so regular and we know the approximate location of the minor groove, we can use that to estimate the position of each nucleotide in a stem. The estimated positions will hereafter be referred to as *virtual residues* and will be important in creating energy functions for excluded volume (Section 8.6) and junction closure (Section 8.7). Given a stem with a start vector S , an end vector E , a length of l base pairs, and an initial twist vector T_s , a twist per nucleotide of t , an angular offset for each nucleotide from the center of the minor groove of o , then the position P_s of the n 'th base pair along the stem is:

$$P_s = S + n(E - S)/l \quad (2)$$

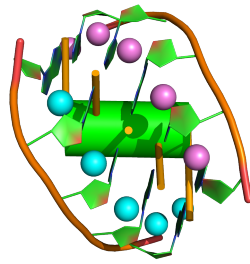


Figure 11: The positions of the virtual residues shown as cyan and magenta balls for the right and left strands, respectively.

The position of the center of the minor groove there is:

$$V = (E - S) \times T_s$$

$$P_m = P_s + T_s * \cos(t * n) + V * \sin(t * n)$$

Because the nucleotides are offset by ϕ radians from the center of the minor groove, the positions on the left and right strands are, respectively.

$$P_{n_l} = P_s + T_s * \cos(t * n + \phi) + V * \sin(t * n + \phi)$$

$$P_{n_r} = P_s + T_s * \cos(t * n - \phi) + V * \sin(t * n - \phi)$$

An illustration of where the virtual residues are positioned is shown in Figure 11.

8.5 Virtual atoms are interpolated atom positions

Similar to how we calculate the positions of *virtual residues*, we can calculate the positions of *virtual atoms*. These positions are where we would expect to see the backbone atoms of a helix, if it were an ideal helix. To calculate these positions, we create an artificial coordinate system consisting of the helix axis $E - S$ and the vector toward the minor groove at each base pair position in the stem, P_m . By using the ribosome as a template structure, we calculate the average position for each backbone atom within this coordinate system. These average positions can then be used to calculate *virtual atom* positions for each nucleotide in a stem (Figure 12).

8.6 Avoiding excluded volume

Of all the energy terms in all the structure prediction suites, perhaps the most common is a *clash* energy term. This term, intended to quantify

These positions could be improved by shifting the left and right sides up or down along the stem axis by some fixed amount.

The clash energy term is intended to approximate the effects of the Lennard-Jones potential which models the interaction of two atoms at a close distance to each other.

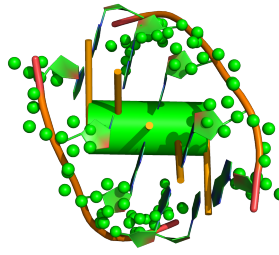


Figure 12: The positions of the virtual atoms shown as green balls. They don't exactly overlap with the real positions of the backbone atoms due to the deviation of the stem from the *average* stem and the nucleotides from the *average* nucleotide.

excluded volume is included to prevent predicted structures from containing sterically impossible overlapping atoms. Because we don't track individual atoms, we use the positions of the virtual residues as an initial filter for where there may be clashes. If we find virtual residues that are within 10\AA of each other, the virtual atoms for the stems containing them are calculated and checked for clashes (i.e. unbonded atoms that are within 1.8\AA of each other). An energy is then calculated by multiplying the number of clashes by a sufficiently large number (10000, in our case) to ensure rejection of the structure in the sampling step.

8.7 Maintaining junction integrity

ERNWIN builds structures by placing stems relative to each other, along the minimum spanning tree of the skeleton graph [87] (Chapter 6). Because multiloops are circular, *breaks* are implicitly inserted at the longest single stranded segment. These *breaks* are sections for which parameters are not sampled because they are implied by the parameters of all the shorter segments in the cycle. The broken segment may, however, end up being longer than can be physically spanned by a single stranded RNA chain. To determine if this is the case, we created a heuristic that takes as input the length of the broken segment in nucleotides and calculates how long (in \AA) the physical segment can be.

We created this heuristic by simulating the construction of thousands of multiloop regions of varying lengths. We then reconstructed the stems adjacent to the multiloop regions and tried to bridge broken segment with an all-atom backbone RNA chain. The bridging was performed by aligning the start of the multiloop RNA chain to the previous stem and performing cyclic coordinate descent [27] along the torsion angles of the multiloop chain to try and align its last nucleotide to the next stem. If the

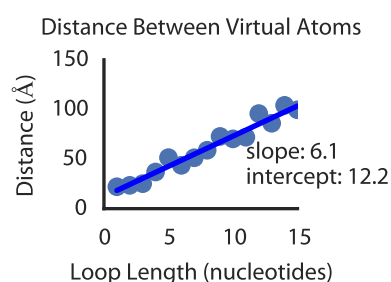


Figure 13: The distance spanned between the O₃' and P atoms of the last nucleotide of the first stem and first nucleotide of second stem adjacent to the closed loop, respectively (Figure 16, *virtual atom length* label). Points show the maximum distance spanned by a given loop length (in nucleotides).

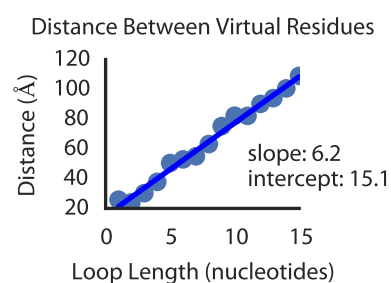


Figure 14: The distance spanned between the last virtual residue the first stem and first virtual residue of second stem adjacent to the closed loop, respectively (Figure 16, *virtual residue length* label). Points show the maximum distance spanned by a given loop length (in nucleotides).

RNA chain spans the broken multiloop segment length, its last nucleotide will align to the first nucleotide of the second stem with a low *RMSD*. To see how far a given fragment can reach, we plotted the maximum distance that could be spanned by multiloop sections of varying nucleotide lengths. A distance can be spanned if the multiloop fragment can be aligned to the last nucleotide of the previous stem and the first nucleotide of the next stem. The last nucleotide of the multiloop is considered aligned if its *RMSD* to the first nucleotide of the next stem is less than 0.1 Å after performing CCD. The nucleotides being aligned are not counted in the nucleotide length of the fragment. We tabulated these results and display them in Figures 13, 14, and 15.

The results show an expected linear relationship between the length of the multiloop segment in nucleotides, and the length between the two stems that it can span. The deviations from the linear regression are due to the fact that we allow rotations of the ϵ and ζ torsion angles of the

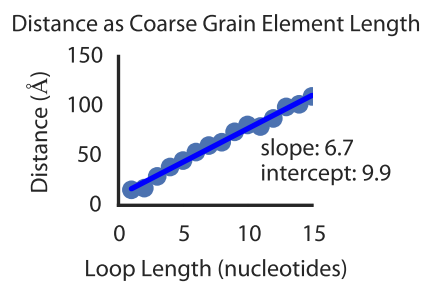


Figure 15: The maximum length of a coarse-grain multiloop element spanned by a fragment of a given nucleotide length (Figure 16, *CG element length* label).

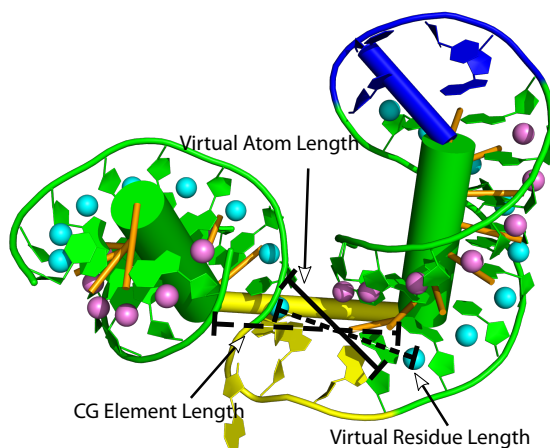


Figure 16: We tested three measures for quantifying the length of a multiloop segment. The *CG element length* measure spans the distance between the start and end of the two stems flanking the multiloop segment. The *virtual residue length* measure spans the distance between the last virtual residue of the first stem and the first virtual residue of the second stem. The *virtual atom length* measures from the O3' atom of the last residue of the first stem to the P atom of the first residue of the second stem.

last nucleotide of the first stem and the α , β , γ torsion angles of the first nucleotide of the second stem, when performing CCD.

Structure prediction programs generate models which generally do not have the same arrangement of atoms as the real (solved by X-ray crystallography or NMR) structure. This chapter will attempt to answer the question of *how we measure the difference between two models*. How can we quantify how *similar* one atomic model is to another? If two macromolecules can be perfectly superimposed on top of each other, they can be considered effectively identically. If they are superimposed on top of each other such as to minimize the distances between equivalent atoms, then we can calculate their deviation as the sum of the squares of the distances between equivalent atoms. This measure, the RMSD, is a simple, universal way to quantify the similarity between two models. The remainder of this chapter will introduce the interaction network fidelity (INF), the RPF, and the ACC measures. These measures allow us to quantify the difference between two atomic models and thus evaluate how well different methods can predict tertiary structures. Better predictions will be closer to the known structure, whereas poorer predictions will be further. Effective methods of measuring model deviation come in different flavors and rely on different properties of the molecular models to measure these differences. Some of their strengths and weaknesses are described in the following sections.

9.1 RMSD and dRMSD

The traditional way to measure the difference between two protein or RNA models is to compute the RMSD between the positions of the atoms after superimposing them [83]. A similar measure, the root mean square deviation of the distances between atoms (dRMSD) quantifies how much the inter-atomic distances differ between two molecules. This measure, while strongly correlated with the traditional RMSD measure is insensitive to stereochemical properties of the molecules [31]. While both approaches are widely used, they offer little insight into the nature of the deviations of the two molecules. To counter this, the INF and RPF measures have been used to quantify which noncanonical interactions are correctly predicted and which sections of the molecule are predicted to be near to each other.

Due to the coarse-grain nature of our model, the exact interactions are difficult to ascertain. As such we rely on the more traditional RMSD measure, computed on the virtual residues of the coarse-grain model. When comparing predictions obtained from ERNWIN to other structure

As interactions (typically hydrogen bonds) determine many structural and functional properties of the molecule, it is fruitful to measure which are correctly and incorrectly predicted.

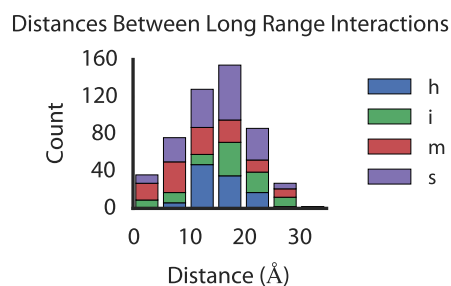


Figure 17: The distances between elements which have interactions in the crystal structure of the large ribosomal subunit (PDB ID: 1S72 [93]), where h = hairpin, i = interior loop, m = multiloop, s = stem.

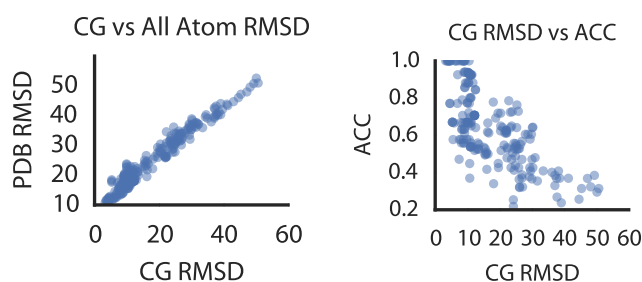


Figure 18: The strong correlation between the all-atom RMSD and the CG RMSD in the plot on the left indicates that the CG RMSD is a suitable proxy for the all-atom RMSD when comparing CG models. The plot on the right shows that the ACC is less correlated to the RMSD measure.

prediction programs, we thread a coarse-grain model onto their representation (be it an all-atom or a different coarse-graining) and calculate the RMSD between the virtual residues of the CG representations. The large correlation between the RMSD between the all-atom models and CG models (see Figure 18) indicates that this is indeed a valid measure and a suitable proxy for the more traditional all-atom RMSD.

9.2 Interaction network fidelity

The INF abandons the notion of using atomic positions to describe an RNA molecule in favor of the interactions that occur in the molecule. Each RNA can be described as a set of canonical Watson-Crick and noncanonical base pair interactions. These interactions can be annotated using programs such as MC-ANNOTATE [56] or RNAView [194]. Given annotations for the native (reference) and predicted (model) structures

we can estimate the true positives (present in both sets), false positives (present in the model, but not the reference), true negatives (absent in both reference and model) and false negatives (absent in the model, but present in the reference) and calculate the [MCC](#):

$$\text{MCC} = \sqrt{\text{PPV} \times \text{STY}} \quad (3)$$

$$\text{PPV} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|} \quad (4)$$

$$\text{STY} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|} \quad (5)$$

The interaction network fidelity between two structures A and B, is calculated as $\text{INF}(A, B) = \text{MCC}(A, B)$ [132]. While this measure is reasonable for all atom models where the emphasis is on modelling minute details correctly, it is intractable for our model where very precise atomic locations are not known and thus noncanonical (or even canonical) pairing can not be assigned. In lieu of the [INF](#), however, the [RPF](#) provides a way to quantify model similarity based on the number of potential contacts that are made between non-connected secondary structure elements.

9.3 The [RPF](#) measure

The [RPF](#) measure quantifies how many elements are placed at within a certain distance of each other [76, 77]. It defines two elements as being correctly positioned if the distance between them is less than some threshold δ . True positives are instances where two elements which are close together in the native structure are predicted to be close together in the model, true negatives when two elements which are far away are predicted to be far away, false positives when two elements are positioned close when they should be far and false negatives when two elements are far away when they should be close. These values can then be used to calculate the recall and precision. The recall and precision, in turn, can be used to calculate an F-score indicating the overall accuracy of the model.

Using the F-score alone, however, provides no measure of its discriminatory power. It is possible that a model has a high F-score regardless of the prediction method used. A small model, for example, where all residues are within the distance threshold will always have a high F-measure. This doesn't say anything about the quality of the model. To mitigate this, the CASP experiments use the discriminatory power (DP) score, which compares the F-measure of the predicted model to one generated using a freely rotating peptide chain [76, 77]. When reporting the [RPF](#) score in CASP experiments, the evaluators are actually referring to the DP score calculated from the recall, precision and F-measure ([RPF](#)).

9.4 The ACC

Our model's coarse-grain representation precludes the use of the INF as a measure of model quality. We simultaneously assume a defined secondary structure for the helical regions and neglect the exact conformations of other regions. This makes it impossible to count the number of noncanonical interactions our model might predict. We can, however, borrow from the RPF measure and calculate which elements are within a certain distance of each. This provides us with a set of *adjacencies*, that indicate which elements are near to each other. By calculating if two elements are correctly predicted to be adjacent (i.e. their closest points are less than $\delta \text{Å}$ apart) we can obtain true positive, false positive, true negative and false negative values from which to calculate the MCC. We will call our measure, as calculated from this confusion matrix, the ACC and use it to measure the similarity between structures.

Which elements, then, should we consider to be adjacent? How do we pick a value for δ which excludes adjacencies arising from local structure, while retaining those from long range interactions? To begin, we examined the distances between elements that are known to interact through at least one hydrogen bond (as annotated by MC-ANNOTATE). These elements were extremely rarely at a distance greater than 30Å apart, measured between the nearest points on each coarse grain element (Figure 17). Of those, only 16 out of 194 were further than 25Å apart. At the same time, the number of pairs greater than some distance, d , from each other begins to rapidly increase at $d = 25 \text{Å}$ (Figure 19, left). Around this point, the posterior probability that two elements interact when they are a distance d from each other decreases to nearly 0 (Figure 19, right green). With this in mind, we feel justified in calling elements that are within 25Å of each other *adjacent* and using them to calculate the similarity between two models using the ACC.

Using all pairs of elements that are within 25Å of each other as adjacencies is problematic, however, because it incorporates many elements that are necessarily close to each other by virtue of the secondary structure they share. We would like to exclude these elements in order to increase the discriminatory power of our measure. How far away, then, should two nucleotides be in the secondary structure graph (nucleotides as nodes, backbone and base pair bonds as links) in order to be counted in the ACC? Two nucleotides separated by a base pair or a backbone will always be within a certain short distance of each other. How far will they be if separated by two backbone or base pair connections? What about ten? Ideally, we want to pick a secondary structure distance beyond which the fraction of elements within a distance δ of each other, out of all pairwise distances, drops to a reasonably low value.

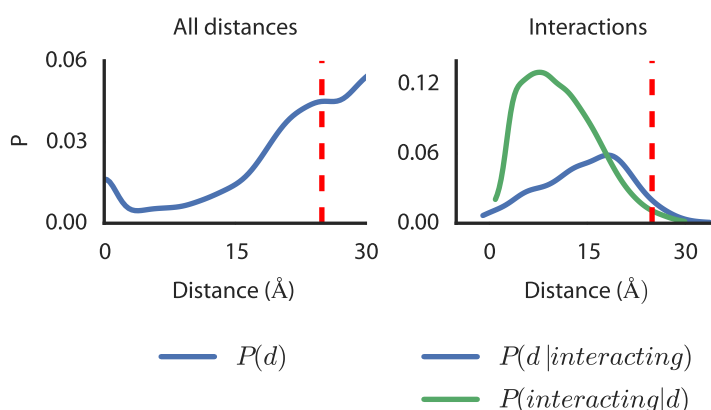


Figure 19: Kernel density estimates of the probability of seeing two elements at a given distance from each other (left), the probability of seeing two elements at a given distance from each when they are known to interact (right, blue) and finally the probability that two elements are interacting when they are a given distance from each other (right, green), as calculated by Bayes' formula. The vertical red dashed lines indicate the distance (25Å) we chose as a cutoff for consider elements as being *adjacent*.

We can also look at the relationship between graph distance (how many backbone or base pair bonds need to be crossed in order to get from one nucleotide to another) and the number of element pairs within 25Å of each other and see that the slope sharply decreases after a distance of about 10 links (Figure 20). This is consistent with A-form helices which have a rise of between 2.3Å and 2.8Å per base pair for DNA and RNA, respectively [182, 173] (See Figure 21). Figure 20 shows that the fraction of all pairs that are less than 16 links apart is around 0.2 and that the total number of pairs of elements which are closer than 25Å begins to plateau at around 10 links. We thus chose 16 links as the cutoff below which we do not consider adjacencies. This will lead to less discrimination among smaller structures but hopefully less noise among larger structures.

The graph distance is calculated as the number of backbones or base pairs that need to be traversed in order to reach one nucleotide from another. Two adjacent nucleotides thus have a distance of one. Two paired nucleotides also have a distance of one.

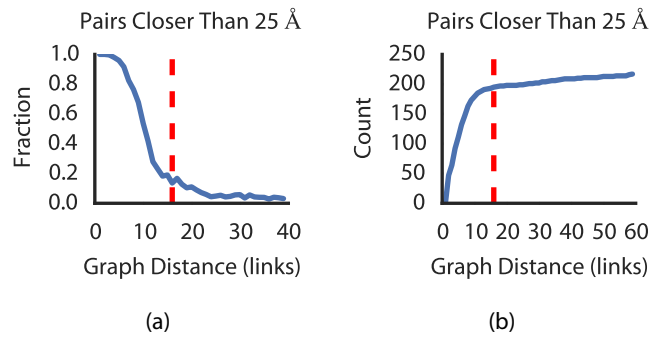


Figure 20: The distribution of pairwise distances between coarse-grain elements as a function of how far apart they are in the secondary structure. Plot (a) shows the fraction of all distances which are less than 25Å while (b) shows the total number closer than 25Å. The vertical dashed red line marks a secondary structure distance of 16 basepair or backbone links (that separate two elements), which we use for calculating ACC values.

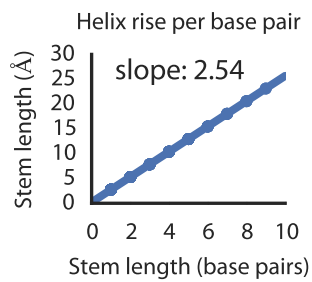


Figure 21: Data from the ribosome crystal structure confirms a rise of 2.54 nucleotides per base pair in helices of varying lengths.

10.1 *How to sample from a probability distribution*

Our goal in predicting ensembles of RNA structures is to draw a sample of structures that represents the true distribution present in the cell. What does this distribution look like and how do we sample from it? These two questions are inextricably linked and the answer to the former greatly influences how we must address the latter. This is true not only because knowledge of the properties of a distribution is necessary in order to sample from it, but because our uncertainty of the nature of RNA structures in solution can yield more than a single probability distribution describing their characteristics.

In our case, we have at least two different distributions that we wish to sample from: the local structure (X), which is simply a tabulation of fragments of known structures and the global structure which can be described by terms such as the Radius of Gyration (ROG). We can ask the more concrete question of how to sample structures which have a local structure conforming to the local structure model $Q(X)$ and at the same time have the properties defined by the global model $P(X)$? The answer lies in the reference ratio method and adaptive rejection sampling [65, 179, 17].

For the remainder of this section, the examples will use the Metropolis-Hastings algorithm for sampling. This algorithm can be used to sample from some distribution $p(x)$ when pulling samples from another distribution $q(x)$ by evaluating an energy equivalent to $E = -\log\left(\frac{p(x)}{q(x)}\right)$. The Metropolis-Hastings sampling algorithm operates on a very simple principle: if the energy of the current value is lower than the energy of the previous value, accept it. Otherwise, accept it with a probability proportional to the difference in energy between the current and previous values:

This yields samples from the distribution $p(x)$ rather than the distribution $q(x)$, which is useful, because it allows us to sample from distributions which don't allow direct sampling. This works fine when we know $q(x)$. In reality, however, $q(x)$ may be unknown. If we are trying to sample structures with a particular ROG, what do we use as the denominator in our energy formulation?

To answer this question, let's consider a simpler example. Suppose that $q(x)$ yields arrays (e.g. of length 5) of numbers drawn from $N(0, 1)$ (where $N(\mu, \sigma)$ is the normal distribution with a mean, μ , and a standard deviation, σ). This will be the *local* distribution. This distribution implies

Listing 1: Metropolis-Hastings algorithm

```

prev_value := sample_from_q()
while true
  curr_value = sample_from_q()
  if energy(curr_value) <= energy(prev_value):
    yield curr_value
    prev_value = curr_value
  else
    if random() <= exp(energy(prev_value) -
                      energy(curr_value)):
      yield curr_value
      prev_value = curr_value
    else:
      yield prev_value
end;

```

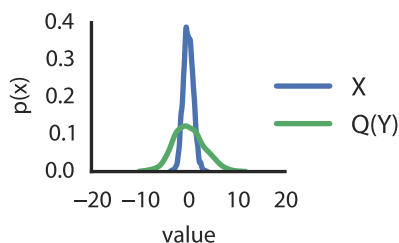


Figure 22: The distribution of individual numbers sampled in each array (the *local* distribution, X) along with the distribution of their sums ($Q(Y)$).

a *global* distribution, $Q(Y)$ over the sums of the values in the arrays (Figure 22):

But what if we actually want to sample arrays which have sums distributed according to $N(2, 3)$ (this being our *target distribution*, $P(Y)$)? How do we accomplish this? The simplest way is to run a sampling with no energy, which we can use to estimate $Q(Y)$, fit a probability distribution to it, and use that to formulate an energy function where $E(X) = -\log(P(Y)/Q(Y))$. This works well when the target $P(Y)$ has a narrower distribution (i.e. lower standard deviation) than the reference ($Q(Y)$) (Figure 23 a), but can lead to artifacts when it has a broader distribution (i.e. greater standard deviation) than the reference (See Figure 23 b).

While Hamelryck et al [65] mention the potential (no pun intended) of using this method iteratively to get nearer and nearer to the target distribution, we take another tack and continuously update an estimate, $Q_e(Y)$, of the background distribution $Q(Y)$ every n iteration steps and

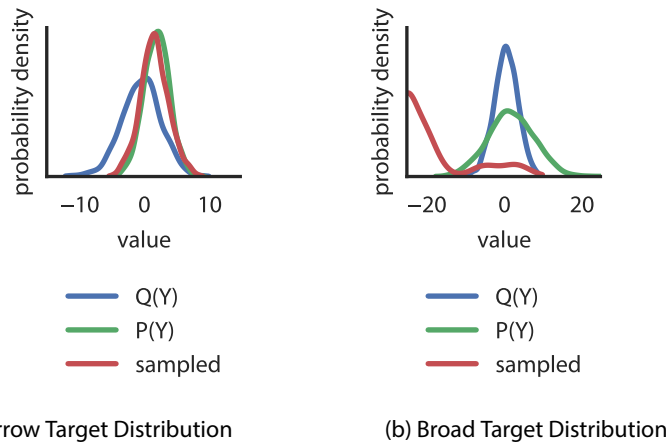


Figure 23: Using the reference ratio method to sample from a coarse-grain variable Y with a distribution of $P(Y)$. A narrow $P(Y)$ leads to excellent sampling using the estimation of $Q(Y)$ from the initial, energy-free sampling (a), whereas a broad $P(Y)$, leads to an undesirable bi-modal sampling of values that does not conform to $P(Y)$.

accept or reject new values with a probability equal to $P_{\text{accept}} = k \frac{P(Y)}{Q_e(Y)}$. This process, known as adaptive rejection sampling [17], allows us to sample from a complex probability distribution of coarse-grain values while drawing values from a different underlying proposal distribution. As can be seen in Figure 24, with this technique we can sample values from our target distribution even though it is wider than the underlying proposal distribution.

10.2 Distribution approximation using kernel density estimates

The previous sections make ample use of the term *probability distribution*. What does this term actually mean in our context? How do we estimate a probability distribution from a set of data? Perhaps it's best to start with an example. Imagine a room containing seven people: three are about 174 cm (or 173.8 cm, 174.2 cm and 174.3 cm, to be precise) two are 180 cm (179.9 cm and 180.0 cm) and one is 160 cm exactly. What are the chances that we go outside and see a person who is 170 cm tall? To answer this question, we have to make a few assumptions:

- The people in the room are representative of the people in the population outside of the room.
- We don't know anything about how heights are distributed, except what we can infer from the people in the room.

The values that one would expect to see when rolling dice are uniformly distributed between 1 and 6.

The word distributed here refers to how often we would expect to see data points with a particular value.

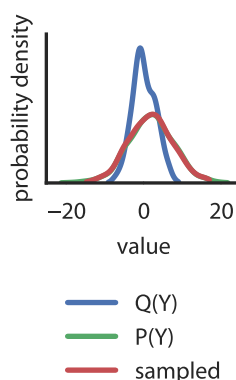


Figure 24: Using adaptive rejection to sample from a probability distribution significantly wider than the the proposal distribution. The edges of the sampled curve are cut off due to the lack of proposed samples in that region.

- When we ask “*What are the chances that we see someone who is 170cm tall?*”, we are really asking “*What are the chances of seeing someone who is greater than 169.5 cm tall and less than 170.5 cm tall?*”

The number of heads one would expect to see after flipping a coin 100 times is binomially distributed.

The heights of people are generally normally distributed (more average sized people, less extremely tall or extremely short people).

Without having seen anybody outside of our little room, assumption 1 is reasonable enough. Of course, it is possible that we already have some idea of what humans look like, such as “*humans are always less than 300 cm tall*”. This would be a *prior* assumption and we’ll assume for the sake of this explanation that we lack it. Although we’ll come to this point later when describing potential uses of our method given experimental data about an RNA molecule, for the moment we assume we have no other information apart from the 7 people in the room.

