Université de Montréal

A new paradigm for the folding of ribonucleic acids

par
Marc Parisien

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des arts et des sciences
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en informatique

Octobre, 2009

Université de Montréal
Faculté des arts et des sciences

Cette thèse intitulée:

A new paradigm for the folding of ribonucleic acids

présentée par:

Marc Parisien

a été évaluée par un jury composé des personnes suivantes:

Nadia El-Mabrouk
président-rapporteur

François Major
directeur de recherche

Serguei Chteinberg
membre du jury

Daniel Herschlag
examinateur externe

Nadia El-Mabrouk
représentant du doyen

RÉSUMÉ

De récentes découvertes montrent le rôle important que joue l'acide ribonucléique (ARN) au sein des cellules, que ce soit le contrôle de l'expression génétique, la régulation de plusieurs processus homéostasiques, en plus de la transcription et la traduction de l'acide désoxyribonucléique (ADN) en protéine. Si l'on veut comprendre comment la cellule fonctionne, nous devons d'abords comprendre ses composantes et comment ils interagissent, et en particulier chez l'ARN. La fonction d'une molécule est tributaire de sa structure tri-dimensionnelle (3D). Or, déterminer expérimentalement la structure 3D d'un ARN s'avère fort coûteux. Les méthodes courantes de prédiction par ordinateur de la structure d'un ARN ne tiennent compte que des appariements classiques ou canoniques, similaires à ceux de la fameuse structure en double-hélice de l'ADN. Ici, nous avons amélioré la prédiction de structures d'ARN en tenant compte de tous les types possibles d'appariements, dont ceux dits non-canoniques. Cela est rendu possible dans le contexte d'un nouveau paradigme pour le repliement des ARN, basé sur les motifs cycliques de nucléotides ; des blocs de bases pour la construction des ARN. De plus, nous avons dévelopées de nouvelles métriques pour quantifier la précision des méthodes de prédiction des structures 3D des ARN, vue l'introduction récente de plusieurs de ces méthodes. Enfin, nous avons évalué le pouvoir prédictif des nouvelles techniques de sondage de basse résolution des structures d'ARN.

Mots clés : ARN, prédiction de structure, comparaison de structure, évaluation de structure, appariements non-canoniques.

ABSTRACT

Recent findings show the important role of ribonucleic acid (RNA) within the cell, be it the control of gene expression, the regulation of several homeostatic processes, in addition to the transcription and translation of deoxyribonucleic acid (DNA) into protein. If we wish to understand how the cell works, we first need to understand its components and how they interact, and in particular for RNA. The function of a molecule is tributary of its three-dimensional (3D) structure. However, experimental determination of RNA 3D structures imparts great costs. Current methods for RNA structure prediction by computers only take into account the classical or canonical base pairs, similar to those found in the well-celebrated DNA double helix. Here, we improved RNA structure prediction by taking into account all possible types of base pairs, even those said non-canonicals. This is made possible in the context of a new paradigm for the folding of RNA, based on nucleotide cyclic motifs (NCM): basic blocks for the construction of RNA. Furthermore, we have developed new metrics to quantify the precision of RNA 3D structure prediction methods, given the recent introduction of many of those methods. Finally, we have evaluated the predictive power of the latest low-resolution RNA structure probing techniques.

Keywords: RNA, structure prediction, structure comparison, structure evaluation, non-canonical base pairs.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DNA	DeoxyriboNucleic Acid
NCM	Nucleotide Cyclic Motif
NOE	Nuclear Overhauser Effect
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
RNA	RiboNucleic Acid
RMSD	Root Mean Squared Deviation

I dedicate this thesis to my parents,
Pauline and Aurèle.

ACKNOWLEDGMENTS

First, I would like to thank my research director, Dr. François Major, who gave me a place within his lab, and gave me a chance to become a scientist, whatever that may be. Second, I thank all those I had the occasion to exchange ideas with during my long journey. Third, all the research funding agencies: les Fonds Québéçois de la Recherche sur la Nature et les Technologies (FQRNT), the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institutes of Health Research (CIHR), the department of computer science and operations research and la faculté des études supérieures of the university of Montréal, and Dr. Damien D'Amours, for their financial support. And finally, my beloved, Janice John, for her TLC.

CHAPTER 1

INTRODUCTION

1.1 Subject introduced

Ribonucleic acids (RNA) are one of the three major classes of bio-polymers within the cell, along with proteins and deoxyribonucleic acids (DNA). It is even speculated that RNA was the primary product from which life originated; hence the *RNA world* scenario [1, 2, 3], following the discovery of the auto-catalytic property of RNA [4, 5, 6]. A vast literature exists concerning the *RNA world*, so let us guide the curious reader to interesting and recent works, in particular on the origin of the genetic code and the translation machinery [7], on the evolution of the 23S molecule [8], the acquisition of functions by fortuitous ligations [9], self-sustained replication [10], and the unavoidable folding of random RNA sequences [11].

It has been long believed that RNA served only as a temporary step in the decoding of DNA into proteins, a process named transcription, as depicted in the central dogma of molecular biology, elaborated by Francis Crick [12, 13]. However, recent discoveries attribute broader and vital roles for RNA in the development and maintenance of the cell, e.g. the translation of RNA coding genes into proteins by the ribosomal complex (see for instance [14, 15, 16]), control of protein expression levels by RNA interference (see [17, 18, 19, 20, 21]), molecular recognition by *aptamers* and *riboswitches* (see [22, 23]), etc. Hence, one then speaks about *non-coding* RNA (for example, see [24, 25, 26, 27]), in which more than 600 families are counted for in the Rfam (*RNA families*) database [28, 29, 30]) (see also RNAdb [31]). The comprehension of the cell's working will necessitate the characterization of its components, and in particular of its coding and non-coding RNAs, which defines the field of ribonomics [32, 33].

The structure of a bio-polymer is the guarantor of its function. Indeed, it suffices for the bio-polymer to position properly in 3-D space a few key chemical groups in order for it to accomplish its function, for which it has been selected to perform. Hence, in order to study the function of a molecule, it is preferable to know its 3-D structure. The famous experience conducted by Anfinsen and colleagues demonstrated that the sequence of a bio-polymer is enough to encode the structure, and that the native structure is of minimal free energy [34]. Although this experience has been carried out using a protein, the same concepts apply also to RNA [35]. This then gives us hope to predict the 3-D structure of RNAs given their sequences only.

Structure determination can be done via high-resolution experiments, like X-ray crystallography (X-ray) [36] or by nuclear magnetic resonance spectroscopy (NMR) [37]. Although precise, these methods make use of unique equipments and specialized operators. It is without saying that they impart a high cost to the determination process (an estimate of a hundred thousand dollars per structure solved by X-ray). The time span needed to resolve a structure is often measured in years, while robotized, high-throughput sequencing produces thousands of sequence fragments in a few seconds [38]. The number of solved 3-D structures in the *Protein Data Bank* (PDB [39]) has been increasing only modestly in the past years [40]. Structure determination can also be done at medium- or low-resolution, by chemical or enzymatic probing [41], microchips [42], fluorescence [43], for instance, but they also require particular measuring instruments and a trained personnel.

That said, we will look at the role of computers for RNA 3-D structure prediction from sequence, for its ease of use and its low cost per prediction. We will also summarize the current approaches, highlighting their strengths and weaknesses. We will then propose a new paradigm in which it will be easier to go from sequence to 3D structure. This paradigm will be put to use to formulate two structural models: apical RNA transport and cap-independent translation element. We will also propose new metrics to compare two RNA models. Finally, we will discuss about the information content of recent low-resolution structure probing methods.

1.2 Organization of chapters

In this section we will list all chapters and their content.

1.2.1 Introduction

Chapter 2 introduces all the concepts needed to understand and appreciate the work done in this thesis. It presents what a ribonucleic acid is and its multiple representations.

1.2.2 The first article

In chapter 3 we will try to put in context what has been done in the first article, specially pressing on the need to develop a new paradigm.

Chapter 4 is the first article and the core of the thesis. It presents the MC-Fold and MC-Sym pipeline, along with MC-Cons, and applications to RNA structure predictions. This paper has been published as:

Parisien M, Major F.
The MC-Fold and MC-Sym pipeline infers
RNA structure from sequence data.
Nature 2008 **452**:51-55.
(c) 2008 Nature Publishing Group.

Since the paper has only two authors, my contribution to it is unquestionable. Because of the limitations on the manuscript's length, we also published an accompanying text, which is the subject of chapter 5. In it you will find all the details of the pipeline, including the ins and outs of the three computer programs presented, and MC-Fold's mathematical model.

To give an idea of the accomplishment laid out in the paper we unfold the paper's time-line. It is without saying that the year 2007 has been nerve wrecking for me, as the first drafts of the manuscript weren't completed before seven months after Nature's initial interest in our work. It must also be said that we previously had submitted a manuscript, which focused on MC-Fold only, to the Proceedings of the National Academy of Sciences of the USA. In that paper we stated that tertiary structure predictions would be easier if we started from secondary structures which featured non-canonical base pairs, predicted from MC-Fold. Unfortunately, that statement had not been followed by a good standing proof, and was therefore not accepted for publication. The creation of the pipeline followed, with a second publication attempt, now in Nature. At this point, the output of MC-Fold, that is the secondary structure, could serve as input to MC-Sym, enabling RNA 3D structure prediction from sequence.

1.2.3 The second article

Because many 3D prediction methods have recently been proposed [44, 45, 46, 47, 48, 49], and that their results do not always look like RNA but, rather, more like a blob of atoms, it was then necessary to develop a new measure to compare a model with the target solution structure. The *de facto* standard measure is the root mean squared deviation (RMSD) [50, 51], and has served well the protein structure prediction community, because in proteins it is the backbone that drives the structure [52, 53, 54], and side-chains associate less specifically than in RNA [55]. However, it is not able to pick up the minute details of nucleobase interactions (pairing and stacking) (see [56] for a discussion of the use of RMSD for RNA modeling), hence the need for a more suited deviation measure for RNA, which is the subject of chapter 6, and has been published as:

Parisien M*, Cruz JA*, Westhof E, Major F.
New metrics for comparing and assessing discrepancies
between RNA 3D structures and models.
RNA 2009 **15**:1875-1885.
(c) 2009 RNA Society.
* Equal contribution.

My contribution to that paper is equal to that of Dr. Cruz; we participated to half of the figures and tables of the article. We also participated in the elaboration of the manuscript. Me and Dr. Major initiated the work here, in Montréal, so this is why I am the first author, and Dr. Major the last author. Dr. Westhof is a world-renowned RNA structure expert and crystallographer. The work in that paper should pave the way for a critical assessment in RNA structure prediction.

1.2.4 The third article

Admittedly, nuclear magnetic resonance spectroscopy (NMR) and X-ray crystallography (X-ray) are the methods of choice to resolve at the atomic level the conformation of an RNA molecule. However, these methods come with great cost and complicated lab equipment, and apply not to all RNAs in all environments. Recently, several low-resolution RNA structure probing methods have been developed. Three of these methods are gel-based: hydroxyl radical footprinting [57], methidiumpropyl-EDTA [58], multiplexed hydroxyl radical cleavage (MOHCA) [59], and thus are accessible to any lab in the world. A more specialized experiment than X-ray, but fast and cheap, is the small-angle X-ray scattering (SAXS) [60]. All these methods provide clues on the conformation of the RNA is solution. However, unlike NMR and X-ray, these low-resolution experiments yield structural constraints which are noisy and vague.

Given that the MC-Fold and MC-Sym pipeline approach provides all-atoms RNA models, we used it to generate sets of three-dimensional RNA structures to challenge these low-resolution data at identifying the native fold adopted by the RNA. The results of our findings are summarized in this paper and in chapter 8, which at the time of writing is under peer-review:

Parisien M, Major F.
Determining RNA three-dimensional structures
using low-resolution data.
(submitted)

In that paper I identified RNA molecules which had been the subject of at least two different probing methods. I also generated the decoy sets, and measured the performance of the experimental data on these sets. I also made a first draft of the manuscript, including all figures and tables. Me and Dr. Major looked at the results and discussed them. I also made additions to our pipeline web site to accommodate these types of experimental data.

1.2.5 Exercises in RNA modeling

What is the use of developing tools if we don't use them! Already, our web site has been put to use in RNA modeling by members of our lab, more specifically, in A-to-I editing recognition sites in RNA double helices [61], and in the transcription attenuation mechanism in bacteria [62]. Here, we provide two RNA modeling attempts. We start from the sequences and any other experimental evidences collected. These modeling exercises show how complicated it is to obtain sound, viable and credible RNA models, despite the RNA's apparent simplistic folding rules (it's not just Guanosine with Cytosine and Adenosine with Uracil). The modeling details can be found in chapter 9. These models are now in the hands of the labs that have interest in them, such that they are now currently being challenged and perfected. They should eventually find their way in molecular and structural biology studies.

1.2.6 The MC-Fold and MC-Sym pipeline web site

Part of the work I done while being a Ph.D. student is the creation and maintenance of the MC-Fold and MC-Sym pipeline web site, which can be found here:

`http://www.major.iric.ca/MC-Pipeline/`

The web site features the main entry points to these computer programs:

- **MC-Fold**; `http://www.major.iric.ca/MC-Fold/` [45]
- **MC-Sym**; `http://www.major.iric.ca/MC-Sym/` [45]
- **MC-Cons**; `http://www.major.iric.ca/MC-Cons/` [45]
- A dot-bracket rendering service
- An automatic MC-Sym script generator
- An Interaction Network Fidelity calculator [63]

The complete web site and its use is described in details in a user's guide, which can be found here:

```
http://www.major.iric.ca/MC-Pipeline/manual.pdf
```

CHAPTER 2

WHAT IS A RIBONUCLEIC ACID?

This chapter will describe what is a ribonucleic acid (RNA). It is mandatory for the good understanding of the rest of the thesis. We switch the avaricious reader toward the reference work of Dr. Saenger on RNA structure [64].

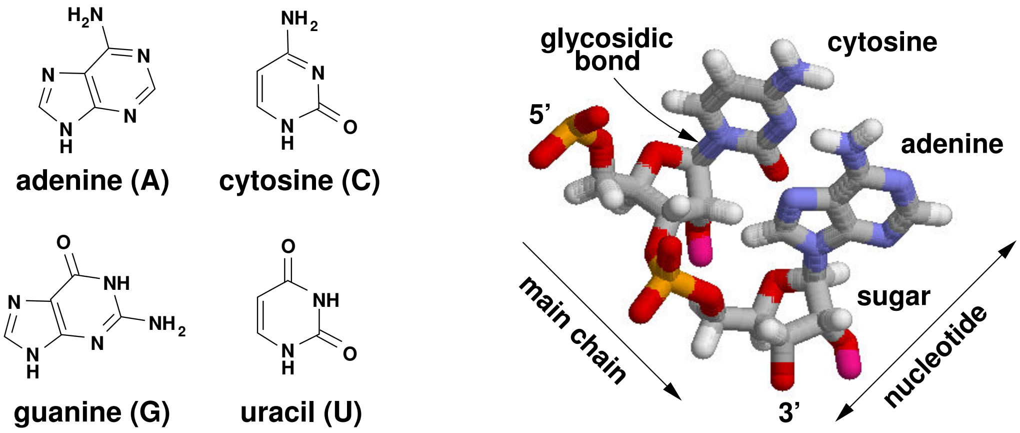
2.1 Nucleobase, nucleotide

A ribonucleic acid is a molecule composed of repeated fundamental units; the nucleobases. Nucleobases come in four types; adenine (A), cytosine (C), guanine (G) and uracil (U). Each nucleobase is attached to its sugar, a five-membered ring, via the glycosidic bond to form nucleotides. These nucleotides are linked together, in a linear fashion, by a cordon, the main chain, that runs from the 5' end to the 3' end. Figure 2.1 shows these different concepts.

2.2 Base pair

The ribonucleic acid chain folds on itself such that the nucleobases associate with one another through hydrogen bonds [65, 66], the base pair. The pairing of two nucleobases is done in a specific fashion, which depends on the types of the nucleobases implicated. In order to describe these pairings, Drs. Leontis and Westhof (LW) proposed a nomenclature [67], which we adopt here. The reader should note that other nomenclatures have been submitted, notably the one used by Saenger [64], the one from Gutell's group [68], and a refined version of the LW by Major's group [69].

Figures 2.2a and 2.2b show the three faces of a nucleobase that can interact with others: the *Watson-Crick* (W) face, the *Hoogsteen* (H) face, and the *Sugar* (S) face.



(a) Nucleobases.

(b) Di-nucleotides.

Figure 2.1: Ribonucleic acid. **(a)** The four nucleobases; adenine (A), cytosine (C), guanine (G) and uracil (U). These images come from the web site newworldencyclopedia.org. **(b)** Two nucleobases with their sugars, nucleotides, linked together by the main chain, which runs from the 5' end to the 3' end. The glycosidic bond attaches the nucleobase to its sugar. Colors used indicate the atom types: white and pink, hydrogen; grey, carbon; red, oxygen; yellow, phosphate and blue, nitrogen.

Hence, to depict a base pair, it suffices to invoke the interacting faces, for example *W/W*. Faces are tagged in this way: the *Sugar* face is the one on the side of the glycosidic bond. The *Hoogsteen* face is opposed to the *Sugar* face. Finally, the *Watson-Crick* face is the most remote to the glycosidic bond. We invite the reader to grasp the difference between the *Sugar* faces of purines (adenine and guanine: figure 2.2a) and that of pyrimidines (cytosine and uracil; figure 2.2b).

The relative orientation of the two glycosidic bonds is added to the description of the base pair. The orientation is either *cis* or *trans*. Figure 2.2c shows an example of each.

Some base pairs are called canonical, following the discovery of the structure of DNA by Drs. Watson and Crick [70]; the nucleobases associate by face complementarity; guanine with cytosine, and adenine with uracil (or thymine in DNA). This unambiguous pairing scheme allows for the perfect duplication of DNA [70]. Figure 2.2d shows the Watson-Crick base pairs in RNA, which are of the type *cis W/W* in the LW nomenclature (not all *cis W/W* are canonical though). Other base pairs, called non-canonicals, also exist in RNA. Many catalogs have been compiled, notably by, Lemieux and Major [69], the Fox group [71, 72], Gutell's group [68], and Olson's group [73].

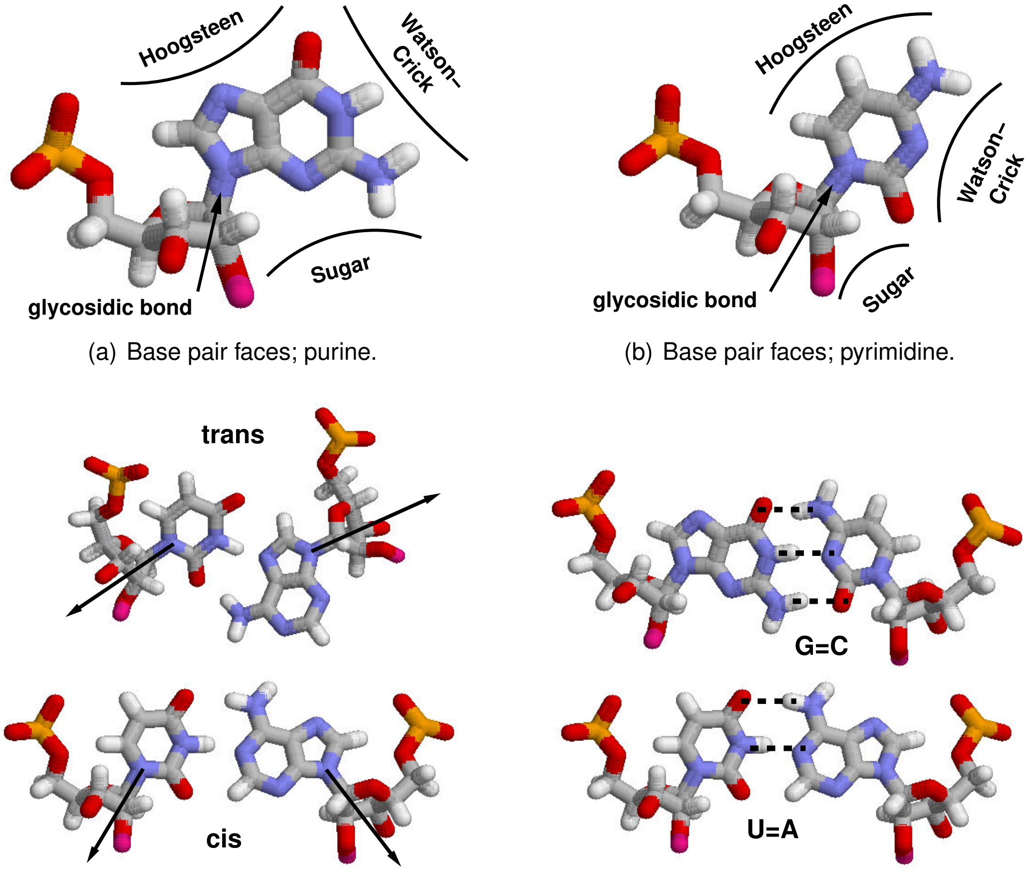
It is interesting to mention here the concept of isosteric base pair substitution [74, 75]. Indeed, the different types of base pairs occupy respective volumes, and position the glycosidic bonds in particular manners, such that in an RNA sequence alignment, given the observed co-variations, it is possible to predict the types of base pairs [76, 77, 78]. For instance, notice the resemblance in size, and the distance between and the position of the two phosphate atoms, between the G=C and U=A base pairs in figure 2.2d. These two basepairs can be superimposed on one another at a surprizing degree of coincidence, despite their different atomic content.

2.3 Double helix

When base pairs stack on one another, for example the pair of nucleotides $(i+1, j-1)$ on top of the pair (i, j), this yields a double helical structure, like the one shown in figure 2.3a. Because base pairs are asymmetric with respect to the glycosidic bonds (figure 2.3c) (i.e. no other in-plane rotations are equivalent), the double helix displays two grooves, the major and the minor, whose difference is notable when the double helix is rendered as a volume (figure 2.3b); the major groove is narrow but deep, while the minor groove is large but shallow. The RNA double helix is mostly found in the A-form (about eleven base pairs per turn) (DNA is mostly found in the B-form; 10.5 base pairs per turn, attributable to thymine) [64, 79]. Interestingly, a particular suite of base pairs and nucleotides can confer to RNA the shape of the DNA double helix [80].

2.4 Folding

The folding of the RNA on itself is a complex phenomena, subject of many studies. Grossly, the folding proceeds in two successive (overlapping) steps: the formation of



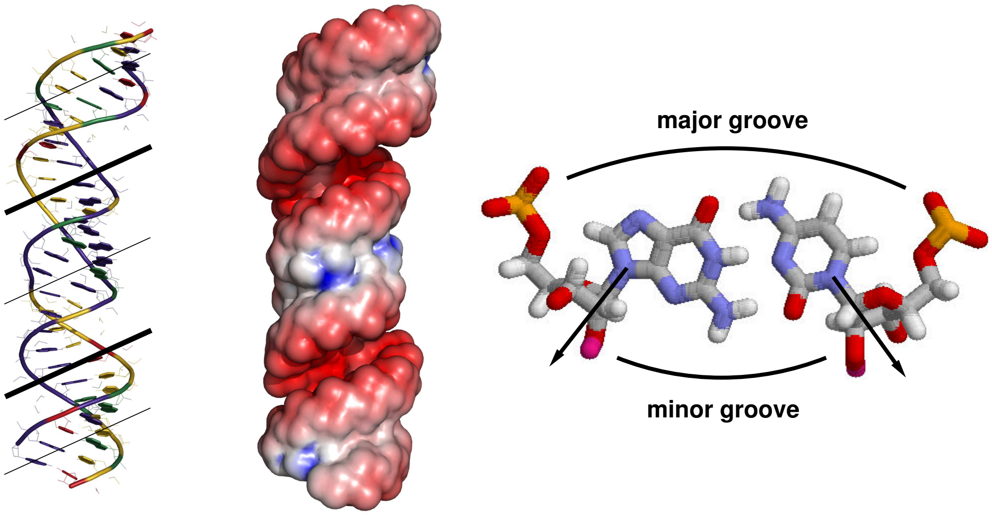
(a) Base pair faces; purine.

(b) Base pair faces; pyrimidine.

(c) Relative orientation.

(d) Watson-Crick base pairs.

Figure 2.2: Base pairs. **(a)** The three faces of a purine (adenine or guanine), according to the Leontis-Westhof (LW) nomenclature. The faces are *Watson-Crick* (W), *Sugar* (S) and *Hoogsteen* (H), all defined with respect to the glycosidic bond. **(b)** The three faces of a pyrimidine (cytosine or uracil), according to the LW nomenclature. **(c)** The relative orientation of the glycosidic bonds determines the *cis* or *trans* type of pairing. The upper base pair is an U=A *trans W/H* while the bottom base pair is an U=A *cis W/W*. **(d)** The canonical Watson-Crick base pairs; The G=C pair (upper) and the U=A pair (lower). Hydrogen bonds are highlighted by dashed lines. The Watson-Crick base pairs are of the type *cis W/W* in the LW nomenclature, since their Watson-Crick faces are interacting, and the relative orientation of their glycosidic bonds are in *cis*. Color scheme is the same as in the previous figure.



(a) Mesh rendering (b) Volume rendering (c) Origin of grooves

Figure 2.3: The RNA double helix (PDB file 1D4R). **(a)** Mesh rendering of a double helix. Nucleotides are colored in this fashion: adenine, green; cytosine, yellow; guanine, violet; uracil, red. Slanted bars indicate the double helix' groove: thin bars for minor grooves, and thick bars for major grooves. **(b)** Atomic volume rendering, showing the water accessible surface area. The electrostatic potential is displayed using a color gradient, from -5 KT/e (red) to $+1$ KT/e (blue). **(c)** Origin of the grooves. Because of the asymmetry of base pairs with respect to the glycosidic bonds (arrows), the face on the side of the bonds is in the minor groove, the opposite face in the major groove.

the secondary structure, i.e. the formation of the double helical stems, then the spatial organization of these double helices and their stabilization by the establishment of long-distance contacts (base pairs and/or base stacks) [81, 82]. The last phase requires the presence of positive ions (mono- and di-valent species), to counteract the negative charges on the main chain. Figure 2.4 sketches a simplified scenario for the folding of an RNA.

Many research groups have for focus the study of RNA folding, in particular the groups of Bevilacqua [83], Herschlag [84], Sosnick [85], Thirumalai [86], Tinoco [35, 87], Woodson [88], and forgive me for all others I forget! Also of intense study are the subjects of electrostatics [89, 90], water [91], specific and coordinated ionic sites [92], the ionic cloud [93] alternative folding intermediates and paths [94, 95, 96], contextual base pair types [97, 98], to name a few.

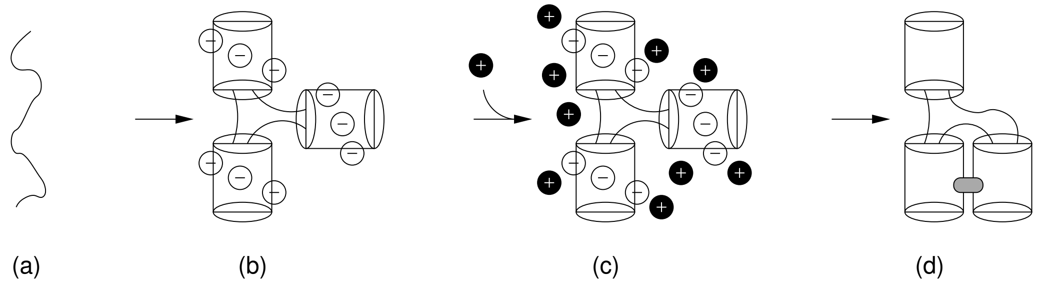


Figure 2.4: Folding of RNA. **(a)** the poly-nucleotide collapses on itself to form double helices (tubes). **(b)** These double helices repel one another because of the negative charges in the main chain. **(c)** The addition of positively charged ions, cations, allows the structure to relax. **(d)** Formation of long range contacts which stabilize the fold, in grey.

2.5 Primary, secondary and tertiary structures

A ribonucleic acid can be represented in many ways. In reality, the molecule is defined by the position (and speed) of each of its atoms, most likely in movement because of the thermal agitation; the three-dimensional (3-D) or tertiary structure. Many experimental methods have for goal to probe or unveil the atomic positions, in particular X-ray crystallography (electron density maps point to absolute atom positions), and nuclear magnetic resonance spectroscopy (NMR) (relative atomic positioning using the nuclear Overhauser effect (NOE) [99]), have the highest resolution of structure determination.