Assumption 2 is vague, but important. If we knew how the heights were distributed, then we could fit that distribution to our data points, calculate a probability and be done with this example. Because we don’t know how they are distributed, we have to find a distribution as well as to estimate parameters for it. The data points could be uniformly distributed from the smallest value (160 cm) to the largest (180 cm). They could be normally distributed with a given mean and standard deviation or they can follow any number of other distributions. As we don’t know, however, we need to start somewhere else. A good first step would be to examine our data by looking at a histogram of the heights.

Given our small sample size, this reveals little. If we used the probabilities in the histogram, we could assume that the chances of finding a person who is 170cm tall is 0, and this is highly unlikely. What we need to do is to *smooth* the probability distribution using a *kernel*. To do this we generate a probability density curve, that is equal to a sum of *kernel*

functions (normal distributions, in our case) centered at each point in our data set. These distributions are then combined and normalized to yield a continuous probability distribution over the entire range of (known) data, in a process known as kernel-density estimation [158].

But how wide should we make the distributions centered on each point? This is a topic of broad contention. For the sake of brevity, we'll simply note that we use a heuristic called Scott's rule [158] to select a bandwidth (distribution width and normalization factor) which would minimize the error of the kernel density estimate if the underlying data were normally distributed. The next sections will begin to describe how we use such kernel density estimates (KDEs) to create energy functions for use with our coarse-grain model.

10.3 *Target distribution describing global RNA structure*

The previous two sections (10.1 and 10.2) introduced the concept of sampling and probability distribution estimation using KDEs. These two methods provide us with powerful tools for sampling not only the tertiary structure of RNA, but for any other macromolecule as well. The two major necessities, a proposal distribution and a target distribution generate arbitrary structures and evaluate how well they conform to our expectation of the global structure, respectively. The remainder of this section will explain the concept of a target distribution, and how we go about estimating it.

To root the analogy from the previous section in reality, consider the simple measure of the radius of gyration of a molecule. This measure indicates roughly how wide a molecule is. Due to intramolecular interactions, macromolecules tend to be more compact than expected by chance. Previous prediction tools [169, 82, 35] seek to simply minimize the radius of gyration in predicted structures. This, especially in the case of RNA, which is not as compact as proteins, becomes counter-productive after a certain point. RNA structures which are too compact are actually less likely than moderately compact ones (Chapter 6, Figure A.3).

So how do we determine what the ideal radius of gyration is? We don't, after all, have access to all RNAs in solution. What we have is the equivalent of our proverbial room of people from the previous example. This room is populated by all of the structures that have been solved. To avoid over-fitting and to reserve ourselves an ample quantity of the test structures for evaluating our prediction method, we only use the large ribosomal subunit of the *H. marismortui* ribosome (PDB ID: 1S72 [93]), to extract statistics about global RNA structure. This 2922 nucleotide structure has a plethora of structural which we can use to extract statistics about how an RNA molecule should look. The 23s ribosomal subunit does have one disadvantage when used as the sole source of information about

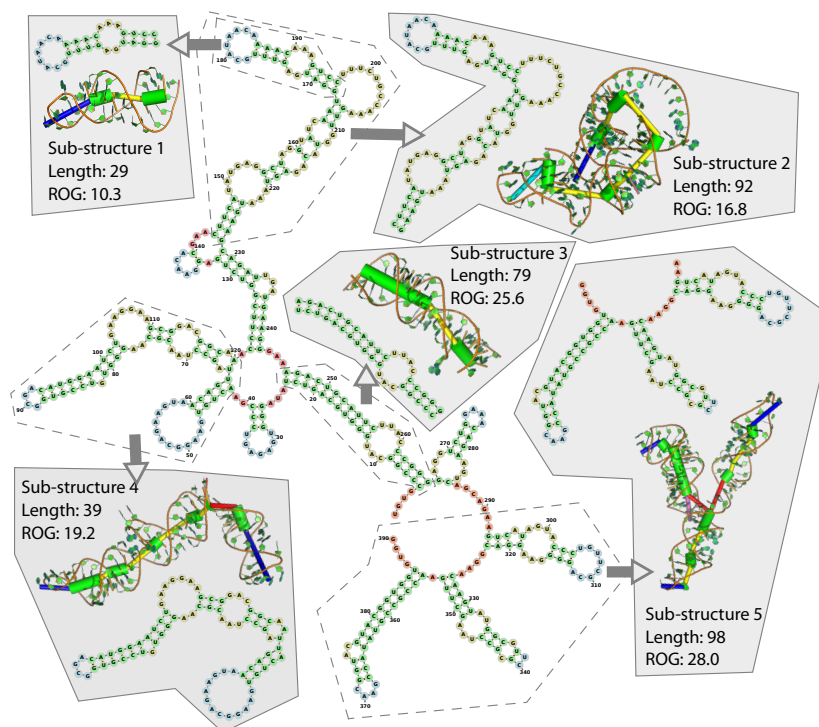


Figure 25: Creating five structures out of one by taking substructures of a solved 3D atomic model. The structure of the Group II Intron (PDB ID: 4FAW [111, 14]), has been randomly divided into five substructures containing linked secondary structure elements. The corresponding nucleotides have been taken from the PDB file and used to calculate the ROG of that fragment.

the distribution of radii of gyration among RNA molecules. Namely, it is just one molecule, with one radius of gyration, and thus should only count as one data point. This is hardly adequate for fitting a probability distribution.

Fortunately, given the size of the ribosome, we can employ some trickery to squeeze extra data out of it. The ribosome itself has a complex secondary structure [139]. If we take substructures containing continuous regions of secondary structure, we can decompose it into a multitude of smaller structures from which to extract statistics about the distribution of radii of gyration. Figure 25 shows an example of a large structure which is partitioned into substructures with unique statistics. This list of ROG_s becomes our list of person heights, to which we fit a KDE and obtain a probability distribution. This is called our *target distribution*, because it is our best guess as to how the ROG_s of RNAs are distributed. When we predict the structure of any similarly sized (in nt) RNA, we will try and sample structures with ROG_s conforming to this distribution.

Because we are predicting structures for a sequence with a given length, it is necessary to condition our target distribution on the length of the input sequence. Thus, if we are predicting the structure for a sequence of length 76, we would try and obtain a target distribution of ROG values that would be expected of a structure of length 76. We do this by taking a subset of our substructures which have a sequence length greater than l and less than u such that there are at least 100 substructures within this range. These values are chosen such that we have a reasonable number of substructures from which to build our target distribution.

A null model to compare it to for a structure X would be the distribution of radius of gyration that we obtain by sampling from our proposal distribution with no energy function. The results, as seen in [87] (Chapter 6, Figure A.3), clearly indicate that substructures from the ribosome are more compact than structures generated using only a local structure model.

Another quality control measure for how well our target distribution approximates the population of known RNAs is to see where the ROG values of the crystal structures fall on the curve. Figure A.6 in [87] (Chapter 6) shows, among other things, the target distribution for the ROG for all of the structures in our test set (Appendix). As seen in the figure, the native values tend to fall just short of the peak of the target distribution. While this is not ideal, it's expected insofar as our method of selecting substructures from the ribosome selects them randomly based on their secondary structure. It does not take into account long-range interactions. A potential improvement could omit substructures involved in a long-range interaction with an element that is not part of that selected secondary structure.

The same approach can and is used to build up target distributions for two other coarse-grain measures: the loop-loop interaction distance, and the expected number of A-minor interactions. How these are calculated is described in the methods of [87], which is added as an addendum to this thesis in Chapter 6.

10.4 Sampling local RNA structure

The local structure of an RNA molecule is defined by the torsion angles of its individual nucleotides. In our coarse-grain representation, torsion angles are replaced by inter-helical orientation parameters. Building a model using such parameters taken from solved structures should, naturally, yield models with native-like local structure. The global structure, as described by the distances between distant elements, however, will not necessarily be consistent with what we know about real structures. As an analogy, we can imagine building a railroad. If all of the curves are laid out such that they resemble real curves on a local scale (i.e. the correct

Future work could establish a more statistically sound method of choosing how many substructures are necessary to provide an adequate estimate of the parameters of a loop.

turn radius, the correct inclination, etc...), but the direction and extent of all of the curves along a certain path are not coordinated, the railroad will not reach its destination. Nevertheless, when building such a railroad, having suitable curves is essential to ensuring its functionality in conveying trains. Similarly, with RNA structures, the importance of the local conformation is just as important as the global fold. Even if the correct long-range contacts are made within the structure, if the local sections are unrealistic, the predicted structure will be unrealistic. To ensure that the local structural elements are realistic, they are sampled from the corpus of known structures and from predicted all-atom models. Such an approach has its roots in the traditional fragment-assembly methods of structure prediction commonly used for proteins [169].

By sampling realistic parameters for each of the local structure elements, we are creating conformations containing realistic local structure throughout the entire model. Creating structures in such a manner is effectively sampling directly from the distribution of local structures found in RNA molecules. No structure need be rejected due to its unfavorable local conformation. Moreover, since the statistics are sampled from a compilation created from a non-redundant set of structures, more common conformations will be found more often than less common conformations, mirroring the distribution of conformations found in real structures. There is one large exception in the case of multiloops. As each section of a multiloop is sampled independently and multiloops are cyclical so one of the segments necessarily remains unsampled. We attempt to remedy this problem by excluding sampled structures which contain sterically impossible multiloop conformations (Section 8.7). Because of this, the distribution of parameters of the closed loop segment will not necessarily match those in the proposal distribution. A more stringent approach would require the creation a continuous distribution of conformations for the closed loop and employing accept-reject sampling to ensure that its sampled conformations match the target distribution. Implementing this is left as an exercise for the future.

10.5 *The proposal distribution consists of fragments of known structure*

To get a tabulation of which parameters (as described in Section 8.2) are available for sampling, we calculated the inter-helical angles for all loops in the *H. marismortui* large ribosomal subunit (PDB ID: 1S72 [93]) and stored them in a flat file. As there are many loop sizes that are not present in the large ribosomal subunit, we also created another ~40K models of random sequences using ROSETTA [35, 36] and extracted their helix and loop parameters to use in our database of known structures. Even with the supplementary structures, there are loop sizes which are not present in our database of known parameters. In these cases, when an

unrepresented loop size is encountered (e.g. 19 unpaired nucleotides on one strand and 16 unpaired on the other, for an interior loop), we find the nearest (as measured by euclidean distance) loop sizes for which we have parameters, such that we have a total of 50 available parameters and sample from that composite set.

10.6 *Sampling of loop parameters is conditioned on the loop's size*

We select loop parameters based on the size of the loop. All of the local structures for interior loops of size (m, n) are pooled together for use in the creation of new structures. Multiloop segments are likewise categorized according to the number of unpaired nucleotides between two consecutive stems. Since each section of a multiloop is sampled independently of the others, each section length is defined by only one number corresponding to the number of unpaired nucleotides. While this approach guarantees reasonable local tertiary structure, it is insensitive to the sequences of the internal loop and multiloop segments. In cases where a well-defined sequence-dependent motif is present in the real structure, we will unnecessarily explore conformational space inconsistent with the motif. We discuss a potential solution to this problem in Section 11.3.

10.7 *Motif-based loop parameter sampling uses sequences to find matching motifs*

An improved approach takes into account sequence information and proposes fragments based on the motif that is present at that location. Ideally we would insert fragments containing a sequence identical to that present in the target element. But what happens when no such sequence is known to exist? Many motifs have a similar local structure but vary slightly in their sequence. Such an approach is liable to miss many close matches by virtue of its stringency. A better approach may be to search for sequences which are close. How then, does one define close? Which and how many mismatches are allowed? Fortunately these questions have been addressed by the creation of the RNA 3D Motif Atlas [138] which catalogs and clusters motifs present in interior loops according to how similar they are to each other. This serves as a knowledge base which can be queried using a sequence to find corresponding motifs. Searching for known fragments using the sequence should constrain the potential number of candidates and narrow the conformation space of the structure being predicted. An illustration of how much our predictions can be improved by constraining interior loops to their correct orientations is shown in Chapter 11, Table 6. Results for predictions using predicted motifs are shown in Chapter 11, Table 5.

10.8 Long-range interaction energy format

To create knowledge-based potentials based on long-range interactions, we first need a tabulation of the occurrences of these interactions in the known structures. For every such interaction, we need to know the identities of the two interacting elements, their distance and the type of interaction that they are a part of. The format is as follows:

```
element1 element2 distance seq1 seq2 interaction_type ...
```

Where `element1` and `element2` are the identifiers of the two elements (i.e. `h1` and `s1`). The distance is the distance between the two closest points of their coarse-grain representations. `seq1` and `seq2` are the sequences of the elements (without the adjacent nucleotides). The interaction type denotes the classification of the interaction (i.e. *a-minor*, *kissing-hairpin*, *tetraloop-tetraloop-receptor*, etc).

10.9 A-minor motif energy

While there are a variety of long-range interactions that are possible, the majority fall into the category of A-minor motif [127, 193]. In A-minor interactions an unpaired adenine is positioned inside or next to the minor groove of a canonical helix in a such a manner that at least one hydrogen bond is formed between the sugars of the Watson-Crick base pair and the donor adenine nucleotide [127]. In the most common type of A-minor interaction, the type I, the sugar edge of the adenine base is positioned snugly within the minor groove of the acceptor Watson-Crick base pair. This leads to an energetically favorable interaction stabilized by at least three hydrogen bonds. In cases where there is more than one bulged adenine, the A-minor interaction can be repeated sequentially to form an even more stable interaction. Such interactions form the basis for the ribose-zipper and the tetraloop-tetraloop receptor interactions [25].

To incorporate A-minor interactions into our training data set we first need an annotation of where they occur. Two approaches exist for this task: a predictive framework that uses features of the 3D structure to predict whether a triple of bases will form an A-minor interaction [167] or a general method which can search tertiary RNA structures for motifs based on the specified base pairing or the similarity to a known motif [153]. We chose the latter variant simply due to its wider applicability and our familiarity with its use. The results revealed the presence of 224 Type I, 77 Type II, 38 Type II and 24 Type o A-minor interactions in the structure of the *H. marismortui* 23S ribosomal RNA (PDB ID: 1S72 [93]). The results are stored in the usual long-range interaction listing format (see beginning of this subsection).

Type	Nucleotides Involved
o	A104 - A957 - U1009
I	A521 - G1364 - C637
II	A520 - G1363 - C63
III	A519 - U1362 - A639

Table 1: The nucleotides from the 23S large ribosomal subunit of *H. marismortui* which were used as templates for the different A-minor interactions.

Type	Hairpin	Interior Loop	Multiloop
o	0.07	0.02	0.01
I	0.28	0.20	0.11
II	0.31	0.22	0.09
III	0.13	0.05	0.02
All	0.44	0.25	0.11

Table 2: The percentage of secondary structure elements involved in A-minor interactions as found in the large ribosomal subunit (PDB ID: 1S72 [93]).

The A-minor motif tabulation is compiled by running a geometric FR3D [153] search for each of the four types of A-minor motifs on the *H. marismortui* ribosome (PDB ID: 1S72 [93]). The output of the FR3D search is converted to the coarse-grain graph nomenclature to create the tabulation of known interactions. The interactions that were used as templates for the geometric search were obtained from [193, 127] and correspond to nucleotides in the *H. marismortui* ribosome structure (PDB ID: 1FFK [10], Table 1).

44% of all hairpins, 25% of all interior loops and 11% of all multiloops are involved in A-minor interactions (Table 2). Of more interest is the proportion of interacting elements based on the type of element and how many adenines it contains (Table 3). The number of occurrences with an interaction is in parenthesis.

As expected, the percentage of elements involved in an interaction increases with the number of adenines present in the element. Hairpin loops are always the most likely to form A-minor interactions and the most represented in terms of interacting elements as a whole (albeit by a small margin). One should be aware of the fact that these statistics are extracted from the ribosome structure which contains a very large number of elements within it. This gives each element ample opportunity to find an interaction partner. In smaller molecules, the presence of a tetraloop containing an 'AAA' may mean nothing due to its steric inability

Adenines	Hairpin	Interior Loop	Multiloop
0+	0.39 (28)	0.25 (25)	0.16 (23)
1+	0.44 (28)	0.30 (24)	0.26 (23)
2+	0.53 (20)	0.46 (17)	0.40 (14)
3+	0.70 (7)	0.53 (7)	0.64 (7)

Table 3: The percentage of secondary structure elements involved in any type of A-minor interaction depending on how many adenines are in the element, as counted in the large ribosomal subunit (PDB ID: 1S72[93]). The numbers in parantheses are the absolute counts.

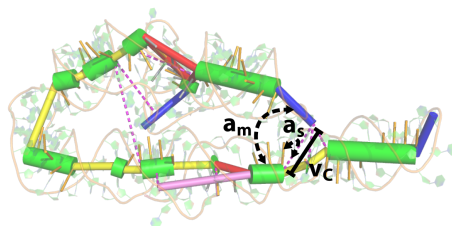


Figure 26: The parameters of an A-minor interaction.

to bend in such a manner as to form an interaction. It may also be present in order to facilitate an interaction with another molecule.

We examined the interactions within the coarse-grain structure of the ribosome. We would expect elements that are involved in A-minor interactions to be closer to each other than pairs of other elements. We would expect them to be at a certain angle to each other and to the A-minor groove. The calculation of these values is shown in Figure 26. The closest distance between the two interacting elements is d , the angle between the receptor stem and the vector between the two closest points (v_c) is α_s and the angle between v_c and the minor groove of the receptor stem is α_m .

We would expect to see different values for these parameters for elements which are involved in A-minor interactions $P(I|d, \alpha_m, \alpha_s)$ in comparison with all pairs of elements, $P(d, \alpha_m, \alpha_s)$. The tabulation of known interactions gives us an estimate of the probability of seeing a set of parameters given a known interaction, $P(d, \alpha_m, \alpha_s|I)$. Using Bayes' theorem, we can express the relationship between these terms as follows:

$$P(d, \alpha_m, \alpha_s|I) = \frac{P(I|d, \alpha_m, \alpha_s) * P(d, \alpha_m, \alpha_s)}{P(I)} \quad (6)$$

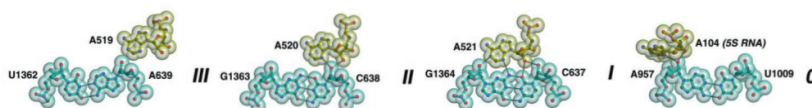


Figure 27: The four types of A-minor interactions commonly found in RNA structures. The figure is reproduced from [127].

Rearranging the terms:

$$P(I|d, \alpha_m, \alpha_s) = \frac{P(d, \alpha_m, \alpha_s|I) * P(I)}{P(d, \alpha_m, \alpha_s)} \quad (7)$$

If we want to calculate the probability that some element e is involved in an interaction, then we need to sum over all other elements which are not its neighbor. We limit all of the elements to those within 30\AA simply because no A-minor interactions have been observed for elements which are more than 30\AA apart. The probability that two elements are separated by a set of parameters, $P(d, \alpha_m, \alpha_s)$ is estimated by tabulating the probabilities for all non-connected pairs of elements which are within 30 angstroms of each other. The probability distributions for $P(d, \alpha_m, \alpha_s)$ and $P(d, \alpha_m, \alpha_s|I)$ are both estimated using a Gaussian kernel density estimate.

For each adenine-containing hairpin loop, we calculate the expected number of A-minor interactions that is involved in ($E_h(I)$), by summing over the probability that such a hairpin forms an interaction with every other stem:

$$E_h(I) = \sum_{s \in \text{stems}} P_{s,h}(I|d, \alpha_m, \alpha_s) \quad (8)$$

The frequency of the number of expected A-minor interactions forms the target distribution. More information and an illustration of the target distribution of the A-minor energy (as well as the other energies we use: the radius of gyration and loop interaction) are shown in [87] (Chapter 6).

10.10 Conclusion

This chapter introduced the idea of sampling from a probability distribution and ended with the construction of such a distribution of expected A-minor interactions. It showed how sampled local values (arrays of length five containing numbers distributed according to $N(0, 1)$, or parameters for loops in a coarse-grain RNA model) can imply a distribution over another property of these values (the sums of the length 5 arrays or

the expected number of A-minor interactions in adenine-containing hairpin loops), and introduced the use of the Metropolis-Hastings algorithm, the reference ratio method and adaptive rejection sampling as ways to approximate the distribution of the implied property (the target distribution) when sampling from a proposal distribution. These methods form the basis of the RNA structure prediction program ERNWIN, described in more detail in [87] (Chapter 6).

RESULTS

11.1 *Local structure model quality evaluation*

At the heart of the ERNWIN structure prediction package is the local structure model (Chapter 8). This describes the orientations of adjacent helices relative to each other. The first helix defines a coordinate system with its axis and twist vectors (Figure 9). In this coordinate system, the direction and twist of the second helix can be defined using three parameters and its starting position can be defined using another three as described in Section 8.2 and [9]. Hairpin loops can be defined using three parameters to indicate their offset from the end of a stem (Section 8.3). Thus our coarse-grain RNA model can be parameterized by its inter-helical orientations and terminal loop positions.

Of course with an unknown RNA, the structure is unknown and so are the parameters. We attempt to predict the structure by sampling inter-helical and loop parameters and then evaluating the energy of the resulting structure. Low-energy structures should correspond to native-like conformations whereas high-energy structures should deviate from the native fold. To sample the native structure, the parameters of the native structure need to be available for sampling from our proposal distribution. If a structure contains a kink-turn, for example, the parameters for the kink turn need to be present in the proposal distribution in order for the native structure to be constructible.

To test how well the proposal distribution is able to construct the native structure, we use an energy which is directly tied to the similarity of the model to the native structure, as measured by the [RMSD](#). This gives us an idea of what the best possible structure that can be sampled from a given proposal distribution is. Naturally, if every parameter of the native structure is in the proposal distribution and we have infinite time to sample every possible structure, then we would expect the best possible sampled structure to be equivalent to the native (i.e. [RMSD](#) of 0 (zero) Å). In practice, however, the presence of local minima in the energy landscape (even with such a *perfect* energy) makes it difficult to sample the exact native structure. Nevertheless, better proposal distributions should yield better structures.

So how accurate is the best structure we can create using just the proposal distribution? That is, given our model of the local structure and sampling procedure, what is the best global structure that we can construct? To answer this question, we will create an energy that only looks

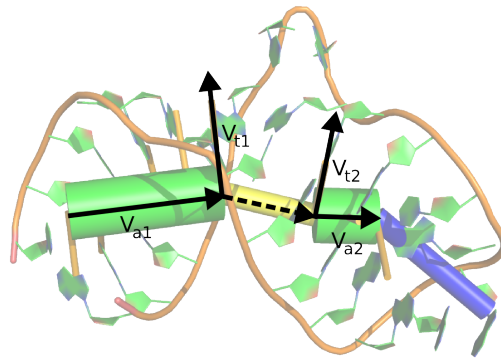


Figure 28: How the helices are defined. The vectors V_{a1} and V_{a2} trace the axes of the two stems. The vectors V_{t1} and V_{t2} denote the end and start *twist* vectors of helices 1 and 2, respectively.

at the deviation from the native **RMSD**. If the **RMSD** value of the best sampled structure is very small, then we can assume that we have adequate coverage of the local structure space and that our efforts need to focus on the global structure. If the generated structures deviate from the real structure, then we can assume that either our local structure distribution is inadequate and lacks the parameters necessary to assemble an optimal global structure or our sampling strategy fails to find the lowest energy structure or that both the proposal distribution and the sampling procedure are inadequate.

We can test how well our sampling strategy performs by spiking the proposal distribution with parameters from the real structure. This is shown in the columns labeled *exact* and *spiked* in Table 4. The *exact* column shows the lowest **RMSD** structure obtained by sampling using a proposal distribution consisting solely of parameters from the structure being predicted. In most cases, we were able to obtain the native structure. The few exceptions where the lowest **RMSD** is greater than 0 occur when two loop parameters are interchangeable within a structure. If a structure has, for example, three interior loops with two unpaired nucleotides on each side, each with different parameters, then there are $3! = 6$ ways to assign their parameters. Our sampling procedure obviously can not assign them perfectly even in the presence of a perfect energy function.

At the other end of the spectrum, we can try and assemble structures using a proposal distribution consisting of parameters obtained from artificial models constructed using ROSETTA (Table 4, column *artificial*). In this case, it's unlikely that the correct parameters are present in the proposal distribution. Nevertheless, we would hope that some of the artificially generated structures resemble real structures closely enough

that we can use their parameters to build native-like models. This works fairly well for smaller structures, but becomes less reliable for larger ones. To test whether this was a sampling problem or simply a matter of lacking the right parameters, we spiked the collection of artificially generated parameters with parameters from real structures (including the ones being predicted).

The results (Table 4, column *spiked*) indicate that a lack of suitable parameters is only part of the problem. For smaller structures, we can retrieve models nearly identical to the native. For larger structures, we can, in most cases, generate models that are better than those created using the artificial proposal distribution. Failing to retrieve the native structure when the necessary parameters are present can be symptomatic of two problems: we fall into some low energy basin and have trouble escaping or the proposal distribution is so biased toward incorrect parameters that sampling using the Metropolis Hastings criterion without simulated annealing leads to a stationary distribution that never samples the native structure.

We can observe evidence that the proposal distribution contains an excess of unsuitable parameters by looking at the distribution of inter-helical angles for parameters from real structures as compared to those from artificially generated structures. Section 10.5 describes how artificially generated structures have many more interior loop parameters defining parallel helical orientations rather than antiparallel, as compared to real structures. How much of an effect this has on our sampling procedure can be clarified by employing different sampling methods. Simulated annealing should force structures toward the global minimum free energy, while replica exchange Monte Carlo can help overcome local minimum energy barriers. In both cases, a longer simulation time should allow for a more thorough exploration of the conformational space. In Section 11.3, we show that using predicted motifs can help improve overall tertiary structure predictions. Constraining some interior loop parameters using predicted motifs can further constrain the conformational space and speed up the process of finding native-like conformations.

11.2 Comparison to the ROSETTA inter-helical statistics

As described in Section 10.5, our catalog of inter-helical angles is augmented by data extracted from predicted all-atom structures generated by ROSETTA. To see how well the inter-helical angles sampled by ROSETTA correspond to those in native crystal structures, we plotted circular histograms of the orientation of the second stem in relation to the first in Figure 29. The results indicate a sizable discrepancy between the inter-helical angles in artificially generated structures and those seen in native structures for certain loop sizes. When used as a proposal distribution,

PDB ID	length	exact	spiked	artificial
3FO4	63	0.00	0.39	2.28
3R4F	66	0.00	0.36	1.56
4PQV	68	0.00	0.69	1.76
1YFG	69	0.00	0.53	1.02
1FIR	69	0.00	1.16	2.42
1KXK	70	0.00	0.88	2.16
1Y26	71	0.00	0.03	2.62
2TRA	73	0.00	0.22	2.16
3CW5	75	0.00	0.06	2.39
2HOJ	78	0.00	0.94	2.12
4P5J	83	0.00	0.4	3.03
3T4B	83	0.00	2.05	1.85
4FRG	84	0.00	1.6	4.38
4LVZ	89	0.00	0.79	2.33
3GX5	94	0.00	1.06	5.81
4L81	96	0.00	1.21	5.24
2QBZ	153	0.00	4.04	7.07
1U9S	155	0.08	3.3	11.79
1GID	158	0.00	3.25	5.72
4GXY	161	0.65	5.55	7.57
3DoU	161	0.00	3.63	8.25
3DIR	172	0.00	2.6	8.08
4P8Z	188	0.01	6.64	7.25
4P9R	189	0.00	6.26	6.64
4GMA	192	0.04	5.41	13.72
3DHS	215	0.19	5.17	9.72
1X8W	247	0.02	6.52	11.13
2A64	298	0.25	9.68	11.01
Average:	125	0.04	2.66	5.4

Table 4: Sampling using a *cheating* energy designed to push structures toward the conformation closest to the native. The *length* column lists the length (in nucleotides) of each structure. The *exact* column contains the *RMSDs* of the best structure predicted using parameters only from the structure being predicted. The *spiked* column lists the *RMSDs* of the best structures predicted using parameters from artificially generated structures along with parameters from real crystal structures. The *artificial* column is the same, but the parameters used were only from artificially created structures.