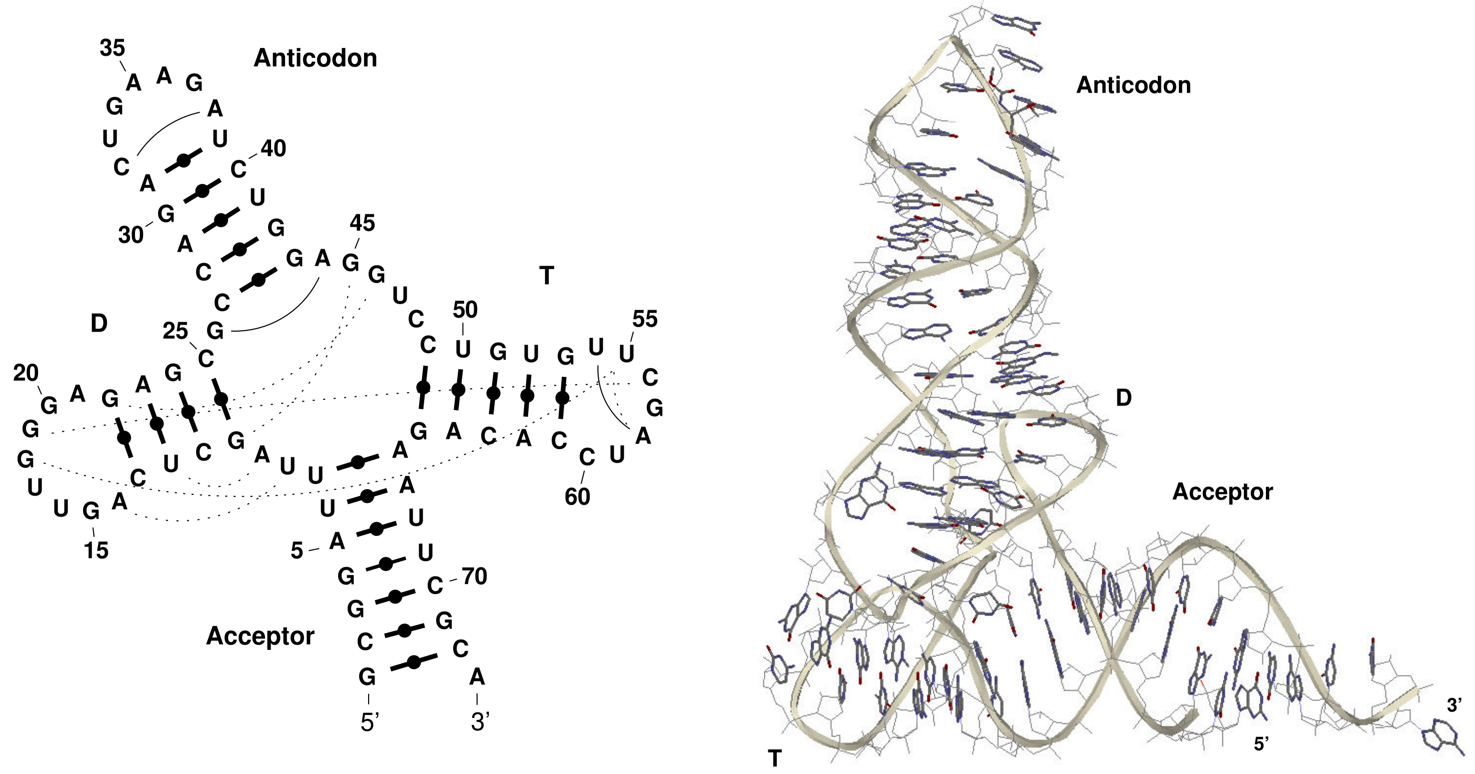
The sequence of nucleotides composing the molecule, the primary structure, is nowadays easily obtained at low cost, using the modest means of the electrophoresis migration theory [38, 100].

The secondary structure puts forward the organization of the sequence into double helices, and is mostly comprised of Watson-Crick base pairs. Although the secondary structure is settled early in the folding of the RNA [35], it can be called to change during the course of folding [101, 102, 103], or in the presence/absence of a ligand for *riboswitches* [22, 23]. This representation seems to be the panacea of RNA structure prediction because of its symbolic nature (the four nucleobase types) and the yes/no state of their base pairing status (sequence of length N yields a N^2 Boolean grid), at the contrary of the 3-D structure in which each atomic position must be given in a numeric form (a sequence of length N, with approximately 30 atoms per nucleotide, yields an N$\times 30 \times \mathbb{R}^3$ list).

In most theoretical studies, the secondary structure is considered to be a simple graph. That is, suppose the base pairs (i, j) and (k,l). Then, if we have $k \leq j$, we also enforce that $i \leq k \leq l \leq j$ [104]. This restrictive definition is useful only to simplify the algorithms of secondary structure prediction methods (for example Mfold [105]), or to extract a few properties (for example see [106, 107]). In reality, RNA structures often contain pseudo-knots [108] (which a 97% majority are of the H-type [109]), hence giving way to non-simple, secondary structure graphs.

The tertiary structure can therefore be seen as the spatial disposition of the secondary structure (the double helices in 3-D space). Figure 2.5 shows these different levels of RNA structure representation. For the tRNA fold, the tertiary structure is stabilized by base triples [110] and long-range base pairs [111]. As much as 25% of the base pairing energy is attributable to tertiary pairing in the tRNA [110], explaining in part why tRNA sequences are elusive to accurate secondary structure prediction, and calls for a more expressive system for symbolic RNA structure representation and computation (as in [112] for example). Other tertiary structure stabilization schemes can be found, like the ribose zipper [113], the A-minor motif [114], tetraloop-helix [115, 116], kissing loops, etc (higher order structure [117, 118, 119]). (see [120] for a recent survey of these motifs).

Local folds in tertiary structures also participate in the stabilization of the molecule [121, 122, 123]: kink-turn [124], U-turn [125], A-platform [126], UNCG tetraloop [127], GNRA and CUYG tetraloops [128], lonepair triploop (including the T-loop) [129, 130, 131], helical coaxial stacking [132, 133], etc. and are amenable to prediction from sequence (see, for instance [77, 134, 135]).



(a) Primary and secondary structures.

(b) Tertiary structure.

Figure 2.5: Primary, secondary and tertiary structures of RNA. The chosen molecule is the *Yeast* phenylalanine transport RNA (tRNA-phe) (PDB code 4TRA). **(a)** The primary structure is the suite of nucleotides comprising the molecule, which proceeds from the 5' to the 3' end, commonly called the sequence. Here, the sequence would be 5'-GCGGA...GCA-3'. The secondary structure puts forward the organization of the sequence into double helices, here shown with dotted thick bars for the canonical base pairs (Watson-Crick), and with continuous thin lines for non-canonical base pairs (for example the nucleotides 26-44). The dashed thin lines highlight base triples (for example 9-12-23), and long-range base pairs (for example, 18-55). For tRNA, its secondary structure adopts the famous cloverleaf shape, given these four double helices: *Acceptor*, *D*, *T* and *Anticodon*. **(b)** The three-dimensional (3-D) or tertiary structure, from the atomic coordinates. The cordon indicates the main chain. The nucleobases are rendered thicker. The spatial organization of the four double helices is shown; the *Anticodon* and the *D* arms are juxtaposed, just like the *T* and *Acceptor* arms. This arrangement gives the tRNA structure an L-shape.

CHAPTER 3

THE NEED FOR A NEW PARADIGM

In this chapter we will position the work done here with respect to current knowledge. We will discuss the problematic of RNA structure prediction and present existing approaches that tackle this problem. We will also state the thesis.

3.1 Problematic

The RNA folding problem [35, 136, 137], the specialized version of the more general bio-polymer folding problem, is enunciated as this: predict the tertiary structure from the primary structure. This folding problem still stands despite decades of research (early computational studies, see [138, 139]), our access to immense computational power (the Guinness world record holder folding@home project [140]), or our keenness at attacking the problem (consider these recent attempts [45, 54]). The core of the problem lies in the Levinthal paradox: how does a bio-polymer fold in a biologically relevant time scale when the conformational search space seems unbounded [141, 142]. Even though the folding of a bio-polymer is known to be hierarchical [35, 141], the conformational search space still remains large, even in a reduced representation [47, 143, 144]. The problem also lies in the delicate balance of the forces acting upon folding [145], and for RNA, the contribution of water and the ionic cloud [146, 147, 148].

We already exposed the urgent need to solve this problem, as sequences are produced at the genomic scale, while 3-D structures are solved one-at-a-time, notwithstanding the protein structure initiative project (http://www.nigms.nih.gov/Initiatives/PSI) and the Critical Assessment of Techniques for Protein Structure Prediction (CASP) (http://predictioncenter.org/), which have no equivalent in the RNA field yet. Structure prediction is useful, particularly because of its broad impact on humanity (functional inference to drug screening) [149].

Classical secondary structures, i.e. secondary structures deprived of helical, non-canonical base pairs, are often too underdetermined to efficiently provide a projection for tertiary structures, leading to a gap in RNA structure prediction [150]. As a matter of fact, non-canonical base pairs provide stabilizing energy to the final fold (see, for instance [110, 151, 152, 153]), as well as restrict the conformational search space by reducing the degrees of freedom of base paired nucleotides. The consideration of non-canonical base pairs in an RNA structure prediction method is assuredly a welcomed feature. Indeed, non-canonical base pairs are found throughout solved RNA structures [154], and often participate in small, recurrent motifs [121, 123].

3.2 Nucleotide Cyclic Motif

With the ever increasing number of solved RNA 3-D structures in the PDB, the notion that RNA is effectively built from small recurring blocks gains appeal [121, 122, 123, 155, 156]. Indeed, recent advances in structural annotations [56, 69, 157], combined to the mathematical analysis offered by graph-centered algorithms [156, 158], and applied to large ribosomal RNA structures (the 30S [159] and the 50S [160]), confirm this modular view of RNA architecture, and forecast its many potent applications [161, 162, 163].

The use of 3-D fragments from the PDB is not new, and perhaps the most eloquent example of such use is made in the computer program ROSETTA [164, 165, 166, 167], which enables atomic precision 3-D structure prediction of proteins from sequence [168]. And, for the RNA counterpart: FARNA [44]. The idea behind the fragment assembly scheme is that local van der Waals and electrostatic interactions are implicitly captured in these 3-D fragments, hence the imprint of the sequence on the 3-D structure [169, 170, 171].

Hornton's algorithm applied to graphs of structural annotations of RNA 3-D structures gives a basis of shortest cycles composing the graphs [156, 158], as shown in Figure 3.1. These shortest cycles can then be viewed as the fundamental building blocks of RNA structure. It has been found that too many types of these cycles occur, such that their subsequent use for structure prediction is prohibitive (Major F, personal communication). An astute downsize of the cycle's types entails the nucleotide cyclic motifs (NCM [45]), as in Figure 3.2, by abstracting base pair types and neglecting base stacking interactions. NCMs are recurrent, compact, generally composed of few nucleotides, sometimes recognized as motifs, and captures base pair isostericity. NCMs are single- and double-stranded motifs that can be welded together as suites to form complete hairpins (see Figure 5.10). Then, the RNA structure is simply an assembly of these hairpins.

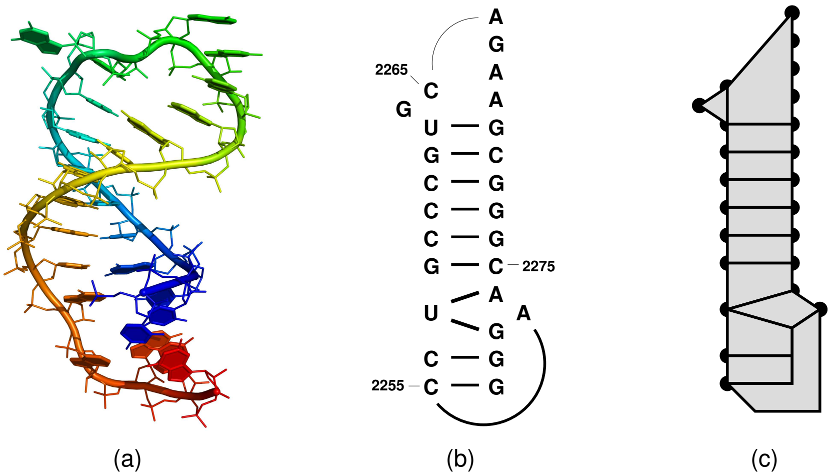
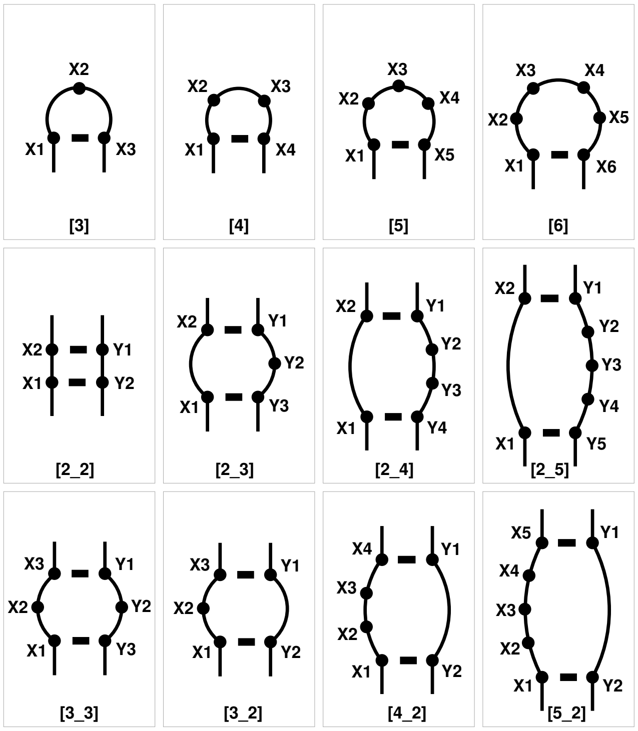


Figure 3.1: Minimum cycles basis. **(a)** 3-D structure of hairpin 2255-2280 of PDB file 1FFK, colored from blue at the 5'-end to red at the 3'-end. **(b)** Its graph of relations, where thick lines represent base pairs. Accounted relations are backbone connectivity, base pairing and base stacking. **(c)** Its corresponding basis of minimum cycles. A total of twelve cycles are highlighted in shaded areas.

Here, we capitalize on the hierarchical nature of the RNA folding process in a divide-and-conquer manner; from the primary structure predict the secondary structure (the MC-Fold computer program), then, from the secondary structure, predict the tertiary structure (the MC-Sym computer program). Both prediction steps are unified under a single and simple model which makes use of a first order object, the nucleotide cyclic motif (compare NCM Fusion 1 and NCM Fusion 2 in Figure 5.1). Our thesis is that the Protein Data Bank currently holds enough information on NCMs, such that statistics on them are able, in return, to predict the native state of an RNA sequence. Furthermore, NCMs ease the prediction of base pair types in 3-D modeling through the NCM



(a)

NCM	Occurrence
3	5 / 2%
4	82 / 39%
5	49 / 23%
6	77 / 36%
2_2	1606 / 89%
2_3	66 / 4%
2_4	21 / 1%
2_5	5 / <1%
3_3	23 / 1%
3_2	47 / 3%
4_2	8 / <1%
5_2	3 / <1%

(b)

Figure 3.2: Nucleotide cyclic motifs (NCM). **(a)** Shown are those that occur most frequently in the PDB. Dots represent nucleotides, thick lines base pairs, and thin lines nucleotide backbone connectivity. Top row displays single-stranded NCMs, while bottom rows double-stranded NCMs. **(b)** Occurrences of the various NCMs.

reconciliation process (depicted in Table 5.3).

NCMs enable a new paradigm in which the contributions of all base pair types are counted for, and not just the canonical ones. Furthermore, the two-stranded versions of NCMs are efficient to model RNA double helical stems, in contrast with a single-stranded fragment merging approach, like FARNA [44], which requires large computational resources [59]. NCMs effectively bridges the gap in RNA structure prediction (to follow comments in [150]). Chapters 4 and 5 discuss in details this new paradigm and its power.

3.3 The master equation

Having identified the Nucleotide Cyclic Motif (NCM) as a first order object for RNA folding, we now seek to devise a mathematical model which would have a predictive value following an analysis of these NCMs from a structural database. We stress here that it is only a model, and many others could be put forth based on the same objects.

Unquestionably, one would like to obtain the simplest model (to abide to the principle of Occam's razor). However, there are no known algorithm to obtain such a model, and we are therefore left with our human intuition (and I would even say "inspiration"). Without a doubt, in order to devise a model for RNA folding, one has to first inspect RNA 3D structures, and that on many planes of representations (cartoons, balls-and-sticks, molecular envelopes, base pairs, base stacks, graphs of tertiary interactions, etc). Only then one can recognize what is like (and what is unlike) an RNA structure.

We have decided to sketch a pseudo-potential energy function, or scoring function, based on statistics extracted from the Protein Data Bank. The main idea here is that if Nature uses some NCMs more often than expected, then one reasonable assumption would be to attribute the use of these NCMs to their thermodynamic stability. We start with the well-celebrated Boltzmann equation:

$$\Phi(\text{structure}|\text{sequence}) = -\text{RT} \cdot \ln \Psi(\text{structure}|\text{sequence})$$

where $\Psi(\text{structure}|\text{sequence})$ is the probability of observing the structure given a sequence. This equation allows us to evaluate the likelyness of many alternative structures for the same given sequence, to sort them according to their scores, and to obtain the one of minimum free energy. The $-\text{RT}$ factor is simply a coupling constant allowing to convert probabilities into energies, thus $\Phi(\text{structure}|\text{sequence})$ represents the energy of a structure given its sequence.

To proceed further, we kindly ask the reader to refer to Figure 5.10, as it defines cycles, or NCMs, junctions, hinges and base pairs. We also lay down the master equation, which is a breakdown of the previous equation into four parts:

$$\Psi(\text{structure}|\text{sequence}) =$$
$$\underbrace{\Psi(\text{NCMs}|\text{sequence})}_{\text{I}} \cdot \underbrace{\Psi(\text{junctions}|\text{NCMs})}_{\text{II}} \cdot \underbrace{\Psi(\text{hinges}|\text{junctions})}_{\text{III}} \cdot \underbrace{\Psi(\text{pairs}|\text{hinges})}_{\text{IV}}$$

where each of the four parts will now be further discussed.

When one looks at the twelve most recurrent NCMs (Figure 3.2), the first question that comes to mind is are there sequences that prefer particular NCMs? For a fact, GNRA sequences are preferred in tetraloops (Figure 3.2, NCM type [4]), and UNCG sequences to a lesser degree. Capturing the fitness of a sequence in an NCM is the goal of part I of the master equation. Hence, given many possible decompositions of a structure into NCMs, we would like to score better those that feature the GNRA sequences into NCMs of type [4].

We now turn our attention to the suites of NCMs, where individual NCMs are welded together to form junctions (see Figure 5.10). A notable feature of RNA 3D structures is the predominant occurence of NCM type [2_2] (Figure 3.2b). Single-bulged nucleotides, as properly described by NCMs types [2_3] and [3_2], are rare (and multiple-bulged rarer). In order to quantify the many different suites of NCMs a sequence can adopt, we encoded in a Markov chain the probabilities of observing an NCM of type i followed immediately by another NCM of type j, defining a junction (i, j) (see Figure 5.10). Markov chains are data structures appropriate to capture the correlation of an event (here NCM types) on others. We adopt the simplest of the chains which is of order one, meaning that the probability of a junction (i, j) at NCM j only depends on the type of the previous NCM i (but could also depend on the previous two ones). This enables us to score better a structure that features many adjacent NCMs of type [2_2] against one which would be made of, say, three consecutive NCMs of type [2_3]. This is what part II of the master equation tries to encode.

By far the most complicated computation is the one of part III of the master equation. Simply put, it tries to guess the type of a base pair common to two consecutive NCMs, an hinge (Figure 5.10). Base pair types are those defined by Leontis and Westhof, which are described by their interacting edges. Borrowing from an analogy to analog electric circuits, consider many electrical paths from a single source to a single destination. The question here would be how the electric current would be divided between the paths given the electric properties of each path (resistance, capacitance, inductance, etc). For the problem at hand, our electrical source and destination would be two consecutive NCMs. The various electrical paths would then be the many different types of base pair that the sequence could choose from. In Figure 5.10 one could ask what type of base pair is most likely found between C13 and G17 given that they are at the hinge of NCM types i (a type [5] NCM) and j (a type [2_3] NCM). One has to consider the possible *a priori* types, that is, what a C=G base pair 5' of a type [5] NCM prefers, then what a C=G base pair 3' of a type [2_3] prefers, then check if at least one type is common which would "convey" the current. A detailed computation of a hinge is given in Table 5.3.

Finally, part IV scores the base pair, regardless of the type (taken into account in the previous term). Hence, a C=G base pair has a better score than a C=C base pair, because the latter occurs less frequently in the PDB.

Further instructions on how statistics from the Protein Data Bank are used to derive actual scores for each individual parts of the master equation are discussed in section 5.1.5. We feel that the master equation sums up the many aspects that are consequent of the use of Nucleotide Cyclic Motifs as first order objects for the folding of RNAs. Perhaps there are other significant phenomena that slipped by our concerns. If so, we would gladly like to be informed of these, so that they could be accounted for in a refined model.

3.4 Existing approaches

The most prominent approach to RNA structure prediction is the use of the thermodynamics theory in secondary structure prediction. Its main concern is the evaluation of the folding free energy, as the RNA collapses on itself to attain its native fold. The latest model which encompasses this view is the INN-HB model (Individual Nearest Neighbor - Hydrogen Bond) [172]. Computer programs that use this model are Mfold [105, 173], Sfold [174, 175, 176], RNAfold [177, 178], RNAsubopt [179], and RNAstructure [180], to name the most widely known programs. Addition of the persistence length concept has been figured out [181, 182]. Even though this model seems the most promising because of its physico-chemical formulation, the thermodynamics data amassed since the 70'(for example [183]) is still not enough to nail an RNA fold from the sequence [137, 184, 185], even by using structure probing data [180]. This is due to the fact that the contribution of non-canonical base pairs is only counted for single mismatches, (which we tried to address in this work), and because of the under-characterization of hairpin heads (which we also address, although in the thermodynamics paradigm some heads seem more stable than the model predicts [186]) and multi-branched loops (too many possibilities to address, even though some groups have tackled them, see [187, 188]), which are often extrapolated from a few data points. McCaskill's algorithm allows for an estimate of a base pair probability [189]. An iterative optimization method that combines both experimental and knowledge-based data supersedes the Turner99 thermodynamics model and thus now serves as the state-of-the-art RNA secondary structure prediction [190].

The enumeration of secondary structures has been proved to belong to the NP-complete time-complexity class of problems, when considering general-case models of pseudoknots [191]. For secondary structures without pseudoknots, obtaining the minimum free energy structure, by filling up a dynamic programming table, takes time $O(N^3)$, for a sequence of length N [105, 192]. Approximations on pseudoknots can cut the computation time to $O(N^6)$ [193], or even $O(N^5)$ or $O(N^4)$ on simpler pseudoknots ([191, 194], respectively). For pseudoknot-free secondary structures, an upper bound of $O(\Psi(N) \cdot N^2)$, where $\Psi(N)$ is almost constant, has recently been proposed [195]. Given that the MFE structure may not be the native fold, and thus sub-optimal structures have to be enumerated (by backtracking in exponential time), further optimizations to obtain the MFE may appear obsolete (see [196, 197] for recent developments). Sub-optimal structure enumeration have been proposed by Zuker [198] (which uses a distance criterion that ensures the generation of non-related structures), and by Schuster's group [179] (systematic and exhaustive enumeration). The later implements the Waterman-Byers algorithm [199, 200]. At this point, we wish to point the reader to these excellent reviews in RNA secondary structure prediction: [201, 202]. For RNA structure prediction from multiple-sequence analysis, we suggest the thorough review of Drs. Gardner and Giegerich [203].

Knowledge-based approaches have also been devised, notably MC-Fold [45] (this work), CONTRAfold [204], a remake of the thermodynamics table for threading [205] or enhanced prediction [206], and structure refinement [207]. Attempts to concoct physico-chemical approaches to the RNA folding problem in 3-D have also been made, noticeably by Cao and Chen [144], and by Dokholyan's group [47]. Of interest, intermediate folding during transcription (due to pausing [102]) is considered in the kinwalker computer program [103].

It is not at all clear what is the best approach for RNA structure prediction. On one hand, secondary structure prediction is really fast for hundreds or thousands nucleotides, compared to all-atoms simulated folding (the Folding@home cluster, a world-wide network of about 250,000 **active** CPUs, has been used for the study of a twelve nucleotides hairpin [208]!). However, the former lacks the excluded volume effects, such that predicting structures with pseudoknots is a tricky business, often resulting in physically impossible folds. Coarse-grained nucleotide representation (like [47, 48, 144]) allows for faster RNA collapse, but they still rely heavily on the secondary structure, and thus are not fully *ab initio*. The RNA folding landscape has been proved quite complex for even a small hairpin [86, 209, 210], with multiple-paths for larger RNA structures [84, 96]. Chen highlights the complicated nature of RNA folding and what still needs to be done [147].

3.5 Further on

Although the new RNA folding paradigm has been enunciated and detailed, we pursued further along the path of RNA structure prediction, and the next studies can be also viewed as part of the new paradigm.

Because RNA side chains interact in highly specific manners (via hydrogen bonds), we sought to improve the RNA structure prediction field by shifting the measure of modeling success toward a distance metric which takes into account the specificity of RNA base pair interactions. The RMSD has been shown to be a deficient measure of base interaction reproducibility [56] and modeling success [63], at least for RNA, despite its ubiquitous usage in the sister protein folding field. Symbolic annotations of the interactions between nucleobases can instead be used to compare a 3D model with the solution structure. Hence, it suffice to reproduce the interactions within the solution structure to declare that the model is correct, and not necessarily obtain low RMSD values. Discrepancies between the model and the reference structure can also be scrutinized by comparing the differences in the annotations. Since the introduction of this new distance metric we realized that much needs to be done in order to produce credible RNA 3D models, despite our recent advances in treating non-canonical base pairs.

The RNA folding paradigm has also been extended to accommodate new RNA structure probing methods and their data. Hence, RNA models produced using our MC-Fold and MC-Sym pipeline can be scored against a variety of low- and high-resolution experimental data, in a effort to produce better RNA models. By generating various decoy sets which makes use (or not) of long-range base pairing interactions, we investigated the power of many low-resolution data to identify the native fold.

CHAPTER 4

ARTICLE 1

The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data

Marc Parisien and François Major

Institute for Research in Immunology and Cancer,
Department of Computer Science and Operations Research,
Université de Montréal,
PO Box 6128, Downtown station,
Montréal, Québec, H3C 3J7, CANADA

Parisien M, Major F.
Nature 2008 **452**:51-55.

4.1 Abstract

The classical RNA secondary structure model considers A•U and G•C Watson-Crick as well as G•U wobble base pairs. Here we substitute it for a new one, in which sets of nucleotide cyclic motifs define RNA structures. This model allows us to unify all base pairing energetic contributions in an effective scoring function to tackle the problem of RNA folding. We show how pipelining two computer algorithms based on nucleotide cyclic motifs, MC-Fold and MC-Sym, reproduces a series of experimentally determined RNA three-dimensional structures from the sequence. This demonstrates how crucial the consideration of all base-pairing interactions is in filling the gap between sequence and structure. We use the pipeline to define rules of precursor microRNA folding in double helices, despite the presence of a number of presumed mismatches and bulges, and to propose a new model of the human immunodeficiency virus-1 -1 frame-shifting element.

4.2 Introduction

The number of RNAs found to be involved in non-coding cellular roles is increasing rapidly and persistently [2, 29], and many RNA transcripts of unknown function have recently been detected in eukaryotic cell maps [211]. RNAs can be grouped into families that share structural features and function. Therefore, unravelling the structure provides crucial insights into the way in which RNA works. However, producing RNA high-resolution structures by X-ray crystallography and NMR spectroscopy is slow compared to sequencing, thus creating an important gap between the number of known tertiary (three-dimensional, 3D) structures [39] and that of sequences [212].

In the search for an effective RNA structure-determination approach, we examined different theoretical schemes and studied their relative merit to attain our goal. Hope came from the fact that secondary structures would provide enough structural constraints to automate 3D building [150]. A secondary structure describes the stems of RNA - crucial building blocks that form when two complementary regions of the sequence base pair and adopt a double-helix structure. A legitimate approximation of secondary structures considers stems that consist of A•U and G•C Watson-Crick base pairs as well as G•U wobble base pairs. These base pairs are called ’canonicals’.

Secondary structures can be derived from a sequence by using a combination of free-energy minimization [213] and covariation analysis [214]. However, the presence of a few key non-canonical base pairs blurs predictions, because they contribute energies and complicate covariation interplay [202]. Even when experimental data are considered (for example, enzymatic or chemical probing), selecting the native amongst many suboptimal secondary structures remains elusive [180]. More importantly, secondary structures deprived of non-canonical base pairs are neither adequate to determine 3D structures nor sufficient to faithfully align sequences of the same family [77, 150, 215]. Recent attempts to replace thermodynamics by statistical scores resulted in either similar [205] or only slightly improved [204] predictive power. Furthermore, empirical scoring of 3D structures applies only to very short sequences and requires covariation data [44]. Taken together, these shortcomings and increasing needs for RNA genome-wide annotation prompted us to develop a new approach.

We extended the classical rationale underlying RNA structure prediction by incorporating all base pairs. To do so, we introduced a new first-order object to represent nucleotide relationships in structured RNAs: the nucleotide cyclic motif (NCM). The NCMs became apparent to us from an analysis of the X-ray crystallographic structure of the 23S ribosomal RNA of *Haloarcula marismortui* [156]. Adjacent NCMs share common base pairs - a property providing enough base-pairing context information to derive an effective scoring function and making possible the use of the same algorithm for predicting secondary and tertiary structures.

We propose a new RNA-structure-prediction method based on NCMs, implemented as a pipeline of two computer programs: MC-Fold and MC-Sym (Supplementary Fig. 5.1). We illustrate the predictive power of the pipeline by reproducing experimentally determined 3D structures from a single sequence, building 3D structures of precursor microRNA (pre-miRNA) that are compatible with Dicer docking, and proposing a new 3D structure of the human immunodeficiency virus (HIV-1) cis-acting -1 frame-shifting element. In practice, judicious pipeline predictions from a single sequence are expected for fragments of up to approximately 150 nucleotides.

4.3 Folding single sequences

We evaluated the predictive power of MC-Fold by comparing the base pairs in the lowest-energy (best) predicted structure of each sequence with those found in experimental hairpin loop structures (Table 4.1). Compared to the thermodynamic approach, MC-Fold predicts over 6% more canonical base pairs, despite a lower positive predictive value, concurrently makes less false positives and negatives, and obtains a higher Matthews correlation coefficient ratio (MCCR) (see Supplementary Table 5.1). In addition, the optimal solution for each hairpin includes more than 60% of the non-canonical base pairs, and this number goes up to more than 80% if the top five solutions are considered. The low rate of false negatives is a prerequisite for building 3D structures.

We evaluated the predictive power of the MC-Fold and MC-Sym pipeline by analysing and comparing the best predictions for thirteen experimental 3D structures (Table 4.2). Eleven of the thirteen examples rank first (that is, match the lowest-energy structure). Eight of the thirteen examples have MCCRs of 100% (average = 98.2%). Seven of the thirteen examples combine first rank and 100% MCCRs. The average root mean squared deviations [51] (r.m.s.d.) of the thirteen examples when optimally superimposed on their corresponding experimental structures are near 2 ångströms (Å) (Fig. 4.1). The nucleotides that increase the r.m.s.d. are those with more degrees of freedom, that is, those not involved in base-pairing interactions (see, for instance, nucleotides A14 and U16 in the iron-responsive element (IRE) hairpin loop in Fig. 4.1a). Another source of high r.m.s.d. is the presence of false positives and negatives. For instance, the telomerase RNA domain IV has an MCCR of 94% and a r.m.s.d. of 3.3 Å (Fig. 4.1b). Interestingly, the false positive and the false negative are made in the hairpin loop. The NMR structure has a single-nucleotide bulge, A22, which stacks inside the helix on the 5’ side of a CUAU tetra-loop. The C23•U26 base pair that closes the tetra-loop is stabilized by a single hydrogen bond, and the two bases are perpendicular to each other.

Although relatively stable, these features are rather rare and might be induced by particular experimental conditions or structure resolution methods. The NMR hairpin loop is less stable than the penta-loop proposed by the MC-Fold and MC-Sym pipeline.

When the experimental structure does not correspond to the lowest-energy structure, it is generally due to the formation of extra base pairs in the latter. The lowest-energy structure is often referred to the 'ground state'. The base pair formation/disruption phenomenon is known to be dependent on conformational changes induced by cofactors [216], which are difficult to represent in any scoring scheme. Consequently, polymorphic structures are found in MC-Fold's suboptimal solutions.

The conserved sequence of the *Deinococcus radiodurans* and *Escherichia coli* 23S rRNA helix 40 (ref. [217]) contains an interior loop, $\begin{smallmatrix} 5'-\mathrm{CU}\mathbf{AA}\mathrm{G}-3' \\ 3'-\mathrm{GA}\boldsymbol{A}\mathbf{G}\mathrm{C}-5' \end{smallmatrix}$, the structure of which differs whether it is solved by NMR or by X-ray crystallography. The NMR conditions favour the formation of a non-canonical A•A/A•G base-pair tandem, which MC-Fold ranks first (shown in bold above). The X-ray crystallographic structure is bound to a protein that possibly induces the disruption of the A•A non-canonical base pair and the apparition of a single bulged-out A (shown in bold-italic above), which MC-Fold ranks fifth.

The 'on' and 'off' conformational states of the cytoplasmic eukaryotic rRNA A site [218] contains an interior loop, $\begin{smallmatrix} 5'-\mathrm{CGC}-\mathrm{U}-3' \\ 3'-\mathrm{A}\mathit{A}\mathbf{AAG}-5' \end{smallmatrix}$, the structure of which differs whether the ribosome is active in protein translation (on) or not (off). X-ray crystallographic data of the *Homo sapiens* A site reveal these two distinct structures [218]. The on state has two unpaired A nucleotides that bulge out of the main helix (shown in bold above), whereas only one A, 3' of the two bulges in the on state, is unpaired in the off state (shown in italic above). MC-Fold ranks the on state as sixth and the off state as fourth.

Multi-branched RNAs are made of more than two helical stems that are joined by a multi-branch loop. We used the pipeline to reproduce the 3D structure of a pre-catalytic conformation of the hammerhead ribozyme [115] (Fig. 4.1c), as well as that of the recent NMR structure of the *Xenopus laevis* 5S rRNA bound to zinc fingers [219] (Fig. 4.1d). When more than two stems are selected by MC-Fold, the coaxial energies are computed and accounted for in the final score (Supplementary Methods). The key base pairs to project properly the hammerhead in 3D space are located near the multi-branch: the three base pairs at the bottom of stem II and the C•C base pair in stem I. The C•C base pair is particularly important to avoid coaxial stacking between stem I and stem III.

Finally, inserting a stem that creates a nested structure generates a pseudo-knot, as shown in the structure of the yellow leaf virus [220] (Fig. 4.1e). In this model, a false positive non-canonical A•A base pair is made at the bottom of the upper stem. Nevertheless, the closest generated model shares 2.7 Å of r.m.s.d. when optimally superimposed on the NMR structure.

4.4 Folding human precursor microRNAs

When we submitted the pre-miRNA sequences of let-7c, mir-19a and mir-29a, our predictions were almost identical, and were similar to the A-RNA double helix (Fig. 4.2). In fact, we did not find any pre-miRNA sequence in mirBase [221] that could not be folded in the double helix (data not shown), despite an overrepresentation of U•U and U•C mismatches. The double helix offers a fixed and stable reference to the scissile phosphates that are cleaved by the Drosha complex upstream of the pre-miRNA [222], as well as by Dicer near the terminal loop [223].

The pre-miRNA double helix of let-7c (Fig. 4.2a) is bulge-free and presents to Dicer the expected docking surface [223], despite the non-canonical base pairs. In the 3D structure of mir-29a (Fig. 4.2b), the unpaired C23 nucleotide stacks inside the helix, acting as a lever to push the scissile phosphate of A26 into its proper position. Finally, in the 3D structure of mir-19a (Fig. 4.2c), the two unpaired nucleotides, A56 and U57, form a bulge behind the docking surface, and hence do not interfere with Dicer binding. These strict 3D structural restraints should further help in distinguishing between RNA stem-loop structures that can be processed by Dicer.

The presumed microRNA mismatches, in fact, adopt a geometry isosteric to Watson-Crick base pairs [74]. Their energies are less than that of canonical ones, which may facilitate the unwinding of the double helix and loading of the mature miRNA into the RNA-induced silencing complex (RISC). Interestingly, we find very few G•A mismatches in the miRNA region interfacing Dicer because their propensity for the sheared conformation is not isosteric to Watson-Crick base pairs. The sheared geometry distorts the backbone path of the double helix and, thus, might interfere with Dicer binding. The natural selection for non-canonical base pairs increases the diversity of possible pre-miRNA sequences, while increasing target specificity and, simultaneously, decreasing off targeting.

4.5 Folding using probing data

As shown above, MC-Fold does not always rank experimental and activated structures first. Reaching these structures is nevertheless of principal importance. Here we show how experimental data can be incorporated to restrain the conformational space of MC-Fold to identify such induced structures.

For example, a recent study investigated the yeast transfer RNA$^{\text{Asp}}$ structure by selective 2'-hydroxyl acylation and primer extension (SHAPE) [224]. SHAPE data reveal the flexible and constrained nucleotides, subject to experimental conditions. The tRNA sequence tested is deprived of the modified nucleotides, and has been shown to adopt the cloverleaf structure [225]. The top MC-Fold prediction of this tRNA$^{\text{Asp}}$ sequence is a hairpin, not a cloverleaf.

If we introduce high- and medium-flexibility SHAPE constraints (Supplementary Fig. 5.2), MC-Fold generates cloverleaf structures and ranks the native one sixth. The D-stem-loop sequence of this tRNA has a positive folding free energy under the thermodynamic model. Amongst the solutions, one includes a correct D-stem base-pairing registry (MCCR of 100%) and the A14•A21 base pair, whereas all other solutions base pair U13 with A21. The A14 inflexibility demonstrated by SHAPE is thus sufficient to discriminate the native amongst all solutions.

Similarly, by introducing dimethyl sulphate (DMS) data [180], all known canonical (with the exception of G56•C28) and all non-canonical base pairs of the E. coli 5S rRNA are captured in the correct *in vivo* Y-shaped topology (Supplementary Fig. 5.3a). Interestingly, the MC-Fold optimal solution (Supplementary Fig. 5.3b) of sub-sequence 16 to 69 has a marked resemblance to the *in vitro* structure that was probed by chemical modifications [226] and by NMR [227]. The latter suggests further a bias towards structures in the ground state.

4.6 Consensus structural assignments

Sequences that are functionally related are another source of structural data. Consider the IRE, a hairpin loop found in the 3' untranslated region of the ferritin and transferrin receptor mRNAs. IREs are involved in maintaining iron homeostasis in vertebrate cells by acting as post-transcriptional factors [228]. MC-Fold and MC-Sym best predictions of several IRE sequences reveal the base pairs and nucleotides found to be involved in receptor binding (Fig. 4.1a): an upper stem of six base pairs [229], a single unpaired nucleotide 3' of the hairpin-loop-flanking base pair [230] and a single (V) or double (W) bulge in the 5' strand of the stem.

MC-Cons computes a structural assignment, that is, it assigns to each sequence the structure that maximizes the overall sum of pairwise structural similarities. The structural assignment returned when 30 IRE sequences available at Rfam (RNA families database) [28] and their top ten MC-Fold predictions are input to MC-Cons (see Supplementary Methods) reveals two IRE subclasses (Supplementary Fig. 5.4), corresponding to both helix-bending motifs, V and W. The two subclasses have been shown to be important in selective repressor binding, in particular to the human iron responsive protein 2 (ref. [231]). Similarly, using ten yeast tRNA sequences, MC-Cons identified the cloverleaf structure for each sequence (see Supplementary Fig. 5.5).

Multiple-sequence and low-resolution data can be used in combination. Using fourteen 5S rRNA *E. coli* sequences and DMS data, the *in vivo* 5S rRNA structure is captured (Supplementary Fig. 5.6). In this case, a high rate of non-canonical false positives is made in the large hairpin (nucleotides 35-47). This is probably due to the fact that the RNA in the crystal structure is bound to the ribosomal complex, in which the large hairpin makes several contacts with the ribosomal protein L5. However, the consensus structural assignment of the E. coli 5S rRNA sequences without DMS probing data predicts the *in vitro* structure (Supplementary Fig. 5.7). Similarly, the Selenocysteine Insertion Sequence (SECIS) structure is also predicted using seven sequences and various RNase probing data [232] to block the base pairing of 13 out of 151 nucleotides (Supplementary Fig. 5.8).

4.7 Modelling HIV-1 frame-shifting element

HIV-1 is known for encoding two proteins, pol and gag-pol [233], using the same mRNA and a -1 cis-acting frame-shifting mechanism owing to the formation of a structure downstream of the slippery sequence [234, 235]. Recent NMR data suggest that this structure could be a hairpin loop with an asymmetric bulge of three nucleotides, 5'-GGA-3'. In a first study, clear NMR signals were obtained by modifying the sequence to include GC base pairs in the lower stem, therefore introducing a coerced registry. In a second study, the native sequence was used. However, it was extracted from the mRNA so that the lower stem was also constrained. Besides, when MC-Fold is run with both NMR sequences, the best solutions match the structures obtained by NMR.

However, using 50 randomly selected sequences out of the 753 reported in Rfam, a single and different structure makes the consensus assignment amongst these sequences. The principal difference between the new structure and those obtained by NMR is in the bulge: MC-Fold predicts a double-A bulge, 5'-AA-3', instead of a 5'-GGA-3' bulge (Fig. 4.3). This is a minor difference, but several arguments support it (see Supplementary Discussion).

4.8 Discussion

Our results highlight the fact that for effective RNA structure predictions, dealing with all base-pairing types in both secondary and tertiary structures is of the utmost importance. A difference between our study and other recent attempts is the use of a first-order object based on NCMs, which incorporates more base-pairing context-dependent information; this suggests that it is key in scoring secondary structures.

The lowest free-energy states determined by MC-Fold often differ from active and experimental states. Furthermore, solving the consensus structural assignment using MC-Cons occasionally predicts such ground rather than active structures (for example, *in vitro E. coli* 5S rRNA). However, we showed that few low-resolution experimental data could be introduced to bias the search towards experimental and *in vivo* structures (for example, tRNA, *in vivo* 5S rRNA and a SECIS element). Predicting both ground state and induced fit structures for the same sequence is a strong indication that MC-Fold predicts correct structures, as well as structures that are accessible to any given sequence. This is reflected in the high rate of false positives when compared to experimental structures.

Thanks to the pipeline, RNA modelling is now more accurate and simpler than ever. The secondary structures generated by MC-Fold are more informative than those deprived of non-canonical base pairs and include very few false negatives. Producing 3D models consistent with these secondary structures is now a straightforward and accessible-to-all online activity. This should translate into keener RNA function hypotheses and less experimental work to verify them.

4.9 Methods summary

The three algorithms and the scoring function are fully described in the Supplementary Information. A web service of the three algorithms has been made publicly available on the Internet at http://www.major.iric.ca. The protocols to produce secondary and tertiary structures using the website are described elsewhere (submitted).

4.10 Acknowledgements

We thank P. Thibault for updating MC-Sym and P. Gendron for helping us with the Condor and web services. We thank D. D'Amours, M.-F. Gaumont-Leclerc and V. Lisi for making suggestions to improve the manuscript. We thank D. H. Mathews and E. Westhof for discussions about MC-Fold. This project was supported by grants from the Canadian Institutes of Health Research (CIHR) and from the Natural Sciences and Engineering Research Council (NSERC) of Canada. M.P. holds Ph.D. scholarships from the NSERC and the Fonds Québécois de la Recherche sur la Nature et les Technologies. F.M. is a member of the Centre Robert-Cedergren of the Université de Montréal.

Author Contributions Both authors were involved in every aspect of the research. M.P. programmed MC-Fold and MC-Cons.

Predicted base pairs (%)	Zipper (lower bound)	RNAsubopt (thermodynamics)	MC-Fold (NCM)
$\text{PPV} = \frac{\text{TP}}{(\text{TP}+\text{FP})}$	59.6	**91.8**	83.4
$\text{STY} = \frac{\text{TP}}{(\text{TP}+\text{FN})}$	74.1	74.8	**89.9**
$\text{Matthews} = \sqrt{\frac{\text{TP}}{(\text{TP}+\text{FN})}\frac{\text{TP}}{(\text{TP}+\text{FP})}}$	66.5	82.9	**86.6**

Table 4.1: Best predictions over 2,093 base pairs (1,784 canonical base pairs) in 264 hairpin loops extracted from 182 different PDB structures. Columns show programs and rows show coefficients. Zipper is a program that implements a greedy algorithm that folds a sequence from bottom-up using exclusively tandems of base pairs. This gives a lower bound on the predictive power. RNAsubopt implements the current thermodynamics model and enumerates systematically all suboptimal structures. The numbers of nucleotides in these hairpin loops vary from 8 to 35 base pairs (average = 19.6). The best value for each row is shown in bold. FN, number of false negatives; FP, number of false positives; TP, number of true positives; Matthews, Matthews correlation coefficient ratio; PPV, positive predictive value; and STY, sensitivity.

RNA (PDB code)	Size	Rank	Matthews	r.m.s.d.
	(nucleotides)		(%)	(Å)
Hairpins				
Loop E (430D†)	29	1	100	1.7
IRE (1NBR)	29	1	100	2.4
Classical swine fever virus IRES domain III (2HUA⋆)	40	4	100	2.6
RNA thermometer (2GIO⋆)	29	1	100	1.7
Eel UnaL2 LINE 3' element (2FDT⋆)	36	1	100	2.0
Telomerase RNA domain IV (2FEY⋆)	43	1	94	3.3
RNase P RNA P4 (2CD1⋆)	27	2	96	2.1
GNYA tetraloop (2EVY⋆)	14	1	100	1.8
U2 snRNA (2O33⋆)	20	1	100	2.0
Group II intron branchsite (2AHT⋆)	27	1	96	1.9
Y-shape				
Hammerhead ribozyme (1NYI†)	36	1	100	2.7
5S rRNA (2HGH⋆)	47	1	96	2.9
Pseudo-knot				
Yellow leaf virus (2AP5⋆)	18	1	94	2.7

Table 4.2: The MC-Fold and MC-Sym pipeline is applied to single sequences. Three different RNA topologies were tested: hairpin, multi-branch (Y-shape) and pseudo-knot. The best predictions (best MCCRs) are reported. The r.m.s.d. values were calculated over all heavy atoms. The average MC-Fold real time for all but one sequence is 7 s on a typical workstation processor (AMD Athlon 64, 2.2 GHz); real time for the 5S rRNA sequence is 143 s. The best 3D models are selected amongst all models generated using a probabilistic search over a period of 12 h.

⋆ A recent NMR structure.

† X-ray crystallographic structure.

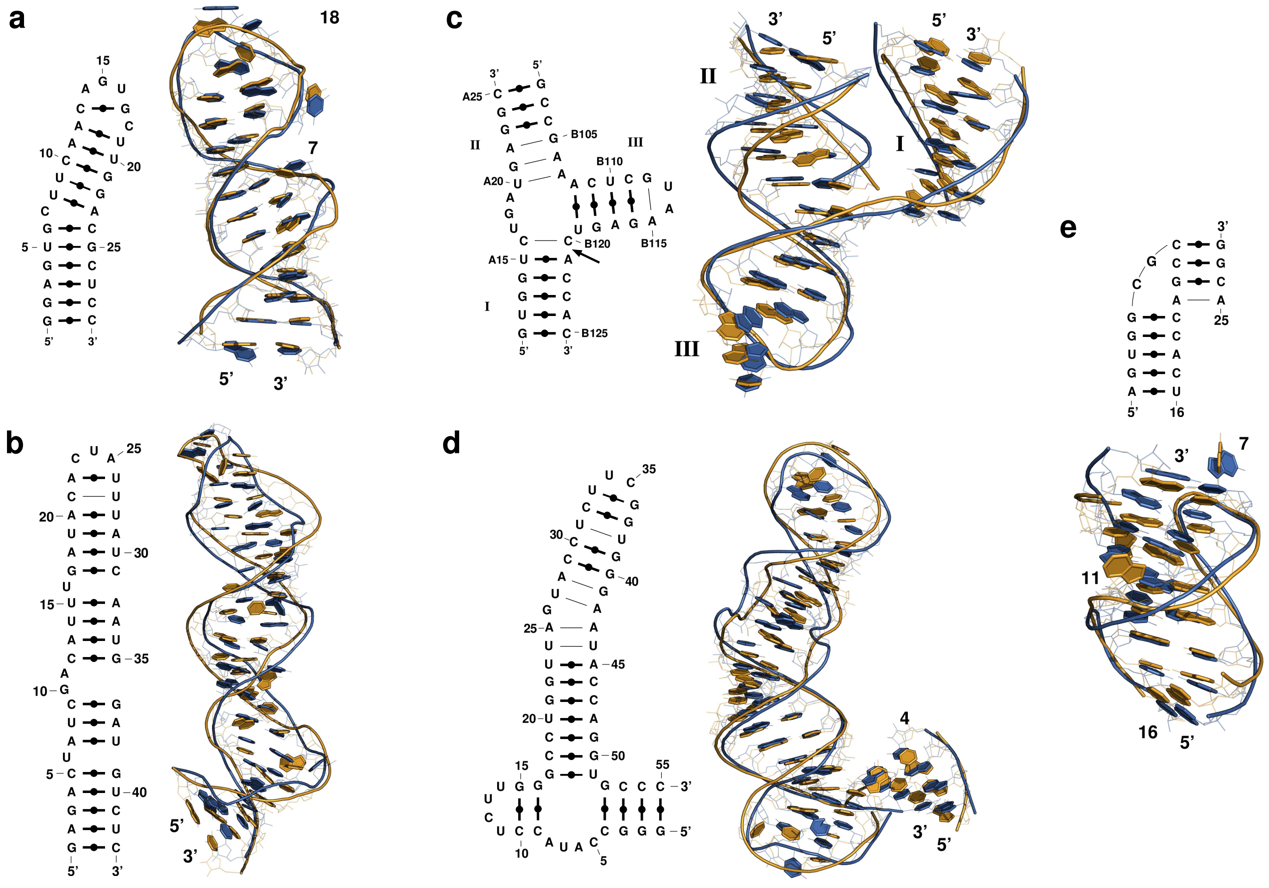


Figure 4.1: A selection of 3D structures predicted from sequence. The canonical (bold lines, black dots) and non-canonical (non-bold lines) base pairs predicted by MC-Fold are shown on the left of the 3D structures. The closest structure (minimum r.m.s.d.) over all heavy atoms is shown (blue) is superimposed on its respective experimental structure (gold). **a**, IRE. **b**, Telomerase RNA domain IV. **c**, Pre-catalytic conformation of a hammerhead ribozyme. The arrow points the cleavage site. **d**, Subdomain of the X. laevis 5S rRNA. **e**, Yellow leaf virus pseudo-knotted element.

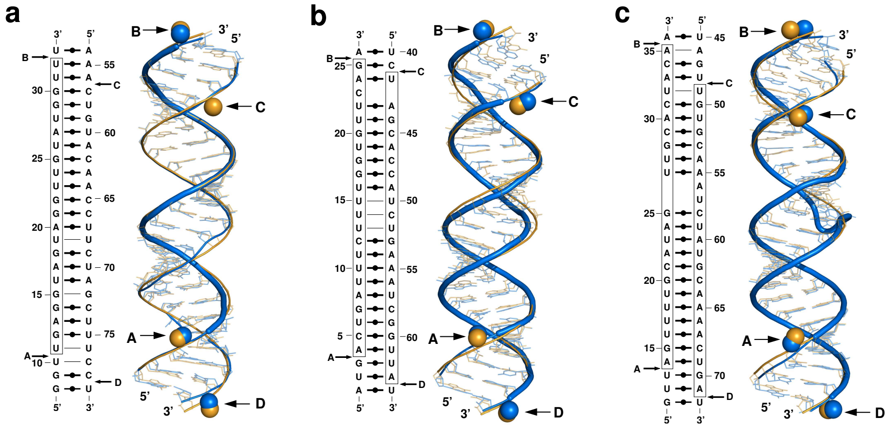


Figure 4.2: A selection of pre-miRNA 3D structures. The predicted structures (blue) are optimally superimposed on a theoretically generated A-RNA double helix (gold). For each pre-miRNA, the nucleotides that form the mature miRNA are shown inside boxes. The spheres represent the scissile phosphate atoms: Drosha complex cleavage (A and D) and Dicer (B and C). **a**, Human let-7c. **b**, Human miR-29a. **c**, Human miR-19a.

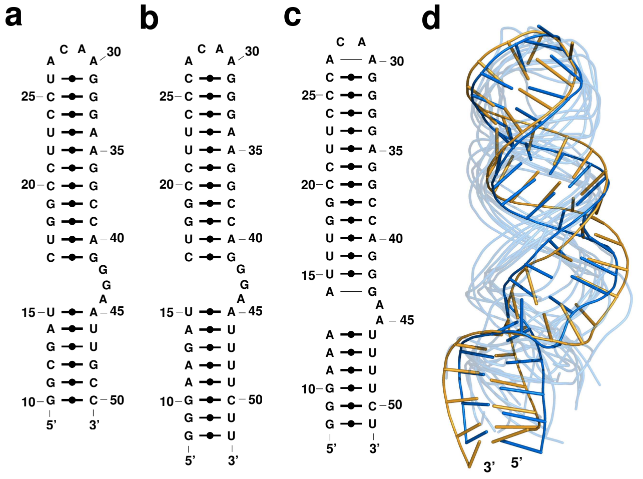


Figure 4.3: HIV-1 -1 frame-shifting-element models. **a**, Secondary structure of the first NMR study (Protein Data Bank, PDB, code 1ZC5). **b**, Second NMR study (PDB code 1Z2J). **c**, MC-Cons best secondary structure prediction. The sequence used is EMBL AJ535040.1 from patient PT747 that expresses the pol and gag proteins. **d**, MC-Sym tertiary structures. The closest model (minimum r.m.s.d. of 3.4 Å all heavy atoms) is shown in blue, optimally superimposed on the NMR structure resolved in the second study (gold), as well as ten representative structures of the conformational space of MC-Sym (light blue).

CHAPTER 5

ARTICLE 1; SUPPL. INF.

The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data

Marc Parisien and François Major

Institute for Research in Immunology and Cancer,
Department of Computer Science and Operations Research,
Université de Montréal,
PO Box 6128, Downtown station,
Montréal, Québec, H3C 3J7, CANADA

Parisien M, Major F.
Nature 2008 **452**:51-55.

5.1 Methods

5.1.1 RNA-Select

Using a simple pair-wise Smith-Waterman sequence comparison, we grouped together the RNA 3-D structures that have similar sequences. The most recently solved structure for each group was selected to form RNA-Select (Table 5.2)

5.1.2 NCM database

The NCM database contains lone-pair loops up to six nucleotides (including the flanking lone base pair; see Fig. 5.1 "output1") and double-stranded NCMs up to eight nucleotides (including both flanking base pairs). For lone-pair loops, we use the syntax "L-<sequence>", where L is the length of the loop and <sequence> is the sequence. Therefore, the NCM database contains 4 types and 5440 different lone-pair loop NCMs: 64 3-loops (3-AAA, 3-AAC, ... 3-UUU); 256 4-loops (4-AAAA, 4-AAAC, ... 4-UUUU); 1024 5-loops; and, 4096 6-loops. For double-stranded NCMs, we use the syntax "L1_L2-<sequence>", where L1 is the length of the 5'-strand, L2 is the length of the 3'-strand, and <sequence> is the sequence. Therefore, the NCM-database contains 15 types and 407808 different double-stranded NCMs. The 2_2-<sequence> NCMs represent the 256 base pairing tandems: 2_2-AAAA, 2_2-AAAC, ... 2_2-UUUU. The 3_2-<sequence> represents 1024 5'-strand single-nucleotide bulges, and the 2_3-<sequence> the 1024 3'-strand single-nucleotide bulges. Similarly, the 4_2-<sequence> represents 4096 5'-strand double-nucleotide bulges, and so on; 2_4 (4096 NCMs), 5_2 and 2_5 (2 x 16384 = 32768 NCMs), 6_2 and 2_6 (2 x 65536 = 131072 NCMs), 3_3 (4096 NCMs), 3_4 and 4_3 (2 x 16384 = 32768 NCMs), 3_5 and 5_3 (2 x 65536 = 131072 NCMs), and 4_4 (65536 NCMs). Because there are so many NCMs, the database is built in a just in time fashion, i.e. instances of the NCMs are built as the MC-Fold | MC-Sym pipeline needs them.