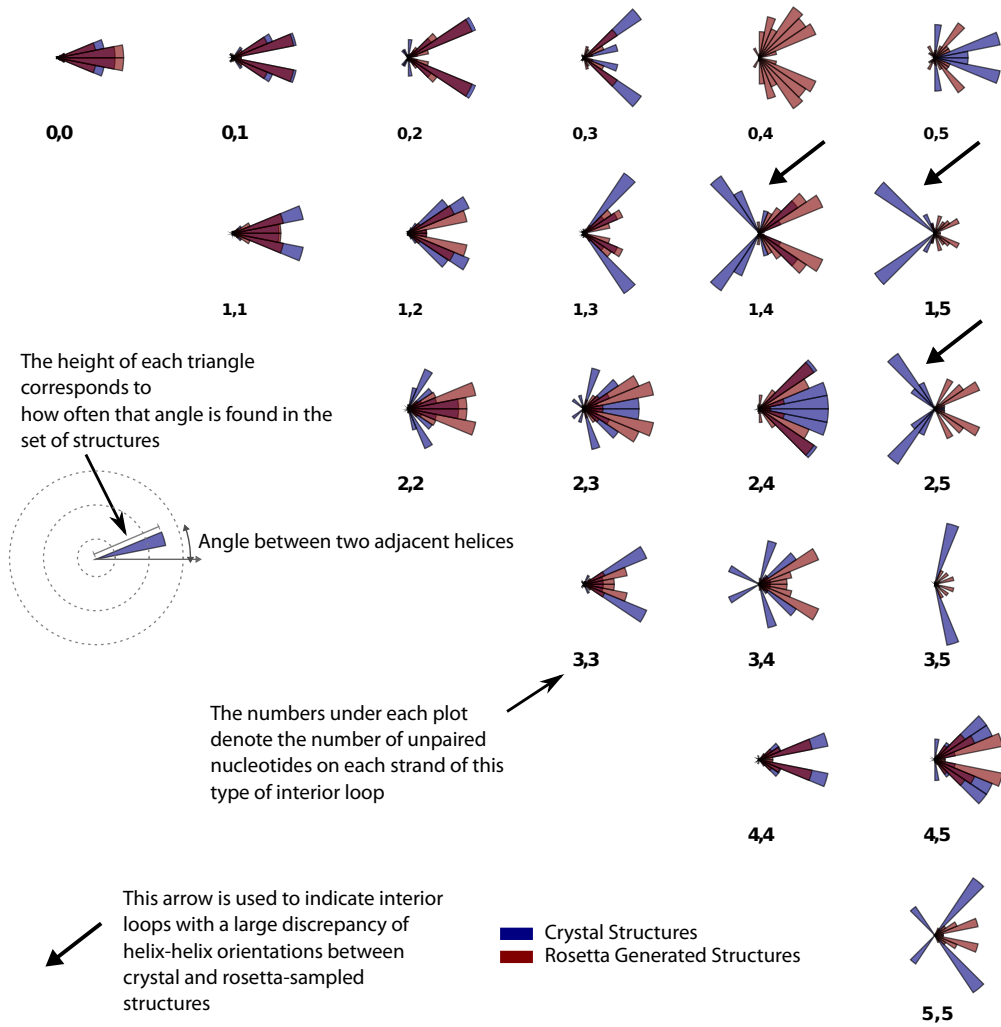


Figure 29: A comparison of the inter-helical angles for adjacent helices from native structures (blue) and structures created using Rosetta (red). The numbers below each plot indicate the number of unpaired nucleotides on each strand. In this data set, the sequences folded by Rosetta do not correspond to the sequences of the native structures so some discrepancy is expected.

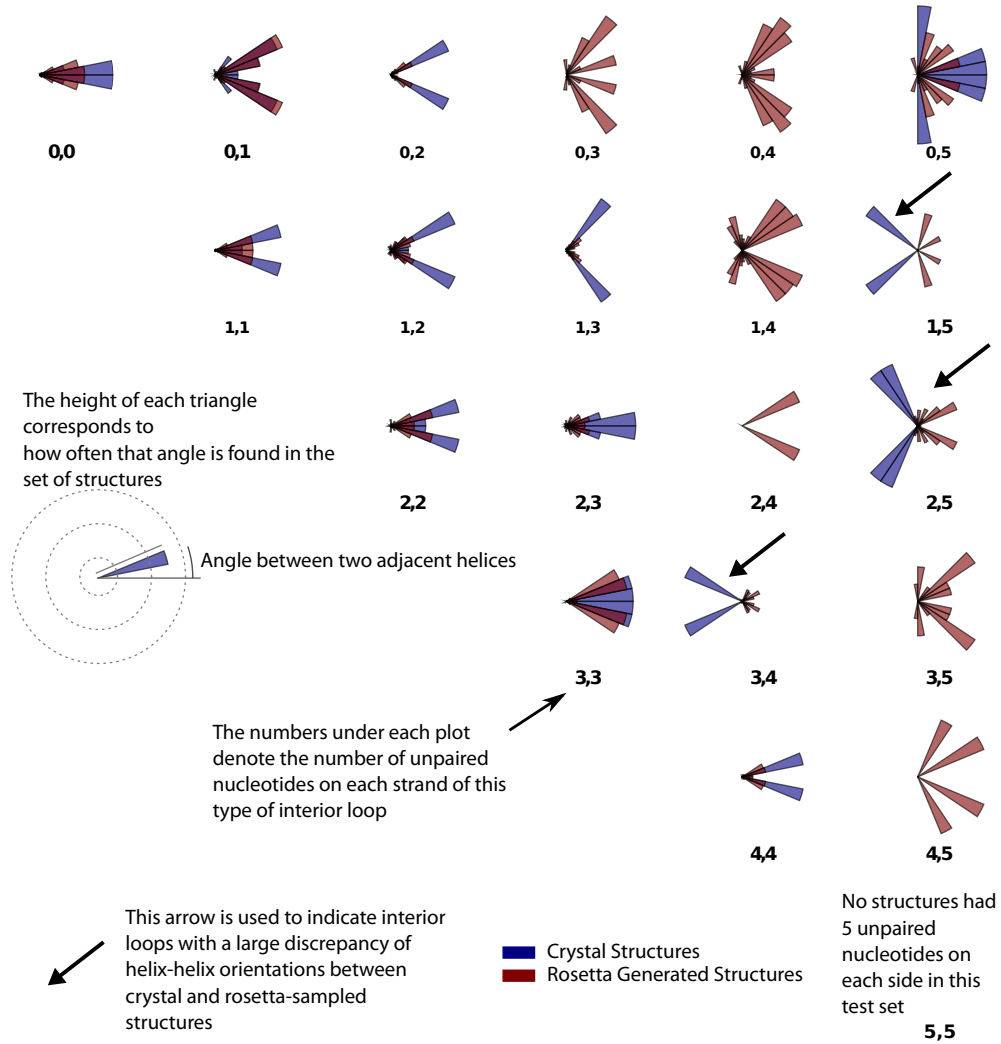


Figure 30: A comparison of the inter-helical angles for adjacent helices from native structures (blue) and structures created using Rosetta (red). The numbers below each plot indicate the number of unpaired nucleotides on each strand. The sequences of the crystal structures were used as inputs for Rosetta, so the distributions should, in the ideal case, be identical. Rosetta, however, seems to be slightly deficient at sampling antiparallel orientations.

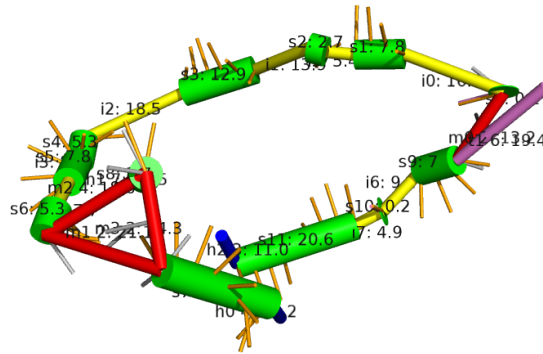


Figure 31: The best structure predicted for the Group I intron (PDB ID: 1GID [28]) using the cylinder-perloop-rog energy.

such a set of parameters would make it difficult to sample native-like structures due to the sparsity of proverbial building blocks required to create them. Fortunately, because our proposal distribution is augmented with data from real structures, it should be possible to generate native-like structures.

It is important to note that the structures built using ROSETTA were composed of random sequences. It is possible that these random sequences were not amenable to creating the antiparallel orientations which are enriched in native conformations. A better comparison would require the generation of structure with sequences identical to the native structures and comparing the distribution of those inter-helical angles as shown in Figure 30. This comparison of predictions to the native structures shows that even in such a fair comparison, ROSETTA still under-samples antiparallel orientations.

11.3 Local structure constraints from JAR3D

While our current approach uses interior loop parameters predicted without consideration of their sequence, in this section we describe using JAR3D [198] to predict motifs for each interior loop. JAR3D is a tool which uses the sequence of an interior loop to predict which 3D motifs it could be found in. For each interior loop, we obtain a set of motif predictions and for each motif prediction, a set of 3D structures in which the motif can be found. We then extract helix orientation parameters from these structures and use those as our proposal distribution for that loop in the prediction process. It is possible that JAR3D finds no matches. In this case the parameters for the loop are not constrained.

Structure Name	Structure Length	Interior Loops	ACC Original	ACC JAR3D	ACC Change	RMSD Change
1GID	158	8	0.67	0.79	-0.12	-10.07
1X8W	242	10	0.64	0.71	-0.07	-9.71
4GMA	192	8	0.68	0.71	-0.04	-3.79
3T4B	83	2	0.89	0.91	-0.02	-1.92
2QBZ	153	5	0.70	0.72	-0.01	-1.13
3DoU	161	7	0.75	0.76	-0.01	-5.52
4GXY	161	10	0.70	0.71	-0.01	0.33
4LVZ	89	2	0.82	0.83	-0.01	-1.74
2TRA	73	0	0.90	0.91	-0.01	-0.80
2HOJ	78	3	0.90	0.91	-0.01	-0.05
4L81	96	2	0.82	0.83	-0.00	-1.98
1Y26	71	1	0.98	0.97	0.00	1.00
3DHS	215	6	0.75	0.74	0.00	0.60
4P5J	83	0	0.86	0.85	0.00	0.15
3CW5	75	0	0.91	0.90	0.01	0.29
1U9S	155	8	0.80	0.79	0.01	-2.01
1KXK	70	4	0.88	0.86	0.02	0.37
4P9R	189	3	0.72	0.70	0.02	-1.16
4PQV	68	0	0.87	0.84	0.03	0.15
3DIR	172	6	0.81	0.76	0.05	0.99
3GX5	94	4	0.88	0.79	0.09	1.38

Table 5: An overview of how adding motif predictions for interior loops affects tertiary structure sampling. ACC values are calculated by considering contacts between coarse-grain elements 30Å from each other at their closest point (Chapter 9, Section 9.4). The displayed ACC corresponds to the average ACC of all sampled structures. A change in ACC does not necessarily correspond to a change in RMSD.

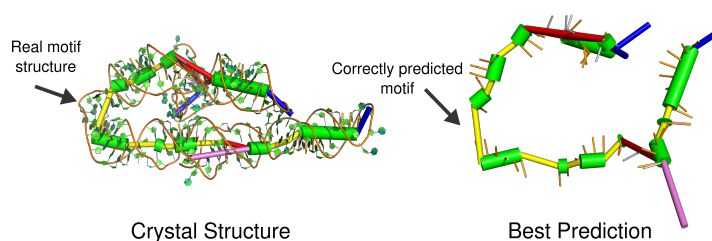


Figure 32: An example of a well-predicted motif using JAR3D. The kink-turn of the Group I Intron (PDB ID: 1GID [28]) was correctly predicted.

Structure Name	Structure Length	Interior Loops	ACC Original	ACC Correct Loops	ACC Change	RMSD Change
2QBZ	153	5	0.70	0.85	-0.15	-8.74
1GID	158	8	0.67	0.79	-0.12	-9.22
1KXK	70	4	0.88	0.98	-0.10	-7.80
1X8W	242	10	0.64	0.73	-0.09	-6.72
3DoU	161	7	0.75	0.84	-0.09	-4.74
4LVZ	89	2	0.82	0.90	-0.08	-5.42
4GXY	161	10	0.70	0.78	-0.08	-2.37
4GMA	192	8	0.68	0.75	-0.07	-5.65
3T4B	83	2	0.89	0.96	-0.07	-4.67
3FO4	63	2	0.93	1.00	-0.06	-0.06
3DIR	172	6	0.81	0.87	-0.06	-4.14
4L81	96	2	0.82	0.87	-0.04	-2.15
3DHS	215	6	0.75	0.78	-0.04	-0.90
2HOJ	78	3	0.90	0.93	-0.02	-0.21
1Y26	71	1	0.98	0.99	-0.02	3.83
4P8Z	188	3	0.73	0.73	-0.01	-3.21
1YFG	69	1	0.90	0.91	-0.01	-0.07
3R4F	66	1	0.80	0.80	-0.01	-1.56
3GX5	94	4	0.88	0.88	-0.00	-0.20
4P9R	189	3	0.72	0.72	-0.00	-0.54
4FRG	84	2	0.88	0.87	0.01	-0.59
2A64	298	6	0.73	0.71	0.02	-2.57
1FIR	69	1	0.93	0.90	0.03	1.05
1U9S	155	8	0.80	0.77	0.03	-2.67

Table 6: An overview of how fixing the parameters of the interior loops to their correct values affects tertiary structure sampling. ACC values are calculated by considering contacts between coarse-grain elements 30Å from each other at their closest point (Chapter 9, Section 9.4). A change in ACC does not necessarily correspond to a change in RMSD.

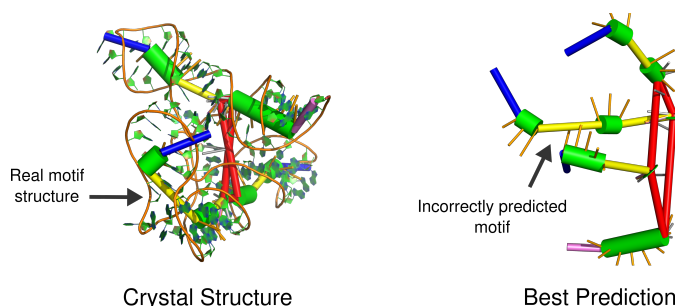


Figure 33: The JAR3D motif prediction for the *T. tencongensis* SAM-I riboswitch (PDB ID: 3GX5 [123]) yielded parameters for the interior loop which were significantly different from those of the native structure.

The results of constraining interior loop parameters to motif predictions made by JAR3D is shown in Table 5. On just over half of the tested structures, motif-constrained sampling yielded better predictions. The most improved examples, two different Group I intron structures (PDB IDs: 1GID [28] and 1X8W [62]), relied on the correct prediction of an kink-turn motif toward the middle of the structure (Figure 32). On the other end of the spectrum, the incorrect prediction of a motif in the structure of a SAM-I riboswitch (PDB ID: 3GX5 [123], Figure 33), led to a significant deterioration in the quality of the predicted structure. The damage caused by cases of incorrect motif prediction can be ameliorated by not strictly limiting the potential parameters for a loop to those of the predicted motif. Ideally, the predicted motif should simply be given a larger weight in the database of potential parameters for that loop. The influence of the global energy function should then steer the sampling toward structures containing the correct parameters rather than the poorly predicted motif.

11.4 Hypothetically perfect interior loop information

After sampling tertiary structures using interior loop parameters constrained by JAR3D predictions, one may wonder how much of an improvement is possible if the correct parameters of all of the interior loops were known. Such an experiment would show the upper bound on the improvement in prediction quality that can be obtained by augmenting tertiary structure prediction with interior loop motif prediction. The outcome of this experiment is shown in Table 6. The vast majority of predictions improved when the sampling was constrained to the correct interior loop conformations. The most improved examples, a M-Box riboswitch aptamer domain and a group I intron (PDB IDs: 2QBZ [33] and 1GID [28], Figure 34, top), have prominent interior loops that are respon-

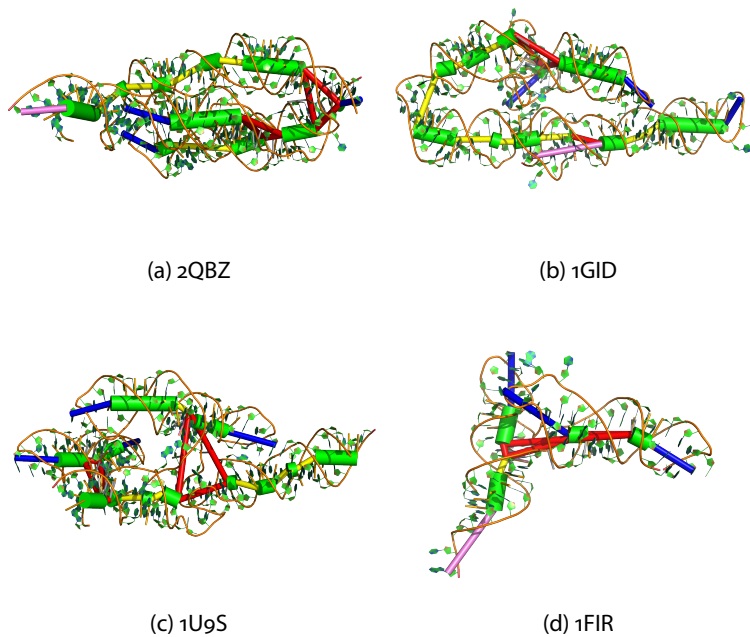


Figure 34: The most (2QBZ, 1GID) and least (1U9S, 1FIR) improved structures when using correct interior loop parameters.

sible for being acceptors to two A-minor interactions and containing a kink-turn motif, respectively. The two least improved examples, ribonuclease P and a HIV-1 reverse transcription primer tRNA (PDB IDs: 1U9S [95] and 1FIR [13], Figure 34, bottom), have smaller and fewer interior loops, respectively. In these two structures, the multiloop regions play a much larger role in the overall structure than the interior loop.

While predicting interior loop parameters may yield some improvement in the prediction of the entire tertiary structure, the results presented here suggest that in many structures this is not enough. Future work should also focus on obtaining correct parameters for multiloops as well. The work of Laing et al. [98], RNA Junctions as Graphs (RNAJAG), has shown that helical orientations in multiloops can be predicted with reasonable accuracy given the lengths of the unpaired nucleotide sequences separating the helices. Such work could be used to narrow down our set of inter-helical multiloop parameters to those consistent with its predictions. If, for example, RNAJAG, predicts two helices in a multiloop as being antiparallel, we would only sample roughly antiparallel parameters for that multiloop segment. A more sophisticated, sequence-specific approach could extend JAR3D to multiloop segments. Any such method, which constrains multiloop parameters should serve to shrink the conforma-

tional space, thus reducing the amount of resources required to explore it and return an accurate structure prediction.

11.5 *Imperfect secondary structure*

The prediction of tertiary RNA structure without first obtaining its secondary structure is like trying to tie a rope into a knot starting from its individual fibers. It is much more efficient and effective to determine the secondary structure and to use that as a stepping stone to the tertiary structure than to try and predict it using the crude and computationally heavy methods designed for tertiary structure prediction. The secondary structure is a useful prerequisite for creating a model of the tertiary structure. When benchmarking tertiary structure prediction methods, the correct secondary structure is usually supplied as input [142, 196] (along with, occasionally, information about long-range interactions [91, 82]). In theoretical terms, this is acceptable since tertiary structure prediction methods should first work with the correct secondary structure before hoping to succeed with inaccurate secondary structure as input. In practical terms, this metric is at best incomplete and at worst misleading. Often, the true secondary structure of a query sequence is not known, but rather computed using one of the available secondary structure prediction programs [105, 113].

The effect of imperfect secondary structure should be particularly onerous in programs, such as RNACOMPOSER, which rely on identifying concrete secondary structure patterns to select parameters with which to build tertiary structures. It should be less punishing in methods which are resistant to small changes in the boundaries of the sequences used for motif searches, such as ERNWIN. Nevertheless, we would expect the prediction accuracy to degrade as a function of the degree of identity between the given and real secondary structures.

When discussing imperfect secondary structure, it is important to distinguish between two different types of errors. The first, random insertions and deletions of base pairs, is easy to simulate and can be used to test the general robustness of a particular method to the accuracy of the secondary structure. The second is derived from biases in secondary structure prediction programs. This can produce large sections of correct secondary structure next to other completely incorrect sections. A trivial example is the prediction of a small additional stem within a large internal loop which turns it into a multiloop. This may involve three or four superfluous base pairs in one location and the absence of real base pairs in another. This is a more realistic scenario and likely more difficult for tertiary structure prediction programs to handle.

Yet another pitfall can arise when the structure is determined using chemical probing methods but due to their inaccuracy some small incon-

sistencies and idiosyncrasies remain in the input structure. To explore this, we can find structures which are energetically close to the real secondary structure but not exactly equivalent. This precludes the presence of, for example, open base pairs in long stems, but might lead to structures with slightly different interior or multiloop dimensions. To evaluate this possibility, we listed the local minima within the folding landscape using the BARRIERS program [49] and selected the secondary structure closest to the native using the RNADISTANCE program. These were then used as input to the tertiary structure prediction programs.

We tested ERNWIN, RNACOMPOSER and ROSETTA (FARFAR) using both methods, as well as with the real secondary structure and show the results in Figure 35. While ERNWIN and ROSETTA (which fail to yield predictions for many structures) show a modest increase in accuracy when given the correct secondary structure, the performance of RNACOMPOSER seriously degrades when presented with inaccurate secondary structure information. Because it relies on loop topologies to match parameters, when the secondary structure is different, it matches completely different parameters and yields poor predictions.

One may argue that given the overall poor performance of ROSETTA and ERNWIN for larger structures, it makes little sense to worry about incorrect secondary structure. Nevertheless, there is a slight improvement in the prediction of larger tertiary structures when calculated using the correct secondary structure. Furthermore, this example demonstrates the fragility of relying on exact loop topology (as with RNACOMPOSER) when creating tertiary structure models. An ideal prediction method should be resistant to such changes by allowing either some flexibility in the input secondary structure or allowing more variability in the choice of parameters used to assemble the all-atom model.

One may also argue that tertiary structure prediction should only be undertaken when there is high confidence in the secondary structure. Covariance analysis and probing experiments can be used to verify the input secondary structure before it is used for tertiary structure prediction. To simulate such cases, we should only consider secondary structures which are close to the native. In Figure 35, the dashed lines single out two cases where the predicted secondary structures are only 4 or 5 base pairs from the real ones. In both of these cases, the quality of the prediction generated by RNACOMPOSER is significantly worse for the predicted secondary structure. The quality of ERNWIN's prediction does not degrade, but it is much poorer than RNACOMPOSER's to begin with. These results should serve to remind users of the fragility of tertiary structure prediction. People wishing to employ it should either be extremely certain of the secondary structure they use as input, or use multiple similar inputs to gauge how much the outputs vary and thus form an opinion about how trustworthy the results are.

It is curious why the secondary structure is so poorly predicted for some of the larger molecules in Figure 35, given that they are monomeric and not protein bound.

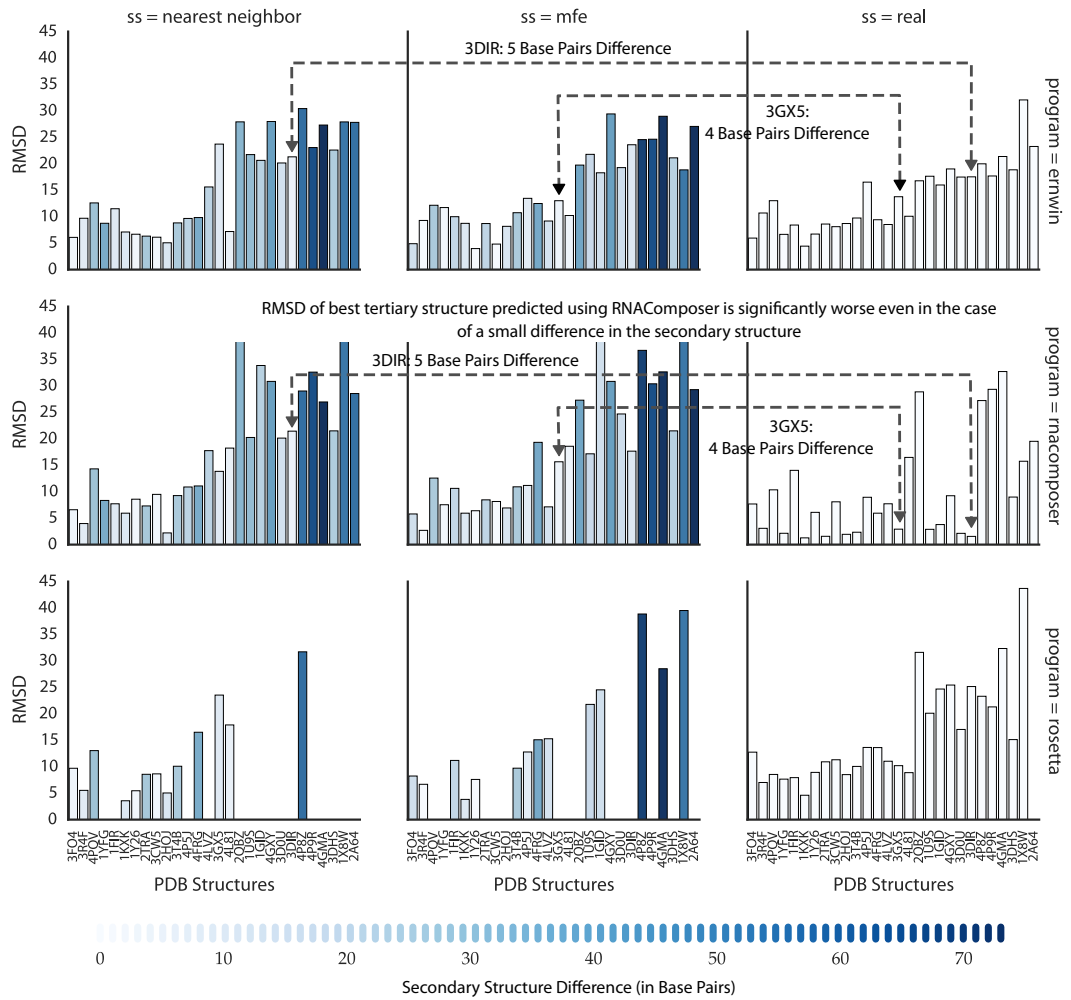
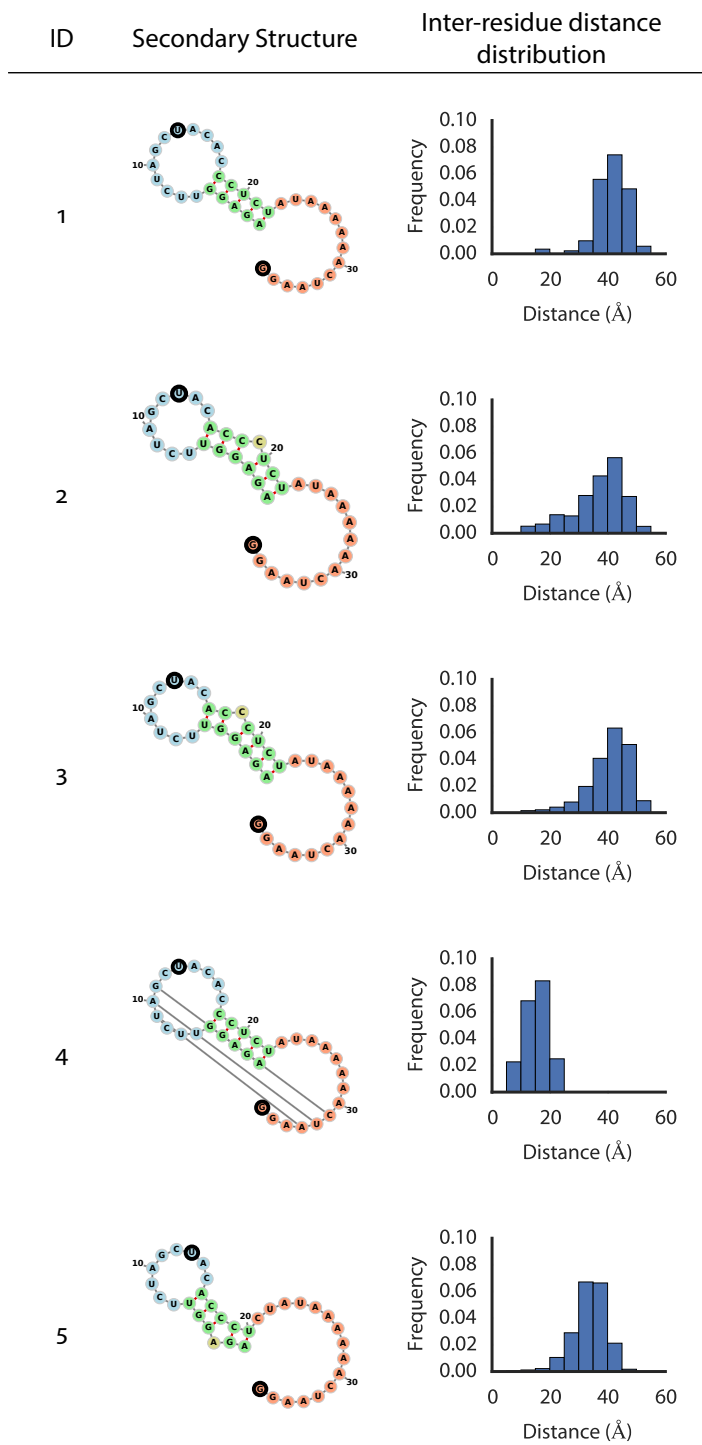


Figure 35: The RMSDs for predicted tertiary structures using predicted secondary structures as input. The colors indicate how different the predicted secondary structures are from the native. The dashed lines connect equivalent benchmark structures which have similar native and predicted secondary structures to show that even small differences in the input secondary structure can lead to large discrepancies in the predicted tertiary structure.

11.6 *Interpreting FRET data*

Continued on next page

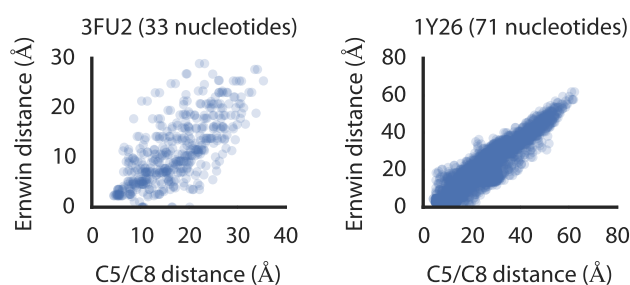


Figure 36: The correlation between C5/C8 atom distances and the Ernwin approximated distance taken from crystal structures.

Table 7 – continued from previous page

ID	Coarse-Grain Structure	Description
6		No structures generated without a stem present.
7		

Table 7: Distributions of inter-nucleotide distance as calculated by Ernwin for different secondary structures of a single sequence. The two nucleotides between which a distance is calculated are highlighted in black. The pseudoknotted structure (4) has markedly shorter inter-nucleotide distances, likely due to the constraint on the structure that the pseudoknot imposes.

Experimental techniques such as Förster resonance energy transfer (FRET) [80] measure when pairs of nucleotides are close to each other. This can help confirm structure models by providing evidence that predicted inter-nucleotide distances are correct. With FRET, the measured efficiency is inversely proportional to the distance between fluorophores covalently attached to atoms on the bases of two nucleotides in a structure.

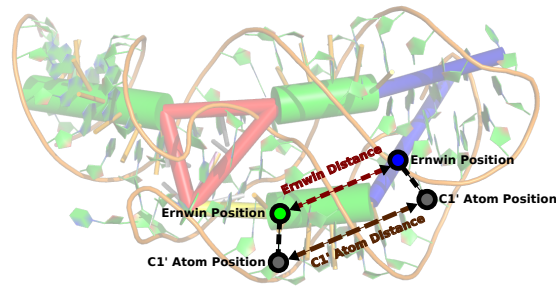


Figure 37: An illustration of how Erwin calculates inter-residue distances. Nucleotide positions are interpolated onto the coarse-grain elements and distances are calculated between the interpolated values.

Recently, Hecker et al. (submitted) created a tool called FRETTRANSLATOR which uses ERNWIN as part of a pipeline to assign secondary structures to a sequence based on experimental data from FRET experiments. A FRET experiment yields a set of energy transfer efficiencies over a period of time (100 seconds in this case). The efficiency is strongly proportional to how close two fluorophores are to each other. A high energy transfer efficiency indicates that the two fluorophores are close to each other, whereas a low efficiency indicates that they are far from each other. Each molecule has two fluorophores (each a moderately large molecule) covalently linked to either the C5 or C8 atom of a nucleotide. When we observe a high energy transfer efficiency, then the two fluorophores must be physically close to each other. This also implies that the two nucleotides that they are attached to are also physically close to each other. These experiments thus yield a clear time-resolved source of structural information. Unfortunately, this information is essentially a binary *close* vs not close assessment of the distance between the two labeled nucleotides.

Since fluorophores are not included in the structure prediction process, their position is approximated as the position of their point of attachment (the C5 or C8 atoms of the base, see Appendix v). In our coarse-grain model, however, the positions of individual atoms are unknown. We thus approximate where they might be and show that distances measured using this approximate position are well correlated with distances measured using the positions of the points of attachment (the C5 or C8 atoms of the base, Figure 36 and 37).

When there is a variation in FRET values over the course of an observation period, the underlying cause could be both variations in the tertiary as well as the secondary structure of the RNA. Hecker et al. created a hidden Markov model (HMM) to simulate such a process and to assign

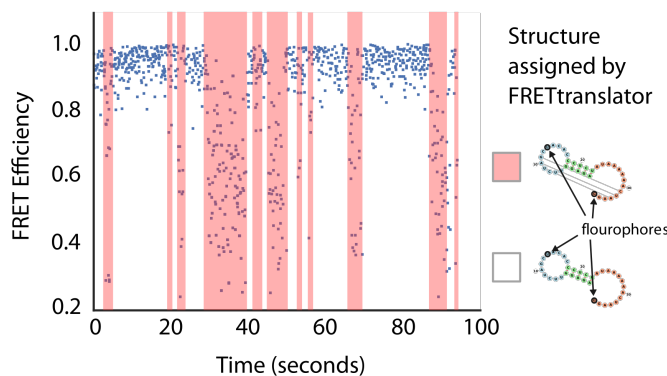


Figure 38: A summary of the structures assigned by FRETtranslator along with the measured FRET efficiencies. The pseudoknotted structure (white background) is assigned as the hidden state more often than the nested structure (light red background). The fluorophores between which the FRET efficiency is measured are attached to nucleotides 13 and 36 (highlighted in black in the legend). High efficiency is indicative of a small physical separation between the two fluorophores, as would be expected from a pseudoknotted structure. Figure adapted from [68].

secondary structures to different points in time. They began by calculating a series of secondary structures corresponding to local minimum in the 2D folding energy landscape. The potential values of the hidden nodes of the HMM correspond to these RNA secondary structures while the possible observations are FRET efficiencies. The observation probabilities for each hidden state (secondary structure), are then calculated according to the distribution of the inter-nucleotide distances sampled using ERNWIN. As seen in Table 7, the tertiary structures sampled for a pseudoknotted secondary structure (ID: 4) place nucleotides 13 and 36 (highlighted in black) closer to each other than the tertiary structures sampled for the other secondary structures. This means that a hidden node having this structure as its value will be more likely to emit high FRET values. The transition probabilities between the hidden states are calculated from the differences in free energy of its secondary structure's values.

Using Viterbi decoding [50], FRETTRANSLATOR can assign hidden states to sequences of observed FRET efficiencies [68]. The results (Figure 38), show that the sequence analyzed (a riboswitch, PDB ID: 3FU2 [92]) vacillates between two secondary structures, #1 and #4 from Table 7. The dominance of the pseudoknotted structure is consistent with the crystal structure in the PDB (ID: 3FU2). This is a promising first step toward using FRET data to analyze RNA folding dynamics and is currently being tested on larger sequences. It takes advantage of the fact that ERNWIN produces a variety of tertiary structures for a given secondary structure, hopefully making the approach more tolerant to inaccuracies in the ter-

tiary structure prediction. At the same time, it is not immune to such inaccuracies and would benefit from any improvement in the quality of the predictions produced by ERNWIN. Future work should seek to use inaccurately predicted distances to help diagnose errors in the energy function used to sample structures. In the meantime, we hope that this application provides a simple example of why all-atom models are not necessarily required to interpret experimental evidence.

SUMMARY AND FUTURE WORK

Nearly a decade since the field of RNA tertiary structure prediction began to seriously emerge, there is no clear solution to the problem. The different approaches described in the introduction along with our contribution represent many small steps along the path to a useful tertiary structure prediction method. They have combined techniques from machine translation to game theory to attempt to uncover and emulate a process ubiquitous in nature since the beginning of life. The progress has been steady, but meandering. With the exception of certain motif-finding and replacement methods [148, 142], the prediction of atomic, or even nearly atomic-level coordinates remains elusive. In the face of this stagnation, we can ask two broad questions:

1. How can we advance the state of the art to improve prediction?
2. How can we utilize the current techniques for further scientific discovery?

The remainder of this chapter will outline some specific avenues of exploration that can either help improve the current techniques or use their output to help answer concrete biological questions.

12.1 *Covariation analysis for long-range constraints*

One of the underlying themes of our work on sampling tertiary structures has been efficiently exploring conformational space. We've done this by picking parameters describing large fragments consisting of entire secondary structure elements and using knowledge-based energy functions to direct sampling toward structures containing native-like global folds. We've touched upon possible improvement by using predicted motifs (Section 11.3) to constrain the potential parameters of certain loop regions. An orthogonal approach involves the direct prediction of long-range tertiary interactions. By analyzing the coevolution of nucleotides in families of related RNA sequences, we may be able to obtain information about which are located near each other in the tertiary structure.

Such an approach, as described by [114], has been successfully used to drastically improve the quality of protein structure prediction by constraining the relative positions of residues. While this method grows in feasibility and utility with the explosion in sequences available, care must

be taken to avoid false locality predictions from transitive correlations between residues. In other words, locality predictions must not be assigned when two residues are correlated with each other only because they both correlate to a third. The direction of some RNA-based long-range interactions (i.e. A-minor, ribose zipper, etc ...) can provide hints as to which correlations result from bona fide interactions, rather than from transitive effects.

It requires little imagination to see how such an approach could be applied to RNA structures. Locality information could be encoded as energy terms which force particular elements toward each other. Using ERNWIN, we could explore the conformations compatible with these constraints as well as to combine them with other motif-based constraints to further narrow the conformational space to be searched.

12.2 Using *SHAPE* data to improve predictions

One of the greatest current limitations to structure prediction is the lack of adequate test and training sets. The corpus of known 3D RNA structures contains less than 3500 instances. Many of those are point-mutation variations of the same structure. Many of the rest are just small fragments that don't contain any information about long-range interactions or large scale structural patterns. To supplement these structures, we can use data from *SHAPE* experiments [117] which provide information about which nucleotides are flexible and which are immobile. Flexible nucleotides tend to be located in unpaired regions, whereas fixed nucleotides tend to be paired. This difference naturally reflects not only the secondary structure of a particular RNA molecule, but also its long-range interactions. There are regions without canonical base pairs which are nevertheless fixed due to a noncanonical interaction with a distant partner.

Using ERNWIN, we can iterate over all possible conformations and generate statistics about which elements are expected to be paired. Potential interactions can be cross-referenced with the *SHAPE* data for verification. For example, the glycine riboswitch from *F. nucleatum* (http://rmdb.stanford.edu/repository/detail/GLYCFN_SHP_0006) contains one A-rich tetraloop which is weakly reactive, in contrast to the others. This is an indication that it may be involved in some A-minor interaction within the cell. A controlled analysis would require a comparison between the reactivity of elements which are known to be involved in long-range interactions and those which are known to be free. This information can help restrict predicted conformations to those that agree with experimental data. Structural constraints can be encoded as energy terms and used to guide sampling towards experimentally supported conformations. The use of probing data can be a great boon to the advancement of RNA struc-

ture prediction. To succeed, however, it will require careful calibration and thoughtful integration into our sampling framework.

12.3 *ERNWIN for multimeric structures*

RNA readily forms double stranded duplexes consisting of two separate molecules. Such intermolecular pairing is implicated in a wide variety of functional regulatory mechanisms ranging from RNAi [1], to Staufen-mediated decay [133]. RNA duplexes form double helices which are recognized by dsRNA binding proteins just like single-molecule helices. Predicting binding motifs in dimers would therefore be just as useful as predicting them in monomers. To this end, ERNWIN could provide an ideal framework for efficiently sampling multimeric RNA structures. Adding this support would require minor changes in the coarse-grain representation of RNA and the introduction of *breaks* in the structure to indicate where the distance and orientation parameters need not be constrained by the joining strand.

Even if this were implemented, prediction of multimeric RNA structure would suffer from the same paucity of accurate predictions that currently plagues single-molecule predictions. Combined with probing data and motif prediction, however, it could provide valuable insights to experimentalists working with RNA dimers.

12.4 *Guiding energy for multiloop construction*

The current implementation of multiloop sampling replaces a single loop at a time and rejects the newly chosen fragment if it creates a loop which cannot be closed. A better alternative may be to resample multiple segments of a multiloop at once using a *guiding energy*. Once one segment is replaced, the probability of accepting a subsequent segment would depend on whether it makes the loop easier to close. Segments pointing toward the start of the loop would be accepted with a higher probability than segments pointing away. This would ideally lead to a higher acceptance probability for the replacement of multiloop segments due to the fact that multiple segments can be replaced at once, overcoming barriers in the energy landscape.

12.5 *Kink turn evolution*

Secondary structures evolved under various selective pressures. Functional elements were selected to remain unchanged whereas unimportant regions evolved at a faster rate. The sizes of the interior loops can determine the range of orientations available to their adjacent helices. An

interior loop with one unpaired nucleotide on each strand will never turn into a kink turn. It must first grow enough to gain the flexibility necessary to fold into the kink turn motif. Once it extends to a reasonable size, the sequence can evolve to facilitate the formation of a kink turn. We can trace the progression of these events by looking at ribosome structures from different organisms [140]. Using tertiary structure prediction, we can determine when an interior loop gains the ability to bend enough such that two far-away elements form a long-range interaction.

The following quotation [146], indicating that the helices associate prior to the formation of the tertiary interactions, suggests that the presence of the kink turn is necessary for the initial assembly step. Without the ability to bend in the middle, the helices would not be able to associate to form the long-range interaction necessary for their function.

Base pairing of the ribozyme core requires 10-fold less Mg^{2+} than stable tertiary interactions, indicating that assembly of helices in the catalytic core represents a distinct phase that precedes the formation of native tertiary structure.[146]

We can run constrained folding simulations to determine which elements are most responsible for the structure of the molecule. If the RNA does, in fact, first condense to a roughly globular shape before all of the tertiary interactions are formed [190], then there must be a series of *hinge* points that allow it to contort itself into a condensed state. To determine which are the key hinges in a molecule we can recursively determine which interior / multiloops contribute the most to its correct 3D folding. Mutational analysis can verify the necessity of the hinges for the creation of a functional molecule [5]. Such information can corroborate previous studies which offer hypotheses about which elements of the ribosome evolved first [140, 18].

12.6 Tetraloop (TR) - TR receptor (TRR) evolution

TR - TRR interactions are well known for their stabilizing effect on the tertiary structure of RNAs. In order to function, however, both the TR and the TRR need to be present and they need to be oriented such that both can interact with each other. In other words, the secondary structure that separates the two elements needs to be flexible enough to allow them to fold over and make the necessary connection. This flexibility, in turn, requires the presence of either flexible multiloops or flexible interior loops. A classic case of this interaction is the Group I intron, whose function requires the presence of both a reverse-kink-turn motif and a TR-TRR interaction [5]. Determining which evolved first can be imagined as a chicken or the egg problem. Predicting which structures have the potential

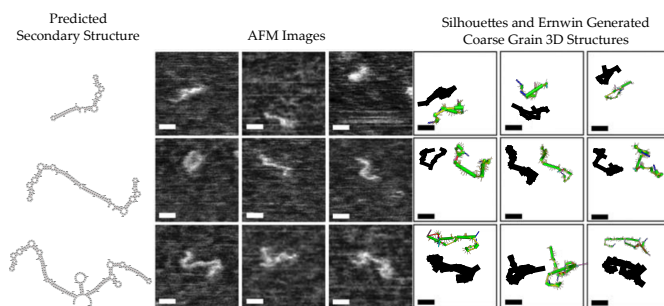


Figure 39: An example of [AFM](#) imaging of RNA structures. The actual images produced by the microscope are in the middle. On the left is the predicted [MFE](#) secondary structure for the sequences. On the right is the best hand-picked 3D structure matching the [AFM](#) image. Its silhouette is shown in black and the coarse-grain model is shown next to it in color. Most are close, but not exact matches to the blurry image visible in the [AFM](#) output. [AFM](#) images are adapted from [137]. Scale bars are 10nm long.

to fold back and bring hairpins near to stems could potentially identify precursors to TR-TRR interactions that have not evolved yet and shed light on the mechanism of both secondary and tertiary RNA structure evolution.

12.7 Using known long-range interactions as constraints

Section 11.4 showed how much our prediction could be improved by using known local structure. A complementary test would involve using a long-range distance constraint to show how much knowledge of interactions would improve the predictions. Done in reverse and iteratively removing constraints, we could show which interactions are necessary for holding particular elements together. For example, if a molecule has an active site located between two elements, we can ask which interactions would need to be disrupted to destroy the active site. One could envision potential applications wherein a complementary RNA is introduced to disrupt an interaction leading to the knockdown of a particular function.

12.8 Corroborating [AFM](#) images

In addition to the methods described in Section 2.3, [AFM](#) can also be used to detect which RNA structures are found in solution [108, 53, 137]. [AFM](#) uses a mechanical cantilever to scan over a sample and detect changes in the contour of the surface. An RNA on a slide perturbs the cantilever and registers as a signal on the microscope. These signals are converted to a visual representation and reported as images (Figure 39, middle). These

images, however, are merely 2D projections of tertiary structures. It would be useful to be able to match predicted 3D structures to the silhouettes visualized using AFM.

This would help to identify not only the tertiary structures, but also the secondary structures that are adopted by the RNA molecules in the solution. If a secondary structure can not fold into a tertiary structure which would have a 2D projection seen in the image, then that secondary structure is likely not in the solution. Figure 39 shows the manual assignment of tertiary structures to imaged RNAs. While most are close, some (e.g. the bottom left) do not perfectly match the 2D image. This presents the interesting question of whether the predicted secondary structure really is present in the solution if none of its potential tertiary structures match the AFM image. Conversely, seeing many silhouettes which can be readily assigned to tertiary structures is indicative that the predicted secondary structure really is the one present in the solution.

The example application shown in Figure 39 relied on manual comparison of the silhouettes of the 3D models to the AFM images. A more methodical approach would automate this process and check multiple projections of multiple tertiary structures to try and find the best match. Confounding this process is the question of which elements are really visible in the AFM image. Do only double-stranded regions show intense signals? What about nearly double-stranded interior loops? Such a method will need to be calibrated on images of known structures. Ideally, an RNA for which a crystal structure is known would be imaged using AFM and used to show that tertiary structure assignment from microscopy is truly feasible. To conclude this section, it is necessary to mention that techniques such as cryo-EM microscopy are already used to generate low-resolution tertiary structure models of RNA [67, 8]. Using ERNWIN for tertiary structure assignment using AFM microscopy would serve as an orthogonal approach which has the advantage of being able to consider different tertiary conformations of the same secondary structure as well as to account for the distortion that may be induced by the action of the cantilever used to probe the structure.

12.9 Summary

This chapter has listed a number of avenues for further research into both improving the accuracy of ERNWIN as well using it to answer biological questions. We can improve the accuracy of our method by incorporating information about long-range interactions obtained from covariation analysis. The limits to the potential improvement of predictions can be tested by providing the correct long-range constraints and seeing how they affect the tertiary structure sampling. Such hypothetical testing can perhaps even shed light on which long-range interactions are most important

in terms of bringing the tertiary structure toward a native conformation. On the experimental front, incorporating data such as SHAPE probing can provide additional constraints on the potential conformations. In terms of sampling, the inclusion of a guiding energy can help ERNWIN more efficiently and effectively sample multiloop conformations and thus lead to faster exploration of the conformational space.

In terms of applications, tertiary structures could be used to investigate and hypothesize about the evolution of kink-turns and long-range interactions (e.g. tetraloop - tetraloop receptor). For such scenarios, knowledge of the true tertiary structure is not required. Rather, the relevant information lies in the potential tertiary structures. What are the secondary structure and sequence conditions that need to be fulfilled in order for different elements to be able to form long-range interactions? Such predictions can be tested by phylogenetic analysis to see when long-range interactions start being conserved.

Finally, we propose the use of ERNWIN for the corroboration of AFM imaging experiments. Simple extensions to the current prediction methods can be tailored to find the coarse-grain structure most likely to be compatible with an RNA in an AFM image. This can be used to form hypotheses about not only the tertiary structures present in solution, but also the secondary structures that are implicit within them. Such tools, we hope, will both ease as well as quantify the analysis of AFM imagery and encourage its use to analyze RNA structures in solution.

Part IV

SECONDARY STRUCTURE VISUALIZATION

While visualization tools that interactively display RNA secondary structure already exist, our goal was to make the process even easier and to embed it within other applications. The development of JavaScript frameworks such as D3.js along with the ubiquity of the world wide web have allowed for the unprecedented dissemination of visual representations. D3.js, in particular, has simplified the composition of complex visual depictions allowing developers to integrate different layouts and stylistic elements into one coherent application. In the latter part of my PhD work, my colleagues and I took past work on laying out RNA structures and combined it with modern frameworks for display, dissemination, interaction and animation to create a number of utilities that, we hope will:

1. Simplify the lives of researchers by allowing them to effortlessly create RNA secondary structure diagrams online (FORNA [86], Chapter 13)
2. Enable enhanced interactive sharing of RNA secondary structures by providing an RNA container that can easily be embedded into web pages and augmented with auxiliary information about color or long range interactions (FORNAC [86], Supplementary Material, Chapter 13).
3. Augment the existing dot-plot representation of predicted base pairs to show the secondary structures associated with those base pairs (The dot-struct plot, Chapter 14).
4. Illustrate cotranscriptional RNA folding by showing the structures that a nascent transcript can form at different time points in its transcription (DR. FORNA, Chapter 15).
5. Diagnose the quality of the predictions generated by our 3D structure prediction tool ERNWIN (RNA long range adjacency plots, Chapter 16) by displaying which parts of the secondary structure are correctly predicted to be near each other.

FORNA AND FORNAC: SUPER SIMPLE SHAREABLE
SECONDARY STRUCTURES

13.1 *See, simplify and share with FORNA*

As described in the introduction (Chapter 5), there is already a wide variety of tools for drawing RNA secondary structure diagrams. None, however, provide a simple viewer that can be accessed using a web browser with no dependencies and none provide the key feature of extracting secondary structure from 3D PDB files. We set out to rectify both of these deficiency by creating a JavaScript application for drawing RNA secondary structures. While it requires a server-side application (written in python) to calculate an initial layout, this is not something that the end user needs to install, maintain, or even really be aware of [86]. To satisfy user demand for a front-end browser-based secondary structure layout, we also created a pure JavaScript layout container that can be instantiated without the need for a back-end server [86].

Having encapsulated MC-ANNOTATE's [56] tertiary structure annotation functionality in an easy-to-use python library, we gave users the opportunity to input PDB files to FORNA and see the secondary structure of the RNA molecules they contain. This makes it incredibly easy to get a simplified representation of the intricate 3D structure of an RNA. It also allows us to see which RNA structures are present as monomers, dimers or multimers as well as to show any proteins present and where they interact with the RNA molecules. This improves our ability to screen for monomeric structures which do not interact with proteins for use as benchmarks in our tertiary structure prediction program, ERNWIN [87] (Part iii). As we prepared our manuscript, Antczak et al published an online tool [4] which served as an interface to existing PDB annotation and RNA visualization tools and let users convert PDB files to secondary structure diagrams. While this exposed some overlapping functionality with FORNA, it still lacked the ability to show multimeric structures, as well as protein interactions.

The work on FORNA is summarized in a publication in Bioinformatics [86] which is reproduced in Chapter 7. It describes most of the functionality that our online tool provides for the scientific community. Not only does FORNA provide novel features not found in any other program, it also implements already existing functionality (such as the simple drawing of structures) in a much more accessible package. Users need only

a web browser and an RNA sequence in order to create aesthetic and descriptive secondary structure diagram.

FORNA can show structures of any size, but due to the heavy computational requirements, we recommend displaying structures no longer than around 1000 nucleotides. As with other software tools, we bend some conventions in the hope of achieving the best balance between aesthetics, usability and scalability. Among the more sacred conventions that we eschew is the notion of a straight helix. While base paired regions are, for the most, part linear and of uniform width, they can be bent in our representation. This allows for more flexibility in avoiding overlaps as well as for better packing of large structures. Figure 40 of the 23s ribosomal subunit from *E. coli* below is an example of a 2841 nucleotide secondary structure extracted from a crystal structure (PDB ID: 4V4Q [156]). This illustration is not automatically generated and required a fair amount of user interaction to not only remove overlaps but to arrange the elements to roughly correspond to the typical arrangement of ribosome secondary structure (See Figure 40 and [139]).

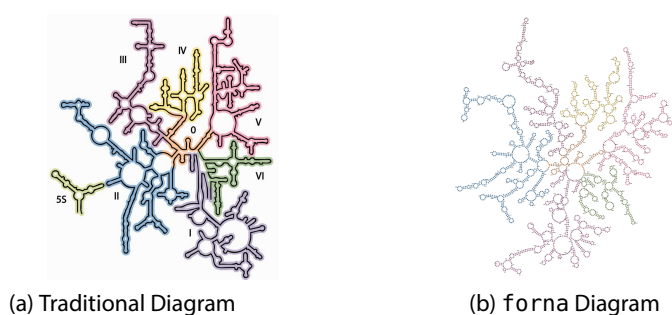
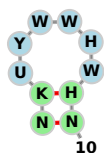
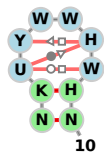


Figure 40: Two examples of a secondary structure diagram of the ribosome. On the left from [139] and on the right drawn by forna.