5.1.3 NCM building

First, we build a database of RNA backbone templates for each NCM: the phosphate groups, riboses, and glycosidic bonds. These correspond to each of the 19 NCM types. Second, we build a database of all possible base pairs: nucleobases and glycosidic bonds. Third, we align the four atoms of the glycosidic bonds of the base pairs with those of the backbone templates. A fit is found if the RMSD measured on the anchor points are within a user-defined precision in Å. Typically, we use values from 0.1 to 1.0 Å (for this study, we used 0.3 for the lone-pair loop and double-stranded NCMs).

5.1.4 MC-Fold structure enumeration

To generate the possible hairpins of a sequence, we first determine a list of initiation sites, which can be assigned lone-pair NCMs. Then, recursively, we match the rest of the sequence to double-stranded NCMs (see Fig. 5.10). Since we consider all possible positions for the initiation sites (even those of more than 6 nucleotides), this assignment process is in $O(N^2)$, where N is the length of the sequence. For each possible hairpin loop, we must find an assignment of approximately N/2 NCMs for the rest of the sequence. Since we have 15 double-stranded NCM types, this process is exponential, in $O(15^{N/2})$. This algorithm enumerates all possible NCM construction exhaustively. The various incompatibilities amongst NCM junctions limit the number of actual constructions, explaining why this algorithm works in practice (see Fig. 5.11).

For multi-branched structures, we use 4 indices: i, j, k, and l, i < j < k < l. We build stem-loops where the lone-pair of the hairpin is located at (j, k), and the last base pair in the stem at (i, l). We store them in a hyper-cube [(i, j) (k, l)]. We keep one (the best energy) stem-loop for each position, E[(i, j) (k, l)]. The time for filling the hyper-cube stays the same as described above, and the process results in a database of stem-loops, which we sort by the i indices. We then fill a dynamic programming table using the following recurrence equation:

$$E(i,l) = \min \begin{cases} E(i+1,l) \\ E(i,l-1) \\ \min\limits_{i<j<k<l} E\left[(i,j)(k,l)\right] \\ \min\limits_{i<p<l} E(i,p) + E(p+1,l) \end{cases} \tag{5.1}$$

The value E(1,N) gives the best possible energy for an assembly of stem-loops. Note the similarity between these recurrence equations and those of Nussinov-Jacobson [192]. In the top equation, nucleotide i is free and in the second equation nucleotide l is free. The third equation is for considering a stem, whereas the last equation is for considering a multi-branch structure. This process is in O(N^4) in time, due to the third equation, and does not consider pseudo-knotted structures. We do not mark the minimum value origins, as we do not need to reconstruct the minimum energy structure at this step.

The dynamic programming table is used to enumerate the sub-optimal solutions. We use the Waterman-Byers algorithm [200], which needs E_{min} = E(1,N), as well as a fraction of the energy, Δ, that limits the sub-optimal solutions considered. The energy of a sub-optimal returned by the algorithm is E, $E_{min} \leq E \leq E_{min} + \Delta$, which is the Waterman-Byers condition.

We solve the problem by backtracking over the stem variables. We pick one, two, three, and so on stems from a list, L, generated *a priori*. In other words, we compute the Cartesian products, {L}×{L}, {L}×{L}×{L}, and so on. We make sure that the selected stems are entirely embedded, i.e. j < i' < l' < k, as well as that they define distinct sequence regions, i.e. (i' > l). Each time a new stem is added, the Waterman-Byers condition is verified. The current energy is added to the minimum energy of the remaining sequence, E(j, k) + E(l, N$\backslash\Omega$), which are both available from the dynamic programming table. Ω is the set of the regions spanned by the previously selected stems (see Fig. 5.12). At anytime, if it is possible to build a structure that will respect the Waterman-Byers condition, then we continue the current construction; otherwise we try the next stem for the current variable or if no more stems are available, we backtrack to the previous stem-variable. This process is exponential and influenced greatly by the Δ value, which determines the probability of satisfying the condition. Haralick and Elliott developed a probabilistic time complexity model of backtracking algorithms in 1980 [236].

For pseudo-knotted structures, we squeeze in an extra stem, B, in a complete secondary structure, such that B creates the ABAB configuration with another stem, A, previously selected in the structure. The ABAB pseudo-knot configuration constitutes the vast majority of pseudoknots (also called H-type) [109]. Several Aalberts and Hodas rules about pseudoknot stem lengths were implemented [109]. The total pseudo-knot energy is that of its constituting stems, including the coaxial stacking contribution. Also, the Waterman-Byers condition must be relaxed to allow for the initial A-A- stem configuration (on which the ABAB pseudo-knot can form). This increases significantly the search space size, and thus computation time.

5.1.5 MC-Fold scoring function

MC-Fold generates a set of sub-optimal structures given a single input sequence. The structures are ranked by their probability of occurrence given the sequence. These scores are transformed in energies by assuming a Boltzmann distribution:

$$\Phi(\text{structure}|\text{sequence}) = -\text{RT} \cdot \ln \Psi(\text{structure}|\text{sequence}) \tag{5.2}$$

where RT has the value 0.606 kcal/mol.

The scoring function accounts for the probabilities of observing the NCMs given the sequence, their junctions, the base pairs in the context of the junctions, and the base pairs themselves, out of any context (Fig. 5.10). As a result, we obtain the following Master equation:

$$\Psi(\text{structure}|\text{seq}) = \tag{5.3}$$
$$\Psi(\text{NCMs}|\text{seq}) \cdot \Psi(\text{junctions}|\text{NCMs}) \cdot \Psi(\text{hinges}|\text{junctions}) \cdot \Psi(\text{pairs}|\text{hinges})$$

When a suite of NCMs is assigned to a sequence, each NCM, c_i, is mapped to a subsequence of the sequence, s_i. The sequence-NCM affinity is evaluated by the first term of the scoring function:

$$\Psi(\text{NCMs}|\text{sequence}) = \prod_{i}^{\text{NCMs}} \Psi(c_i|s_i) \tag{5.4}$$

which can be written as:

$$\Psi(c_i|s_i) = \frac{\Psi(s_i|c_i) \cdot \Psi(c_i)}{\Psi(s_i)} \tag{5.5}$$

using Bayes's theorem. Since $\Psi(c_i|s_i)$, the probability of c_i given s_i, cannot be computed directly, we compute $\Psi(s_i|c_i)$, the probability of observing s_i in c_i, $\Psi(s_i)$, the probability of s_i, and $\Psi(c_i)$, the probability of c_i. The probability of s_i, $\Psi(s_i)$, is the product of the occurrence probabilities of each nucleotide in s_i, or $\Psi_p(s_i)$.

Note that in the PDB we do not find every sequence within each NCM. To avoid null probabilities whenever a sequence cannot be found in a specific NCM, we accept sibling alternative sequences. Each nucleotide in the sequence is allowed the following IUPAC-IUB single-letter code lists: A:[A,R,M,N], C:[C,Y,M,N], G:[G,R,K,N], and U:[U,Y,K,N]. Consequently, a sequence of n nucleotides is represented by 4^n sequences. We call the generalized sequence, gs_i, the sequence that maximizes the ratio of the actual sequence probability within a given cycle on the *a priori* sequence probability:

$$\Psi(c_i|s_i) \propto \max_{g} \frac{\Psi(gs_i|c_i)}{\Psi_{apriori}(gs_i)} \tag{5.6}$$

Here, the maximization of the ratio prevents the over-generalization of the sequence into the degenerate N-only sequence.

For computation speedup, all sequence variations of each cycle were pre-calculated, and their worst probabilities, $\Psi(\text{c}_i|\text{s}_i)$, were arbitrarily assigned a maximum energy of +1.0 kcal/mol; the term $\Psi(\text{s}_i)$ has now been absorbed into the scaling of converting the probability into energy.

The second term evaluates the junction of two cycles, corresponding to a Markov chain of order 1:

$$\Psi(\text{junctions}|\text{NCMs}) = \prod_{(j,k)}^{\text{junctions}} \Psi(\text{junction}_{(j,k)}|\text{NCM}_j \wedge \text{NCM}_k) \tag{5.7}$$

where $\Psi(\text{junction}_{(j,k)}|\text{NCM}_j \wedge \text{NCM}_k)$ is the probability to observe a junction composed of NCM_j followed by NCM_k. The maximum energy associated with the lowest junction probabilities was arbitrarily assigned to +1.0 kcal/mol.

When two NCMs are joined, the base pairing type of the common base pair depends not only on the sequence, but also on the two NCMs. For example, the flanking base pair of a tri-loop must accommodate the sharp turn of the RNA backbone. Thus, the hinge can be scored by:

$$\Psi(\text{hinges}|\text{junctions}) = \prod_{l}^{\text{hinges}} \Psi(\text{hinge}_l|\text{junction}_{(j,k)}) \tag{5.8}$$

where $\Psi(\text{hinge}_l|\text{junction}_{(j,k)})$ is the probability of observing hinge_l at $\text{junction}_{(j,k)}$. Let $\Psi(\text{type}_m|\text{NCM}_j^l)$ be the probability to observe base pairing type m in NCM_j in hinge_l. To consider all base pairing types of the hinge, we must consider all common base-pairing types of NCM_j and NCM_k:

$$\Psi(\text{hinge}_l|\text{junction}_{(j,k)}) = \tag{5.9}$$
$$\sum_m^j \sum_n^k \delta_{m,n} \cdot \Psi(\text{type}_m|\text{NCM}_j^l) \cdot \Psi(\text{type}_n|\text{NCM}_k^l)$$

where δ is the Dirac delta function, which ensures that the joint probabilities are calculated for the common base pairing types only. This computation prevents the incorporation of an invalid base pair in the hinge (see Table 5.3).

Finally, once the hinge has been specified, we must quantify the specific nucleotide association of the base pair. Thus:

$$\Psi(\text{pairs}|\text{hinges}) = \prod_p^{\text{pairs}} \Psi(\text{pair}_p|\text{hinge}_l) \tag{5.10}$$

where $\Psi(\text{pair}_p|\text{hinge}_l)$ is the probability of observing pair$_p$ in the hinge$_l$. The maximum energy has been arbitrarily fixed to +1.0 kcal/mol.

5.1.6 Coaxial stacking energetic contributions

The coaxial stacking between two stems is scored accordingly to the creation of a new 2_2 NCM between the two stems. This NCM is similar to the others of its class, but lacks one phosphodiester linkage, which is substituted by a base stacking interaction. The total energetic contribution of coaxial stacking, therefore, comes from the new NCM itself, i.e. its fitness to the sequence, as well as from the two new junctions (-2.9 kcal/mol). An entropy cost of +2.5 kcal/mol is added for the loss of the phosphodiester linkage. This arbitrary value is a compromise between single and multi-branched structures: low costs favour multi-branched structures; high costs hairpins.

5.1.7 MC-Sym structure generation

Libraries of 3-D fragments corresponding to each NCM are built (see NCM building above). The NCM fusion in MC-Sym is conceptually equivalent to that of MC-Fold, i.e. all possible NCM 3-D fragments are systematically assigned to the sequence. However, since MC-Fold has already assigned a score, no scoring is necessary. The concatenation of two adjacent NCMs is done by optimal superimposition of the two copies of the common base pair in 3-D. Since there are many possible NCM 3-D fragments for each NCM, an exhaustive assignment is prohibitive. Instead, a Las Vegas algorithm is used to explore as many structures as possible in a given period of time, fixed for this study to 12h. The difference between the Las Vegas and the better-known Monte Carlo algorithms is that the former never gives an incorrect result, i.e. all 3-D structures generated by MC-Sym are consistent with the input constraints.

5.1.8 MC-Fold | MC-Sym pipeline

The pipeline is described in Fig. 5.1. Input 1 is a single sequence. MC-Fold performs the NCM fusion 1 and returns a sorted list of possible structures in dot-bracket notations (Fig. 5.13). An MC-Sym input script for any MC-Fold solution can be generated by providing it in the “mask” field of MC-Fold (see Fig. 5.14). This represents Input 2 in the pipeline diagram of Fig. 5.1. MC-Sym is invoked and run for 24 hours, producing atomic-precision 3-D models that satisfy the interactions specified in the script. An RMSD threshold for each NCM merge, an overall atomic clash constraint, a ribose construction threshold, an implicit phosphate restraint, a time limit or a maximum number of models, and a threshold RMSD amongst the models produced parameterize MC-Sym. These values can be edited in the script generated by MC-Fold. However, default values for these parameters are fixed, and the scripts generated by MC-Fold can be submitted to MC-Sym without editing. The output of MC-Sym is a set of 3-D structures in PDB format [39] (Fig. 5.15).

5.1.9 MC-Cons

The algorithm MC-Cons does not find a consensus structure deprived of many base pairs that fit all sequences of an RNA family. Instead, we assign to each sequence one of its suboptimal predictions that globally optimizes the sum of pair-wise similarities. In other words, we look for a global and structural consensus assignment (that may include more than one structures) rather than for a common structure. This is similar to the concept of RNA “shapes” proposed by Reeder and Giegerich [237]. First, a similarity score is computed for each pair of suboptimal solutions and stored in a similarity matrix. This score is largely biased towards structural alignment, rather than sequence alignment. Then, from the similarity matrix, the maximum sum is found by backtracking over all suboptimal solutions. As the sequence-structure space grows exponentially, a cyclic coordinate method [238], where the optimal structure of one sequence is searched while all others are fixed, is used as an optimization heuristic. We then apply hierarchical clustering to unveil the structural features of the consensus assignment.

5.1.10 RNA structure images

The 3-D structures were rendered using PyMOL. The secondary structure were rendered using a modified version of the CONTRAfold renderer [204].

5.2 Discussion

5.2.1 Arguments in favour of a new HIV-1 -1 frameshifting element

First, the double A bulge is conserved across all 753 sequences, suggesting a possible functional role. It can adopt the A-minor motif that can simultaneously kink the structure [235] and dock to any tandem of Watson-Crick base pairs [114]. In comparison, the GGA bulge is found in half of these sequences (Fig. 5.9), substituted by a GAA bulge in the other half. G and A have different chemical groups and, in general, cannot easily be substituted. Second, the flanking base pair above the bulge can either be GA or AA, which are frequent and stable at the end of double-helical stems [239]. Third, the model satisfies enzymatic probing data applied to the native sequence from two studies [234, 240]. Fourth, the model applies to all HIV-1 subtypes, introducing three times less NC base pairs in only one rather than three sites in the NMR model [241]. Fifth, our has lower thermodynamic average energies than the NMR model (-23.0 vs. -21.3 kcal/mol; as computed by the RNAeval program of the Vienna package [178]). Sixth, the model corroborates with recent enzymatic cleavage data that indicate an unpaired nucleotide A45 [242].

Predicted base pairs (%)	Zipper (lower bound	RNAsubopt (thermodynamics)	CONTRAfold (machine learning)	MC-Fold (NCM)
False positives	50.2	**6.7**	7.5	12.9
False negatives	25.9	25.2	26.9	**10.1**
True positives	74.1	74.8	73.1	**89.9**
Canonicals	75.6	88.4	86.3	**94.7**
Non-canonicals	**64.9**	N/A	1.4	62.1
Matthews = $\sqrt{\frac{TP}{(TP+FN)}\frac{TP}{(TP+FP)}}$	66.5	82.8	81.4	**86.6**

Table 5.1: Comparison of the predictive power of three approaches. The predictions of three approaches are compared over 1968 base pairs (1665 Watson-Crick) in 264 hairpins extracted from 182 different PDB structures. Zipper implements a greedy algorithm that folds a sequence from bottom-up using exclusively tandems of base pairs. This gives us a lower bound on the predictive power. RNAsubopt implements the current thermodynamics model and enumerates exhaustively all suboptimal solutions. For each approach, the best predicted structures are analyzed. In each row, the best value is shown in bold. By increasing the number of sub-optimal solutions to 5, the Matthews coefficient ratios go up to 93.1 (99.1% of the canonical base pairs) and 87.7 (97.3% of the canonical base pairs), respectively for MC-Fold and RNAsubopt. Interestingly, MC-Fold's ratio reaches 92.2 when the top 2 solutions are analyzed (RNAsubopt 86.3).

104D	124D	157D	168D	170D	176D	17RA	1A34	1A4T	1A51
1A60	1A9N	1AFX	1AJF	1AJT	1AL5	1AM0	1APG	1ATO	1ATV
1ATW	1AUD	1AV6	1B23	1B36	1B7F	1BAU	1BGZ	1BJ2	1BMV
1BN0	1BR3	1BVJ	1BYJ	1BYX	1BZ2	1BZT	1C0A	1C0O	1C2Q
1C4L	1C9S	1CK5	1CQ5	1CSL	1CVJ	1CX0	1CX5	1D0T	1D0U
1D4R	1D6K	1D9H	1DDL	1DDY	1DFU	1DQF	1DRR	1DUH	1DUL
1DUQ	1DXN	1DZ5	1E4P	1E7K	1E95	1EBR	1EC6	1EFO	1EFS
1EFW	1EHZ	1EJZ	1EKA	1EKD	1EKZ	1ELH	1ESH	1ET4	1EUY
1EVP	1EXD	1EXY	1F27	1F5G	1F5U	1F6U	1F6X	1F6Z	1F7U
1F84	1F85	1F8V	1F9L	1FEQ	1FEU	1FG0	1FHK	1FIX	1FL8
1FMN	1FNX	1FQZ	1FUF	1FYO	1G1X	1G2E	1G2J	1G3A	1G4Q
1G70	1GKW	1GSG	1GTF	1GTN	1GUC	1H0Q	1H2C	1H2D	1H38
1H3E	1H4S	1HC8	1HJI	1HLX	1HO6	1HOQ	1HS1	1HS2	1HS3
1HS4	1HS8	1HWQ	1HYS	1I2X	1I2Y	1I3X	1I3Y	1I46	1I4B
1I5L	1I6U	1I7J	1I9F	1I9K	1I9V	1I9X	1ICG	1IDV	1IE1
1IK1	1IK5	1IKD	1IL2	1IVS	1J1U	1J4Y	1J6S	1J8G	1J9H
1JBR	1JBT	1JID	1JO7	1JOX	1JTJ	1JTW	1JU7	1JUR	1JZC
1JZV	1K1G	1K2G	1K4A	1K4B	1K5I	1K6G	1K6H	1K8S	1KAJ
1KD3	1KFO	1KH6	1KIS	1KKS	1KNZ	1KOC	1KOD	1KOS	1KP7
1KPD	1KPY	1KQ2	1KUO	1KUQ	1KXK	1L1C	1L1W	1L2X	1L3Z
1L8V	1L9A	1LDZ	1LMV	1LNT	1LPW	1LUU	1LUX	1LVJ	1M5K
1M5L	1M82	1M8V	1M8W	1M8X	1M8Y	1MDG	1ME0	1ME1	1MFJ
1MFK	1MFY	1MHK	1MHM	1MIS	1MJI	1MMS	1MNX	1MSY	1MT4
1MUV	1MV1	1MV6	1MWG	1MY9	1MZP	1N1H	1N35	1N38	1N53
1N66	1N77	1N7A	1N8X	1NA2	1NAO	1NB7	1NBK	1NBR	1NC0
1NEM	1NTA	1NTQ	1NTS	1NTT	1NUJ	1NXR	1NYB	1NZ1	1O15
1OKF	1OLN	1OO7	1OOA	1OQ0	1OSU	1OSW	1OW9	1P5M	1P5N
1P5O	1P79	1PBL	1PGL	1PJY	1PVO	1Q29	1Q75	1Q8N	1Q93
1Q96	1Q9A	1QBP	1QC0	1QC8	1QD3	1QES	1QET	1QF6	1QLN
1QU2	1QWB	1R2P	1R3E	1R3O	1R3X	1R4H	1R7W	1R7Z	1RAW
1RC7	1RFR	1RGO	1RKJ	1RLG	1RMV	1RNA	1RNG	1RNK	1ROQ
1RPU	1RXA	1S03	1S2F	1S76	1S9L	1SA9	1SAQ	1SDR	1SDS
1SER	1SI3	1SJ3	1SLP	1SYZ	1SZY	1T0D	1T0E	1T28	1T2R
1T4L	1T4X	1TFN	1TFW	1TJZ	1TLR	1TOB	1TTT	1TUT	1TXS
1U0B	1U2A	1U3K	1U6P	1U8D	1U9S	1ULL	1UTD	1UUD	1UUU
1UVJ	1UVK	1UVL	1UVN	1VFG	1VOP	1VQ7	1WKS	1WNE	1WPU
1WRQ	1WSU	1WTS	1WWD	1WWE	1WWF	1WWG	1XHP	1XJR	1XMQ
1XOK	1XP7	1XPE	1XPF	1XSG	1XSH	1XV0	1XV6	1XWP	1XWU
1Y26	1Y27	1Y39	1Y3O	1YFG	1YFV	1YG3	1YMO	1YN1	1YNC
1YNE	1YSV	1YTU	1YTY	1YVP	1YYK	1YYW	1YZ9	1Z2J	1Z30
1Z31	1Z43	1Z7F	1ZBI	1ZC5	1ZCI	1ZDJ	1ZDK	1ZE2	1ZEV
1ZFV	1ZIF	1ZIG	1ZIH	1ZJW	1ZL3	1ZX7	1ZZ5	205D	216D
219D	246D	247D	255D	259D	280D	283D	28SP	2A0P	2A1R
2A43	2A8V	2A9X	2AB4	2AD9	2ADC	2ADT	2AO5	2ASB	2ATW
2AU4	2AWE	2AWQ	2AZ0	2B3J	2B6G	2BBV	2BE0	2BGG	2BH2
2BJ6	2BNY	2BS0	2BS1	2BTE	2BX2	2C06	2C4Y	2C4Z	2C50
2C51	2CHJ	2CSX	2D17	2D18	2D1A	2ERR	2ES5	2ESI	2EUY
2EZ6	2F4X	2F88	2F8K	2FK6	2FMT	2FQN	2FRL	2FZ2	2G1W
2G8F	2G92	2GBH	2GM0	2TOB	2TPK	2TRA	310D	315D	332D
333D	353D	354D	361D	364D	377D	393D	397D	398D	3PHP
402D	404D	405D	409D	413D	418D	419D	420D	421D	422D
429D	430D	433D	435D	438D	439D	464D	466D	468D	469D
470D	471D	472D	479D	484D	485D	5MSF	6MSF	7MSF	8DRH
8PSH									

Table 5.2: RNA-Select. The 531 PDB codes corresponding to the X-ray crystallographic and NMR structures.

Number of occurences	Probability of appearance (%)	Base pairing type
5' G-A base pair of NCM$_i$	$\Psi_{m\|i}(\text{type}_m\|\text{NCM}^l_i)$	
72	0.889	S/H anti trans
3	0.037	S/W anti cis
3	0.037	S/W para trans
2	0.025	W/H anti trans
3' G-A base pair of NCM$_j$	$\Psi_{n\|j}(\text{type}_n\|\text{NCM}^l_j)$	
161	0.821	S/H anti trans
29	0.148	W/W anti cis
2	0.010	H/W para cis
2	0.010	W/B anti cis
2	0.010	W/H anti trans
G-A base pair of hinge$_l$	$\sum_m^j \sum_n^k \delta_{m,n} \cdot \Psi(\text{type}_m\|\text{NCM}^l_j) \cdot \Psi(\text{type}_n\|\text{NCM}^l_k) = 0.731$	
	0.889 × 0.821 = 0.730	S/H anti trans
	0.037 × 0.000 = 0.000	S/W anti cis
	0.000 × 0.148 = 0.000	W/W anti cis
	0.025 × 0.010 $\approx$ 0.001	W/H anti trans

Table 5.3: Hinge scoring. The score of a GA hinge, l, at the junction of NCMs i (4-GAGA) and j (2_2-CGAG) is 0.731: the sum of the products of the probabilities of appearance, Ψ, of the GA base pairing types, m and n, found in the instances of the two NCMs in RNA-Select, independently of the junction. The sheared GA base pair (S/H anti) validates the hinge created by the junction of the two cycles since it is the most frequent among all possible base pairs (probability of 0.730). The Watson-Crick/Hoogsteen is another valid option, but is less likely to appear in this context (probability of 0.001).

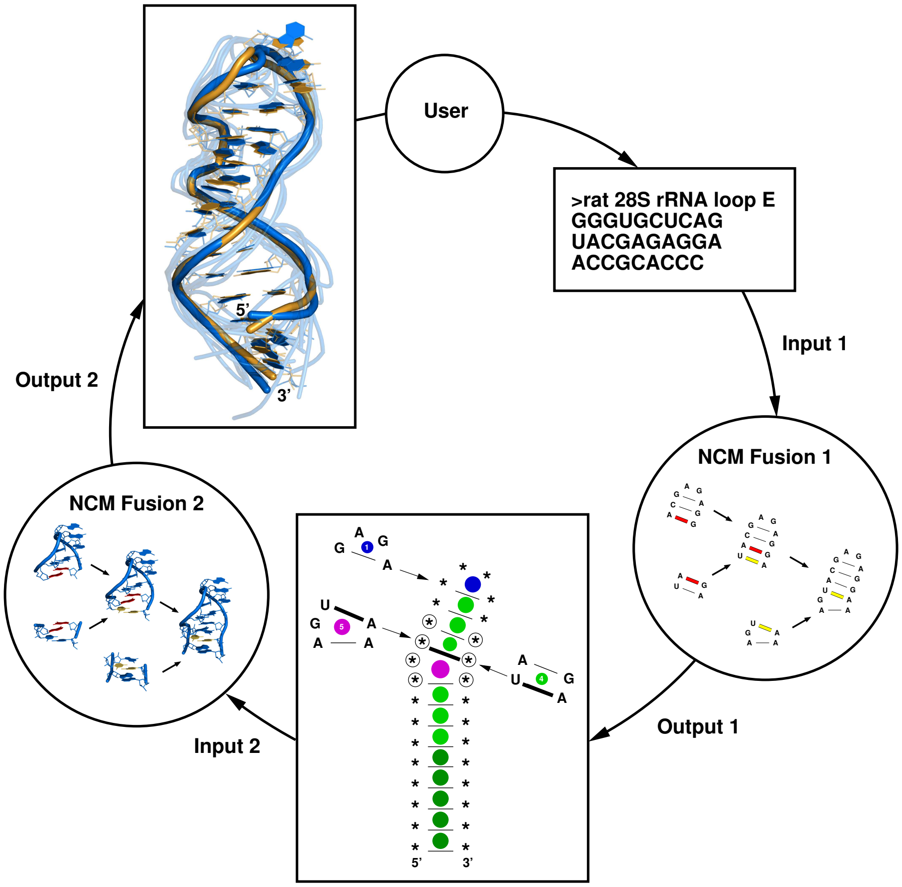


Figure 5.1: The MC-Fold | MC-Sym pipeline applied to the rRNA loop E. Input 1: Sequence of the rat 28S rRNA loop E. NCM Fusion 1: MC-Fold. Two adjacent NCMs share a common hinge base pair (red and yellow). Output 1/Input 2: The optimal assignment contains 13 NCMs (circles), 14 base pairs (lines), and 29 nucleotides (stars). The three main NCM types are shown: blue) lone-pair loops (GAGA tetraloop; NCM #1); green) base pair tandems (dark green indicates canonical tandems); and, purple) bulge and interior loops, an extension of the base pair tandem. The NC UA hinge base pair (bold line) is common to NCMs #4 and #5, which combination forms the sarcin/ricin motif. Each stem-loop is one chain of NCMs. Since the output of MC-Fold can be a multi-branch or pseudo-knotted structure made of more than onoe hairpin, the output is a set of chains of NCMs. NCM Fusion 2: MC-Sym. Output 2: The closest prediction (blue) that shares 1.8 Å of RMSD and a representative sampling of structures (light blue) are shown optimally superimposed on the rat 28S rRNA X-ray crystallographic loop E structure (gold).

```
>tRNA ASP
GCCGUGAUAGUUUAAUGGUCAGAAUGGGCGCUUGUCGCGUGCCAGAUCGGGGUUCAAUUCCCCGUCGCGGCGC
.................xx.............xxxx.................x...................  High
...............xx..x................xx................xx.................  Medium
(((((((..(((((......)))))((((((.......))))))...((((((...)..))))))))))))..  native - 2TRA  RNK TP FP FN Mthw
(((((((((((((.......))))))(((((.......)))))....((((((...)..)))))))))))).. -59.75 ( -1.25)   1 18  6  6  75.0
(((((((.(((((.......)))))((((((.......))))))...((((((...)..))))))))))))).. -59.73 ( -0.84)   2 19  5  5  79.2
(((((((.(.(((.......))))(((((((.......)))))))..((((((...)..))))))))))))).. -59.47 ( -0.71)   3 19  5  5  79.2
(((((((..((((.......))))(((((((.......)))))))..((((((...)..))))))))))))).. -58.72 ( -0.79)   4 19  5  5  79.2
(((((((.(.(((.......))))(((((((.......)))))).).((((((...)..))))))))))))).. -58.69 ( -0.71)   5 19  5  5  79.2
(((((((..(((((......)))))((((((.......))))))...((((((...)..))))))))))))).. -58.46 ( -0.71)   6 24  0  0 100.0**
(((((((...(((.......)))((((((((.......)))))).))((((((...)..))))))))))))).. -58.25 ( -0.71)   7 19  5  5  79.2
(((((((...((((......))))(((((((.......)))))))..((((((...)..))))))))))))).. -58.23 ( -1.07)   8 23  1  1  95.8
(((((((.((((((......))))))(((((.......)))))....((((((...)..))))))))))))).. -58.02 ( -1.14)   9 23  1  1  95.8
(((((((...(((.......)))((((((((.......)))))))).((((((...)..))))))))))))).. -57.93 ( -0.71)  10 19  5  5  79.2
```

Figure 5.2: MC-Fold predictions for the yeast tRNA^{ASP}. The top ten structures generated by MC-Fold for the Yeast tRNA under SHAPE constraints are shown. The native structure (PDB file 2TRA) ranks 6th (Matthews coefficient ratio of 100%). The numbers in parenthesis represent the energy contributions of the coaxial stacking. The nucleotides marked with a dot under the “High” SHAPE constraints have an 8 kcal/mol penalty if found paired; 4 kcal/mol for “Medium”. Nucleotide 47 is absent. Real time = 159 seconds.

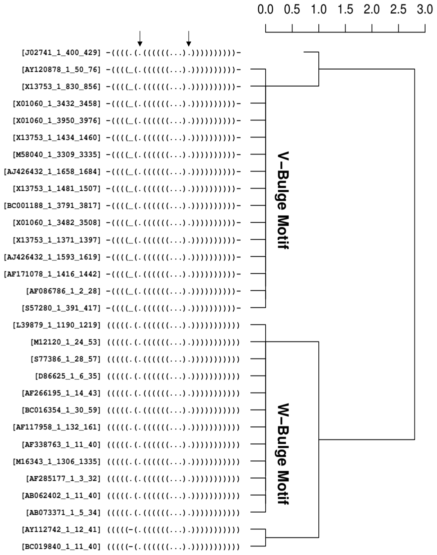
```
>E.coli 5
1   5   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95  100  105  110  115  120
|   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
UGCCUGGCGGCCUUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
..............x.............................x...........................................................................  strong DMS
......................................x............xx.....x......x......................................................  moderate DMS
(((((((((((....((((((((....(((((((.............))))..)))...)))))).)).(((((((((((((((((...)))))))))))))))))..))))))))))).  native (2AW4)   RNK TP FP FN Mthw
(((((((((((((..(((((((((((..((((((((.(.....).))))))..))))).)))))).)).(((((((((((((((((...)))))))))))))))))))))))))))))))). -110.44 ( -2.02)   1 42  8  1 90.6
(((((((((((((..(((((((((((..(((((((((......).))))))..))))).)))))).)).(((((((((((((((((...)))))))))))))))))))))))))))))))). -110.08 ( -2.02)   2 42  8  1 90.6
(((((((((((.(..(((((((((((..((((((((.(.....).))))))..))))).)))))).))((((((((((((((((((...)))))))))))))))))))))))))))))))). -109.35 ( +0.00)   3 42  8  1 90.6
((((((((((((...(((((((((((..((((((((.(.....).))))))..))))).)))))).))((((((((((((((((((...)))))))))))))))))))))))))))))))). -109.29 ( -1.26)   4 42  8  1 90.6
(((((((((((.((.(((((((((((..((((((((.(.....).))))))..))))).)))))).)).(((((((((((((((((...)))))))))))))))))))))))))))))))). -109.01 ( +0.00)   5 42  8  1 90.6
```

(a)

```
20   25   30   35   40   45   50   55   60   65   70
|    |    |    |    |    |    |    |    |    |    |
aGCGCGGUGgUcCCacCUGAcccCAUGCCGaacUCAGaaGUGaAaCGCCGUAGCg
((((((((((((((((((((((((.....))))))))))).)))))))))).))) -45.05 ( +0.00)
```

(b)

Figure 5.3: MC-Fold predictions for the E. coli 5S rRNA. **a**. The top 5 structures generated by MC-Fold for the *E. coli* 5S rRNA under DMS constraints are shown. The native structure (PDB file 2AW4) is not predicted (best Matthews coefficient ratio 90.6%). The numbers in parenthesis represent the energy contributions of the coaxial stacking. The nucleotides marked with a dot under the “strong” DMS reactivity have an 8 kcal/mol penalty if found paired; 4 kcal/mol for “moderate”. Real time = 2131 seconds. **b**. The optimal MC-Fold solution of the 16-69 *E. coli* 5S rRNA subsequence. The NC base pairs are shown using lowercase letters.



(a)

EMBL Number	Sequence / Structure	Rank
Consensus	((((((.((((((...).)))))))))))	
AY112742_1_12_41	GUcCUGCUUCAACAGUGCUUGAACGGaAC	(1st)
BC019840_1_11_40	GUcUUGCUUCAACAGUGUUUGAACGGaAC	(1st)
AF086786_1_2_28	-UcGUUCGUCCUCAGUGCAGGGCAACaG-	(1st)
S57280_1_391_417	-UcGUUCGUCCUCAGUGCAGGGCAAUaG-	(5th)
AF171078_1_1416_1442	-UgGUUCGUCCUCAGUGCAGGGCAACaG-	(1st)
AJ426432_1_1593_1619	-AUUAUCGGGAGCAGUGUCUUCCAUAAU-	(1st)
X13753_1_1371_1397	-AUUAUCGGGGGCAGUGUCUUCCAUAAU-	(3rd)
X01060_1_3482_3508	-AUUAUCGGAAGCAGUGCCUUCCAUAAU-	(1st)
BC001188_1_3791_3817	-AUUAUCGGGAACAGUGUUUCCCAUAAU-	(3rd)
X13753_1_1481_1507	-AUUAUCGGGGACAGUGUUUCCCAUAAU-	(3rd)
AJ426432_1_1658_1684	-UAUAUCGGAGACAGUGAUCUCCAUAUG-	(1st)
M58040_1_3309_3335	-UAUAUCGGAGaCAGUGAcCUCCAUAUG-	(1st)
X13753_1_1434_1460	-UAUAUCGGAGGCAGUGACCUCCAUAUG-	(1st)
X01060_1_3950_3976	-UGUAUCGGAGACAGUGAUCUCCAUAUG-	(1st)
X01060_1_3432_3458	-UUUAUCAGUGACAGAGUUCACUAUAAA-	(1st)
X13753_1_830_856	-UUUAUCAGUGACAGCGUUCACUAUAAA-	(1st)
AY120878_1_50_76	-GgUCGCGUCAACAGUGUUUGAUCGAaC-	(1st)
Consensus	(((((.(.((((((...).)))))))))))	
AB062402_1_11_40	uUUCCUGCUUCAACAGUGCUUGGACGGAAc	(1st)
AB073371_1_5_34	uCUCCUGCUUCAACAGUGCUUGGACGGAGc	(1st)
AF285177_1_3_32	GUUCCUGCUUCAACAGUGCUUGGACGGAAC	(1st)
M16343_1_1306_1335	GUUCCUGCgUCAACAGUGCUUGGaCGGAAC	(2nd)
AF338763_1_11_40	UUaCCUGCUUCAACAGUGCUUGAACGGcAA	(1st)
AF117958_1_132_161	ucUCUUGUUUCAACAGUGUUUGGACGGAac	(21th)
BC016354_1_30_59	ucUCUUGCUUCAACAGUGUUUGGACGGAac	(2nd)
AF266195_1_14_43	AUUCUUGCUUCAACAGUGUUUGAACGGAAU	(1st)
D86625_1_6_35	GUUCUUGUUUCAACAGUGAUUGAACGGAAC	(1st)
S77386_1_28_57	GUUCUUGCUUCAACAGUGAUUGAACGGAAC	(1st)
M12120_1_24_53	GUUCUUGCUUCAACAGUGUUUGAACGGAAC	(1st)
L39879_1_1190_1219	GUaCUUGCUUCAACAGUGUUUGAACGGaAC	(1st)
J02741_1_400_429	-aUCUUGCUUCAACAGUGUUUGGACGGAa-	(1st)

(b)

Figure 5.4: Clustering and aligned IRE sequences. **a.** The results of a hierarchical clustering of the predicted structures identified by MC-Cons using inputs from MC-Fold. Each sequence is identified by its EMBL identifier, and as found in the Rfam database. A structural distance of 0 indicates identical structures. The IRE sequences are clearly grouped in their respective structural class: the V-bulge (above) and the W bulge (below). The W bulge is recognized in the bracket notation by the typical "((.(.((", whereas the V bulge is recognized by "((. ((". The arrows indicate the C involved in IRE function. MC-Cons determines the IRE consensus assignment in about 10 minutes. **b.** The alignment was made according to consensus structures identified by MC-Cons. The sequences are divided in two groups: the V-bulge (up) and the W-bulge (down). The non-canonical base pairs are highlighted using lowercase letters.

```
>tRNA-ASN
GACUCCAUGGCCAAGUUGGUUAAGGCGUGCGACUGUUAAUCGCAAGAUCGUGAGUUCAACCCUCACUGGGGUCGCCA
(((((((..((((.........))))((((((((...))))))))....((((((...)..)))))))))))).... (    2nd)
...............xx..xx.............xxx..........x.............................
 >tRNA-GLY
GCGCAAGUGGUUUAGUGGUAAAAUCCAACGUUGCCAUCGUUGGGCCCCGGUUCGAUUCCGGGCUUGCGCACCA
(((((((..(((((.....)))))((((((((...))))))))..((((((...)..))))))))))))....     (    2nd)
........x......x..x.............xxx......................................
 >tRNA-ILE
GGUCUCUUGGCCCAGUUGGUUAAGGCACCGUGCUAAUAACGCGGGGAUCAGCGGUUCGAUCCCGCUAGAGACCACCA
(((((((..(((((.......)))))((((((((...))))))))....((((((...)..)))))))))))).... (    5th)
........x......xx..xx.............xxx..........x.............................
 >tRNA-LYS
UCCUUGUUAGCUCAGUUGGUAGAGCGUUCGGCUUUUAACCGAAAUGUCAGGGGUUCGAGCCCCCUAUGAGGAGCCA
(((((((..(((((......)))))((((((((...))))))))....((((((...)..))))))))))))....  (    1st)
...............xx..x.............xxx..........x.............................
 >tRNA-MET
GCUUCAGUAGCUCAGUAGGAAGAGCGUCAGUCUCAUAAUCUGAAGGUCGAGAGUUCGAACCUCUCCUGGAGCACCA
(((((((..(((((......)))))((((((((...))))))))....((((((...)..))))))))))))....  (    5th)
...............x.................xxx..........x.............................
 >tRNA-THR
GCUUCUAUGGCCAAGUUGGUAAGGCGCCACACUAGUAAUGUGGAGAUCAUCGGUUCAAAUCCGAUUGGAAGCACCA
(((((((..((((........))))((((((((...))))))))....((((((...)..))))))))))))....  (    1st)
...............xx..x.............xxx..........x.............................
 >tRNA-TRP
GAAGCGGUGGCUCAAUGGUAGAGCUUUCGACUCCAAAUCGAAGGGUUGCAGGUUCAAUUCCUGUCCGUUUCACCA
(((((((..(((((.....)))))((((((((...))))))))....((((((...)..))))))))))))....   (    1st)
........x......x..x.............xxx..........x.............................
 >tRNA-ALA
GGGCGUGUGGCGUAGUCGGUAGCGCGCUCCCUUAGCAUGGGAGAGGUCUCCGGUUCGAUUCCGGACUCGUCCACCA
(((((((..(((((......)))))(((((((.....)))))))....((((((...)..))))))))))))....  (   60th)
........x......x...x.............xxxx.........x.............................
 >tRNA-ARG
UUCCUCGUGGCCCAAUGGUCACGGCGUCUGGCUACGAACCAGAAGAUUCCAGGUUCAAGUCCUGGCGGGGAAGCCA
(((((((..(((((......)))))((((((((...))))))))....((((((...)..))))))))))))....  (   48th)
........x......x..x..............xxx..........x.............................
 >tRNA-ASP
UCCGUGAUAGUUUAAUGGUCAGAAUGGGCGCUUGUCGCGUGCCAGAUCGGGGUUCAAUUCCCCGUCGCGGAGCCA
(((((((..(((((......)))))((((((.......))))))...((((((...)..))))))))))))....     2TRA    TP FP FN Mthw
(((((((..(((((......)))))(((((((.....)))))))...((((((...)..))))))))))))....   (    5th) 24  1  0 98.0
...............x..x..............xxxx......................................
>tRNA-GLU
UCCGAUAUAGUGUAACGGCUAUCACAUCACGCUUUCACCGUGGAGACCGGGGUUCGACUCCCCGUAUCGGAGCCA
(((((((..(((((......)))))((((((((...))))))))...((((((...)..))))))))))))....   (   55th)
...................x.............xxx.......................................
 >tRNA-HIS
GGCCAUCUUAGUAUAGUGGUUAGUACACAACAUUGUGGCUGUUGAAACCCUGGUUCGAUUCUAGGAGGUGGCACCA
((((((((..(((((......)))))(((((((.....)))))))...((((((...)..)))))))))))).)..  (   13th)
................x..xx.............xxxx......................................
 >tRNA-PHE
GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA
(((((((..((((........))))(((((((.....)))))))....((((((...)..))))))))))))....    4TRA    TP FP FN Mthw
(((((((..(((((......)))))((((((.(....)))))))....((((((...)..))))))))))))....  (  336th) 23  2  1 93.9
...............xx................xxxx.......................................
>tRNA-VAL
GGUUUCGUGGUCUAGUCGGUUAUGGCAUCUGCUUAACACGCAGAACGUCCCCAGUUCGAUCCUGGGCGAAAUCACCA
(((((((..(((((.......)))))((((((((...))))))))....((((((...)..)))))))))))).... (   11th)
........x......x...xx.............xxx..........x.............................
```

Figure 5.5: Consensus structural assignment for yeast tRNA sequences. The yeast non-mitochondrial tRNA sequences are from the September 2004 edition of the compilation of tRNA sequences and sequences of tRNA genes database. The modified nucleotides in MC-Fold are treated like their canonical counterparts. The modified nucleotides that cannot adopt the A-RNA helix are constrained. For each tRNA, the anticodon nucleotides are unpaired. The positions marked with 'x' are either modified nucleotides that cannot form the A-RNA helix (unpaired), or anticodon nucleotides. The average real time to fold each tRNA sequence is 223.6 sec. MC-Cons determines the consensus structural assignment in about 53 minutes.

```
>E.coli 1
UGCCUGGCGGCCGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..))))))))))).   (   2nd)
>E.coli 2
UGCCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCUCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..))))))))))).   (   5th)
>E.coli 3
UGCCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..))))))))))).   (   3rd)
>E.coli 4
UGUCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGACAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..))))))))))).   (   3rd)
>E.coli 5
UGCCUGGCGGCCUUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
..........xx..x.............................x...........................................................................   strong DMS
..................................x..xx............xx.....x......x......................................................   moderate DMS
(((((((((((....((((((((....(((((((.............))))..)))...)))))).)).(((((((((((((((((...)))))))))))))))))..))))))))))).   native (2AW4)  TP FP FN Mthw
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..))))))))))).   (   5th)       41  4  2 93.2
(((((((((((((((((((((((....(((((((.............))))..)))...)))))).))(((...))).((((((((...)))))))).....)))))..))))))))))).   Mathews et al. 33  8 10 78.6
>E.coli 6
UGUCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGACUCCCCAUGCGAGAGUAGGGAACUGCCAGACAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..))))))))))).   (   3rd)
>E.coli 8
UGCCUGGCGGCCUUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGGCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..))))))))))).   (   5th)
>E.coli 10
UGUCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCUCAUGCGAGAGUAGGGAACUGCCAUGCAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..))))))))))).   (   2nd)
>E.coli 11
UGCCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAUCA
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..)))))))))))... (   3rd)
>E.coli 14
UGCCUGGCGGCCGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGACAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..))))))))))).   (   2nd)
```

Figure 5.6: MC-Cons consensus assignment for the *in vivo E. coli* 5S rRNA. The ten sequences were obtained from the 5S ribosomal RNA database. Each sequence was submitted to MC-Fold. The top 100 structures for each sequence were then submitted to MC-Cons. The *E. coli* sequence #5 is the same as used by Mathews and colleagues (*Proc Natl Acad Sci USA*. **101**, 7287-7292, 2004). For each consensus structure, the MC-Fold rank is shown in parenthesis. MC-Fold average real time = 925.6 sec. MC-Cons real time = 2151 sec.

```
>E.coli 1
UGCCUGGCGGCCGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).)))))))))))).   (12th)
>E.coli 2
UGCCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCUCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).)))))))))))).   ( 1st)
>E.coli 3
UGCCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).)))))))))))).   ( 1st)
>E.coli 4
UGUCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGACAU
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).)))))))))))).   ( 1st)
>E.coli 5
UGCCUGGCGGCCUUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
(((((((((((.(.((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).)))))))))))).   ( 6th)
>E.coli 6
UGUCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGACUCCCCAUGCGAGAGUAGGGAACUGCCAGACAU
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).)))))))))))).   ( 1st)
>E.coli 10
UGUCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCUCAUGCGAGAGUAGGGAACUGCCAUGCAU
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).)))))))))))).   ( 1st)
>E.coli 11
UGCCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAUCA
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).))))))))))))... ( 1st)
>E.coli 14
UGCCUGGCGGCCGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGACAU
((((((((((.(..((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...)))))))))))))))))..))))))))))).   (64th)
```

Figure 5.7: Unconstrained MC-Cons consensus assignment for the *E. coli* 5S rRNA. The nine sequences were obtained from the 5S ribosomal RNA database. Each sequence was submitted to MC-Fold. The top 100 structures for each sequence were then submitted to MC-Cons. The consensus structure resembles that deduced from structural probing in solution and computer modelling by Brunel et al. (*J Mol Biol.* **221**, 293-308, 1991). For each consensus structure, the MC-Fold rank is shown in parenthesis. MC-Cons real time = 508 sec.

```
>Se1
CCCAGAUGAUGGCUUCACUGCUUGAUGGG
((((...(((((((....)))))))))))   (  4th)
....x..........xx............
>Se3
CCCAGAUGAUGCUUUAUCAGGCGGAUGGG
((((...(((((((....)))))))))))   (  1st)
....x.........x..............
>Se5
CCCAGAUGAUAGUGAGGCGCGGCUUGAUGGG
((((...(((((((.....).)))))))))) ( 14th)
....x.........xxx..............
>Se6
CCCAGAUGAUAGUAAGGCGCGGCUUGAUGGG
((((...(((((((.(...)))))))))))) (  3rd)
....x.........x................
>Se7
CCCAGAUGAUCCGACGCGCUUUGGUGAUGGG
((((...((((((((....)))))))))))) (  4th)
....x.............x............
```

Figure 5.8: MC-Fold predictions for the SECIS element. Positions marked with 'x' have high reactivity to single-stranded enzymatic probing, and are penalized by 8 kcal/mol if they are found base-paired in MC-Fold solutions. The nucleotides that participate in the formation of the K- turn motif are shown in bold. MC-Cons real time = 23 seconds.

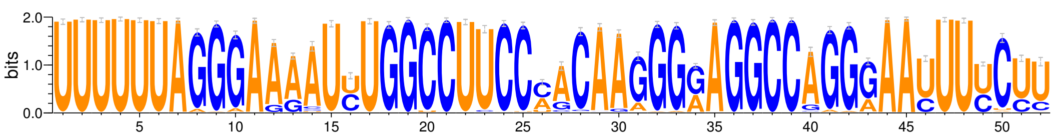


Figure 5.9: The sequence variations observed in 753 HIV-1 frame-shifting elements. The sequences were obtained from Rfam. The slippery sequence is located in positions 1-7. The G(G|A)A bulge in the NMR model is located at positions 42-44. The AA bulge in our model is located at positions 44-45. The drawing was made with WebLogo (http://weblogo.berkeley.edu/logo.cgi).

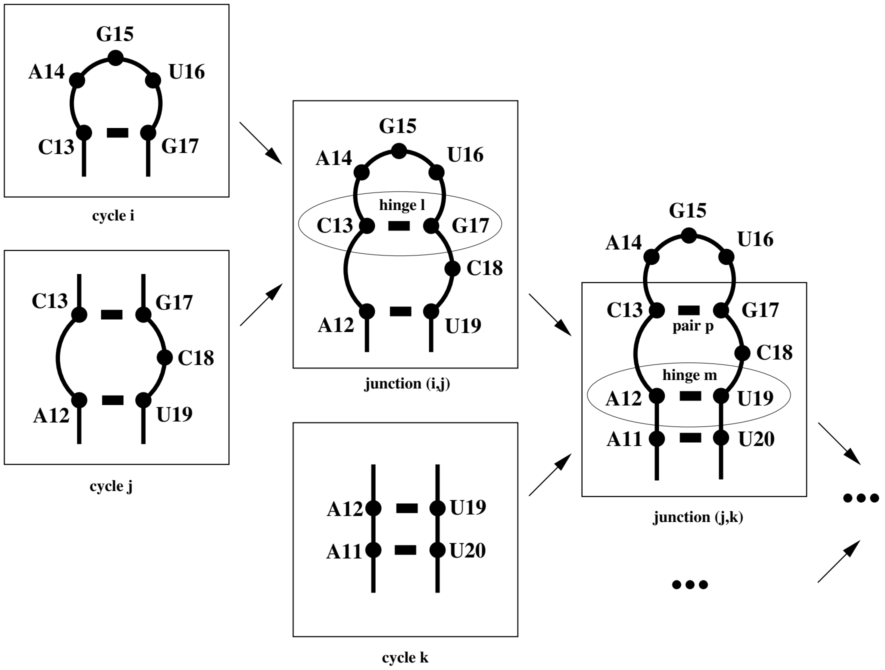


Figure 5.10: Cycles, junctions, hinges, and base pairs. The dots represent nucleotides and the thick lines base pairs. Two NCMs (left), i and j, are joined, defining a junction (center above), (i, j), which includes a hinge (center above), l, and corresponding common pair (right), p. The junction (i, j) and the hinge l are valid, and thus a new NCM (center below), k, can be added. The arrows indicate the formation of junctions. The hinges are highlighted using ovals. The sum needed to compute the score resulting from this particular combination are:

1. Ψ(5-NCM|"CAGUG"), the probability of observing a 5-NCM given "ACGUU".
2. Ψ(2_3-NCM|"ACGUU"), the score of assigning a 2_3-NCM to "ACGUU".
3. Ψ(junction$_{(i,j)}$|5-CAGUG,2_3-ACGUU), the probability of observing a junction between 5-CAGUG and 2_3-ACGUU.
4. Ψ(CG|junction$_{(i,j)}$), the probability of observing a CG base pair in junction$_{(i,j)}$.
5. Ψ(CG|hinge$_l$), the probability of observing a CG base pair given hinge$_l$.

Table 5.3 shows how the hinge probability is determined when a base pair tandem is added to a GNRA tetraloop.

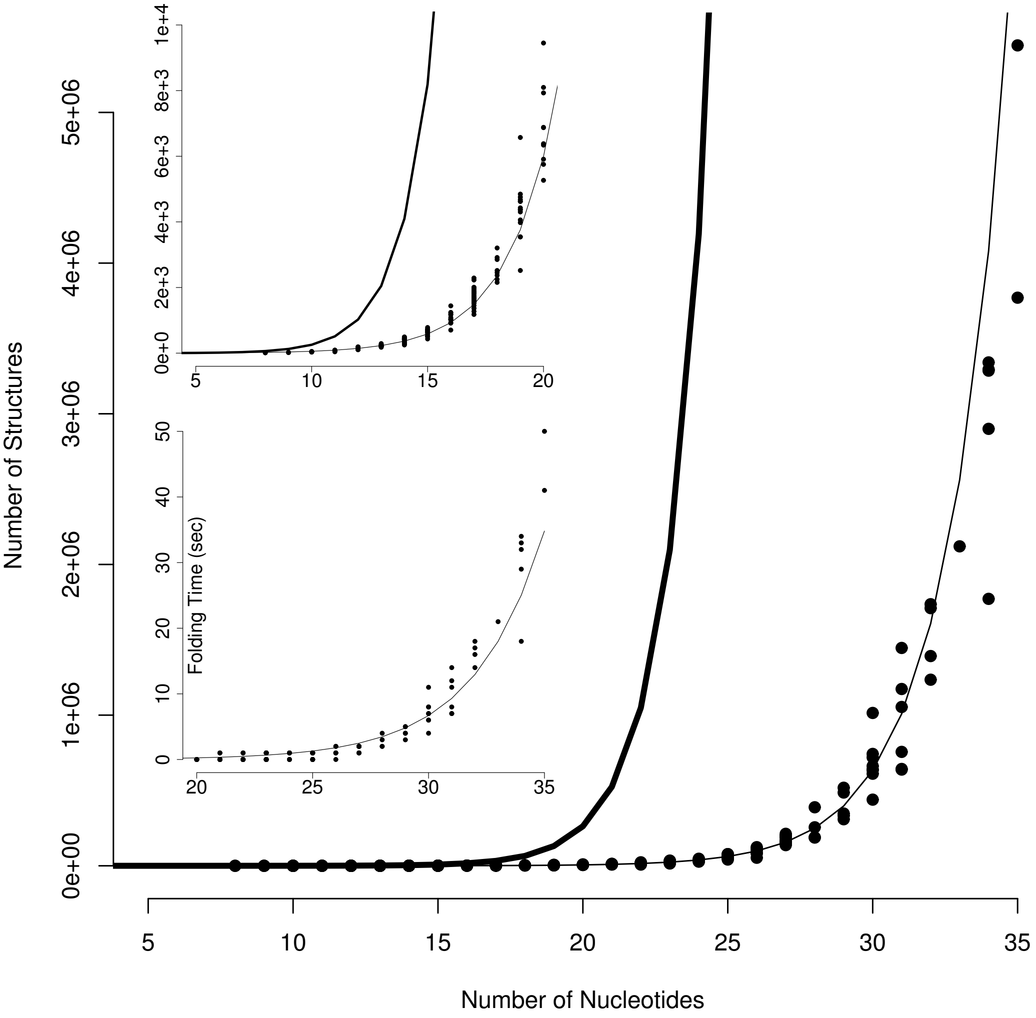


Figure 5.11: Number of structure vs. sequence lenght. The number of secondary structures generated by MC-Fold versus the length of hairpin sequences. Each dot represents one hairpin. The curve for hairpins of 1 to 20 nucleotides is zoomed (inset above). The thick line shows the theoretical number of structures approximated by an exponential least square fit. The time required to compute the hairpin structures is proportional to the number of generated structures (inset below).