Canonical base pairs



Noncanonical base pairs added

13.2 Potential improvements

Between the instant secondary structure drawing, custom coloring, PDB structure extraction and the multitude of additional features, FORNA provides an immediately usable tool for generating secondary structure diagrams. Even so, there are four major improvements that could be implemented to increase the breadth of its utility, make it more accessible as a custom application, to simplify the process of rearranging the layout, and to create even more beautiful diagrams.

DISPLAY NONCANONICAL INTERACTIONS The current version of FORNA does not have the capability to display noncanonical in-

teractions. Adding these interactions would make it suitable to at least three different potential applications.

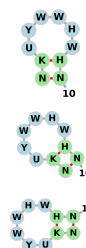
The JavaScript display container, FORNAC, could be coupled to databases of noncanonical motifs [138], providing an interactive interface where only a static one currently exists. It could further be coupled to the output of a web service which provides predictions of noncanonical motifs, should one ever be created out of existing tools such as RNAWolf [199]. Finally, we can leverage our existing infrastructure for parsing PDB files to extract and display the noncanonical base pairs present in solved crystal and NMR structures.

ALLOW ROTATION OF SELECTED NUCLEOTIDES The current implementation allows dragging on single as well as multiple selected nucleotides. This works well for adjustments involving the translations of nodes.

It is often necessary, however, to rotate large sections of a molecule. Such functionality would be a welcome addition to the currently available repertoire of user interactions and ease the disentanglement of larger molecules.

IMPLEMENT DIFFERENT LAYOUT ALGORITHMS FORNA currently uses the NAView [24] (Section 5.3) layout for drawing RNA secondary structure. As described in Chapter 5, there are a variety of different layout algorithms available for arranging nucleotides and base pairs. Adding implementations for one or more of these and allowing users to select which one they wish to use would be helpful in letting them further tailor the diagram to their particular requirements.

OUTPUT TO R2R R2R [184] is a program for drawing aesthetically pleasing RNA diagrams. It generally requires some manual intervention to specify the desired stem orientations (parallel, antiparallel, etc...), loop layouts as well as various other available parameters. One of FORNA's weaknesses is its difficulty in drawing perfectly positioned stems and loops. Combining FORNA's crude layout with R2R's precision and aesthetics could greatly ease the creation of RNA diagrams. A sufficiently motivated diagram creator could roughly lay out their RNA in FORNA, and then export it to an R2R input file for further manipulation and rendering. The result would be a rapidly prototyped RNA diagram, crisply and precisely rendered using the facilities of R2R.



Different rotations of an RNA.

THE DOT-STRUCT PLOT: AN IMPROVED DOT PLOT

This chapter will introduce the dot-struct plot, an extension of the traditional dot plot diagram for showing the probabilities of predicted base pairs. We created the dot-struct plot as an entry to the BioVis 2015 design contest where the stated objective was to create a visualization for displaying ensembles of predicted structures. Our intention was to take the familiar dot plot layout and decorate it with images of the predicted secondary structures. To make it easier to navigate, we made it fully interactive with mouseover effects to show details about specific base pairs and zooming to focus on details. The remainder of this chapter will provide specific details about our motivation and design decisions as well as use cases for the dot-struct plot. The actual entry to the design contest is presented as an addendum at the end of the chapter.

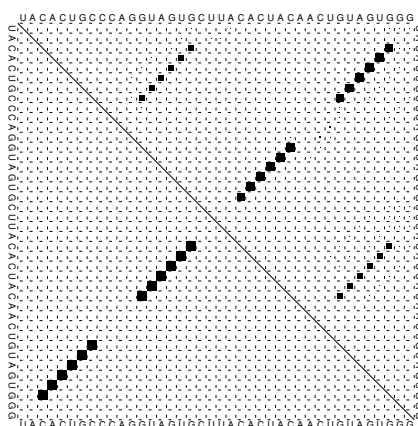
14.1 *Augmenting the dot plot*

Figure 41: A typical dot plot showing the base pairs in the **MFE** structure in the lower left triangle and all the potential base pairs above a probability of 0.0001 in the upper right triangle. The sizes of the square dots are proportional to the pairing probability they represent.

The traditional dot plot (Figure 41) shows the probabilities of suboptimal base pairs that are predicted for a given sequence. As mentioned in Section 2.4, it is possible to predict not only the **MFE** structure but also structures which have a slightly higher energy and thus contain suboptimal base pairs. These suboptimal base pairs can then be assigned probabilities by calculating the partition function over all possible struc-

tures. This is illustrated in the dot plot (Figure 41) where the sides of the rectangle show the sequence and the points indicate the probability of potential base pairs. The plot is divided along its diagonal into a lower left and an upper right half. The lower left half shows the base pairs that are present in the **MFE** structure, while the upper right half contains base pairs that are also present in suboptimal structures.

The dot-struct plot in Figure 41 shows the base pairs of a particular class of RNA structure called a riboswitch [160]. Sequences which encode riboswitches tend to have two very different low-energy structures. Environmental conditions can cause them to switch from one conformation to another. Somebody used to seeing dot plots may be able to infer that this is a riboswitch by noticing one set of large dots bisecting the upper right hand portion of the chart and another set of smaller dots flanking it. From there, however, it takes a bit of mental gymnastics to picture what sort of structure these base pairs correspond to. It is even harder, if not impossible, to compare the free energies of the two structures from which these base pairs came (if one can even deduce which structure they came from). To alleviate this situation, we augmented the traditional dot plot with the diagrams of the structures which contain the listed base pairs (Figure 42).

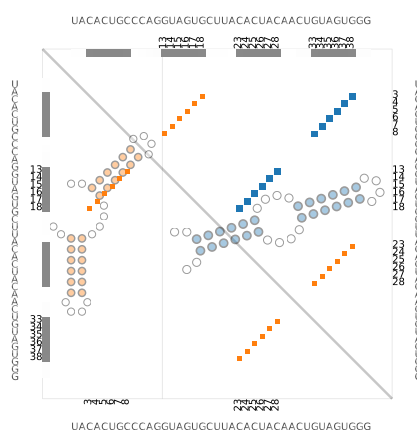


Figure 42: A dot-struct plot of the same sequence depicted in Figure 41

The resulting diagram (Figure 42) makes it possible to see the structures from which the base pairs were derived. The colors of the dots correspond to the lowest energy structure in which those base pairs were found. In Figure 42, the blue base pairs are found in the **MFE** structure, while the orange ones are found in the first suboptimal structure. The secondary structure diagrams are scaled according to their Boltzmann probability in the ensemble of predicted structures. Thus, we see that while the two dominant structures are similar in stability, the **MFE** structure is about twice as probable in the ensemble as the first suboptimal. The dot-struct

plot shown in Figure 42 is relatively simple, having only two suboptimal structures. A multi-stable riboswitch can have more than two significant suboptimal structures with similar free energies (Figure 43). Displaying such an ensemble can lead to a busy dot-struct plot where the structures and dots overlap. To make it easy to distinguish the dots from the structures, we provide an option to separate the dot plot from the treemap of structures (Figure 44).

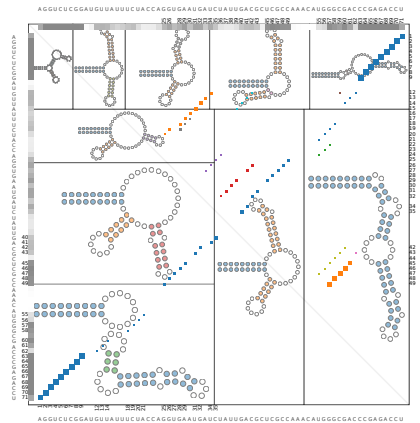


Figure 43: A dot-struct plot of a multi-stable RNA riboswitch displaying 10 suboptimal structures.

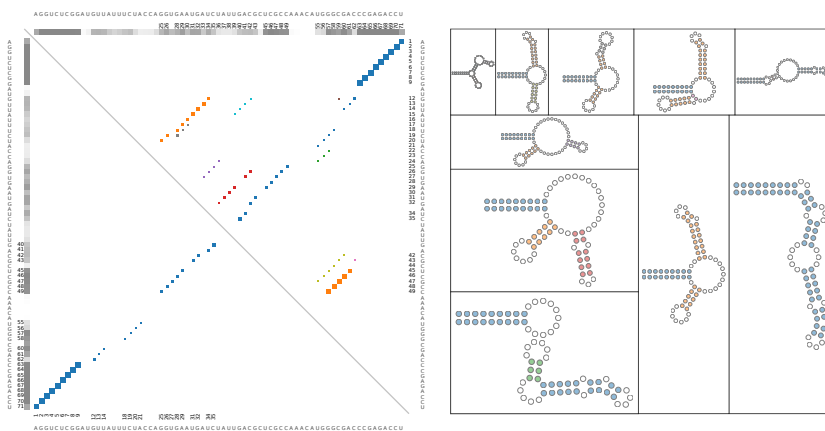
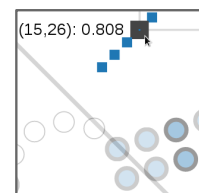


Figure 44: An expanded dot-struct plot of the multi-stable riboswitch shown in Figure 43.

The depictions of the dot-struct plot shown in this thesis (Figures 42,43 and 44) are necessarily static. By design, however, they are fully interactive. Switching between the collapsed (Figure 42 and 43) and expanded view (Figure 44) occurs with the click of a mouse. Hovering over a dot highlights the base pair in every structure where it occurs and displays the numbers of the two participating nucleotides. Conversely, hovering



A zoomed in view of the hover behavior showing the nucleotide numbers and probability corresponding to the highlighted dot and base pair in the structure.

over any nucleotide in a structure diagram highlights that nucleotide in every other structure as well as the dots corresponding to the base pairs it is involved in. The entire display is fully zoomable, allowing the user to enlarge and focus on areas of interest. Many of the other design decisions are documented in our submission of this display to the design contest at the BioVis 2015 meeting in Dublin, Ireland.

14.2 *Design contest submission*

Shortly after we created the FORNA software package for visualizing RNA secondary structure, we saw a call for the BioVis design contest asking for the creation an enhanced dot plot. It was the perfect testing ground for our new implementation. By augmenting the traditional dot plot with structural information, we could shorten the mental leap that the reader has to take to connect base pair probabilities to the real structures they are derived from. Putting it online and coding it in JavaScript using the D3.js framework allowed us to make it interactive. The existence of the JavaScript container we created for FORNA made the entire task an easy exercise in piecing together different components to re-create and augment the existing dot plot. The actual submission is reproduced on the next two pages.

Visualizing Ensembles of Predicted RNA Structures and Their Base Pairing Probabilities

Peter Kerpedjiev and Ivo Hofacker

Abstract—In this design contest submission we present an enhanced version of a traditional RNA dot plot containing a multitude of extra features and data, foremost among which is the inclusion of diagrams for the top Zuker sub-optimal RNA secondary structures. This new design facilitates and eases the interpretation of the dot plot by providing the viewer with an immediate representation of which structures the displayed base-pair probabilities belong to.

1 INTRODUCTION

The traditional RNA dot plot conveys the probability that a particular base-pair is present in the ensemble of predicted structures. This information is presented as a 2D scatter plot, where the size of the rectangular marks is proportional to the probability of a pairing between nucleotide i (on the x-axis) and nucleotide j (on the y-axis). The upper right triangle of the plot displays this information for the ensemble of predicted secondary structures whereas the bottom left displays only the pairs present in the minimum-free energy structure (MFE). The dot plot is useful in conveying to the viewer that some nucleotides may have a propensity to form differing base-pairs. At first glance, it shows whether there are stems which are consistent across the whole ensemble and which nucleotides they encompass.

Beyond this application, however, it becomes difficult (albeit far from impossible) to extract extra information. The key unanswered question, in our opinion, is which structures correspond to the indicated base-pairs? As previously mentioned, the pairs corresponding to the MFE structure are shown in the lower left hand corner. What does this structure look like, however? What about the other base-pairs in the upper right section? Which structures do those correspond to? How many different structures do they correspond to? Which can be found in the same sub-optimal structure?

With these questions in mind, we set about redesigning the dot plot to include actual secondary structure diagrams in the background. The result, shown in Figure 1, gives the viewer an answer to each of the questions posed above and more. It further provides a platform which can be extended to create an interactive tool to ease the exploration of the data presented in the visualization.

2 DESIGN CONSIDERATIONS

Our design was created to answer some basic questions that researchers might ask about an ensemble of predicted RNA structures, as well as to provide some minor improvements to the way the traditional dot plot is laid out. In each section we describe what we did, why we did it, as well as how we feel it could be improved with an interactive version of our design.

2.1 What does the MFE structure look like?

Description: In the traditional use case, one receives a secondary structure diagram representing the minimum free energy structure in one file and the dot plot in another. We strive to unite these two representations by showing the MFE structure in the background of the dot plot. Such an approach is alluded to in a figure in [3], but we go one step further and arrange the MFE structure along with other sub-optimal structures and scale their size according to their expected population in the Boltzmann ensemble of predicted secondary structures.

Motivation: To give the viewer an immediate representation of the MFE secondary structure.

2.2 Which other structures are predicted?

Description: RNA folding, being a kinetic process, leads to the presence of more than one particular structure in solution. We display a subset of these sub-optimal structures, along with the MFE structure, in the background of the dot plot. Based on the energy of each predicted structure, one can calculate its expected weight within the ensemble and use it to scale the size of its secondary structure using a squarified treemap layout [1]. Only structures which correspond to base-pairs with a probability above a certain threshold (see next section) are displayed.

Motivation: The MFE structure can quickly be compared to the other predicted structures in the ensemble in terms of not only structure, but also energy value.

Potential Improvement: Some structures can appear quite small. An interactive version of the plot can enlarge them when one hovers over a base pair belonging to that structure.

2.3 Which structures do the predicted base pairs correspond to?

Description: The upper right hand corner of the dot plot shows all of the potential predicted base pairs above a certain probability value (0.08 in our case, 0.00001 in the traditional dot plot). We chose a higher cut-off due to the simple fact that a lower cutoff would yield points so small as to be virtually indistinguishable without a magnifying glass. Each of the dots is colored to match the color of the best sub-optimal secondary structure containing that base pair. Recall that these structures are displayed in the background of the dot plot.

Motivation: This encoding helps to link the predicted base pairs with the structures they are expected to appear in.

Potential Improvement: Increasing the size on mouse-over, as suggested in the previous section, should help alleviate this issue. Clicking on a structure could also be employed to highlight/enlarge the base pairs belonging to it.

2.4 Which base pairs in a structure are displayed in the dot plot?

Description: The MFE and sub-optimal structures in the background are generated by finding the lowest energy structure given a base pair constraint. Within those structures, we highlight the pairs which, when constrained to being paired, lead to the prediction of that structure. These also correspond to the base-pairs displayed as dots on the dot plot.

Motivation: By highlighting the base pairs in the secondary structure, one can easily see not only how many, but which pairs in a sub-optimal structure are represented in the dot plot.

Potential Improvement: The identity of the base-pairs could be clarified by drawing lines between the secondary structure and the dots when users hover the mouse over the dots.

2.5 Minor improvements

Nucleotide Numbering: We added the positions of the nucleotides to the margins. To avoid clutter, we only add the numbers for nucleotides

• Peter Kerpedjiev (pkerp@tbi.univie.ac.at) and Ivo Hofacker (ivo@tbi.univie.ac.at) are both at the University of Vienna.

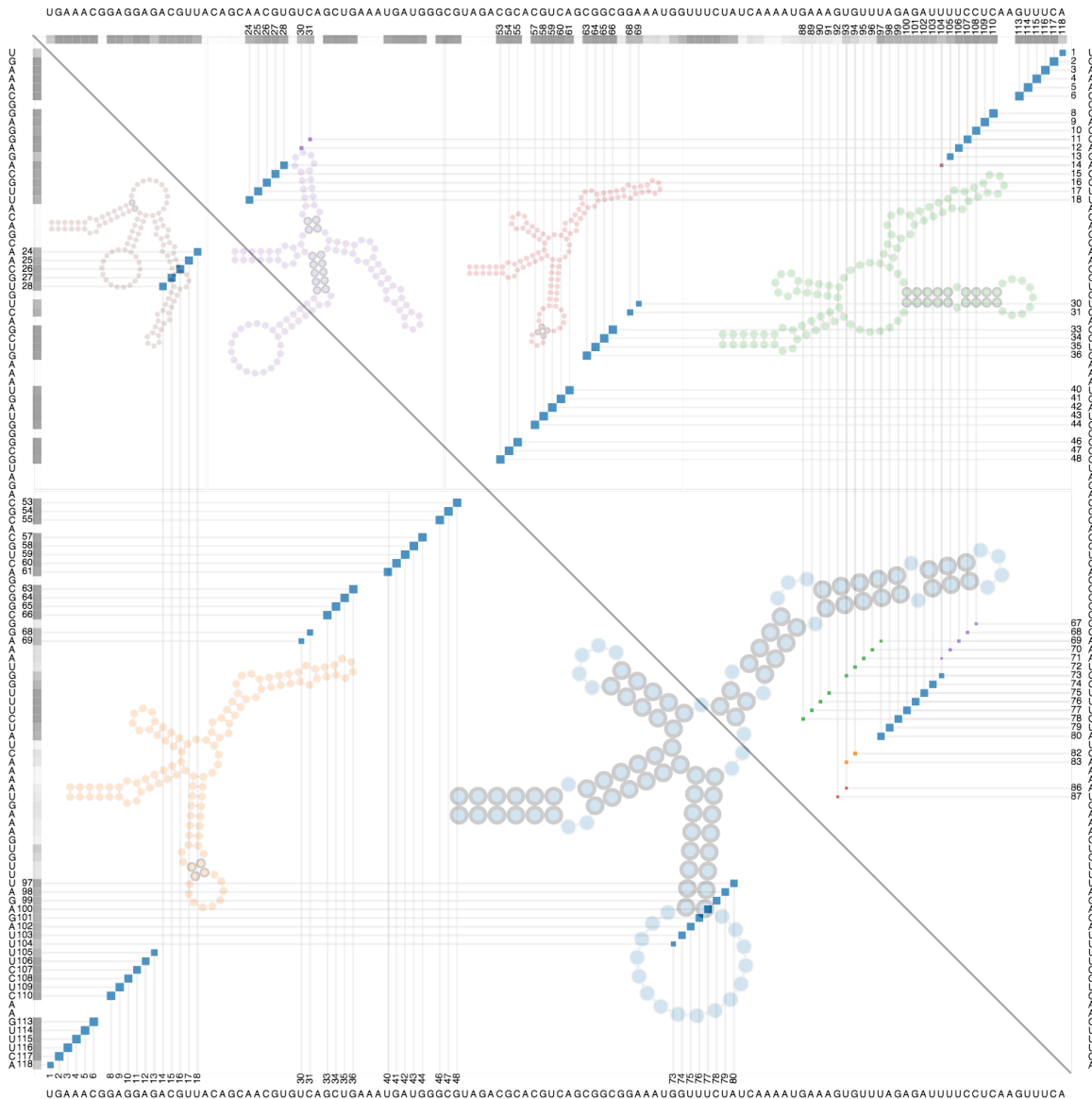


Fig. 1. Sample enhanced dot plot for the Human Highly Accelerated Region 1A provided in the contest data.

that have the potential to be in a base-pair (i.e. have a probability greater than the threshold).

Numbering Guide Lines: To guide the viewer in reading out the identity and position of each paired nucleotide, we have added faint lines from each dot on the dotplot to the numbers of the nucleotides in the margin

Total Pairing Probability: The summed pairing probability for each nucleotide is encoded as colored squares on the upper and left border of the plot. This provides an overview of which nucleotides are likely to be paired in the whole ensemble. It can be used as a comparison with data from probing experiments.

3 GENERATION

The data for the plot is generated by a python script which makes use of the python binding of the ViennaRNA package. The actual plot is rendered in the browser using the D3.js and fornac.js libraries. Such a format makes it easy to add interactivity to the current design.

4 AVAILABILITY

The code for creating this visualization is available at:

<https://github.com/pkerpedjiev/dotstruct>

A higher resolution rendering of Figure 1 can be found at:

<http://www.tbi.univie.ac.at/~pkerp/dotplus/>

ACKNOWLEDGMENTS

We wish to thank Ronny Lorenz for his help with the ViennaRNA package python interface and Stefan Hammer for his tireless work on creating the secondary structure visualization tool **forna** [2] which paved the way for this submission.

REFERENCES

- [1] M. Bruls, K. Huizing, and J. J. Van Wijk. *Squarified treemaps*. Springer, 2000.
- [2] P. Kerpedjiev, S. Hammer, and I. L. Hofacker. **forna** (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *submitted*, 2015.
- [3] R. B. Lyngsø, J. W. Anderson, E. Sizikova, A. Badugu, T. Hyland, and J. Hein. Frnakenstein: multiple target inverse RNA folding. *BMC bioinformatics*, 13(1):260, 2012.

That RNA begins to fold as soon as it starts being transcribed has been known for over three decades [23, 94, 97]. Certain parts of the molecule can form stable secondary structures before the entire molecule is completely transcribed. This can prevent the fully transcribed RNA from adopting its MFE structure by forming so-called kinetic traps, wherein a nascent structure is so stable that unfolding it to form the MFE structure can take an inordinate amount of time. Evidence suggests that rearrangements do, in fact, take place during folding and that mutually exclusive structures are present at different points in the transcription process [94]. This can have many important consequences for the fate of the nascent transcript. Proteins binding a particular motif may find the motif while the transcript is being transcribed but not when it is fully formed. Structures interfering with the action of the transcription machinery may slow down the transcript's own transcription. The appearance of kinetic traps may lead to the transcript being trapped in a suboptimal energy well that makes it difficult for the molecule to attain its MFE structure.

The speed of transcription has an influence on the types of transcripts that will be produced. Slower transcription can lead to more meta-stable intermediates by giving them a chance to form before the latter part of the sequence is transcribed. Faster transcription can interfere with proper folding by favoring other incorrect intermediates. In both cases, it is the sequence itself that determines which intermediates can be formed and which transcription speed is ideal for creating a functional RNA. Transcribing an RNA at an unsuitable speed can result in non-functioning products (as in the case of rRNA [102, 29]). This can be seen more clearly in studies of the Group I intron, where cotranscriptional folding occurs roughly twice as fast as refolding, both of which occur 10-times slower in-vitro than in-vivo [69]. An explanation for this is that cotranscriptional folding favors the formation of various local structures which reduce the conformational search space for the rest of the molecule. Simulations of the likely intermediate structures can give researchers a glimpse at the possible local structures which are formed during transcription.

While there has been ample research into predicting secondary structures and taking cotranscriptional and kinetic folding factors into account [122, 121, 144, 48, 79, 191, 192, 60, 55] there has been little work in the way of displaying the results in a meaningful manner. The graphs reproduced in this chapter range from the arcane (Figure 45) to the sim-

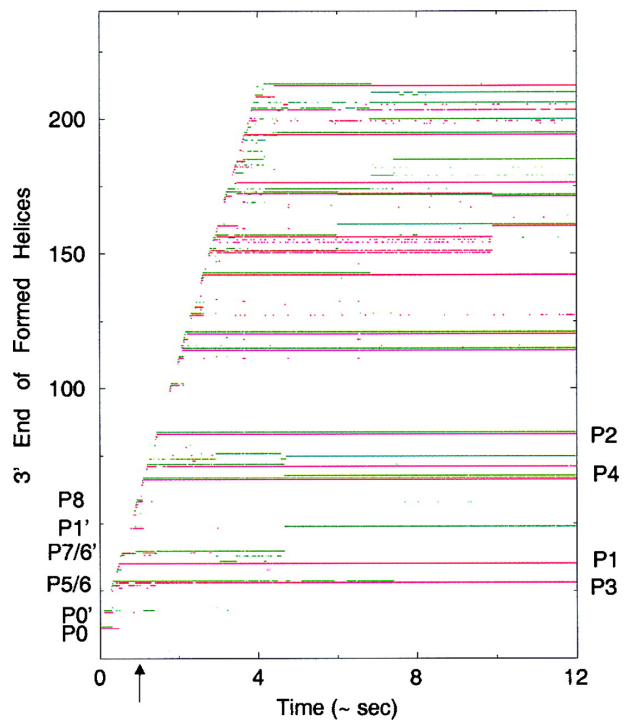


Figure 45: "Chart of helices present in the ribozyme plus attenuator sequence as it folds during and after synthesis for a molecule that folds via the catalytic folding path" [79]. Figure reproduced from [79].

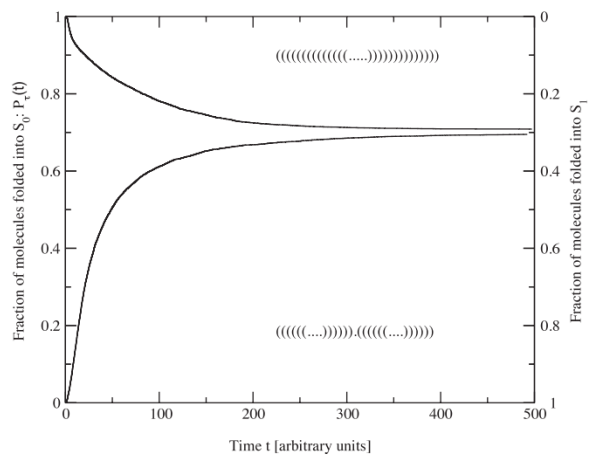


Figure 46: An illustration of the fraction of two conformations in a solution over time [48]. The use of dot-bracket notation to depict the structures is forgivable due to their simplicity. There is, however, no indication of which of the two is S_0 and which is S_1 . Figure reproduced from [48].

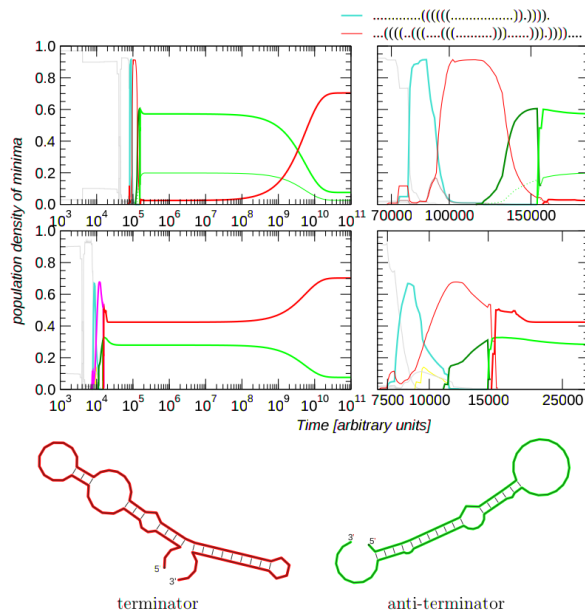


Figure 47: How transcription speed affects the simulated abundance of different RNA transcripts [72]. Slow transcription is on top while fast transcription is shown on the bottom. The two major RNA species are illustrated below the graphs, but no information is provided about the other species represented in the line graphs. Figure reproduced from [72].

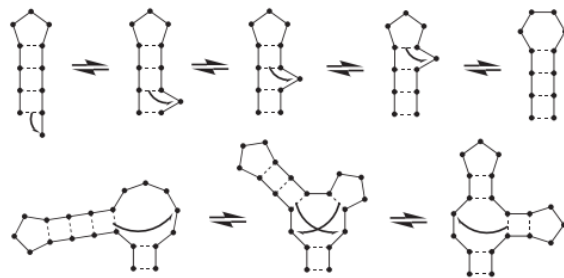


Figure 48: A diagram explaining how loops can travel through double stranded regions and lead to distal rearrangements of the secondary structure. Figure reproduced from [48].