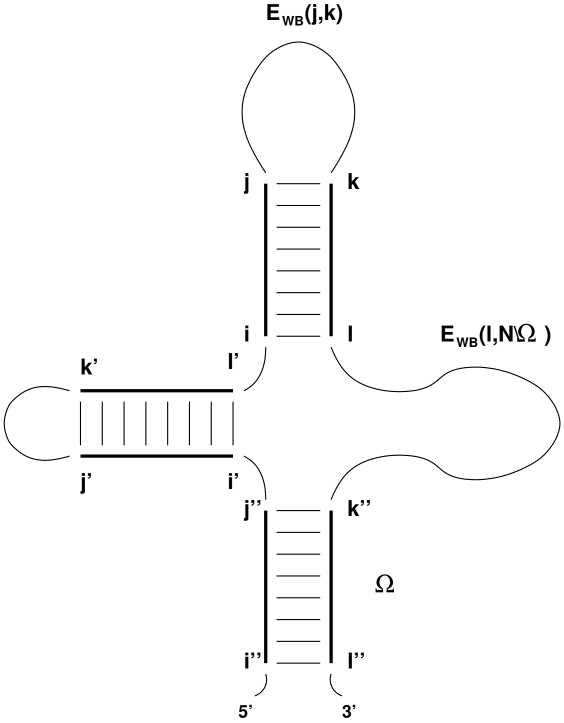
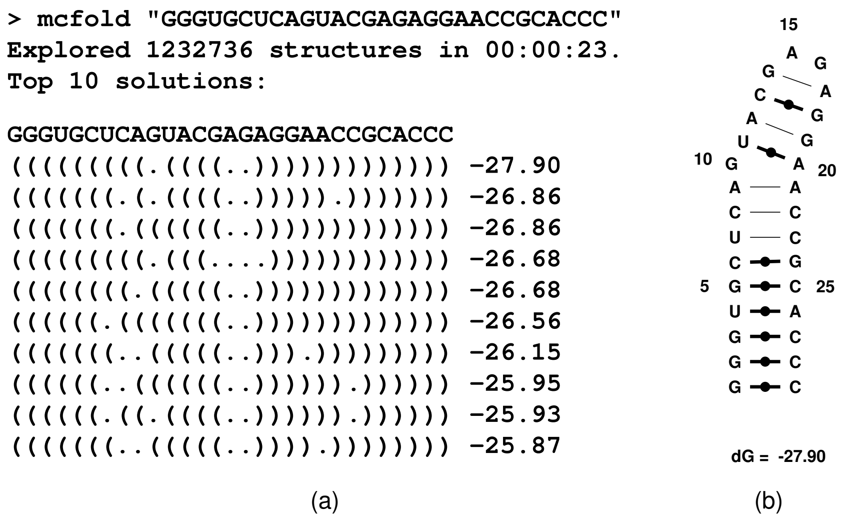


Figure 5.12: Multi-branch construction. Stems are represented by i < j < k < l. $E_{WB}(j,k)$ represents the best energy between positions j to k, as found in the dynamic programming table at entry (j,k). Ω represents the positions that were previously assigned in stems.



```
> mcfold "GGGUGCUCAGUACGAGAGGAACCGCACCC"
Explored 1232736 structures in 00:00:23.
Top 10 solutions:

GGGUGCUCAGUACGAGAGGAACCGCACCC
(((((((((.((((..))))))))))))) -27.90
(((((((.(.((((..))))).))))))) -26.86
(((((((.((((((..))))))))))))) -26.86
(((((((((.(((....)))))))))))) -26.68
((((((((.(((((..))))))))))))) -26.68
((((((.(((((((..))))))))))))) -26.56
(((((((..(((((..))).))))))))) -26.15
((((((..((((((..)))))).)))))) -25.95
((((((.((.((((..)))))).)))))) -25.93
(((((((..(((((..)))).)))))))) -25.87
```

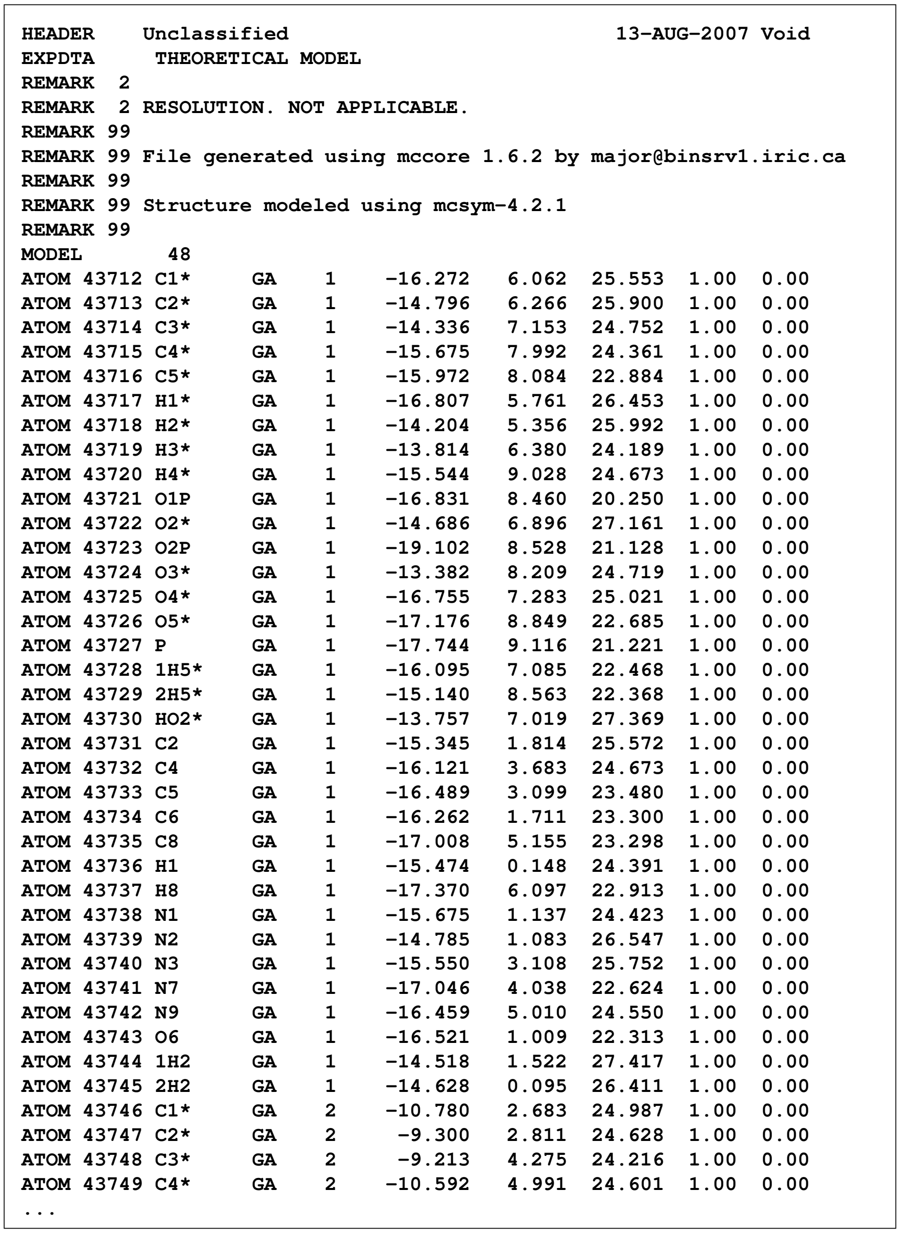
(a)

(b)

Figure 5.13: MC-Fold call and output. **a**. MC-Fold is invoked in a Unix shell with the sequence of the rat 28S rRNA Loop E. The structures are generated, evaluated, and sorted by energies, indicated by the numbers on the right of each solution shown in dot-bracket notation. The number of solutions returned is an option of the program, 10 is the default value. **b**. Secondary structure of the best solution. A dot-bracket can be converted in a secondary structure representation. The dotted lines represent canonical base pairs; the lines non-canonical base pairs.

```
> mcsym IRE.mcc
//========== Sequence ==========
sequence( r A1 GGAGUGCUUCAACAGUGCUUGGACGCUCC )
//             ((((((.((((((...).)))))))))))

//========== NCMs ==========
ncm_01 = library(
         pdb( "MCSYM-DB/5/CAGUG/*.pdb.gz" ) #1:#5 <- A13:A17
         rmsd( 0.1 sidechain && !( pse || lp || hydrogen ) ) )
ncm_02 = library(
         pdb( "MCSYM-DB/2_3/ACGCU/*.pdb.gz" ) #1:#2, #3:#5 <- A12:A13, A17:A19
         rmsd( 0.1 sidechain && !( pse || lp || hydrogen ) ) )
ncm_03 = library(
         pdb( "MCSYM-DB/2_2/AAUU/*.pdb.gz" ) #1:#2, #3:#4 <- A11:A12, A19:A20
         rmsd( 0.5 sidechain && !( pse || lp || hydrogen ) ) )
ncm_04 = library(
         pdb( "MCSYM-DB/2_2/CAUG/*.pdb.gz" ) #1:#2, #3:#4 <- A10:A11, A20:A21
         rmsd( 0.5 sidechain && !( pse || lp || hydrogen ) ) )
ncm_05 = library(
         pdb( "MCSYM-DB/2_2/UCGG/*.pdb.gz" ) #1:#2, #3:#4 <- A9:A10, A21:A22
         rmsd( 0.1 sidechain && !( pse || lp || hydrogen ) ) )
ncm_06 = library(
         pdb( "MCSYM-DB/2_2/UUGA/*.pdb.gz" ) #1:#2, #3:#4 <- A8:A9, A22:A23
         rmsd( 0.1 sidechain && !( pse || lp || hydrogen ) ) )
ncm_07 = library(
         pdb( "MCSYM-DB/3_2/GCUAC/*.pdb.gz" ) #1:#3, #4:#5 <- A6:A8, A23:A24
         rmsd( 0.1 sidechain && !( pse || lp || hydrogen ) ) )
ncm_08 = library(
         pdb( "MCSYM-DB/2_2/UGCG/*.pdb.gz" ) #1:#2, #3:#4 <- A5:A6, A24:A25
         rmsd( 0.1 sidechain && !( pse || lp || hydrogen ) ) )
ncm_09 = library(
         pdb( "MCSYM-DB/2_2/GUGC/*.pdb.gz" ) #1:#2, #3:#4 <- A4:A5, A25:A26
         rmsd( 0.1 sidechain && !( pse || lp || hydrogen ) ) )
ncm_10 = library(
         pdb( "MCSYM-DB/2_2/AGCU/*.pdb.gz" ) #1:#2, #3:#4 <- A3:A4, A26:A27
         rmsd( 0.5 sidechain && !( pse || lp || hydrogen ) ) )
ncm_11 = library(
         pdb( "MCSYM-DB/2_2/GAUC/*.pdb.gz" ) #1:#2, #3:#4 <- A2:A3, A27:A28
         rmsd( 0.5 sidechain && !( pse || lp || hydrogen ) ) )
ncm_12 = library(
         pdb( "MCSYM-DB/2_2/GGCC/*.pdb.gz" ) #1:#2, #3:#4 <- A1:A2, A28:A29
         rmsd( 0.5 sidechain && !( pse || lp || hydrogen ) ) )

//=========== Backtrack ===========
stem_01 = backtrack(
         ncm_01
         merge( ncm_02 0.3 )
         merge( ncm_03 0.3 )
         merge( ncm_04 0.3 )
         merge( ncm_05 0.3 )
         merge( ncm_06 0.3 )
         merge( ncm_07 0.3 )
         merge( ncm_08 0.3 )
         merge( ncm_09 0.3 )
         merge( ncm_10 0.3 )
         merge( ncm_11 0.3 )
         merge( ncm_12 0.3 ) )

// ========= Constraints / Restraints =========
clash                  ( stem_01 1.5 !( pse || lp || hydrogen ) )
ribose_rst             ( stem_01 method = ccm, threshold = 0.2, pucker = C3p_endo )
backtrack_rst          ( stem_01 method = probabilistic )
implicit_phosphate_rst( stem_01 sampling = 90% )

// ========= Search =========
explore(
         stem_01
         option( model_limit = 5000, time_limit = 24h )
         rmsd( 1.2 sidechain && !( pse || lp || hydrogen ) )
         pdb( "Build/IRE" zipped ) )
```

Figure 5.14: MC-Sym input script for the IRE consensus sequence. This script has been generated by MC-Fold. It can be submitted to MC-Sym without any editing. It produces the 3-D structure of the main manuscript Fig. 4.2a, shown superimposed with an NMR structure of the IRE.



```
HEADER    Unclassified                          13-AUG-2007 Void
EXPDTA     THEORETICAL MODEL
REMARK  2
REMARK  2 RESOLUTION. NOT APPLICABLE.
REMARK 99
REMARK 99 File generated using mccore 1.6.2 by major@binsrv1.iric.ca
REMARK 99
REMARK 99 Structure modeled using mcsym-4.2.1
REMARK 99
MODEL       48
ATOM 43712 C1*     GA    1    -16.272   6.062  25.553  1.00  0.00
ATOM 43713 C2*     GA    1    -14.796   6.266  25.900  1.00  0.00
ATOM 43714 C3*     GA    1    -14.336   7.153  24.752  1.00  0.00
ATOM 43715 C4*     GA    1    -15.675   7.992  24.361  1.00  0.00
ATOM 43716 C5*     GA    1    -15.972   8.084  22.884  1.00  0.00
ATOM 43717 H1*     GA    1    -16.807   5.761  26.453  1.00  0.00
ATOM 43718 H2*     GA    1    -14.204   5.356  25.992  1.00  0.00
ATOM 43719 H3*     GA    1    -13.814   6.380  24.189  1.00  0.00
ATOM 43720 H4*     GA    1    -15.544   9.028  24.673  1.00  0.00
ATOM 43721 O1P     GA    1    -16.831   8.460  20.250  1.00  0.00
ATOM 43722 O2*     GA    1    -14.686   6.896  27.161  1.00  0.00
ATOM 43723 O2P     GA    1    -19.102   8.528  21.128  1.00  0.00
ATOM 43724 O3*     GA    1    -13.382   8.209  24.719  1.00  0.00
ATOM 43725 O4*     GA    1    -16.755   7.283  25.021  1.00  0.00
ATOM 43726 O5*     GA    1    -17.176   8.849  22.685  1.00  0.00
ATOM 43727 P       GA    1    -17.744   9.116  21.221  1.00  0.00
ATOM 43728 1H5*    GA    1    -16.095   7.085  22.468  1.00  0.00
ATOM 43729 2H5*    GA    1    -15.140   8.563  22.368  1.00  0.00
ATOM 43730 HO2*    GA    1    -13.757   7.019  27.369  1.00  0.00
ATOM 43731 C2      GA    1    -15.345   1.814  25.572  1.00  0.00
ATOM 43732 C4      GA    1    -16.121   3.683  24.673  1.00  0.00
ATOM 43733 C5      GA    1    -16.489   3.099  23.480  1.00  0.00
ATOM 43734 C6      GA    1    -16.262   1.711  23.300  1.00  0.00
ATOM 43735 C8      GA    1    -17.008   5.155  23.298  1.00  0.00
ATOM 43736 H1      GA    1    -15.474   0.148  24.391  1.00  0.00
ATOM 43737 H8      GA    1    -17.370   6.097  22.913  1.00  0.00
ATOM 43738 N1      GA    1    -15.675   1.137  24.423  1.00  0.00
ATOM 43739 N2      GA    1    -14.785   1.083  26.547  1.00  0.00
ATOM 43740 N3      GA    1    -15.550   3.108  25.752  1.00  0.00
ATOM 43741 N7      GA    1    -17.046   4.038  22.624  1.00  0.00
ATOM 43742 N9      GA    1    -16.459   5.010  24.550  1.00  0.00
ATOM 43743 O6      GA    1    -16.521   1.009  22.313  1.00  0.00
ATOM 43744 1H2     GA    1    -14.518   1.522  27.417  1.00  0.00
ATOM 43745 2H2     GA    1    -14.628   0.095  26.411  1.00  0.00
ATOM 43746 C1*     GA    2    -10.780   2.683  24.987  1.00  0.00
ATOM 43747 C2*     GA    2     -9.300   2.811  24.628  1.00  0.00
ATOM 43748 C3*     GA    2     -9.213   4.275  24.216  1.00  0.00
ATOM 43749 C4*     GA    2    -10.592   4.991  24.601  1.00  0.00
...
```

Figure 5.15: Header of a PDB file generated by MC-Sym.

CHAPTER 6

ARTICLE 2

New Metrics for Comparing and Assessing Discrepancies Between RNA 3D Structures and Models

Marc Parisien[1*], José Almeida Cruz[2*], Éric Westhof[2] and François Major[1]

[1] *Institute for Research in Immunology and Cancer,*
Department of Computer Science and Operations Research,
Université de Montréal,
PO Box 6128, Downtown station,
Montréal, Québec, H3C 3J7, CANADA

[2] *Architecture et Réactivité de l'ARN,*
Institut de Biologie Moléculaire et Cellulaire du CNRS,
Université de Strasbourg,
67084 Strasbourg Cedex, FRANCE

** Equal contribution*

Parisien M, Cruz JA, Westhof E, Major F.
RNA 2009 **15**:1875-1885.

6.1 Abstract

To benchmark progress made in RNA three-dimensional modeling and assess newly developed techniques, reliable and meaningful comparison metrics and associated tools are necessary. Generally, the average root-mean-square deviations (RMSDs) are quoted. However, RMSD can be misleading since errors are spread over the whole molecule and do not account for the specificity of RNA base interactions. Here, we introduce two new metrics that are particularly suitable to RNAs: the deformation index and deformation profile. The deformation index is calibrated by the interaction network fidelity, which considers base-base-stacking and base-base-pairing interactions within the target structure. The deformation profile highlights dissimilarities between structures at the nucleotide scale for both intradomain and interdomain interactions. Our results show that there is little correlation between RMSD and interaction network fidelity. The deformation profile is a tool that allows for rapid assessment of the origins of discrepancies.

Keywords: RNA; structure; comparative analysis; three-dimensional modeling; RMSD

6.2 Introduction

Determining RNA three-dimensional (3D) structures is key in studying RNA function [2]. Physical methods such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are the most common ways for determining RNA 3D structures at high resolution. However, these methods cannot be applied to all RNAs and RNA systems. Alternative methods include interactive modeling [46, 118, 243] and conformational space searching [44, 45, 47, 48].

The development and improvement of alternative methods are highly dependent on what we learn from experimentally resolved structures. In particular, close inspection of rRNA structures revealed the presence of structural motifs that we can recognize from sequence [77]. To assist the production of new knowledge, systematic methods to annotate RNA 3D structures [56, 69, 157, 244], discover and analyze structural motifs [120, 156, 245, 246, 247], and formally represent RNA structures [163, 248] have been developed. This systematization of knowledge generation and integration in ever-improving predictive methods is typical of the postribosomal X-ray crystallographic era. A problem that has been largely neglected, however, is how one can measure quantitatively the improvements brought by new approaches or methods.

The classical index for comparing predictive methods is to benchmark with the average root-mean-square deviations (RMSDs) after optimal superimposition between the modeled RNA 3D structures they produce and their corresponding experimental structures. RMSDs are extremely useful, and obtaining models close to experimental structures is a noble exercise. RMSDs capture the general 3D shape of an RNA, but give little information about its base-pairing and base-stacking patterns, local deviations of the structure, intradomain deformation, or interdomain deviations. Most importantly, RMSDs spread errors over the whole molecule to obtain the best global superimposition so that it is very difficult to localize the origins of the modeling defects and thus to improve the modeling process [56, 249, 250]. RNA molecules have specific structural features, such as a modular and hierarchical architecture of structural elements like helices, hairpins, and single-stranded loops connected by tertiary interactions. In addition, RNA bases associate in well-defined patterns of pairings that usually stack on each other. As modeling and predictive methods are getting increasingly accurate, it is now desirable that their results could be compared based on the reproducibility of these important and specific RNA structural features rather than on global average measurements. Here, we introduce two new RNA 3D structure comparison tools: (1) an RNA 3D structure comparison index, the deformation index (DI), which evaluates and indicates the deviations between two RNA 3D structures with both RMSDs and base interactions; and (2) a deformation profile (DP), which depicts the conformation differences between two models at local, interdomain, and intradomain scales. These new tools provide quantitative measures to compare the accuracy in reproducing the base-base interaction networks of different 3D models, as well as the ability to evaluate the local and global prediction precision and quality of RNA molecules.

6.3 Results

6.3.1 Deformation index

We define the DI as the RMSD between two optimally aligned 3D structures (general shape) divided by the base interaction network fidelity (INF). The INF is computed from the base-stacking and base-pairing annotations of both structures. For practical reasons, we use two automated annotation procedures that have been proposed recently: MC-Annotate [56, 69] and RNAview [157]. Note that the index uses, but is not related to, the annotation programs, which are obviously prone to the quality of the reference structures.

6.3.2 Base-stacking and base-pairing interactions

MC-Annotate detects that two bases stack using the Gabb et al. method [251]. The base-stacking annotation results are described using the Major and Thibault nomenclature [252], which indicates the relative orientation of the two bases. The relative orientation is determined by comparing the direction of the normal vectors of each base, i.e., the rotational vector obtained by a right-handed axis system defined by atoms N1 to N6 around the pyrimidine ring (Fig. 6.1A).

Two possible relative orientations in each base result in four base-stacking types: upward (>>), downward (<<), outward (<>), and inward (><) (see Fig. 6.1B). Two vectors pointing in the same direction (upward and downward) corresponds to the base-stacking type in canonical A-RNA double helices. Upward or downward is chosen depending on which base is referred to first (i.e., A>>B means B is stacked upward of A, or A is stacked downward of B). The two other types are, respectively, inward (A><B; A or B is stacked inward of, respectively, B or A) and outward (A<>B; A or B is stacked outward of, respectively, B or A).

MC-Annotate uses an unsupervised machine-learning approach to detect H-bonds and H-bonding patterns [69], and RNAview uses geometrical constraints [157]. Both programs describe their base-pairing annotations using the Leontis and Westhof nomenclature. Each type describes the interacting edge of the two bases. Three interacting edges are defined: the Watson-Crick edge: ● (*cis*), ○ (*trans*); the Hoogsteen edge: ■ (*cis*), □ (*trans*); and the sugar edge: ◀ (*cis*), ◁ (*trans*) (Fig. 6.1C; [67]). The *cis*/*trans* notation reflects the relative orientation of the backbone according to the median of the plane formed by the two bases. In Figure 6.1C, the base pair is *cis* since the riboses are positioned on the same side of the base-pair plane. When two bases interact by the same edge, only one symbol is used. For instance, a *trans* X-Y Hoogsteen base pair is either written "H/H *trans*" or X □Y. Figure 6.1D lists all possible base-pairing types that are described by this nomenclature.

The DI considers the full set of interactions, i.e., base-stacking and base-pairing interactions defined by the classical two-dimensional (2D) structure (A-U and G-C Watson-Crick and G-U Wobble base pairs that form in the stems); extended 2D structures (the noncanonical base pairs, but that can be represented in the dot-bracket notation); and tertiary structure interactions, such as nonhelical stacking and long-range base pairs. Note that 40% of the interactions in crystallized ribosomal RNAs enter the latter category [75].

6.3.3 Interaction network fidelity

A stacking or pairing interaction, I, involves two distinct nucleotides, N_i and N_j, $i < j$, to form an interaction (N_i, N_j, I), where I is one of the above base-pairing or base-stacking types. The annotation of a 3D structure produces a set, S, of such interactions. Given the two sets of interactions in two distinct RNA structures, we can then compare them using simple set theory operations.

Let S_r be the set of interactions in a reference structure (usually an experimentally resolved structure) and S_m the set of interactions of a modeled structure. The interactions found in the intersection of both sets are true positives, TP = $S_r \cap S_m$. The interactions in S_m that are not present in S_r are false positives, FP = $S_m \backslash S_r$. The interactions absent in

S_m but present in S_r are false negatives, FN = $S_r \backslash S_m$.

The Matthews correlation coefficient (MCC) is estimated by [253]:

$$\begin{aligned}\text{MCC} &= \sqrt{\text{PPV} \times \text{STY}},\\ \text{where PPV (specificity)} &= \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|},\\ \text{and STY (sensitivity)} &= \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|}\end{aligned}$$

When the model reproduces exactly the base interactions of the reference, then |FP| = |FN| = 0, |TP| > 0, and thus MCC = 1. When the model does not reproduce any of the interactions of the reference structure, then MCC = 0, since |TP| = 0.

We define the interaction network fidelity (INF) between structures A and B as the MCC, INF(A,B) = MCC(A,B). We propose a new measure of the resemblance between two structures A and B (for example, a model and its corresponding experimental structure), which is quantified by a deviation index,

$$\text{DI(A,B)} = \text{RMSD(A,B)} / \text{INF(A,B)}$$

Not having an INF, the DI would simply be the RMSD. However, given an INF from 0 to 1, then the RMSD between A and B could either have a large (and even infinite) DI if the two structures share no common interactions (INF = 0), or meaningful RMSD as INF approaches 1 (i.e., the majority of the interactions in A are reproduced in B).

6.3.4 Example: Modeling the rat 28S rRNA loop E 3D structure

Consider the crystal structure of the rat 28S rRNA loop E (PDB code 1Q9A; resolution 1.04 Å;[254]) shown in Figure 6.2A. MC-Annotate (Fig. 6.2B) and RNAview (Fig. 6.2C) were used to compute the base-pairing network of this structure. Since

RNA structure annotation is subject to interpretation and small geometrical variations - for instance, MC-Annotate is stricter than RNAview - we therefore take the intersection of both programs. MC-Annotate also computes the base-stacking network (see Fig. 6.2D).

To illustrate the benchmarking of RNA 3D structure modeling results, we generated loop E 3D structures using MC-Sym ([45]; see Materials and Methods). We generated a decoy of 9847 3D structures, where each structure is at least at 1 Å RMSD from each other. The RMSDs (all atoms but H) between these structures and the crystal structure range from 1.6 Å to 7.8 Å, whereas the INF values range from 0.49 to 0.89 (Fig. 6.3). We note that for a given RMSD threshold, we have a wide range of INF values, and for a given INF threshold, we have a wide range of RMSDs. However, as RMSDs worsen, the INF values also worsen. We note an absence of population in the upper right corner (i.e., high RMSD and high INF values). The Pearson correlation coefficient between RMSD and INF values is P = 0.60 for this particular decoy.

For further analyses, we randomly selected three of the MC-Sym-generated structures. Structure A is located in the upper-left corner of Figure 6.3 and is shown in Figure 6.4A. This structure has good RMSDs (1.64 Å) with the crystal structure, and good INF and DI values, 0.88 and 1.86, respectively. Since RMSDs are averaged values, they do not inform about the maximum modeling error. Therefore, we also report the max RMSD(i,j) (j > i), i.e., the maximum RMSD over any sequence fragment defined by i and j; j > i. If we exclude from the analysis the dangling nucleotide U1 in the crystal structure, the fragment that has maximum RMSD with the crystal structure is C20-C21 with 1.7 Å. This is shown by the fact that C20 and C21 are base paired in the generated structures, as annotated by RNAview, but they have problematic geometries in the crystal structure, as indicated by the absence of annotation by MC-Annotate (Fig. 6.2B).

Structure A contains 29 TP, i.e., 29 of the 30 base interactions (10 base pairs and 20 base stacks) in the crystal structure. Six FP are made: (1, 2) two upward stacking between C3-U4 and C5-C6. Note that in principle these base-stacking interactions make sense since they are located in a stem. They were not detected in the crystal structure by MC-Annotate; (3) a flip of the C20 base around the glycosidic bond creates an inward stacking A19-C20; (4) as assumed in the modeling, A8-C20 now form a base pair (H/W *trans*); (5) the dangling nucleotide U1 in the models is base paired to G27 as a canonical W/W *cis* type; and (6) as assumed in the modeling, U7-C21 now form a base pair (S/H *trans*). Due to the C20 base flip, the upward stacking A19-C20 and C20-C21 are not reproduced, making two FN.

Structure B was selected in the upper right section of Figure 6.3, i.e., it has a good INF (0.88), but a bad RMSD with the crystal structure (3.76 Å). It is shown in Figure 6.4B. If we remove U1, the worst fragment is G2-C20 (19-nucleotides [nt] long) with 3.66 Å. This is shown in Figure 6.4B by a shifted backbone in almost all nucleotide positions. Structure B contains 28 of the 30 base interactions (10 base pairs and 20 base stacks) in the crystal structure. Five FP are made. They are the same as in structure A, but the upward stacking between C5-C6 is absent as in the crystal structure. The two FN due to the C20 base flip are also present in structure B. In addition, no inward base stacking is detected between G9 and G18.

Finally, structure C (Fig. 6.4C) was selected in the lower left region of Figure 6.3, i.e., bad INF (0.71), but relatively good RMSD (2.03 Å). Again, the worst fragment is G2-C20, but its RMSD is now 2 Å. What hurts the RMSD of this model is related to difficulties to reproduce the base triple and the A8-C20 base pair of the crystal structure; typical errors in RNA modeling. In our particular case, it is noteworthy that the bases in the generated base triple have a more planar geometry than those observed in the crystal structure (Fig. 6.4D). As for the A8-C20 base pair, its H/W *trans* type now makes a consensus between MC-Annotate and RNAview. Structure C contains 21 of the 30 base interactions in the crystal structure. Seven FP are made: (1-5) are the same as in structure A; but, in addition, (6) an upward stacking between A16-G17 is detected that was not detected in the crystal structure; (7) the flanking base pair of the GAGA tetraloop, which is changed to a W/H *trans* (S/H *trans* in the crystal). The three FN of structure B are also made in structure C (two are due to the C20 base flip) (Fig. 6.4E). In addition, four upward stackings are not detected between A11-C12, C12-G13, A14-G15, and U4-C5. The outward stacking between G13-G17 and the G9-U10 S/H *cis* base pair are also not detected. The tenth FN is the absence of the S/H *trans* G13-A16 base pair.

6.3.5 Deformation profile

The DP is a distance matrix representing the average distance between a predicted model (PM) and reference model (RM). The DP matrix is obtained by (1) computing all 1-nt superimposition of PM over RM and then (2) for each superimposition, computing the average distance between each base in RM and the corresponding base in PM. Let RM_i and PM_i represent the ith nucleotides of RM and PM respectively, let $\text{SUP}(\text{A}_i,\text{B}_i)$ be the model that results from the superposition of model B over the reference model A, minimizing the RMSDs between all the atoms of the nucleotides A_i and B_i, and let $\text{AVG\_DIST}(\text{A}_i,\text{B}_i)$ be the average distance between all atoms of the nucleotides A_i and B_i. Thus, the deformation profile of PM regarding RM is defined as:

$$\text{DP}_{i,j} = \text{AVG\_DIST}\ [\text{SUP}(\ \text{RM}_i,\ \text{PM}_i\ )_j,\ \text{RM}_j\ ].$$

Figure 6.5 illustrates the process of computing a DP matrix.

Once a pair of nucleotides (PM_i, RM_i) is superimposed, every other pair of nucleotides will be closer or farther depending on how well PM_i predicts RM_i. Those average distances are represented in the ith row of the matrix. Thus, the row average provides information about local similarity regarding the ith nucleotide. For example, an individual row with higher values than the rest of the matrix (Figs 6.6, 6.7, represented as yellow/red rows in the DP matrices) usually means a particularly poorly predicted nucleotide. The jth column of the matrix contains the average atomic distances between the jth nucleotides of PM and RM, for each superimposition. Thus, the column average indicates how the distance between PMj and RMj depends on the overall prediction of all nucleotides. Finally, the main diagonal contains the average atomic distance of each nucleotide, allowing a perspective of individual nucleotide conformation similarity.

An interesting property of DP is the ability to reveal similarity information at several structural scales. The rectangles corresponding to the intersection of two strands indicate the relative similarity between those strands. This way, one can easily apprehend the structural similarity at intradomain (such as between both strands of a helix or the nucleotides of a loop) and interdomain scales (such as between two helices or two loops).

It is worth noticing that values in a DP are not normalized across the whole matrix. Values close to the main diagonal tend to be smaller than values farther away. This is because nucleotide pairs closer from the superimposing pair tend to have smaller average atomic distances than those farther away. Consequently, one should only compare DP values from regions at similar distances to the main diagonal or, obviously, values from DPs of distinct models.

6.3.6 Example: The hammerhead ribozyme

To exemplify the deformation profile, we compared three predicted models of a hammerhead ribozyme with the reference crystal structure (PDB: 1NYI) [255]. We generated a decoy of 9999 3D structures, where each structure is at least at 1 Å RMSD from each other. The RMSDs (all atoms but H) between these structures and the crystal structure range from 2.5 Å to 15.8 Å. Selecting models from decoys is a thorny question. Here, we limited our analysis to a series of structural properties offered by the MC-Pipeline website (see Materials and Methods). We reduced the decoy by performing a five-clustering of the 10,000 models, and selecting one model per cluster that has a small volume (<25,000), a good P-Score (<-15), and to either be bipolar or coplanar (at the >0.7 level) [256]. The “thresholds” were established by comparing each structural property with RMSDs to the crystal structure (Supplemental Fig. 7.1). The selected models and their properties are shown in Table 6.1.

From the modeling results, we further analyzed models 553, 633, and 2698, the resulting DPs of which are pictured in Figures 6.6 and 6.7, and Supplemental Figure 7.2, respectively. The models share 3.4, 12.2, and 4.9 Å RMSDs with the crystal structure, respectively. The helical regions of the models score fairly well and much better than interhelical and interloop regions (Table 6.2). Not surprisingly, nucleotides involved in canonical WC base pairing are better predicted than nucleotides involved in noncanonical base pairs or in loops. The 3- and 2-nt-long single-stranded regions (L1 and L3) present the worst deformation score of all short (<5-nt) contiguous regions (Supplemental Fig. 7.3), except for L3 in model 2698, which was particularly well predicted. The difficulty in predicting L1 and L3 also reflects in the poor prediction of the relative positions of L1 and L3. The main difference between prediction quality among the three models is due to the relative position of helix H1 with respect to the other two helices. Noticeably, the coaxial stacking of helices H2 and H3 was reasonably well predicted in all three models. While model 553 scored well in all helix-helix relative positions, models 633 and 2698 present a displacement of helix H1 regarding H2 and H3. In model 2698, helix H1 is slightly twisted, which significantly penalizes H1×H2 and, to a lesser extent, H1×H3. In model 633, helix H1 has its double-helical axis rotated by half a turn, pointing in the opposite direction of H1 in the reference molecule, which is reflected in the high values of H1×H2 and H1×H3.

6.4 Discussion

So far, the field of 3D structural modeling has been driven by RMSD comparisons. In particular, GDT-TS (global distance test) is a measure that accounts for the number of atoms that are within 1, 2, 4, and 8 Å of the RMSD from a reference structure [257, 258]. A perfect model scores 1.0. Recently, optimal GDT-TS scores of ∼0.35 for a tRNA (∼75 nt) and 0.20 for the P4-P6 domain of a group I intron (∼150 nt) have been reported [48]. In our study, the optimal score for the hammerhead ribozyme (∼40 nt) is 0.68. However, when objectively selecting models from decoys by applying K-clustering, GDT-TS scores of 0.20, 0.06, and 0.60 are obtained, respectively. In comparison, protein structure predictions now reach GDT-TS scores as high as 0.75 on average [259]. These results highlight the fact that there is a need for improved RNA model selection and generation methods.

RMSD-based measures might be a sufficient criterion for modeling protein structures since the backbone trace is indicative of the structure and correct positioning of the side chains [260]. However, RNA structures contain specific patterns of interacting side chains that are characteristic of folded modules and typical to each overall architecture [119]. To evaluate adequately the accuracy of a predicted model, it is key to assess how well such tertiary modules and the non- Watson-Crick base pairs have been reproduced. We show that, in the context of the modeling example we used, the Pearson coefficient between RMSD and INF values (P = 0.6) presents little correlation between the two indexes. Our results further show that RMSDs do not provide information about the quality and fidelity of the base interaction network. Besides, the Pearson coefficient for structures with RMSD $\leq$3.0 Å (P = 0.2) is even weaker. These results point to the potential risk of using averaged values such as RMSD in evaluating the quality of RNA 3D models and, thus, the structure prediction methods that generate them. Besides, if the correlation on a small hairpin RNA example is already low, then it is expected to be even lower on larger RNAs.

Besides, the INF is less subject to variations than RMSD for an RNA under thermal motion [261]. Intrahelical distortions include: collective atomic motion resulting in slight helix twisting that rarely affect base-base interactions (Fig. 6.4B), and relative atomic motion that is handled by discretizing the base-base interactions using symbolic annotation [56, 67, 69, 157]. Interhelical disposition from thermal motion affects the angle between helices, which greatly affects atomic distances and thus RMSD. However, such changes in general concern only a small fraction of the base-base interactions, and thus do not affect much the INF (Table 6.1).

In the structure prediction field, models <3 Å of RMSD from an experimental structure are considered accurate. Our results suggest extreme prudence at this particular value, since in our test case the INF value of such models can be as low as 0.7. In our example, structure C has an INF of 0.71. This structure, despite 21 TP, also had seven FP and 10 FN. If we look between 3 and 5 Å of RMSD, then INF values can be as low as 0.5; with a wider range of INF values (0.5-0.9) located at or near 4 Å of RMSD. Clearly, assessing the quality and accuracy of any given RNA 3D model needs both the RMSD and INF values.

Capturing the dissimilarity between two structures in a single value, as does RMSD, is a practical way of assessing the accuracy of predicted models. However, a single value cannot provide enough information about the shape of the actual structure and the local dissimilarities. Understanding the contribution of individual domain - nucleotides, helices, single-stranded regions - to an overall dissimilarity score demands the intervention of a human expert, which is not compatible with the analysis of dozens or hundreds of candidate models produced by automatic prediction tools. The proposed deformation profile provides a compact representation of RNA model dissimilarities from nucleotide length to intradomain scales and can be used in complement to the DI to fully assess the quality of predicted models.

Consequently, a full quantification of the comparison between two RNA 3D structures should include the overall RMSD, max RMSD(i,j), INF, as well as the DI. If only one value is to be used, then the DI is the most significant one since it reflects the overall features encoded by the RMSD calibrated by the quality of the reproduced interaction network, which is encoded by the INF value. As the size of modeled RNAs increases, the importance of using both quantifiers increases as well since the correlation between RMSD and INF values is expected to decrease. Finally, phosphate or backbone atom-only, as well as canonical base-paired region-only RMSD, should be avoided since they are not indicative of the quality of the produced models, and the field has now made sufficient progress in RNA 3D modeling and prediction methods so that all-atom models are now the gold standard.

6.5 Materials and Methods

6.5.1 Generating MC-Sym decoys

To generate a decoy for the Loop E, we produced an MC-Sym script from the dot-bracket notation supported by the RNAview annotated secondary structure, “((((((((.((((..))))))))))))”. The Dot2Sym program is an MC-Tool to generate MC-Sym input scripts from dot-bracket notations (see Supplemental Information). Note that no base-pairing type information is used, and MC-Sym in such a case attempts all consistent base-pairing types. For the hammerhead ribozyme, we also obtained a first script from Dot2Sym using the following dot-bracket input:

“`((((((...((((((((..))))))))(((((..)))))))))))`”. The script was manually edited and can be found in the provided Supplemental Information. We reduced the 10,000 structure decoys to a list of five models using the five-clustering and the following SQL query:

```
SELECT * FROM BKOiY0dM2m T1 INNER JOIN (SELECT MIN(PScore) AS
minP, Cluster FROM BKOiY0dM2m WHERE ((Bipolar >= 0.7) OR
(Coplanar >= 0.7)) and Volume <= 25000 and PScore <= -15 GROUP
BY Cluster) T2 ON T1.PScore = T2.minP and T1.Cluster =
T2.Cluster WHERE T1.Volume <= 25000
```

See the MC-Sym FAQ (http://www.major.iric.ca/MC-Sym/faq.html), commands.html page generated by MC-Sym, and the MC-Pipeline website for details (http://www.major.iric.ca/MC-Pipeline). The 3D structures were visualized and rendered using Pymol [262].

6.5.2 RMSD

RMSD values were for all-atom but H, as computed using the MC-RMSD program. MC-RMSD is part of the MC-Tools, which are available from the authors.

6.5.3 Deformation profile

All the data processing, PDB file manipulation, and superimposition used to compute the Deformation Profile were done in Python using Bio.PDB (http://biopython.org) and NumPy (http://numpy.scipy.org/). The script to produce DP matrices is available from the authors.

6.6 Acknowledgments

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to F.M. F.M. is a member of the Centre Robert-Cedergren of the Université de Montréal. M.P. holds a Ph.D. scholarship from the NSERC. J.A.C. is supported by the Ph.D. Program in Computational Biology of the Instituto Gulbenkian de Ciência, Portugal (sponsored by Fundacao Calouste Gulbenkian, Siemens SA, and Fundacao para a Ciencia e Tecnologia; SFRH/BD/33528/2008).

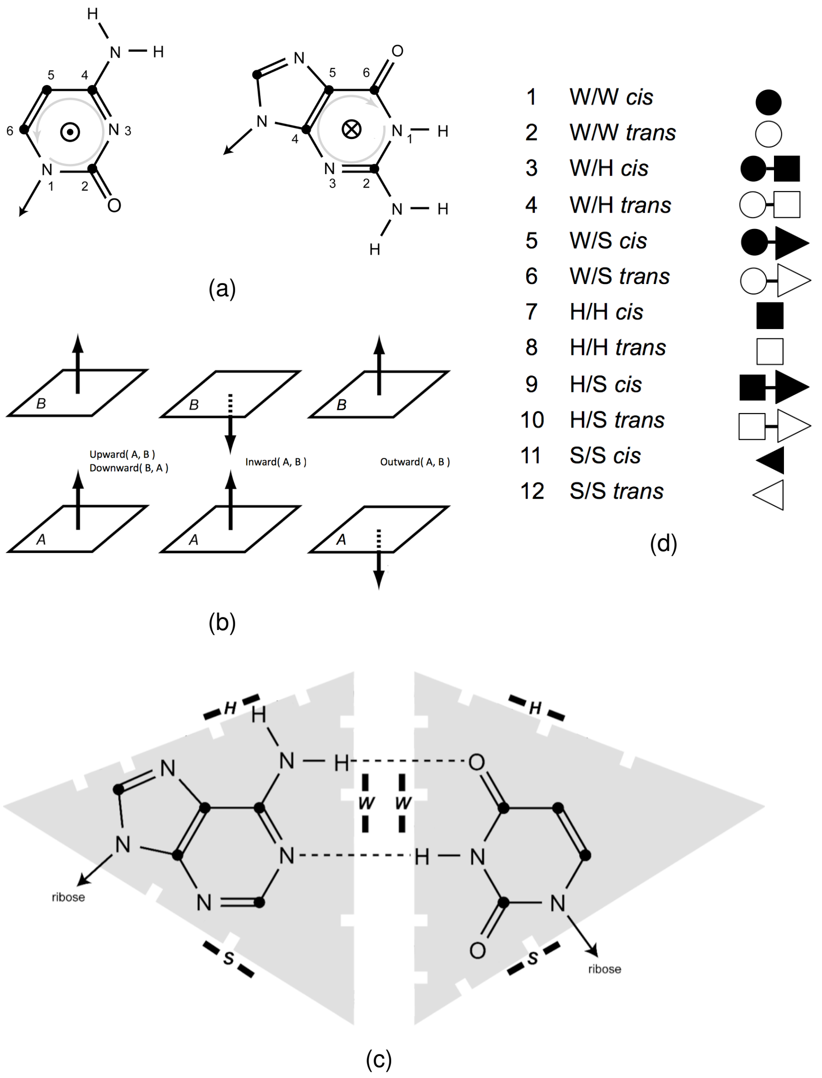


Figure 6.1: Base-stacking and base-pairing nomenclature. **(a)** Normal vectors in pyrimidines and purines. Using a right-handed axis system, the normal vector in the pyrimidine (left) comes out of the paper plane (atom numbers counterclockwise), whereas it is reversed in the pyrimidine ring of the purine (atom numbers clockwise). **(b)** The four base-stacking types. Using the normal vectors (represented by arrows), we can distinguish three types of base stacking. If base A is below base B, the normal vector of A points B, and both normal vectors point in the same direction (left), then base B is stacked upward of A (or symmetrically base A is stacked downward of B). If the normal vectors of A and B point toward each other (middle), then bases A and B stack inward. If the normal vectors flee each other (right), then bases A and B stack outward. **(c)** Base edges. Each base is divided into three edges: the Watson-Crick (W) edge is at the tip of the base and where the chemical groups involved in Watson-Crick base pairs interact; the Hoogsteen (H) edge is on the opposite side of the ribose; and the sugar (S) edge is on the side of the ribose. Here is a *cis* A-U Watson-Crick base pair, and we write W/W *cis* and represent it using the black dot. The fact that any edge in any base can interact with any other edge in a partner results in six different base-base interactions: W/W, W/H, W/S, H/H, H/S, and S/S. Since there are two possible relative orientations of the ribose according to the place formed by the two bases of a base pair, then this nomenclature describes 12 different base-pairing patterns. **(d)** The 12 base-pairing patterns, or types, and their associated symbols.

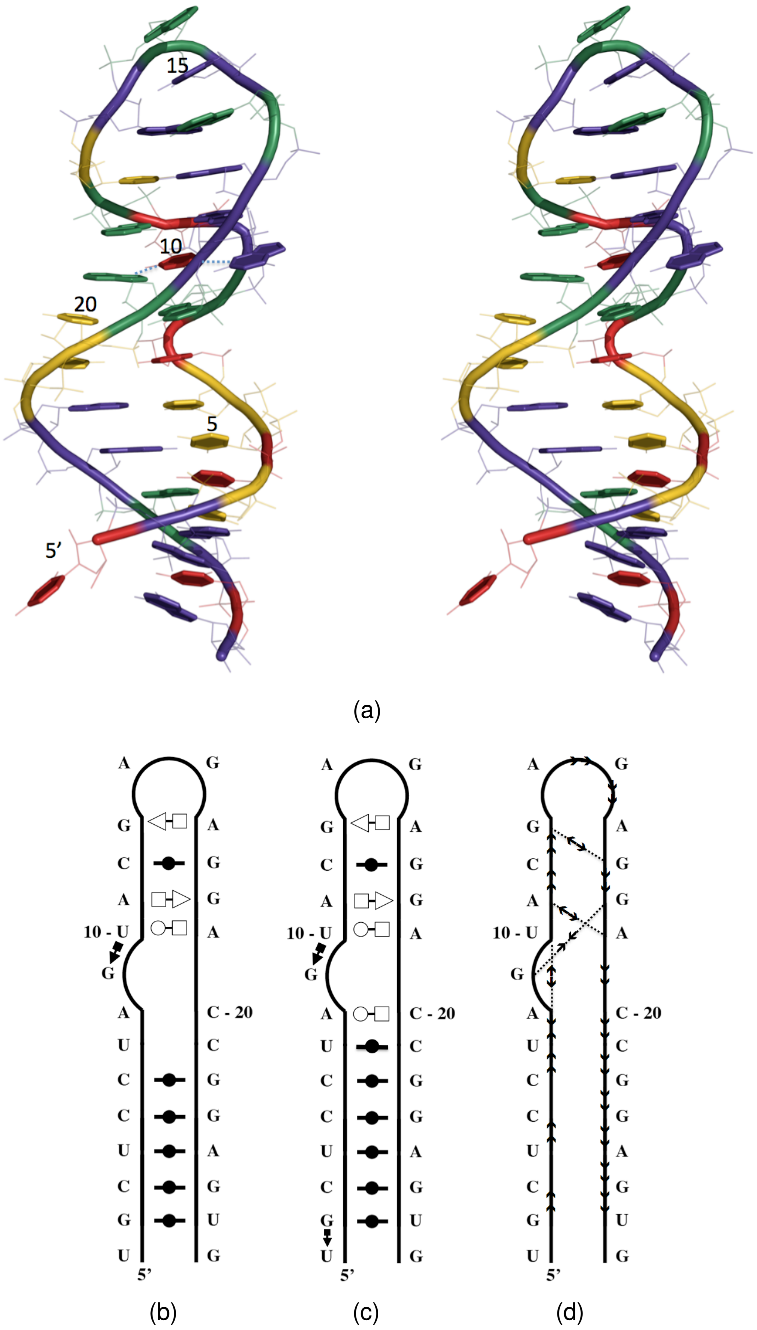


Figure 6.2: The rat 28S rRNA loop E structure. **(a)** Stereoview of the crystal structure (PDB code 1Q9A). (Green) Adenosines, (yellow) cytosines, (violet) guanosines, (red) uracils. The thread through the phosphate atoms is shown using a cylinder. Each base ring is filled and highlighted by thick covalent bonds. The H-bonded bases of the characteristic loop E structure, here the G9-U10-A19 base triple, are linked with dotted lines. Note that U1 in this crystal structure is not paired with G27. The image was generated using Pymol. **(b)** Secondary structure annotated by MC-Annotate. **(c)** Secondary structure annotated by RNAview. **(d)** Stacking annotation.

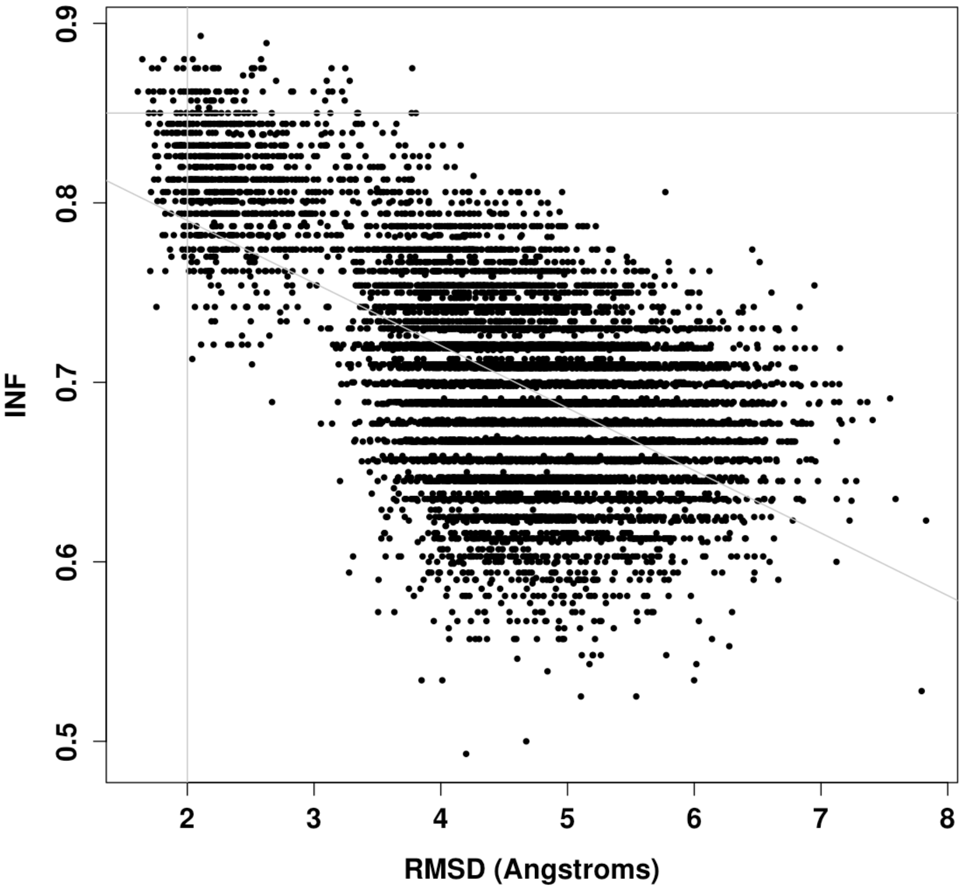


Figure 6.3: Distribution of (RMSD, INF) values. For each MC-Sym generated structure, the RMSD and INF values when compared with the crystal structure are plotted. The oblique line is the linear regression (P = 0.6). The horizontal line is at an INF of 0.85, and the vertical line at 2.0 Å RMSD.

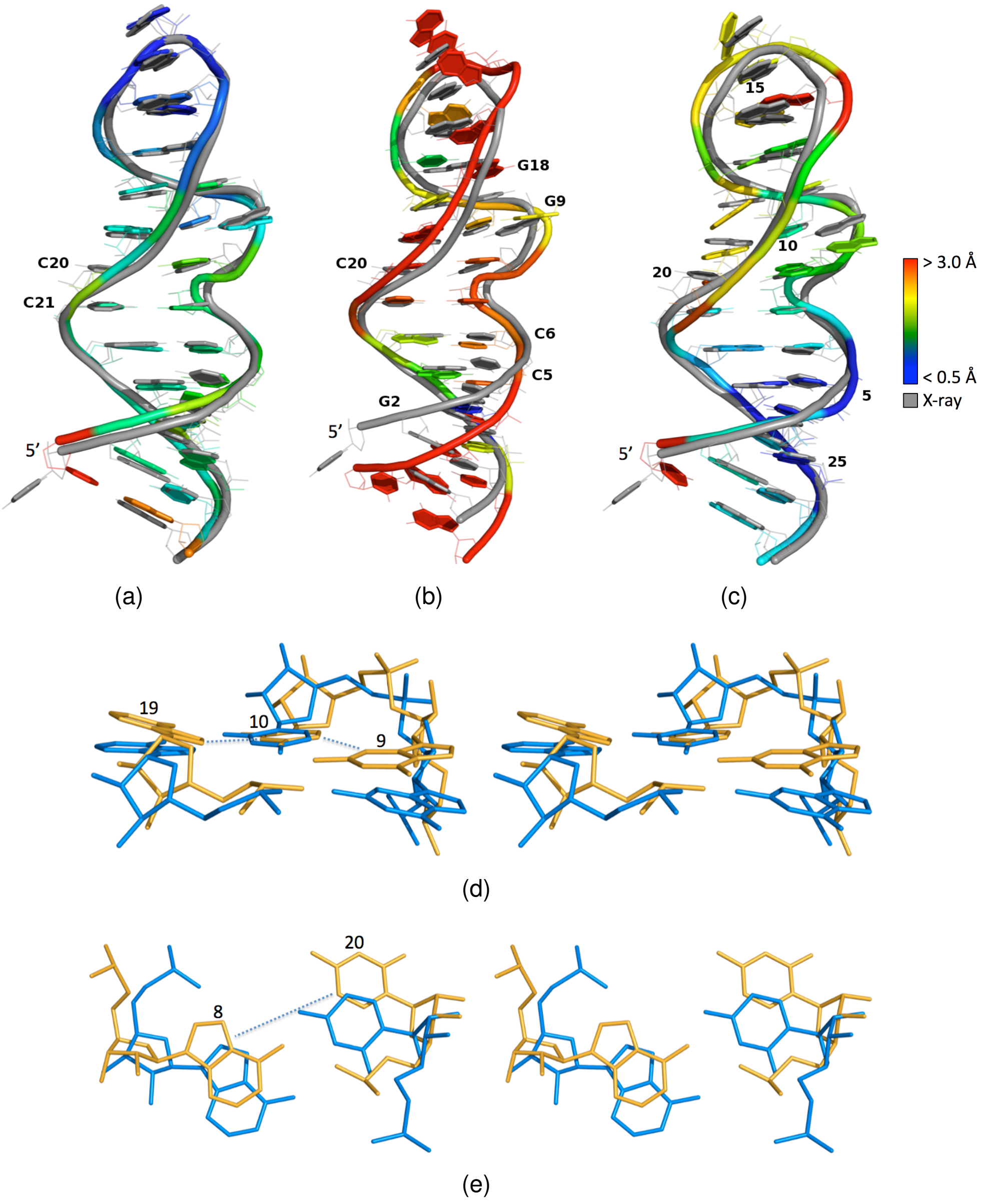


Figure 6.4: Three models of the rat 28S rRNA loop E. The models are shown colored and the crystal structure in gray (PDB code 1Q9A). (Blue) Well modeled regions (RMSD < 0.5 Å), (red) badly modeled regions (RMSD > 3.0 Å). The models were optimally aligned (all atoms but H) with the crystal structure. **(a)** Model with a good INF (0.88; TP 29; FP 6; FN 2) and good RMSD (1.64 Å); DI = 1.86. **(b)** Model with a good INF (0.88; TP 28; FP 5; FN 3), but bad RMSD (3.76 Å); DI = 4.30. Although the geometry of the base pairs is well conserved, the thread through the phosphate atoms is shifted. **(c)** Model with a bad INF (0.71; TP 21; FP 7; FN 10), but good RMSD (2.03 Å); DI = 2.85. The thread through the phosphate atoms is well superimposed, but the base-pairing geometry is wrong. Structural features that lead to a bad INF include: **(d)** base-stacking parameters that differ between the crystal (yellow) and model (blue) structures, such as G9, which shows a high rise in the crystal structure when compared with the model, and A19, for which a tilt can be observed between the crystal and model structures; and **(e)** base-pairing parameters that differ between the crystal and model structures, such as C20, which flips (propeller twist of 180°) between the crystal and model structures.