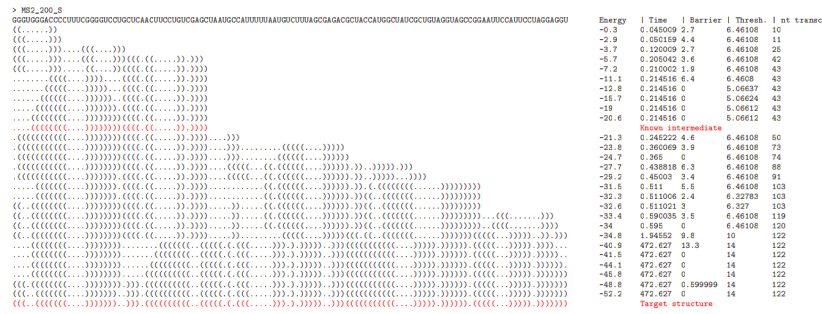


Figure 49: The folding pathway of MS2 A-protein 5'UTR as it is being transcribed (left to right). The red color indicates a *trap* structure described in another publication. The structures are described in dot-bracket notation, making it difficult for an untrained eye to recognize which nucleotides are paired with each other. Time is not scaled linearly with the position of the output. Figure reproduced from [55] with permission according to the license in the Appendix.

ple yet spartan (Figure 46) to the cartoonish yet clear (Figure 48), to the outright overwhelming (Figure 49). Given our implementation of the simple polygonal layout (Figure [163]) as a D3.js layout, we are able create a visualization which arranges and scales multiple RNA diagrams in one window while allowing users to interactively explore the predictions at different points in time.

As an experiment in novel ways to see how the dynamic population of RNA structures in a nascent transcript evolves, we used an as yet unpublished work, DR. TRANSFORMER, (Badelt et al. unpublished) to simulate the structures and their concentrations at a number of time points over the course of transcription. The output of DR. TRANSFORMER could just as well have come from any of the other RNA folding kinetics simulation tools mentioned above and is not specific to its functionality. It simply describes an id, a time, a concentration, a structure and an energy (similar to that shown in Figure 49):

```
id time conc struct energy
1 0.380000 1.000000 ..... 0.00
2 0.3800005 3.396416e-04 .((((((.....)))))) -0.80
1 0.3800005 9.996583e-01 ..... 0.00
3 0.3800005 2.029009e-06 ...((((.....)))... 0.10
```

Within this output, the three key points for downstream analysis are the time, the concentration and the structure. The original output for DR. TRANSFORMER (see Figure 50) displayed all three, but with a number of shortcomings:

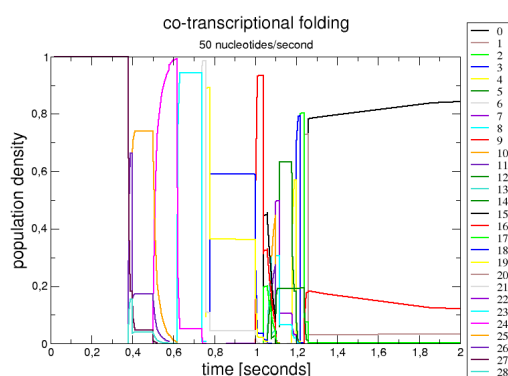


Figure 50: The original output for Dr. Transformer showed the concentration (normalized and displayed as population density, the time (linear scale) and the structure identifier. This plot bears a close resemblance to that of Flamm et al [48] and Hofacker et al. [72] as shown in Figures 46 and 47, respectively. Figure courtesy of Stefan Badelt.

- The structures themselves are not displayed. The user must shift his/her attention to another document to find out what structure 1 or 2 or x actually looks like.
- There is an overwhelmingly large number of lines, making it hard to match colors to structure ids.
- The multitude of lines obscures their path through the plot and makes it hard to read and interpret.

As a solution, we offer an interactive display (dubbed DR. FORNA) to let viewers scroll through the the time series and view the data at each time point individually. In constructing our visualization, we make two major assumptions:

- Researchers are not so interested in structures which are present in miniscule concentrations. This helps remove the clutter in the line chart.
- The dot-bracket notation for structures is not amenable to human interpretation. We therefore replace dot-bracket strings with the equivalent RNA secondary structure diagram, drawn using the *standard polygonal layout* [163] (Section 5.1).

Working under these assumptions, we decided to retain the line chart due to its familiarity to those in the field of RNA kinetics. It gives a quick overview of how the landscape of RNA structures evolves over time. This is augmented with a visual depiction of the structures we expect to find in

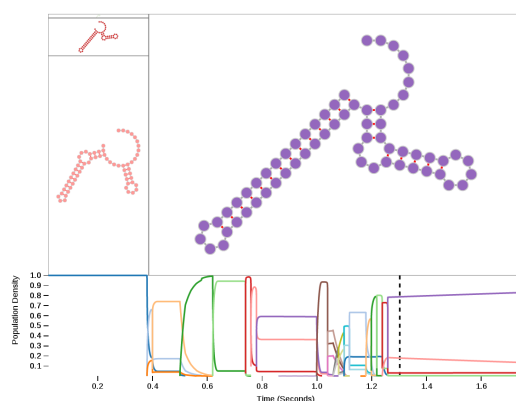


Figure 51: A screenshot of the initial Dr. Forna design. The line chart at the bottom displays the concentrations of various secondary structures over the course of the transcription process, like in Figure 50. The secondary structures in the top two thirds of the plot are the species present at the time marked by the dashed line below.

solution at any particular time point. The areas that the structures occupy are scaled according to their population density (i.e. what percentage of all of the structures that are present in the solution are this structure?). Hovering over any point in the line graph displays a snapshot of the structures present at that point in time. A screenshot of the Dr. Forna display applet can be found in Figure 51.

The current implementation, which requires the user to hover the mouse over the line graph to see the corresponding structure, is useful for exploring the landscape of potential structures. At the same time, it is difficult to convey the structural landscape over the entire course of the simulation in one static output, as would be required for a publication. One possibility, showing a series of screenshots taken at regular intervals is revealing, yet cluttered (Figure 52). Little extra information is gained from the conceptually similar structures shown at the beginning and end of the simulation. Many of the smaller, less populated, structures are difficult to see. The line chart becomes redundant and hard to see when shrunken and repeated. To improve this chart we have come up with a number of ideas for how to better organize and display this data.

LOGARITHMIC TO LINEAR TIME SCALE Currently the time scale is linear from start to finish. The changes in concentration that occur as a consequence of differing energy values, however, happen on a logarithmic time scale. Fusing the two and showing certain portions of the interval on logarithmic scale and others on a linear scale may be more meaningful in terms of displaying the dynamics of the RNA structure population.



Figure 52: A tile of sequential screen captures of the Dr. Forna plotting application. In this busy image, one can see the cotranscriptional folding path of an RNA molecule as time passes and it grows in size.

POPULATION THRESHOLDS Allowing users to set a threshold for how populated a certain structure must be before it is displayed may reduce the clutter and devote more space to the more prominent structures.

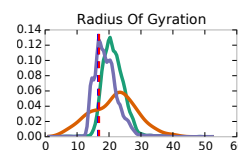
TIMELINE BRUSH It may be that a user is only interested in a particular interval of the simulation. Allowing them to select a start and end time and showing snapshots of the landscape at fixed points within that interval may limit the number of multiples that need to be displayed. This can effectively lead to more screen area for each time point.

CLUSTERING A discerning viewer may notice that many of the structures shown in Figure 52 differ by a single nucleotide or base pair. Such small differences may have a negligible functional relevance and might be omitted. Clustering similar structures may allow us to devote more space to the major rearrangements that take place during the transcription process.

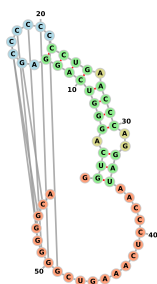
In our paper describing ERNWIN [87], we demonstrated how our tertiary structure prediction program facilitates the sampling of ensembles of structures. What do those ensembles look like? We used histograms of their coarse-grain measures (such as ROG, shown on the right) and RMSDs to characterize the ensemble of sampled structures. None of these diagnostics give us any idea of what these structures actually look like. In the case of a single lowest energy structure the natural solution is to open it in a 3D structure viewer such as PyMOL [154] and to manually inspect it. This is useful for examining the features of a single structure but quickly becomes unwieldy when we are confronted with a multitude of conformations.

How then, can we display a summary of an ensemble of structures that is both meaningful and condensed? To answer this question, we need to ask the complementary question of what is the viewer trying to see? What is it about the ensemble of predicted structures that is relevant to the person doing the analysis? An experimentalist working with cryo-EM data may be interested in 2D projections of the tertiary structures. They may be interested in what a cryo-EM image of a given RNA looks like if it adopts the conformations predicted by a tertiary structure program. This will allow them to compare this simulated image to a real image and to make judgments as to their similarity. An experimentalist working with SAXS data may be interested in the distribution of radius of gyration values.

As a creator of tertiary structure prediction software, I am concerned with how well my method performs. How well does it recapitulate the known structures? As described in Section 9.4 we use the ACC measure to summarize how accurately we predict long-range adjacencies. This condenses the difference between two models to a single number which while useful when benchmarking performance on multiple structures, provides little information about how any two particular structures actually differ. This number furthermore only captures the difference between two models. Given that ERNWIN predicts an ensemble rather than just a single model for any given input secondary structure, an effective diagnostic would display which interactions were correctly or incorrectly predicted for all of the predicted structures. It will be easy to identify what types of elements are involved and which nucleotides they contain in order to go back and diagnose which aspects of our energy function lead to their correct or incorrect adjacency prediction.



Distribution of ROG values in sampled structures.



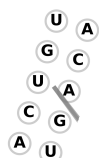
An illustration of the MLV readthrough pseudoknot (PDB ID: 2LC8 [75]) secondary structure as displayed by FORNA.

16.1 Long-range interaction overlay

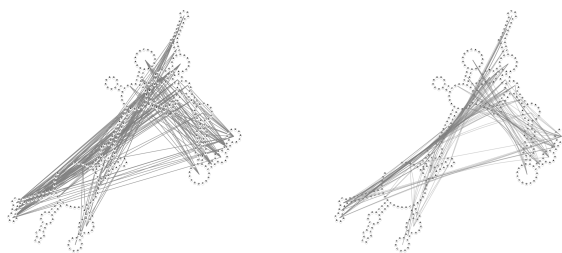
The simplest way to display long-range interactions is to overlay them on top of a 2D diagram (see margin). This structural context gives the viewer easy access to information about the neighboring elements. In the example with the MLV pseudoknot in the margin, one can clearly see that there is an interaction between the hairpin loop and the 3' external loop. To analyze tertiary structures, we can use FORNA's ability to overlay long-range links on top of the secondary structure diagram to display a flattened representation of a tertiary structure prediction. Given our coarse-grain structure representation this presents two difficulties:

1. How do we determine if there is an interaction?
2. What do we actually link when individual nucleotides are not represented in the tertiary structure?

To answer the first question, we need only refer to Section 9.4, which described the calculation of the ACC value between two coarse-grain structures. This section showed how we chose a distance of 25\AA to consider elements sufficiently 'adjacent' as to be potentially involved in a long-range interaction. We also showed that two elements need to be separated by at least 16 base pair or backbone links in order to reduce the influence of local secondary structure on their proximity and be considered a potential long-range interaction. This suggests a natural adjacency criterion for displaying links between distant (in terms of secondary structure) elements. We simply show all links between elements further than 16 nucleotides / base pairs in the secondary structure and closer than 25\AA in the tertiary structure (Figure 53).



Long-range interaction overlay links starting at the centroid of a stem.



(a) Normal Long-Range Overlay

(b) Bundled Long-Range Overlay

Figure 53: Adding connections for all nodes further than 16 nucleotides / base pairs apart which are closer than 25\AA (left). Using force-directed link bundling [73] to clean up the figure and reduce clutter (right).

Because every coarse-grain element consists of a set of nucleotides, Question 2 can be resolved by drawing a link's terminus at the centroid

of the element's nucleotides in the 2D diagram. The exception to this rule are degenerate elements that are inserted as placeholders (e.g. a multiloop segment of length zero that exists between two adjacent stems in a multiloop). If we extend the set of nucleotides which compose a coarse-grain element to include the nucleotides directly adjacent to that element, then every coarse-grain element will be composed of a non-zero number of nucleotides. To link a CG element to something else we simply place the start of the linking line segment at the centroid of its composing nucleotides.

The scheme described here allows us to convey information about coarse-grain element positions in the third dimension directly on top a secondary structure diagrams. This will hopefully provide context and clarify which nucleotides are near to each other and thus have the potential to interact with each other.

16.2 Diagnosing the accuracy of predicted adjacencies

The illustration in the previous section provides a starting point for analyzing coarse-grain tertiary RNA structures in two dimensions. While it can be used independently for exploring the 3D proximity of distal 2D elements, it lacks the key capability of allowing us to diagnose which long-range adjacencies are correctly predicted and which are incorrectly predicted. By coloring correctly predicted adjacencies green and incorrectly predicted ones red, we can pair the 2D representation with diagnostic information about the veracity of the predicted adjacencies.

Figure 54 shows a native (left) and predicted (right) structure in both its original coarse-grain 3D model as well as the 2D representation with the adjacency overlay described in this chapter. The 2D representation offers a quick and easy summary of the tertiary structure models. One can immediately see the correct long-range interaction between the far right hairpin and the 3' unpaired region on the left in the native model. The poorly predicted model, in contrast, contains a lonely 3' unpaired region and incorrect pairings among the other hairpins. Figure 54 includes the 3D structures to show where the adjacencies in the 2D diagrams come from. Without them, a glance at the colored 2D representation with long-range interaction overlay informs the user of the general quality of the prediction (green for good, red for poor), how compact the molecule is (i.e. how many long-range interactions are predicted), and which elements are correctly and incorrectly predicted to be near to each other.

This view is informative for a single structure, but as emphasized, our tertiary structure prediction work produces ensembles of structures. One structure may be well predicted, but is this the case with the others? Are there certain long-range interactions that are consistently correctly or incorrectly predicted? The answers to these questions can help us evaluate



Long-range interaction overlay links starting at the centroid of a hairpin loop.

The `cg_to_fornac.py` script in the `forgi` package generates an HTML document containing the long-range diagnostic diagram.

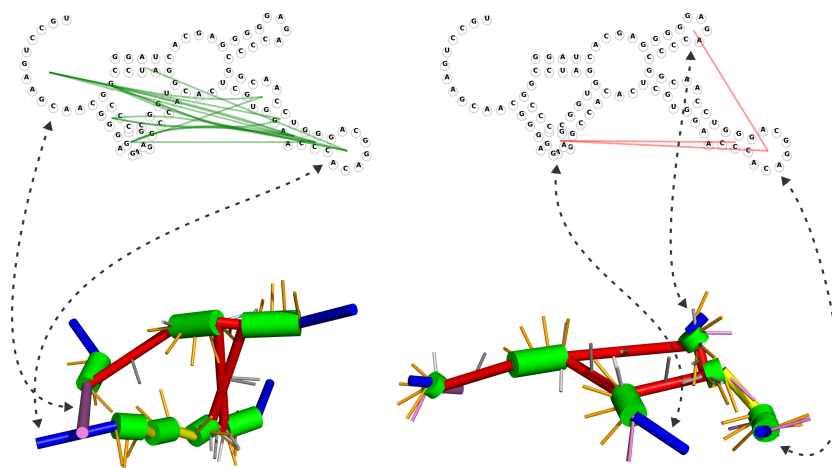


Figure 54: Using color to show correctly and incorrectly predicted adjacencies in the SAM-I/IV riboswitch (PDB ID: 4L81 [176]). The secondary and tertiary structures on the left correspond to the crystal structure whereas those on the right are from the cluster with the lowest ACC. The dashed lines connect the equivalent 2D and 3D elements. Green lines show correctly predicted adjacencies and red lines show incorrectly predicted adjacencies.

the quality of our prediction method, help to design better energy terms to improve it, or to look for biological implications in the predictions of unknown structures.

16.3 Showing multiple predictions using small multiples

The quantity of predicted structures precludes their simultaneous display on a typical computer monitor. We thus wish to select and display some subset which reflects the entire population. One possible way to do this is to simply pick a random subset of structures. Another is to cluster all of the predicted structures into some number of clusters (e.g. the number of structures we wish to display) and then display the centroids. Figure 55 shows examples of both approaches. The native (crystal) structure is shown in the upper left corners of the two sets of structures. The results are superficially similar, but this may simply be attributed to the order in which they are displayed (according to their ACC measure). Both methods of selecting structures reflect the underlying ensemble of coarse-grain models. They both contain two well predicted models and a number of not so well predicted models. We can see from the long-range interaction overlay, that there is a tendency to over-predict hairpin loop-hairpin loop interactions and to under-predict interactions with the 3' unpaired region.

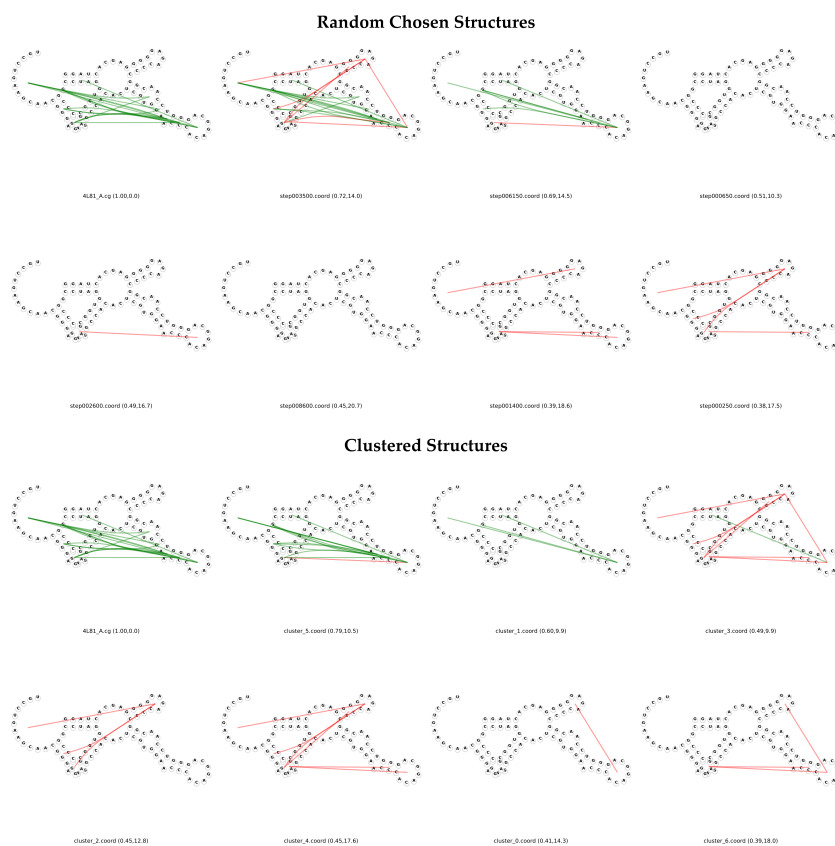


Figure 55: Small multiples of a random sample of structures (top) and the cluster centers of 7 clusters generated using k-means clustering. The native structure is shown in the upper left corner of both the random and clustered structures. The small, barely legible text below each diagram contains the filename of the 3D model it was generated from, its *ACC* and *RMSD*. The interactive version of this plot allows users to zoom in and examine this information in more detail.

16.4 *Summary and further work*

Expanding our RNA secondary structure visualization tool and augmenting it with links between adjacent elements allows us to abstract information from the tertiary structure and place it in two dimensions. This informs us of two important properties of the sampled ensemble: how well we predict structures as whole and where we predict correct and incorrect adjacencies. Combined, these properties show us how to improve our prediction methods. In the case shown in Figure 55, we need an energy term to pull hairpins to the 3' unpaired region. We also need to investigate why the wrong hairpins are brought together and perhaps introduce a check to make sure that a single hairpin is not involved in more than one other interaction. Conversely, one may ask why the interactions predicted in our model don't occur in reality. By zooming in and reading which structure the diagram came from, we can open it in a 3D viewer and examine it in more detail. The overview shown in this view thus enables the viewer to judge the quality of an ensemble of predicted structures and to identify and investigate specific failings (or successes) of the tertiary structure prediction.

This chapter, along with the dot-struct plot (Chapter 14), and the DR. FORNA cotranscriptional folding visualization (Chapter 15) are presented in this thesis to not only demonstrate new ways of looking at multiple RNA structures, but to highlight the flexibility of our JavaScript / D3.js RNA secondary structure display container, FORNAC. Its ability to display supplementary information such as additional links as well as its easy embedding in more complex visualizations make it an ideal container for showing and sharing RNA secondary structure in a relevant context. This, we hope, will advance the long tradition of displaying RNA secondary structure to illustrate the output of a prediction tool, explain a biological process, or to share auxiliary information.

CONCLUSION

This thesis has introduced two complementary tracks in the disciplines of RNA tertiary structure prediction and secondary structure visualization. The former not only introduces a coarse-grain model parameterized by the locations of the helices and hairpin loops, but also an adaptive rejection-based sampling technique for generating ensembles of structures whose features conform to predefined distributions. This enables us, for example, to sample a set of structures whose **ROGs** conform to the **ROGs** observed in solved structures of a similar size. Other features, such as the distance between hairpin loops or the orientation of potential A-minor interaction donors and receptors with respect to each other, can also be measured and sampled. By combining multiple different descriptors of tertiary RNA structure, we can sample ensembles of conformations whose measured features resemble those of real structures.

Conformations on a local level are sampled by assembling fragments of structures solved using X-Ray crystallography or predicted using the high-resolution FARFAR program. This provides a high level of confidence that the sampled loop conformations resemble those of real RNAs. Conversely, our lack of sequence dependence provides room for improvement in constraining the potential parameters available to certain loops in the structure. We have shown that including the correct parameters for just interior loops can markedly increase the prediction accuracy (Section 11.4). While we lack knowledge of the correct parameters, the use of predicted motifs for interior loops also leads to an improvement in prediction quality (Section 11.3). An improved methodology for selecting which motif predictions to include as parameters constraints and how often to constrain loops should alleviate the problem of incorrectly predicted motifs worsening the structure predictions.

When prediction methods yield poor results, their developers try to isolate the cause of the failure. In the case of tertiary structure prediction, two components are always suspected: the energy function and the sampling method. We have tried to improve upon the sampling method by exploring a wider variety of conformations than other methods. We have succeeded, insofar as we can sample a wider range of **RMSD** values than some of the best other methods [87] (Chapter 6). This, however, is but a bitter victory. Our results indicate that even when using a perfect energy function, we still fail to sample native or near native states for some of the larger structures. This is in spite of the fact that there are fragments present in the database which come from the target structures. Clearly, a

more sophisticated sampling approach is necessary. Cheap improvements may be obtained by using replica-exchange, or multicanonical MCMC sampling.

It is more likely, however, that a more thorough and methodical approach will be necessary. Excluded volume and loop closure present two major obstacles to generating valid conformations. Resampling entire multiloop regions (as opposed to just single segments) and adding pseudo-energies to guide loop closure may greatly increase the rate at which new fragments are accepted. An artificial database of fragments can be created starting with only those found in the native structure and gradually expanded to characterize where and why our sampling begins to fail to yield native structures when using a *cheating* energy. Building on top of this, we could use an extremely slow but accurate method to build up a high fidelity database of predicted fragments for use in assembling larger structures.

The second track, secondary structure visualization, has introduced new tools for interactively displaying RNA's secondary structure. The first and foremost of these tools, FORNA [86] (Chapter 7), has freed RNA visualization from any software dependencies barring an internet-connected computer and a web browser. It enables unprecedented flexibility in interactively rearranging not only the layout, but also editing the structure itself. The ability to extract and display the secondary structure of a PDB file rounds out the set of key features that have made this a worthy tool for the exploration and dissemination of RNA secondary structure.

Just as importantly, the JavaScript components for laying out RNA structure can be incorporated into more complex layouts that integrate additional information. Concentration can be encoded in the size of each diagram. Coloring can indicate conservation. The layouts can be dynamically resized to show the relative abundance of various structures over the course of transcription. Overlays can be added to show tertiary structure interactions. At the time of writing, the layouts available were being integrated into a tool to explore the folding landscape of RNAs. The component's ease of use, versatility, and clarity make it an ideal tool to display different characteristics of an RNA molecule within the context of its structure.

We hope that the work on RNA tertiary structure prediction and RNA secondary structure visualization will pave the way for the creation of ever more sophisticated tools to analyze the complex ways that RNA folds and interacts within human cells while simultaneously making it easier to abstract and clarify the output to make it easier to analyze and disseminate.

Part V

APPENDIX

This part of the document contains reference material that may be useful for the reader to refer back to. Included are the names and illustrations of all of the [PDB](#) structures used in benchmarking ERNWIN, as well as the chemical structures of the RNA nucleotides annotated with labels describing their representation in [PDB](#) files. Also included is the license for a figure reproduced in the DR. FORNA chapter [15](#).

EXTRA RELEVANT INFORMATION

PDB files

The line below shows a typical line in a [PDB](#) file describing an RNA molecule.

```
ATOM      73  C5'   U A 106      41.907  43.818  94.541  0.00 82.28      C
```

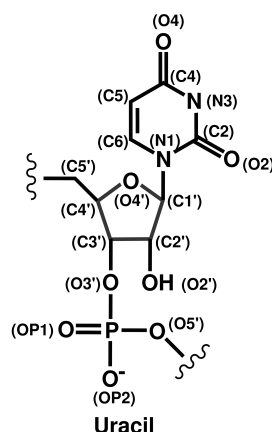
The values on this line are described in the table below:

ATOM	Indicates that this line describes a normal atom.
73	The number of the atom in this model.
C5'	The identity of the atom (See below).
U	The type of the residue (Uracil, in this case).
A	The chain identifier.
106	The residue number, which doesn't necessarily have to start with 1, nor does it have to be sequential over the length of the chain (i.e. some residues can be skipped).
41.907	The x-coordinate of the atom.
43.818	The y-coordinate of the atom.
94.541	The z-coordinate of the atom.
0.00	The occupancy, an indicator of alternate conformations.
82.28	The B-factor, an indicator of flexibility or confidence in the position of the atom. Higher values indicate that that region of the molecule may move around, making it harder to pin down an exact location.
C	The type of molecule.
(not shown)	The charge on the atom

In our calculations, we only use the identity and location values (i.e. the values before the occupancy).

Atoms in [PDB](#) file are identified according to their identity and position within the molecule. They are commonly referenced in literature to identify functional sites. Figure 56 shows the structures and atom names of each of the four different RNA nucleotides.

Pyrimidines



Purines

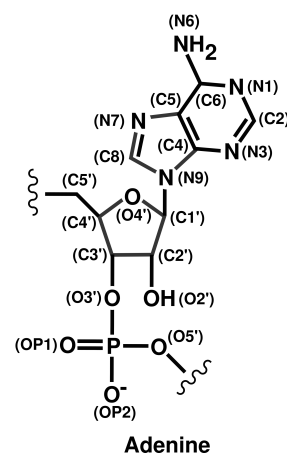
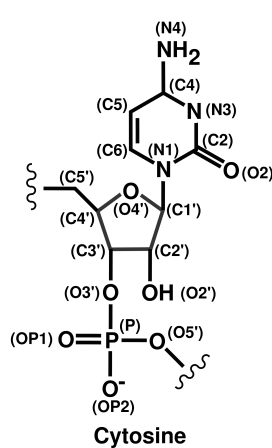
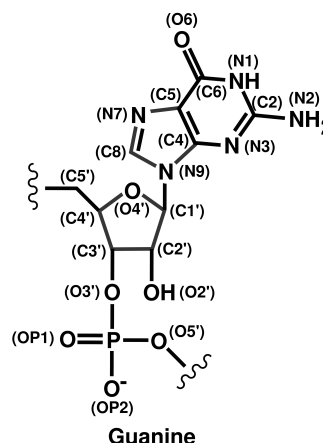


Figure 56: The names of the atoms in each of the four different nucleotides as labeled in a typical PDB file.

Local structure database construction flowchart

In an ideal world, all of the data to reproduce the results of this thesis would be generated using a single command. Unfortunately, in our real world, such a command requires a lot of prerequisites to function. In the interest of brevity, Figure 57 provides a high level overview of how the parameters (statistics) used by ERNWIN are generated.

The structure test set

Table 8 provides a list of all of the structures that were used in testing the performance of ERNWIN. They were obtained by taking all of the structures between 60 and 500 nucleotides long from the BGSU list of non-

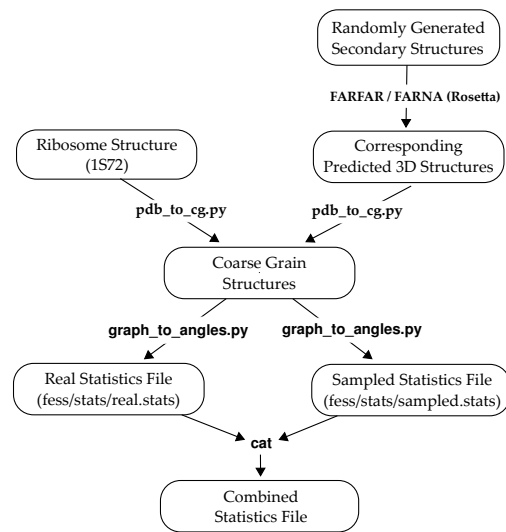
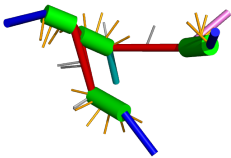
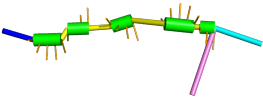
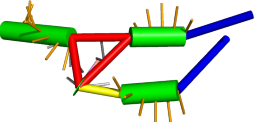


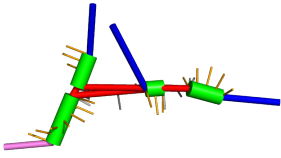
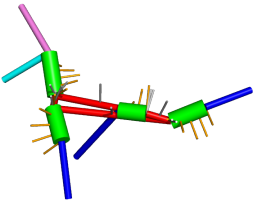
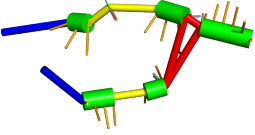
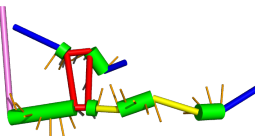
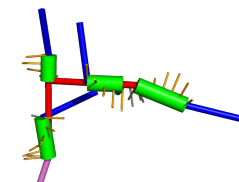
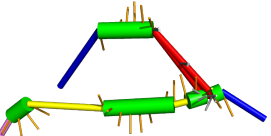
Figure 57: The pipeline for generating the statistics used for sampling coarse-grain structures with Erwin.

redundant RNA structures [101] and filtered to exclude entries containing multiple interacting RNAs and RNAs interacting with proteins.

PDB ID	Length (nt)	Coarse Grain Structure	Description
4PQV	68		XRN1-Resistant Flaviviral RNA [30]
1KXK	70		Domains 5 & 6 of Group II Intron [195]
1Y26	71		Adenine Riboswitch [161]

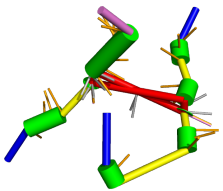
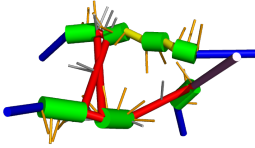
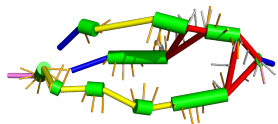
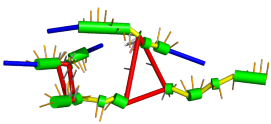
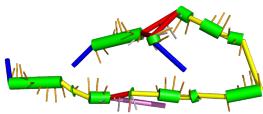
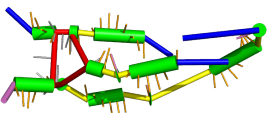
Continued on next page

Table 8 – continued from previous page

PDB ID	Length (nt)	Coarse Grain Structure	Description
2TRA	73		tRNA [185]
3CW5	75		Initiator tRNA [12]
2HOJ	78		Thi-Box Riboswitch [43]
3T4B	83		HCV IRES Pseudoknot Domain [15]
4P5J	83		tRNA Mimicking RNA [32]
4LVZ	89		THF Riboswitch [175]

Continued on next page

Table 8 – continued from previous page

PDB ID	Length (nt)	Coarse Grain Structure	Description
3GX5	94		SAM-I Riboswitch [123]
4L81	96		SAM Riboswitch [176]
2QBZ	153		M-box RNA [33]
1U9S	155		RNase P [95]
1GID	158		Group I Ribozyme Domain [28]
3DoU	161		Lysine Riboswitch (mRNA element) [54]

Continued on next page

Table 8 – continued from previous page

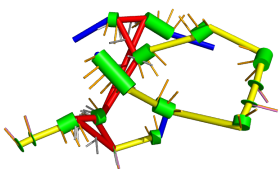
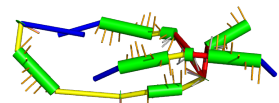
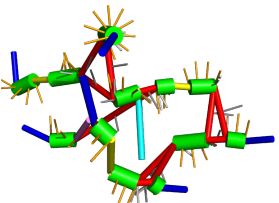
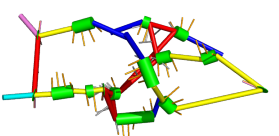
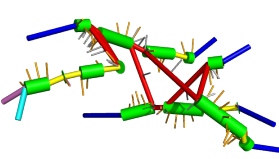
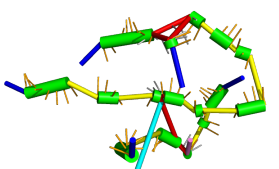
PDB ID	Length (nt)	Coarse Grain Structure	Description
4GXY	161		Adenosylcobalamin Riboswitch [136]
3DIR	172		Lysine Riboswitch [162]
4P9R	189		Lariat Capping Ribozyme [119]
4GMA	192		Adenosylcobalamin Riboswitch [81]
3DHS	215		RNase P [85]
1X8W	247		Tetrahymena Ribozyme [62]

Table 8: A list of all the structures used for benchmarking Erwin. The length of each structure is given in nucleotides (nt).

LICENSES

This chapter contains a copy of the license for the reproduction of Figure 49. All other reproduced figures either require no license (Proceedings of the National Academy of Sciences) or are licensed under a Creative Commons license. The reproductions of the papers in Chapters 6 and 7 are covered under the Creative Commons license.

1/5/2016

RightsLink Printable License

**ELSEVIER LICENSE
TERMS AND CONDITIONS**

Jan 05, 2016

This is a License Agreement between Peter S Kerpedjiev ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard, Langford Lane Kidlington, Oxford, OX5 1GB, UK
Registered Company Number	1982084
Customer name	Peter S Kerpedjiev
Customer address	Pezlgasse 47/21-22 Vienna, Wien 1170
License number	3782550837379
License date	Jan 05, 2016
Licensed content publisher	Elsevier
Licensed content publication	Journal of Molecular Biology
Licensed content title	Folding Kinetics of Large RNAs
Licensed content author	Michael Geis, Christoph Flamm, Michael T. Wolfinger, Andrea Tanzer, Ivo L. Hofacker, Martin Middendorf, Christian Mandl, Peter F. Stadler, Caroline Thurner
Licensed content date	23 May 2008
Licensed content volume number	379
Licensed content issue number	1
Number of pages	14
Start Page	160
End Page	173
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	Figure 4
Title of your thesis/dissertation	Seeing Secondary, Sampling Tertiary
Expected completion date	Jan 2016

1/5/2016

RightsLink Printable License

Estimated size (number of pages)	240
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 USD
VAT/Local Sales Tax	0.00 USD / 0.00 GBP
Total	0.00 USD

[Terms and Conditions](#)

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at <http://myaccount.copyright.com>).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.
3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."
4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.
5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)
6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.
9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

1/5/2016

RightsLink Printable License

10. **Indemnity:** You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. **No Transfer of License:** This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. **No Amendment Except in Writing:** This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. **Objection to Contrary Terms:** Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. **Revocation:** Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation:** This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.

16. **Posting licensed content on any Website:** The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at <http://www.sciencedirect.com/science/journal/xxxxx> or the Elsevier homepage for books at <http://www.elsevier.com>; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu. Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at <http://www.elsevier.com>. All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.

17. **For journal authors:** the following clauses are applicable in addition to the above:

Preprints:

A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles

1/5/2016

RightsLink Printable License

however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
 - o via their non-commercial person homepage or blog
 - o by updating a preprint in arXiv or RePEc with the accepted manuscript
 - o via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
 - o directly by providing copies to their students or to research collaborators for their personal use
 - o for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- after the embargo period
 - o via non-commercial hosting platforms such as their institutional repository
 - o via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.

Policies for sharing publishing journal articles differ for subscription and gold open access articles:

Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.

Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.

Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a [CrossMark logo](#), the end user license, and a DOI link to the formal publication on ScienceDirect.

1/5/2016

RightsLink Printable License

Please refer to Elsevier's [posting policy](#) for further information.

18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.

19. **Thesis/Dissertation:** If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions

You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our [open access license policy](#) for more information.

Terms & Conditions applicable to all Open Access articles published with Elsevier:

Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.

The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.

If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.

Additional Terms & Conditions applicable to each Creative Commons user license:

CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at <http://creativecommons.org/licenses/by/4.0>.

CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at <http://creativecommons.org/licenses/by-nc-sa/4.0>.

CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at <http://creativecommons.org/licenses/by-nc-nd/4.0>. Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission

1/5/2016

RightsLink Printable License

from Elsevier and will be subject to a fee.

Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. Other Conditions:

v1.8

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

BIBLIOGRAPHY

- [1] Neema Agrawal, PVN Dasaradhi, Asif Mohammed, Pawan Malhotra, Raj K Bhatnagar, and Sunil K Mukherjee. RNA interference: biology, mechanism, and applications. *Microbiology and Molecular Biology Reviews*, 67(4):657–685, 2003.
- [2] Paul F Agris, Franck AP Vendeix, and William D Graham. tRNA's wobble decoding of the genome: 40 years of modification. *Journal of Molecular Biology*, 366(1):1–13, 2007.
- [3] Victor Ambros. The functions of animal microRNAs. *Nature*, 431(7006):350–355, 2004.
- [4] Maciej Antczak, Tomasz Zok, Mariusz Popena, Piotr Lukasiak, Ryszard W Adamiak, Jacek Blazewicz, and Marta Szachniuk. RNAdbee - a webserver to derive secondary structures from pdb files of knotted and unknotted RNAs. *Nucleic Acids Research*, 42(W1):W368–W372, 2014.
- [5] Alexandra H Antonioli, Jesse C Cochrane, Sarah V Lipchock, and Scott A Strobel. Plasticity of the RNA kink turn structural motif. *RNA*, 16(4):762–768, 2010.
- [6] Sandro F Ataide, Nikolaus Schmitz, Kuang Shen, Ailong Ke, Shou Shan, Jennifer A Doudna, and Nenad Ban. The crystal structure of the signal recognition particle in complex with its receptor. *Science*, 331(6019):881–886, 2011.
- [7] Sigrid D Auweter, Florian C Oberstrass, and Frederic H-T Allain. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Research*, 34(17):4943–4959, 2006.
- [8] Xiao-chen Bai, Israel S Fernandez, Greg McMullan, and Sjors HW Scheres. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *Elife*, 2:e00461, 2013.
- [9] Maximillian H Bailor, Anthony M Mustoe, Charles L Brooks III, and Hashim M Al-Hashimi. 3D maps of RNA interhelical junctions. *Nature Protocols*, 6(10):1536–1545, 2011.
- [10] Nenad Ban, 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(5481):905–920, 2000.

- [11] Josh Barnes and Piet Hut. A hierarchical $O(N \log N)$ force-calculation algorithm. *Nature*, 324:446–449, 1986.
- [12] Pierre Barraud, Emmanuelle Schmitt, Yves Mechulam, Frédéric Dardel, and Carine Tisné. A unique conformation of the anticodon stem-loop is associated with the capacity of tRNA^{fMet} to initiate protein synthesis. *Nucleic Acids Research*, 36(15):4894–4901, 2008.
- [13] P Benas, G Bec, G Keith, R Marquet, C Ehresmann, B Ehresmann, and P Dumas. The crystal structure of HIV reverse-transcription primer tRNA (Lys, 3) shows a canonical anticodon loop. *RNA*, 6(10):1347–1355, 2000.
- [14] Frances C Bernstein, Thomas F Koetzle, Graheme JB Williams, Edgar F Meyer, Michael D Brice, John R Rodgers, Olga Kennard, Takehiko Shimanouchi, and Mitsuo Tasumi. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Archives of Biochemistry and Biophysics*, 185(2):584–591, 1978.
- [15] Katherine E Berry, Shruti Waghray, Stefanie A Mortimer, Yun Bai, and Jennifer A Doudna. Crystal structure of the HCV IRES central domain reveals strategy for start-codon positioning. *Structure*, 19(10):1456–1466, 2011.
- [16] Eckart Bindewald, Robert Hayes, Yaroslava G Yingling, Wojciech Kasprzak, and Bruce A Shapiro. RNAJunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign. *Nucleic Acids Research*, 36(suppl 1):D392–D397, 2008.
- [17] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.
- [18] Konstantin Bokov and Sergey V Steinberg. A hierarchical model for evolution of 23S ribosomal RNA. *Nature*, 457(7232):977, 2009.
- [19] Linda Bonen and Jörg Vogel. The ins and outs of group II introns. *TRENDS in Genetics*, 17(6):322–331, 2001.
- [20] Wouter Boomsma and Thomas Hamelryck. Full cyclic coordinate descent: Solving the protein loop closure problem in $C\alpha$ space. *BMC Bioinformatics*, 6(1):159, 2005.
- [21] Wouter Boomsma, Kanti V Mardia, Charles C Taylor, Jesper Ferkinghoff-Borg, Anders Krogh, and Thomas Hamelryck. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*, 105(26):8932–8937, 2008.

- [22] Mélanie Boudard, Julie Bernauer, Dominique Barth, Johanne Cohen, and Alain Denise. GARN: Sampling RNA 3D Structure Space with Game Theory and Knowledge-Based Scoring Strategies. *PLoS One*, 10(8):e0136444, 2015.
- [23] John Boyle, George T Robillard, and Sung-Hou Kim. Sequential folding of transfer RNA: a nuclear magnetic resonance study of successively longer tRNA fragments with a common 5' end. *Journal of Molecular Biology*, 139(4):601–625, 1980.
- [24] Robert E Bruccoleri and Gerhard Heinrich. An improved algorithm for nucleic acid secondary structure display. *Computer Applications in the Biosciences: CABIOS*, 4(1):167–173, 1988.
- [25] Samuel E Butcher and Anna Marie Pyle. The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Accounts of Chemical Research*, 44(12):1302–1311, 2011.
- [26] Yanga Byun and Kyungsook Han. PseudoViewer: web application and web service for visualizing RNA pseudoknots and secondary structures. *Nucleic Acids Research*, 34(suppl 2):W416–W422, 2006.
- [27] Adrian A Canutescu and Roland L Dunbrack. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science*, 12(5):963–972, 2003.
- [28] Jamie H Cate, Anne R Gooding, Elaine Podell, Kaihong Zhou, et al. Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science*, 273(5282):1678, 1996.
- [29] Michael Y Chao, Ming-Chung Kan, and Sue Lin-Chao. RNAII transcribed by IPTG-induced T7 RNA polymerase is non-functional as a replication primer for ColE1-type plasmids in *Escherichia coli*. *Nucleic Acids Research*, 23(10):1691–1695, 1995.
- [30] Erich G Chapman, David A Costantino, Jennifer L Rabe, Stephanie L Moon, Jeffrey Wilusz, Jay C Nix, and Jeffrey S Kieft. The structural basis of pathogenic subgenomic flavivirus RNA (sfRNA) production. *Science*, 344(6181):307–310, 2014.
- [31] Fred E Cohen and Michael JE Sternberg. On the prediction of protein structure: the significance of the root-mean-square deviation. *Journal of Molecular Biology*, 138(2):321–333, 1980.
- [32] Timothy M Colussi, David A Costantino, John A Hammond, Grant M Ruehle, Jay C Nix, and Jeffrey S Kieft. The structural

- basis of transfer RNA mimicry and conformational plasticity by a viral RNA. *Nature*, 511(7509):366–369, 2014.
- [33] Charles E Dann, Catherine A Wakeman, Cecelia L Sieling, Stephanie C Baker, Irnov Irnov, and Wade C Winkler. Structure and mechanism of a metal-sensing regulatory RNA. *Cell*, 130(5): 878–892, 2007.
- [34] Kévin Darty, Alain Denise, and Yann Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15):1974, 2009.
- [35] Rhiju Das and David Baker. Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences*, 104(37):14664–14669, 2007.
- [36] Rhiju Das, John Karanicolas, and David Baker. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature Methods*, 7(4):291–294, 2010.
- [37] Peter De Rijk and Rupert De Wachter. RnaViz, a program for the visualisation of RNA secondary structure. *Nucleic Acids Research*, 25(22):4679–4684, 1997.
- [38] Peter De Rijk, Jan Wuyts, and Rupert De Wachter. RnaViz 2: an improved representation of RNA secondary structure. *Bioinformatics*, 19(2):299–300, 2003.
- [39] Feng Ding, Shantanu Sharma, Poornima Chalasani, Vadim V Demidov, Natalia E Broude, and Nikolay V Dokholyan. Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, 14(6):1164–1173, 2008.
- [40] Jean-Pierre Dumas and Jacques Ninio. Efficient algorithms for folding and comparing nucleic acid sequences. *Nucleic Acids Research*, 10(1):197–206, 1982.
- [41] Olivier Duss, Erich Michel, Maxim Yulikov, Mario Schubert, Gunnar Jeschke, and Frédéric H-T Allain. Structural basis of the non-coding RNA RsmZ acting as a protein sponge. *Nature*, 509(7502): 588–592, 2014.
- [42] Sean R Eddy. How do RNA folding algorithms work? *Nature Biotechnology*, 22(11):1457–1458, 2004.
- [43] Thomas E Edwards and Adrian R Ferré-D’Amaré. Crystal structures of the thi-box riboswitch bound to thiamine pyrophosphate analogs reveal adaptive RNA-small molecule recognition. *Structure*, 14(9):1459–1468, 2006.

- [44] Sebastian M Fica, Nicole Tuttle, Thaddeus Novak, Nan-Sheng Li, Jun Lu, Prakash Koodathingal, Qing Dai, Jonathan P Staley, and Joseph A Piccirilli. RNA catalyses nuclear pre-mRNA splicing. *Nature*, 2013.
- [45] Sebastian M Fica, Melissa A Mefford, Joseph A Piccirilli, and Jonathan P Staley. Evidence for a Group II intron-like catalytic triplex in the spliceosome. *Nature Structural & Molecular Biology*, 2014.
- [46] Ivo Fierro-Monti and Michael B Mathews. Proteins binding to duplexed RNA: one motif, multiple functions. *Trends in Biochemical Sciences*, 25(5):241–246, 2000.
- [47] Andrew Fire, 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(6669):806–811, 1998.
- [48] Christoph Flamm, Walter Fontana, Ivo L Hofacker, and Peter Schuster. RNA folding at elementary step resolution. *RNA*, 6(03):325–338, 2000.
- [49] Christoph Flamm, Ivo L Hofacker, Peter F Stadler, and Michael T Wolfinger. Barrier trees of degenerate landscapes. *Zeitschrift für Physikalische Chemie International Journal of Research in Physical Chemistry and Chemical Physics*, 216(2/2002):155, 2002.
- [50] G David Forney Jr. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.
- [51] George E Fox and Carl R Woese. 5S RNA secondary structure. *Nature*, 256(5517):505–507, 1975.
- [52] Jes Frellsen, Ida Moltke, Martin Thiim, Kanti V Mardia, Jesper Ferkinghoff-Borg, and Thomas Hamelryck. A probabilistic model of RNA conformational space. *PLoS Comput Biol*, 5(6):e1000406–e1000406, 2009.
- [53] Rees F Garmann, Ajaykumar Gopal, Shreyas S Athavale, Charles M Knobler, William M Gelbart, and Stephen C Harvey. Visualizing the global secondary structure of a viral RNA genome with cryo-electron microscopy. *RNA*, 21(5):877–886, 2015.
- [54] Andrew D Garst, Annie Héroux, Robert P Rambo, and Robert T Batey. Crystal structure of the lysine riboswitch regulatory mRNA element. *Journal of Biological Chemistry*, 283(33):22347–22351, 2008.

- [55] Michael Geis, Christoph Flamm, Michael T Wolfinger, Andrea Tanzer, Ivo L Hofacker, Martin Middendorf, Christian Mandl, Peter F Stadler, and Caroline Thurner. Folding kinetics of large RNAs. *Journal of Molecular Biology*, 379(1):160–173, 2008.
- [56] Patrick Gendron, Sébastien Lemieux, and François Major. Quantitative analysis of nucleic acid three-dimensional structures. *Journal of Molecular Biology*, 308(5):919–936, 2001.
- [57] Zoubin Ghahramani. Learning dynamic Bayesian networks. In *Adaptive processing of sequences and data structures*, pages 168–197. Springer, 1998.
- [58] Tina Glisovic, Jennifer L Bachorik, Jeongsik Yong, and Gideon Dreyfuss. RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters*, 582(14):1977–1986, 2008.
- [59] Nicholas J Green, Frank J Grundy, and Tina M Henkin. The T box mechanism: tRNA as a regulatory molecule. *FEBS letters*, 584(2):318–324, 2010.
- [60] Alexander P Gultyaev, FHD Van Batenburg, and Cornelis WA Pleij. The computer simulation of RNA folding pathways using a genetic algorithm. *Journal of Molecular Biology*, 250(1):37–51, 1995.
- [61] Peter Güntert, Ch Mumenthaler, and Kurt Wüthrich. Torsion angle dynamics for NMR structure calculation with the new program dyana. *Journal of Molecular Biology*, 273(1):283–298, 1997.
- [62] Feng Guo, Anne R Gooding, and Thomas R Cech. Structure of the *Tetrahymena* ribozyme: base triple sandwich and metal ion at the active site. *Molecular Cell*, 16(3):351–362, 2004.
- [63] Estella M Gustilo, Franck AP Vendeix, and Paul F Agris. tRNA's modifications bring order to gene expression. *Current Opinion in Microbiology*, 11(2):134–140, 2008.
- [64] KB Hall. Interaction of RNA hairpins with the human U1A N-terminal RNA binding domain. *Biochemistry*, 33(33):10076–10088, 1994.
- [65] Thomas Hamelryck, Mikael Borg, Martin Paluszewski, Jonas Paulsen, Jes Frellsen, Christian Andreetta, Wouter Boomsma, Sandro Bottaro, and Jesper Ferkinghoff-Borg. Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PloS One*, 5(11):e13714, 2010.

- [66] Kyungsook Han and Yanga Byun. PSEUDOVIEWER2: visualization of RNA pseudoknots of any type. *Nucleic Acids Research*, 31(13):3432–3440, 2003.
- [67] Yaser Hashem, Amedee Des Georges, Jie Fu, Sarah N Buss, Fabrice Jossinet, Amy Jobe, Qin Zhang, Hstau Y Liao, Robert A Grassucci, Chandrajit Bajaj, et al. High-resolution cryo-electron microscopy structure of the *Trypanosoma brucei* ribosome. *Nature*, 494(7437):385–389, 2013.
- [68] N Hecker, M Kahlscheuer, Kerpedjiev P, M Kucharik, J Gorodkin, P Stadler, IL Hofacker, Walter N, and J Qin. FRETtranslator: translating FRET traces into RNA structural pathways. *Bioinformatics*.
- [69] Susan L Heilman-Miller and Sarah A Woodson. Effect of transcription on folding of the *Tetrahymena* ribozyme. *RNA*, 9(6):722–733, 2003.
- [70] Miyoko Higuchi, Stefan Maas, Frank N Single, Jochen Hartner, Andrei Rozov, Nail Burnashev, Dirk Feldmeyer, Rolf Sprengel, and Peter H Seeburg. Point mutation in an ampa receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature*, 406(6791):78–81, 2000.
- [71] Ivo L Hofacker, Walter Fontana, Peter F Stadler, L Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, 1994.
- [72] Ivo L Hofacker, Christoph Flamm, Christian Heine, Michael T Wolfinger, Gerik Scheuermann, and Peter F Stadler. BarMap: RNA folding on dynamic energy landscapes. *RNA*, 16(7):1308–1316, 2010.
- [73] Danny Holten and Jarke J Van Wijk. Force-Directed Edge Bundling for Graph Visualization. In *Computer Graphics Forum*, volume 28, pages 983–990. Wiley Online Library, 2009.
- [74] Joseph Douglas Horton. A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM Journal on Computing*, 16(2):358–366, 1987.
- [75] Brian Houck-Loomis, Michael A Durney, Carolina Salguero, Nee-laabh Shankar, Julia M Nagle, Stephen P Goff, and Victoria M D’Souza. An equilibrium-dependent retroviral mRNA switch regulates translational recoding. *Nature*, 480(7378):561–564, 2011.
- [76] Yuanpeng J Huang, Robert Powers, and Gaetano T Montelione. Protein NMR recall, precision, and F-measure scores (RPF scores):

- structure quality assessment measures based on information retrieval statistics. *Journal of the American Chemical Society*, 127(6):1665–1674, 2005.
- [77] Yuanpeng J Huang, Binchen Mao, James M Aramini, and Gaetano T Montelione. Assessment of template-based protein structure predictions in CASP10. *Proteins: Structure, Function, and Bioinformatics*, 82(S2):43–56, 2014.
- [78] Danny Incarnato, Francesco Neri, Francesca Anselmi, and Salvatore Oliviero. Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. *Genome Biology*, 15(10):491, 2014.
- [79] Hervé Isambert and Eric D Siggia. Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proceedings of the National Academy of Sciences*, 97(12):6515–6520, 2000.
- [80] Elizabeth A Jares-Erijman and Thomas M Jovin. FRET imaging. *Nature Biotechnology*, 21(11):1387–1395, 2003.
- [81] James E Johnson Jr, Francis E Reyes, Jacob T Polaski, and Robert T Batey. B12 cofactors directly stabilize an mRNA regulatory switch. *Nature*, 492(7427):133–137, 2012.
- [82] Magdalena A Jonikas, Randall J Radmer, Alain Laederach, Rhiju Das, Samuel Pearlman, Daniel Herschlag, and Russ B Altman. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, 15(2):189–199, 2009.
- [83] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.
- [84] Alexei V Kazantsev, Angelika A Krivenko, Daniel J Harrington, Stephen R Holbrook, Paul D Adams, and Norman R Pace. Crystal structure of a bacterial ribonuclease P RNA. *Proceedings of the National Academy of Sciences*, 102(38):13392–13397, 2005.
- [85] Alexei V Kazantsev, Angelika A Krivenko, and Norman R Pace. Mapping metal-binding sites in the catalytic domain of bacterial RNase P RNA. *RNA*, 15(2):266–276, 2009.
- [86] Peter Kerpedjiev, Stefan Hammer, and Ivo L Hofacker. Forna (force-directed RNA): Simple and effective online RNA secondary

- structure diagrams. *Bioinformatics*, 31(20):3377–3379, 2015. doi: 10.1093/bioinformatics/btv372.
- [87] Peter Kerpedjiev, Christian Höner Zu Siederdisen, and Ivo L Hofacker. Predicting RNA 3D structure using a coarse-grain helix-centered model. *RNA*, 21(6):1110–1121, 2015.
- [88] Michael Kertesz, Yue Wan, Elad Mazor, John L Rinn, Robert C Nutter, Howard Y Chang, and Eran Segal. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–107, 2010.
- [89] Namhee Kim, Christian Laing, Shereef Elmetwaly, Segun Jung, Jeremy Curuksu, and Tamar Schlick. Graph-based sampling for approximating global helical topologies of RNA. *Proceedings of the National Academy of Sciences*, 111(11):4079–4084, 2014.
- [90] Sebastian Kirchner and Zoya Ignatova. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. *Nature Reviews Genetics*, 16(2):98–112, 2015.
- [91] Wipapat Kladwang, Christopher C VanLang, Pablo Cordero, and Rhiju Das. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nature Chemistry*, 3(12):954–962, 2011.
- [92] Daniel J Klein, Thomas E Edwards, and Adrian R Ferré-D’Amaré. Cocystal structure of a class I preQ₁ riboswitch reveals a pseudoknot recognizing an essential hypermodified nucleobase. *Nature Structural & Molecular Biology*, 16(3):343–344, 2009.
- [93] DJ Klein, PB Moore, and TA Steitz. The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *Journal of Molecular Biology*, 340(1):141–177, 2004.
- [94] Fred Russell Kramer and Donald R Mills. Secondary structure formation during RNA synthesis. *Nucleic Acids Research*, 9(19):5109–5124, 1981.
- [95] Andrey S Krasilnikov, Yinghua Xiao, Tao Pan, and Alfonso Mondragón. Basis for structural diversity in homologous RNAs. *Science*, 306(5693):104–107, 2004.
- [96] Victor Kunin, Rotem Sorek, and Philip Hugenholtz. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol*, 8(4):R61, 2007.
- [97] Daniel Lai, Jeff R Proctor, and Irmtraud M Meyer. On the importance of cotranscriptional RNA structure formation. *RNA*, 19(11):1461–1473, 2013.

- [98] Christian Laing, Segun Jung, Namhee Kim, Shereef Elmetwaly, Mai Zahran, and Tamar Schlick. Predicting helical topologies in RNA junctions as tree graphs. *PLoS One*, 8(8):e71947, 2013.
- [99] Alexis Lamiable, Franck Quessette, Sandrine Vial, Dominique Barth, and Alain Denise. An algorithmic game-theory approach for coarse-grain prediction of RNA 3D structure. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 10(1):193–199, 2013.
- [100] Sébastien Lemieux and François Major. Automated extraction and classification of RNA tertiary structure cyclic motifs. *Nucleic Acids Research*, 34(8):2340–2346, 2006.
- [101] Neocles B Leontis and Craig L Zirbel. *Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking*. Springer, 2012.
- [102] Birgit TU Lewicki, Tõnu Margus, Jaanus Remme, and Knud H Nierhaus. Coupling of rRNA transcription and ribosomal assembly in vivo: Formation of active ribosomal subunits in *Escherichia coli* requires transcription of rRNA genes by host RNA polymerase which cannot be replaced by bacteriophage T7 RNA polymerase. *Journal of Molecular Biology*, 231(3):581–593, 1993.
- [103] Fan Li, Qi Zheng, Lee E Vandivier, Matthew R Willmann, Ying Chen, and Brian D Gregory. Regulatory impact of RNA secondary structure across the *Arabidopsis* transcriptome. *The Plant Cell*, 24(11):4346–4359, 2012.
- [104] Harvey F Lodish, Arnold Berk, S Lawrence Zipursky, Paul Mutsaers, David Baltimore, James Darnell, et al. *Molecular cell biology*, volume 4. Citeseer, 2000.
- [105] Ronny Lorenz, Stephan HF Bernhart, Christian Hoener Zu Siederdisen, Hakim Tafer, Christoph Flamm, Peter F Stadler, Ivo L Hofacker, et al. ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.
- [106] Xiang-Jun Lu and Wilma K Olson. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31(17):5108–5121, 2003.
- [107] Julius B Lucks, Stefanie A Mortimer, Cole Trapnell, Shujun Luo, Sharon Aviran, Gary P Schroth, Lior Pachter, Jennifer A Doudna, and Adam P Arkin. Multiplexed RNA structure characterization

- with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences*, 108(27):11063–11068, 2011.
- [108] Yuri L Lyubchenko, Luda S Shlyakhtenko, and Toshio Ando. Imaging of nucleic acids with atomic force microscopy. *Methods*, 54(2): 274–283, 2011.
- [109] Paul M MacDonald. bicoid mRNA localization signal: phylogenetic conservation of function and RNA secondary structure. *Development*, 110(1):161–171, 1990.
- [110] Maumita Mandal and Ronald R Breaker. Gene regulation by riboswitches. *Nature Reviews Molecular Cell Biology*, 5(6):451–463, 2004.
- [111] Marco Marcia and Anna Marie Pyle. Visualizing group II intron catalysis through the stages of splicing. *Cell*, 151(3):497–507, 2012.
- [112] Kanti V Mardia and Peter E Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009.
- [113] Nicholas R Markham and Michael Zuker. UNAFold. In *Bioinformatics*, pages 3–31. Springer, 2008.
- [114] Debora S Marks, Thomas A Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature Biotechnology*, 30(11):1072–1080, 2012.
- [115] David H Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911–940, 1999.
- [116] Craig C Mello and Darryl Conte. Revealing the world of RNA interference. *Nature*, 431(7006):338–342, 2004.
- [117] Edward J Merino, 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(12):4223–4231, 2005.
- [118] Ana C Messias and Michael Sattler. Structural basis of single-stranded RNA recognition. *Accounts of Chemical Research*, 37(5): 279–287, 2004.

- [119] Mélanie Meyer, Henrik Nielsen, Vincent Oliéric, Pierre Roblin, Steinar D Johansen, Eric Westhof, and Benoît Masquida. Speciation of a group I intron into a lariat capping ribozyme. *Proceedings of the National Academy of Sciences*, 111(21):7659–7664, 2014.
- [120] Francois Michel and Eric Westhof. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *Journal of Molecular Biology*, 216(3):585–610, 1990.
- [121] AA Mironov and VF Lebedev. A kinetic model of RNA folding. *BioSystems*, 30(1):49–56, 1993.
- [122] Andrei A Mironov, Lyudmila P Dyakonova, and Alexander E Kister. A kinetic approach to the prediction of RNA secondary structures. *Journal of Biomolecular Structure and Dynamics*, 2(5):953–962, 1985.
- [123] Rebecca K Montange, Estefanía Mondragón, Daria van Tyne, Andrew D Garst, Pablo Ceres, and Robert T Batey. Discrimination between closely related cellular metabolites by the SAM-I riboswitch. *Journal of Molecular Biology*, 396(3):761–772, 2010.
- [124] Stefanie A Mortimer, Mary Anne Kidwell, and Jennifer A Doudna. Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics*, 15(7):469–479, 2014.
- [125] Akihiro Nakaya, Kenjiro Taura, Kenji Yamamoto, and Akinori Yonezawa. Visualization of RNA secondary structures using highly parallel computers. *Computer Applications in the Biosciences: CABIOS*, 12(3):205–211, 1996.
- [126] Subba Rao Nallagatla, Christie N Jones, Saikat Kumar B Ghosh, Suresh D Sharma, Craig E Cameron, Linda L Spremulli, and Philip C Bevilacqua. Native tertiary structure and nucleoside modifications suppress tRNA's intrinsic ability to activate the innate immune sensor PKR. *PloS One*, 8(3):e57905, 2013.
- [127] Poul Nissen, Joseph A Ippolito, Nenad Ban, Peter B Moore, and Thomas A Steitz. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proceedings of the National Academy of Sciences*, 98(9):4899–4903, 2001.
- [128] Harry F Noller, JoAnn Kop, Virginia Wheaton, Jürgen Brosius, Robin R Gutell, Alexei M Kopylov, Ferdinand Dohme, Winship Herr, David A Stahl, Ramesh Gupta, et al. Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Research*, 9(22):6167–6189, 1981.

- [129] Ruth Nussinov and Ann B Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences*, 77(11):6309–6313, 1980.
- [130] Ruth Nussinov, George Pieczenik, Jerrold R Griggs, and Daniel J Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied mathematics*, 35(1):68–82, 1978.
- [131] Marc Parisien and Francois Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183):51–55, 2008.
- [132] Marc Parisien, José Almeida Cruz, Éric Westhof, and François Major. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, 15(10):1875–1885, 2009.
- [133] Eonyoung Park and Lynne E Maquat. Staufen-mediated mRNA decay. *Wiley Interdisciplinary Reviews: RNA*, 4(4):423–435, 2013.
- [134] Paul J Paukstelis, Jui-Hui Chen, Elaine Chase, Alan M Lambowitz, and Barbara L Golden. Structure of a tyrosyl-tRNA synthetase splicing factor bound to a group I intron RNA. *Nature*, 451(7174):94–97, 2008.
- [135] Debra A Peattie and Walter Gilbert. Chemical probes for higher-order structure in RNA. *Proceedings of the National Academy of Sciences*, 77(8):4679–4682, 1980.
- [136] Alla Peselis and Alexander Serganov. Structural insights into ligand binding and gene expression control by an adenosylcobalamin riboswitch. *Nature Structural & Molecular Biology*, 19(11):1182–1184, 2012.
- [137] Sonja Petkovic, Stefan Badelt, Stephan Block, Christoph Flamm, Mihaela Delcea, Ivo Hofacker, and Sabine Müller. Sequence-controlled RNA self-processing: computational design, biochemical analysis, and visualization by AFM. *RNA*, 2015.
- [138] Anton I Petrov, Craig L Zirbel, and Neocles B Leontis. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA*, 19(10):1327–1340, 2013.
- [139] Anton S Petrov, Chad R Bernier, Eli Hershkovits, Yuzhen Xue, Chris C Waterbury, Chiaolong Hsiao, Victor G Stepanov, Eric A Gaucher, Martha A Grover, Stephen C Harvey, et al. Secondary structure and domain architecture of the 23S and 5S rRNAs. *Nucleic Acids Research*, 41(15):7522–7535, 2013.

- [140] Anton S Petrov, Chad R Bernier, Chiaolong Hsiao, Ashlyn M Norris, Nicholas A Kovacs, Chris C Waterbury, Victor G Stepanov, Stephen C Harvey, George E Fox, Roger M Wartell, et al. Evolution of the ribosome at atomic resolution. *Proceedings of the National Academy of Sciences*, 111(28):10251–10256, 2014.
- [141] James M Pipas and James E McMahon. Method for predicting RNA secondary structure. *Proceedings of the National Academy of Sciences*, 72(6):2017–2021, 1975.
- [142] Mariusz Popena, Marta Szachniuk, Maciej Antczak, Katarzyna J Purzycka, Piotr Lukasiak, Natalia Bartol, Jacek Blazewicz, and Ryszard W Adamiak. Automated 3D structure composition for large RNAs. *Nucleic Acids Research*, page gks339, 2012.
- [143] Elizabeth A Proctor, Feng Ding, and Nikolay V Dokholyan. Discrete molecular dynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(1):80–92, 2011.
- [144] Jeff R Proctor and Irmtraud M Meyer. CoFold: an RNA secondary structure prediction method that takes co-transcriptional folding into account. *Nucleic Acids Research*, page gkt174, 2013.
- [145] Tomasz Puton, Lukasz P Kozlowski, Kristian M Rother, and Janusz M Bujnicki. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Research*, 41(7):4307–4323, 2013.
- [146] Prashanth Rangan, Benoît Masquida, Eric Westhof, and Sarah A Woodson. Assembly of core helices and rapid tertiary folding of a small bacterial group I ribozyme. *Proceedings of the National Academy of Sciences*, 100(4):1574–1579, 2003.
- [147] Debashish Ray, Hilal Kazan, Esther T Chan, Lourdes Peña Castillo, Sidharth Chaudhry, Shaheynoor Talukder, Benjamin J Blencowe, Quaid Morris, and Timothy R Hughes. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnology*, 27(7):667–670, 2009.
- [148] Vladimir Reinharz, François Major, and Jérôme Waldispühl. Towards 3D structure prediction of large RNA molecules: an integer programming framework to insert local 3D motifs in RNA secondary structure. *Bioinformatics*, 28(12):i207–i214, 2012.
- [149] Nicholas J Reiter, Amy Osterman, Alfredo Torres-Larios, Kerren K Swinger, Tao Pan, and Alfonso Mondragón. Structure of a bacterial ribonuclease P holoenzyme in complex with tRNA. *Nature*, 468(7325):784–789, 2010.

- [150] JD Robertus, Jane E Ladner, JT Finch, Daniela Rhodes, RS Brown, BFC Clark, and A Klug. Structure of yeast phenylalanine tRNA at 3Å resolution. *Nature*, 250:546–551, 1974.
- [151] Magdalena Rother, Kristian Rother, Tomasz Puton, and Janusz M Bujnicki. ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Research*, 39(10):4007–4022, 2011.
- [152] Jodi M Ryter and Steve C Schultz. Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. *The EMBO Journal*, 17(24):7505–7513, 1998.
- [153] Michael Sarver, Craig L Zirbel, Jesse Stombaugh, Ali Mokdad, and Neocles B Leontis. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *Journal of Mathematical Biology*, 56(1-2):215–252, 2008.
- [154] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. August 2010.
- [155] Susan J Schroeder, Jonathan W Stone, Samuel Bleckley, Theodore Gibbons, and Deborah M Mathews. Ensemble of secondary structures for encapsidated satellite tobacco mosaic virus RNA consistent with chemical probing and crystallography constraints. *Biophysical Journal*, 101(1):167–175, 2011.
- [156] Barbara S Schuwirth, 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(5749):827–834, 2005.
- [157] Charles D Schwieters, John J Kuszewski, Nico Tjandra, and G Marius Clore. The Xplor-NIH NMR molecular structure determination package. *Journal of Magnetic Resonance*, 160(1):65–73, 2003.
- [158] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.
- [159] Eric Selker and Charles Yanofsky. A phenylalanine tRNA gene from *Neurospora crassa*: conservation of secondary structure involving an intervening sequence. *Nucleic Acids Research*, 8(5):1033–1042, 1980.
- [160] Alexander Serganov and Evgeny Nudler. A decade of riboswitches. *Cell*, 152(1):17–24, 2013.

- [161] Alexander Serganov, 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(12):1729–1741, 2004.
- [162] Alexander Serganov, Lili Huang, and Dinshaw J Patel. Structural insights into amino acid binding and gene control by a lysine riboswitch. *Nature*, 455(7217):1263–1267, 2008.
- [163] Bruce A Shapiro, Lewis E Lipkin, and Jacob Maizel. An interactive technique for the display of nucleic acid secondary structure. *Nucleic Acids Research*, 10(21):7041–7052, 1982.
- [164] Bruce A Shapiro, Jacob Maizel, Lewis E Lipkin, Kathleen Currey, and Carol Whitney. Generating non-overlapping displays of nucleic acid secondary structure. *Nucleic Acids Research*, 12(1Part1):75–88, 1984.
- [165] Shantanu Sharma, Feng Ding, and Nikolay V Dokholyan. iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, 24(17):1951–1952, 2008.
- [166] Kinneret Shefer, Yogev Brown, Valentin Gorkovoy, Tamar Nussbaum, Nikolai B Ulyanov, and Yehuda Tzfati. A triple helix within a pseudoknot is a conserved and essential element of telomerase RNA. *Molecular and Cellular Biology*, 27(6):2130–2143, 2007.
- [167] Palak Sheth, Miguel Cervantes-Cervantes, Akhila Nagula, Christian Laing, and Jason TL Wang. Novel features for identifying A-minors in three-dimensional RNA molecules. *Computational Biology and Chemistry*, 47:240–245, 2013.
- [168] Naoki Shigi. Biosynthesis and functions of sulfur modifications in tRNA. *Frontiers in Genetics*, 5, 2014.
- [169] Kim T Simons, Charles Kooperberg, Enoch Huang, and David Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225, 1997.
- [170] Jesse Stombaugh, Craig L Zirbel, Eric Westhof, and Neocles B Leontis. Frequency and isostericity of RNA base pairs. *Nucleic Acids Research*, 37(7):2294–2312, 2009.

- [171] Scott A Strobel, Lori Ortoleva-Donnelly, Sean P Ryder, Jamie H Cate, and Eileen Moncoeur. Complementary sets of noncanonical base pairs mediate RNA helix packing in the group I intron active site. *Nature Structural & Molecular Biology*, 5(1):60–66, 1998.
- [172] Makio Tamura, Donna K Hendrix, Peter S Klosterman, Nancy RB Schimmelman, Steven E Brenner, and Stephen R Holbrook. SCOR: Structural Classification of RNA, version 2.0. *Nucleic Acids Research*, 32(suppl 1):D182–D184, 2004.
- [173] P Taylor, F Rixon, and U Desselberger. Rise per base pair in helices of double-stranded rotavirus RNA determined by electron microscopy. *Virus Research*, 2(2):175–182, 1985.
- [174] Pilar Tijerina, Sabine Mohr, and Rick Russell. DMS footprinting of structured RNAs and RNA–protein complexes. *Nature Protocols*, 2(10):2608–2623, 2007.
- [175] Jeremiah J Trausch and Robert T Batey. A disconnect between high-affinity binding and efficient regulation by antifolates and purines in the tetrahydrofolate riboswitch. *Chemistry & Biology*, 21(2):205–216, 2014.
- [176] Jeremiah J Trausch, Zhenjiang Xu, Andrea L Edwards, Francis E Reyes, Phillip E Ross, Rob Knight, and Robert T Batey. Structural basis for diversity in the SAM clan of riboswitches. *Proceedings of the National Academy of Sciences*, 111(18):6624–6629, 2014.
- [177] Donald E Tsai, David S Harper, and Jack D Keene. U1-snRNP-A protein selects a ten nucleotide consensus sequence from a degenerate RNA pool presented in various structural contexts. *Nucleic Acids Research*, 19(18):4931–4936, 1991.
- [178] Douglas H Turner and David H Mathews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, page gkp892, 2009.
- [179] Jan B Valentin, Christian Andreetta, Wouter Boomsma, Sandro Bottaro, Jesper Ferkinghoff-Borg, Jes Frellsen, Kanti V Mardia, Pengfei Tian, and Thomas Hamelryck. Formulation of probabilistic models of protein structure in atomic detail using the reference ratio method. *Proteins: Structure, Function, and Bioinformatics*, 82(2):288–299, 2014.
- [180] Yue Wan, Michael Kertesz, Robert C Spitale, Eran Segal, and Howard Y Chang. Understanding the transcriptome through RNA structure. *Nature Reviews Genetics*, 12(9):641–655, 2011.

- [181] Michael S Waterman and Temple F Smith. RNA secondary structure: A complete mathematical analysis. *Mathematical Biosciences*, 42(3):257–266, 1978.
- [182] James D Watson, Tania A Baker, Stephen P Bell, Gann Alexander, Levine Michael, and Richar Losick. *Molecular Biology of the Gene*. New York: WA Benjamin, Inc., 2 edition, 1970.
- [183] Kevin M Weeks. Advances in RNA structure analysis by chemical probing. *Current Opinion in Structural Biology*, 20(3):295–304, 2010.
- [184] Zasha Weinberg and Ronald R Breaker. R2R-software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics*, 12(1):3, 2011.
- [185] E Westhof, P Dumas, and D Moras. Restrained refinement of two crystalline forms of yeast aspartic acid and phenylalanine transfer RNA crystals. *Acta Crystallographica Section A: Foundations of Crystallography*, 44(2):112–124, 1988.
- [186] Kathryn A Whitehead, Robert Langer, and Daniel G Anderson. Knocking down barriers: advances in siRNA delivery. *Nature Reviews Drug Discovery*, 8(2):129–138, 2009.
- [187] Kay C Wiese, Edward Glen, and Anna Vasudevan. jViz.Rna-A Java tool for RNA secondary structure visualization. *NanoBioscience, IEEE Transactions on*, 4(3):212–218, 2005.
- [188] Carl R Woese and Norman R Pace. Probing RNA Structure, Function, and History by Comparative Analysis. *Cold Spring Harbor Monograph Archive*, 24:91–117, 1993.
- [189] CR Woese, LJ Magrum, R Gupta, RB Siegel, DA Stahl, J Kop, N Crawford, R Brosius, R Gutell, JJ Hogan, et al. Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Research*, 8(10):2275–2294, 1980.
- [190] Sarah A Woodson. Compact intermediates in RNA folding. *Annual Review of Biophysics*, 39:61–77, 2010.
- [191] A Xayaphoummine, T Bucher, F Thalmann, and H Isambert. Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proceedings of the National Academy of Sciences*, 100(26):15310–15315, 2003.

- [192] A Xayaphoummine, T Bucher, and H Isambert. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Research*, 33(suppl 2):W605–W610, 2005.
- [193] Yurong Xin, Christian Laing, Neocles B Leontis, and Tamar Schlick. Annotation of tertiary interactions in RNA structures reveals variations and correlations. *RNA*, 14(12):2465–2477, 2008.
- [194] Huanwang Yang, Fabrice Jossinet, Neocles Leontis, Li Chen, John Westbrook, Helen Berman, and Eric Westhof. Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Research*, 31(13):3450–3460, 2003.
- [195] Lan Zhang and Jennifer A Doudna. Structural insights into group II intron catalysis and branch-site selection. *Science*, 295(5562):2084–2088, 2002.
- [196] Yunjie Zhao, Yangyu Huang, Zhou Gong, Yanjie Wang, Jianfen Man, and Yi Xiao. Automated and fast building of three-dimensional RNA structures. *Scientific Reports*, 2, 2012.
- [197] Guohui Zheng, Xiang-Jun Lu, and Wilma K Olson. Web 3DNA - a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Research*, 37(suppl 2):W240–W246, 2009.
- [198] Craig L Zirbel, James Roll, Blake A Sweeney, Anton I Petrov, Meg Pirrung, and Neocles B Leontis. Identifying novel sequence variants of RNA 3D motifs. *Nucleic Acids Research*, page gkv651, 2015.
- [199] Christian Höner zu Siederdisen, Stephan H Bernhart, Peter F Stadler, and Ivo L Hofacker. A folding algorithm for extended RNA secondary structures. *Bioinformatics*, 27(13):i129–i136, 2011.
- [200] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.

Peter Kerpedjiev

Pezzlgasse 47/21-22
1170 Vienna
Austria

+43 680 401 3012
pkerpedjiev@gmail.com
<http://emptypipes.org/about>

Education

- **University of Vienna** Vienna, Austria
PhD Student, Bioinformatics October 2011 -
Tentative Thesis: *Seeing Secondary, Sampling Tertiary: A parallel journey through the prediction and visualization of RNA tertiary and secondary structure*
- **University of Copenhagen** Copenhagen, Denmark
M.Sc. Bioinformatics September 2009 - August 2011
Thesis: *Using Position Specific Scoring Matrices for Short Read Alignment on a Compressed Genome Index*
- **University of British Columbia** Vancouver, Canada
B.Sc. Computer Science and Biology September 2003 - May 2006

Skills / Interests

Spoken Languages: English (native), Bulgarian (fluent), German (working)

Programming Languages: C, Python, Javascript, C#, R, Java

Databases: PostgreSQL, Oracle, MS SQL Server

Bioinformatics: Short-read alignment, Next Generation Sequencing (NGS), Protein structure prediction, RNA structure prediction, Molecular Modelling

Professional Experience

- **Bayer Technology Services** Leverkusen, Germany
Intern, Systems Biology June 2010 - September 2010
- **Bioinformatics Center, University of Copenhagen** Copenhagen, Denmark
Research Assistant, Short Read Alignment February 2010 - June 2010
- **McKesson Corporation** Seattle, WA (USA)
Software Engineer, Patient Records Group June 2006 - July 2009
- **Carnegie Mellon University** Pittsburgh, PA (USA)
Programmer, Universal Speech Interface Project Summer 2000, Summer 2001

Teaching Experience

- **Max F. Perutz Laboratories** Vienna, Austria
Teaching Assistant, Grundlagen der Bioinformatik June 2014, December 2014, April 2015

Other Experience

- **UBC Bioinformatics Center** Vancouver, Canada
Volunteer Programmer, Atlas Data Warehouse Project January 2006 - May 2006

Open Source Software

- **forna** An online tool for displaying the secondary structure of RNA.
(<http://rna.tbi.univie.ac.at/forna>)

- **forgi** A python library for annotating and manipulating RNA secondary and tertiary structure. (<http://www.tbi.univie.ac.at/~pkerp/forgi/>)
- **bwa-pssm** A short read aligner using position-specific scoring matrices to encode the error probabilities for each base. (<http://bwa-pssm.binf.ku.dk/>)

Scholarships / Awards

- **BioVis Design Contest** Dublin, Ireland
Honorable Mention - "DotStruct - An Improved Dot-Plot" July 2015
- **Novo Scholarship 2011** Copenhagen, Denmark
Support for Master's Thesis Work February 2011 - August 2011

Publications

- Kerpedjiev, P., Hammer S., Hofacker I. (2015). Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, 31(20), 3377-9.
- Kerpedjiev, P., Höner zu Siederdisen C., Hofacker I. (2015). Predicting RNA 3D structure using a coarse-grain helix-centered model. *RNA*, 21.6, 1110-1121.
- Kerpedjiev, P., Frellsen, J., Lindgreen, S., & Krogh, A. (2014). Adaptable probabilistic mapping of short reads using position specific scoring matrices. *BMC bioinformatics*, 15(1), 100.
- Jakobsen, T. H., ... & Bjarnsholt, T. (2013). Complete genome sequence of the cystic fibrosis pathogen *Achromobacter xylosoxidans* NH44784-1996 complies with important pathogenic phenotypes. *PloS one*, 8(7), e68484.

Selected Oral Presentations

- **ISMB / ECCB Bioinformatics Open Source SIG** Dublin, Ireland
Simple, Shareable, Online RNA Secondary Structure Diagrams July 2015
- **Vienna RNA Community Retreat** Retz, Austria
Visualizing Co-Transcriptional RNA Folding April 2015
- **TBI Winter Seminar** Bled, Slovenia
Super Simple Software for Showing Secondary Structure February 2015
- **Herbstseminar der Bioinformatik** Doubice, Czech Republic
RNA Secondary Structure Visualization Using D3.js October 2014

Selected Posters

- **RNA Regulation of the Genome SFB Hearing** Vienna, Austria
From 2D to 3D: RNA Structure Prediction October 2014
- **ÖGBMT Meeting** Vienna, Austria
RNA in 3D: Sequence to Tertiary Structure Using a Simplified Representation September 2014
- **ISMB / ECCB** Berlin, Germany
Adaptable Probabilistic Short Read Alignment Using PSSMs July 2013

Colophon

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both \LaTeX and \LyX :

<http://code.google.com/p/classicthesis/>

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

<http://postcards.miede.de/>

Final Version as of January 14, 2016 (`classicthesis`).