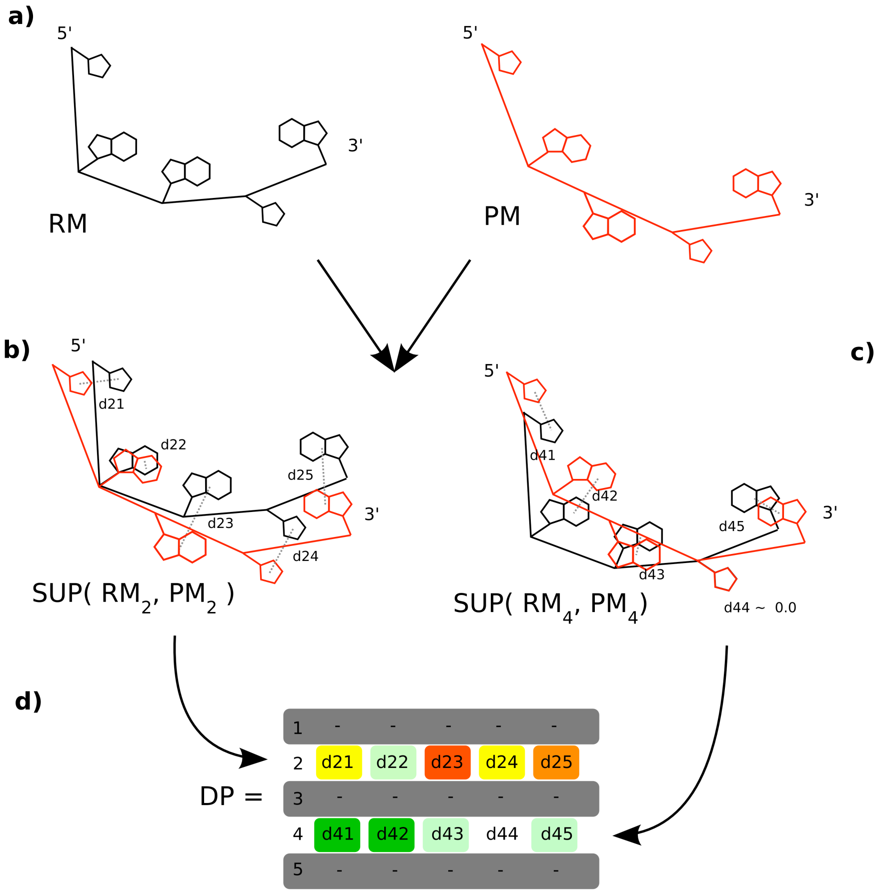
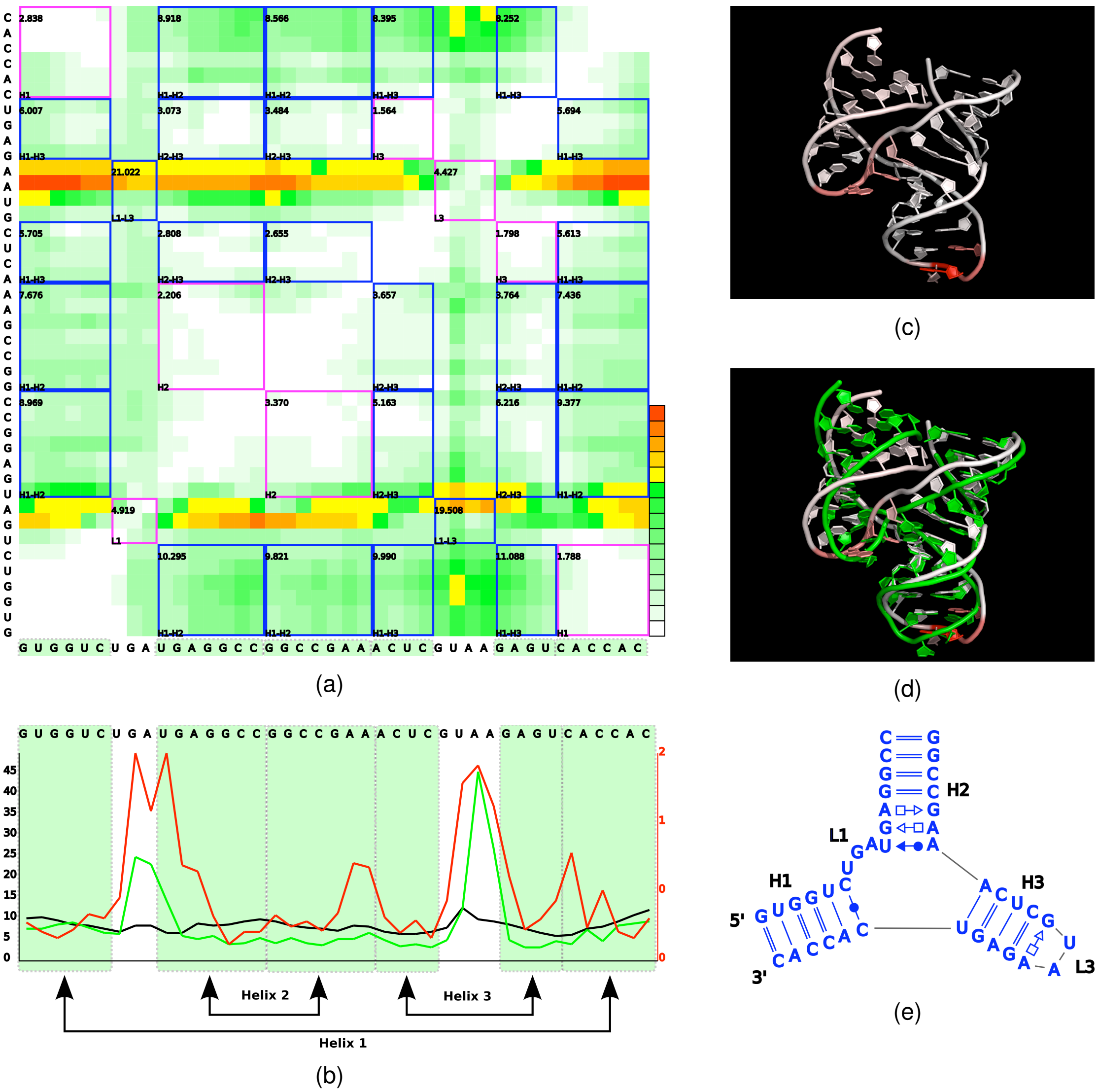


Figure 6.5: Building steps of the deformation profile. **(a)** A predicted model (PM) will be compared with the reference model (RM). After superimposing PM over RM, minimizing the RMSD between nucleotides 2 **(b)** and 4 **(c)**, the average distances between all atoms of corresponding nucleotides is calculated and recorded in DP matrix **(d)**.



(a)

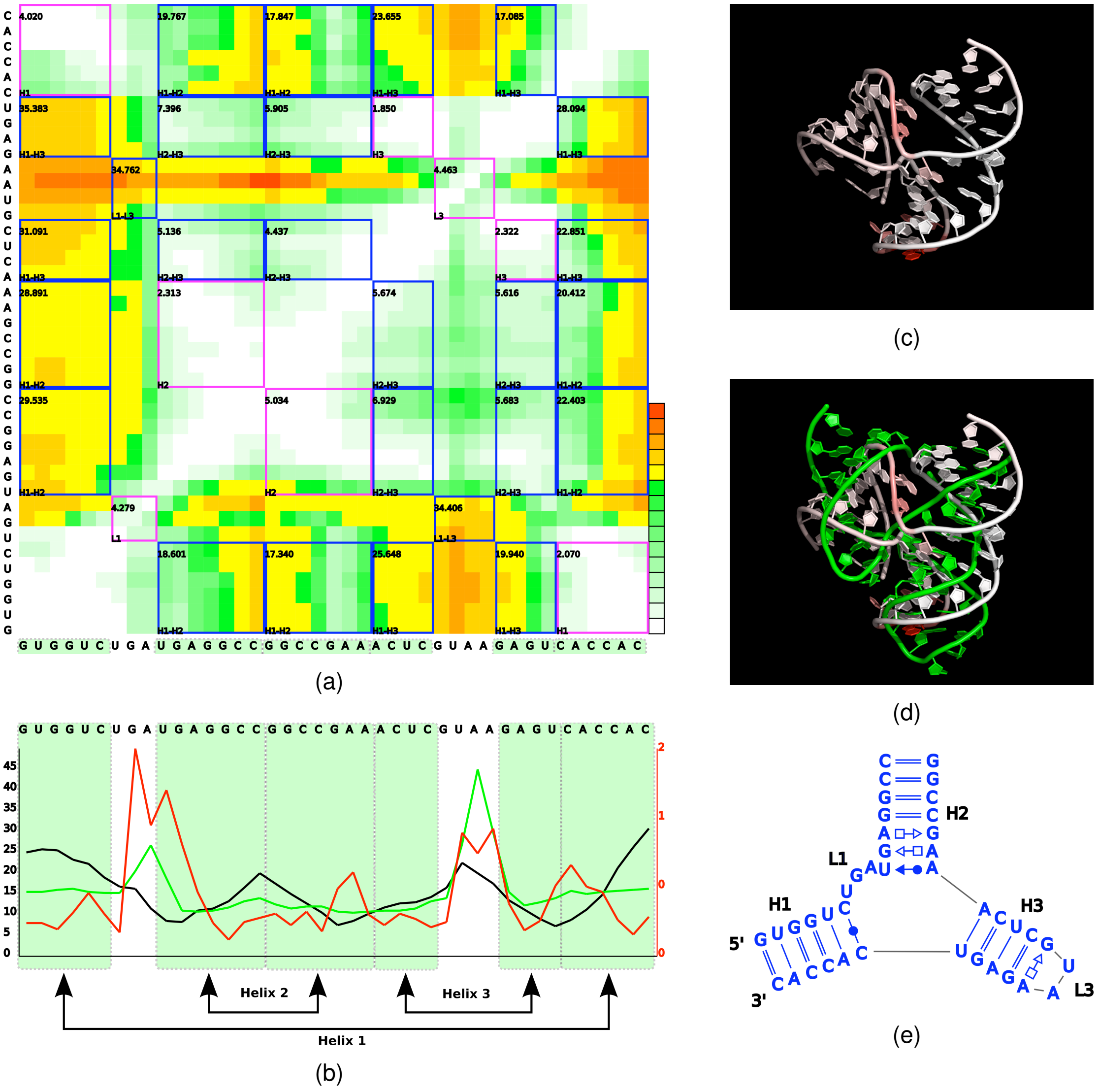
(b)

(c)

(d)

(e)

Figure 6.6: Deformation Profile between predicted model 553 and the hammerhead ribozyme crystal structure. **(a)** DP matrix. Blue and pink squares inside the matrix correspond to intra- and interdomain similarity relationships, respectively. Numbers in the left top corner of each square are the average value of all positions inside the square. Color scale goes from 0 Å (white) to (but not including) 20 Å (dark green) in 10 equal steps and from 20 Å (yellow) to 80 Å (red) in five equal steps. **(b)** Average values of rows (green), columns (black), and main diagonal (red) of the matrix. (Shaded green regions) Helical strands. **(c)** 3D structure of the model. Each nucleotide is colored according to the respective row average value, from minimum (white) to maximum deformation (red) value. **(d)** Superimposition of the model and reference 3D structures. **(e)** Interaction network of the original molecule.



(a)

(b)

(c)

(d)

(e)

Figure 6.7: Same as Figure 6.6 but for model 633.

Model[a]	Bipol[b]	Copl[b]	Rand[b]	RMSD	P-Sc[c]	Vol[d]	INF^{all} [e]	INF^{bp} [e]	GDT-TS[f]	Cluster
553	0.83	0.05	0.11	3.4	-23.6	23,635	0.82	0.90	0.60	2
633	0.80	0.12	0.08	12.2	-26.0	23,861	0.87	0.94	0.15	1
2698	0.81	0.12	0.07	4.9	-21.0	24,900	0.84	0.89	0.38	4
3778	0.84	0.05	0.12	12.2	-20.6	24,338	0.86	0.92	0.15	3
6870	0.76	0.08	0.16	13.9	-16.5	23,599	0.79	0.89	0.09	5

Table 6.1: Structural parameter values for five models of the hammerhead ribozyme.

[a] “Model” represents one model per cluster (Cluster) selected from the results of a “five-clustering”.

[b] Bipolar (Bipol), coplanar (Copl), and random (Rand) are measurements against the RMSD. These parameters describe the field of nucleobase normal vectors, which have been shown to be highly organized in solved RNA structures [256]. A threshold at 0.7 for the bipolar scores corresponds to a low RMSD (see Supplemental Fig. S1).

[c] The P-Score (P-Sc) against the RMSD measures the A-RNA likeliness of the phosphate chain-measured using the probabilities of valence angles of three consecutive atoms and the torsion angles of four consecutive atoms. The probabilities, P, are converted in pseudo-energies, E, using the Boltzmann relation: E = -RT log(P).

[d] Approximated ellipsoidal volume (Vol) against the RMSD. The volume is computed as described by Hao et al. [263]. A threshold at 25,000 corresponds to a low RMSD (see Supplemental Fig. S1).

[e] The INF values over base pairing and base stacking (INF^{all}) and base-pairing interactions alone (INF^{bp}).

[f] Global distance test (GDT-TS) values measure the average percentage of atoms within 1, 2, 4, and 8 Å from the target structure [257, 258]. The higher the value, the better the model compared with the target structure.

Intradomain	Model 553	Model 633	Model 2698
Helix H1	2.31	3.04	3.04
Helix H2	2.79	3.67	3.89
Helix H3	1.68	2.08	2.03
Loop L1	4.92	4.28	4.72
Loop L3	4.43	4.46	1.18

Interdomain	Model 553	Model 633	Model 2698
H1 × H2	8.88	21.85	13.25
H1 × H3	7.59	25.47	9.10
H2 × H3	3.85	5.85	6.49
L1 × L3	20.26	34.54	13.13

Table 6.2: Intradomain and interdomain scores for all helices, loops, helix-helix, and loop-loop combinations. The intradomain score of domain D is the average of all positions (i,j) of the Deformation Profile where both nucleotides i and j belong to D. The interdomain score of domains D1×D2 is the average of all positions (i,j) and (k,l) of the deformation profile where nucleotides i and j belong to D1 and j and k belong to D2.

CHAPTER 7

ARTICLE 2; SUPPL. INF.

New Metrics for Comparing and Assessing Discrepancies Between RNA 3D Structures and Models

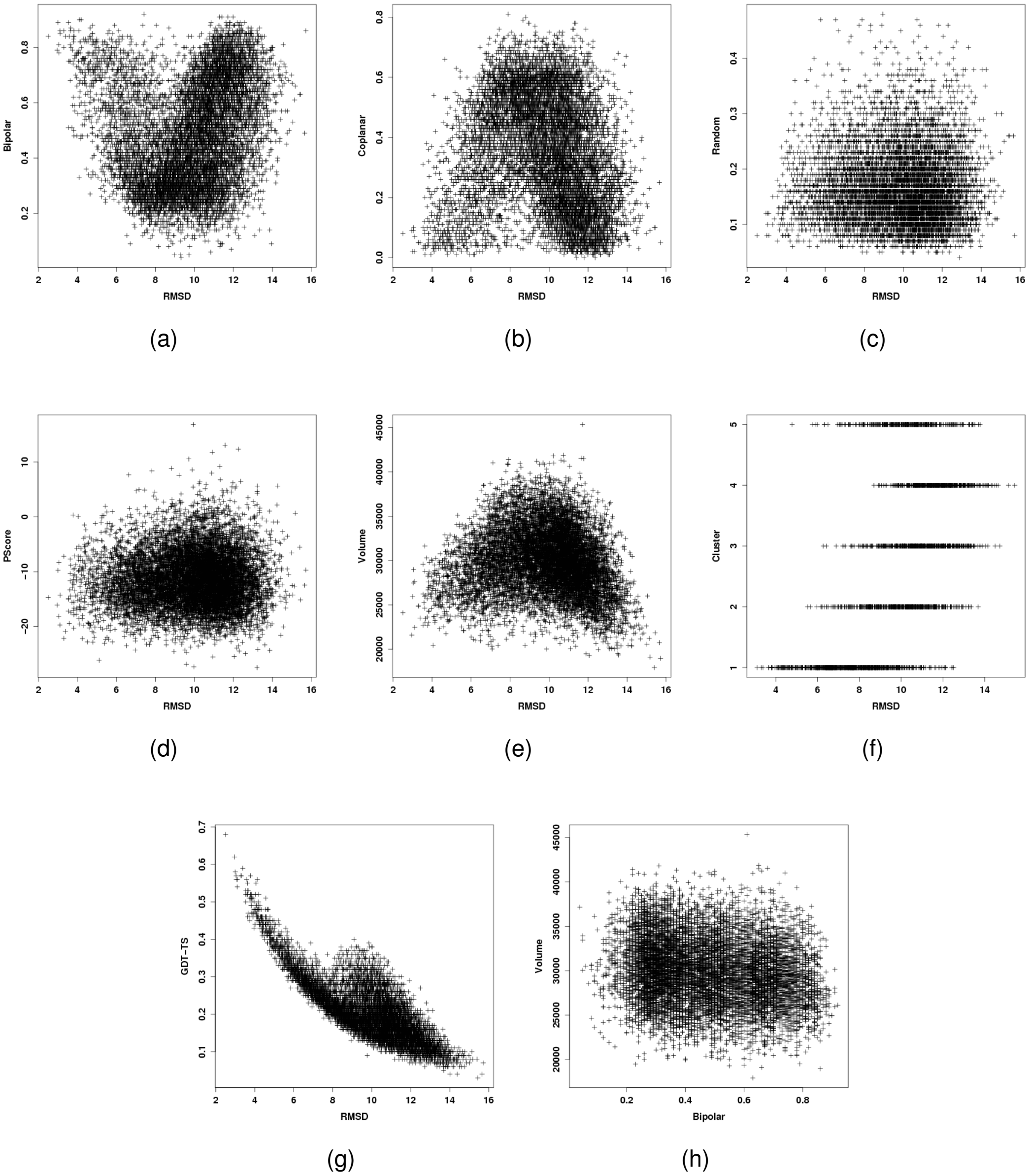
Marc Parisien[1*], José Almeida Cruz[2*], Éric Westhof[2] and François Major[1]

[1] *Institute for Research in Immunology and Cancer,*
Department of Computer Science and Operations Research,
Université de Montréal,
PO Box 6128, Downtown station,
Montréal, Québec, H3C 3J7, CANADA

[2] *Architecture et Réactivité de l'ARN,*
Institut de Biologie Moléculaire et Cellulaire du CNRS,
Université de Strasbourg,
67084 Strasbourg Cedex, FRANCE

** Equal contribution*

Parisien M, Cruz JA, Westhof E, Major F.
RNA 2009 **15**:1875-1885.

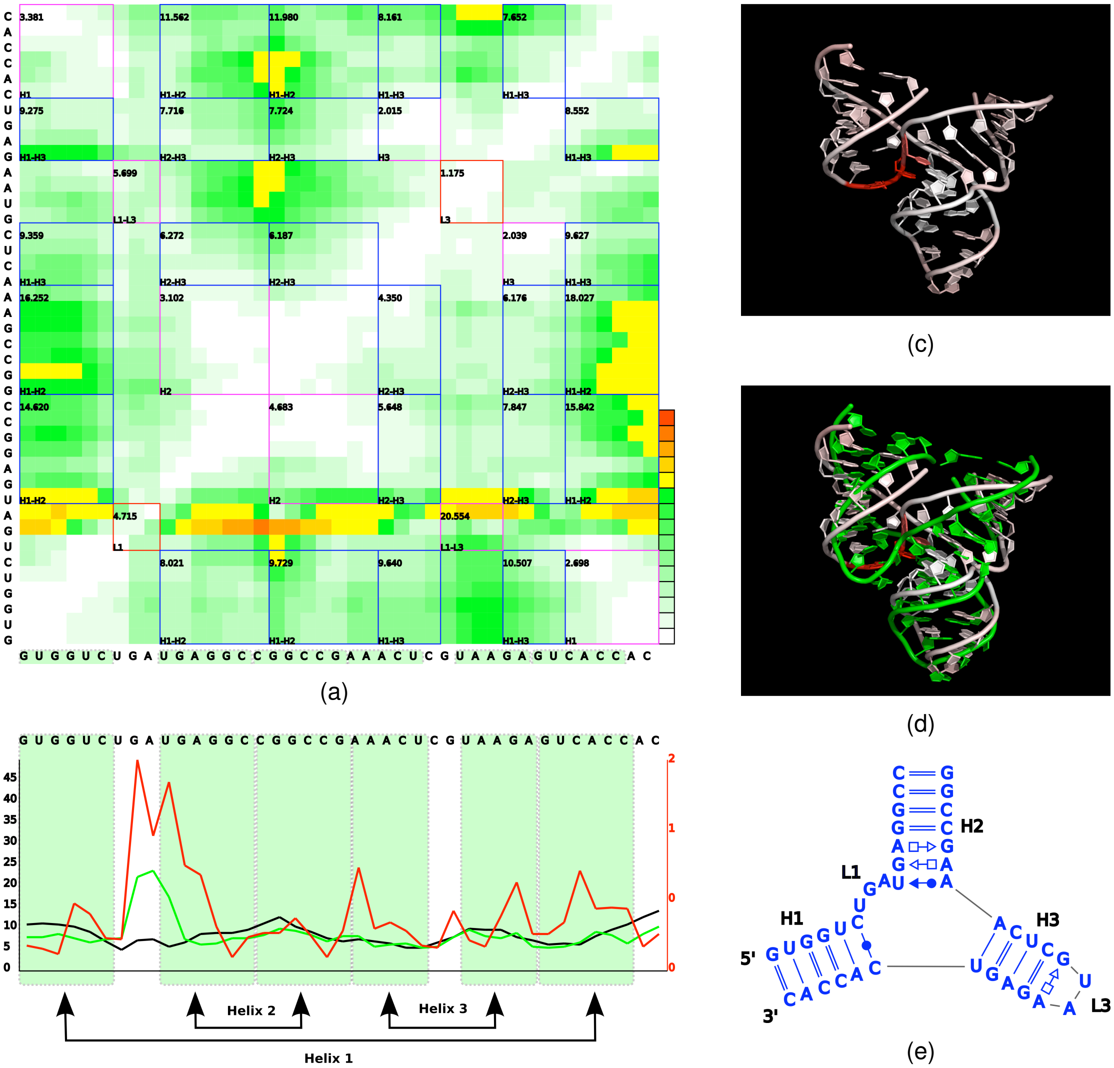


(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure 7.1: Structural parameters for decoys of the Hammerhead Ribozyme. In all figures, each decoy is localised in a two dimensional coordinate system using the plus (+) symbol. **(abc)** Bipolar, coplanar and random measurements against RMSD. These measurements describe the field of nucleobase normal vectors, which has been found to be highly organised in solved RNA structures [256]. An horizontal threshold has been set at 0.7 given the bipolar scores of low RMSD decoys. **(d)** PScore against RMSD. The PScore measures the A-RNA likeliness of the phosphate chain, by evaluating the probabilities of valence angles of three consecutive atoms, and the torsion angles of four consecutive atoms. The probabilities, P, are then converted in pseudoenergies, E, using the well celebrated Boltzmann relation: E = -RT log(P). **(e)** Approximated ellipsoidal volume against RMSD. The volume is computed as descibed by Hao et al. [263]. An horizontal threshold has been set at 25 thousand, given the volumes of low RMSD decoys. **(f)** RMSD ranges within each of the 5-clusters. **(g)** GDT-TS against RMSD. GDT-TS measures the average percentage of atoms of the decoy within 1, 2, 4, and 8 Å from the target structure [257, 258]. The higher the value, the better the decoy compared to the target structure. **(h)** Volume against bipolarity in which there is no correlation.



(a)

(b)

(c)

(d)

(e)

Figure 7.2: Same as Figure 6.7 but for model 2698.

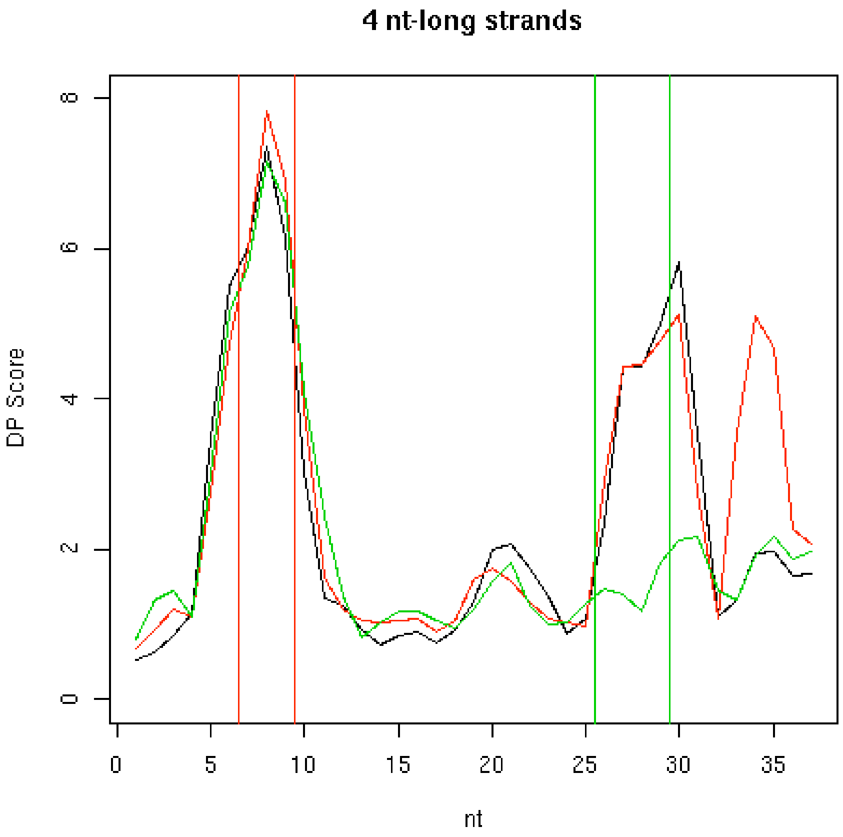


Figure 7.3: Deformation score of all short (<5 nt) contiguous regions.

CHAPTER 8

ARTICLE 3

Determining RNA three-dimensional structures using low-resolution data

Marc Parisien and François Major

Institute for Research in Immunology and Cancer,
Department of Computer Science and Operations Research,
Université de Montréal,
PO Box 6128, Downtown station,
Montréal, Québec, H3C 3J7, CANADA

(submitted)

8.1 Abstract

Despite recent improvements in our understanding of the RNA folding process and structure prediction algorithms, determining RNA three-dimensional (3-D) structures requires the exploitation of experimentally determined structural data. Here, we introduce an alternative method to determine RNA 3-D structures using low-resolution data when high-resolution methods such as X-ray crystallography and NMR cannot be applied. We generated sets of 3-D models of the *Escherichia coli* tRNA^{VAL} and the P4-P6 fragment of the *Tetrahymena thermophila* group I intron using the MC-Fold and MC-Sym pipeline. We then filtered the models using experimental data of single and combinations of hydroxyl radical footprinting (OH), methidiumpropyl-EDTA (MPE), multiplexed hydroxyl radical cleavage (MOHCA), and small-angle X-ray scattering (SAXS) as a measure of their discriminative power to identify the native structures. We show that: 1) The precision of the structures generated by MC-Sym is higher than that of the experimental data tested in the context of this study; 2) OH-derived constraints are precise and discriminate with atomic precision, whereas 3) EDTA-based constraints allow for distinguishing the global shape of RNA structures given their radius of gyration is greater that the probe length; 4) SAXS is a promising approach to high-throughput RNA structure determination. However, unlike EDTA, SAXS-derived constraints cannot be directly exploited to guide the structure generation algorithm.

8.2 Introduction

Ribonucleic acid (RNA) is one of the central molecules of cellular information transfer [13, 264]. Characterising the reactions and knowing how RNA participates increase our comprehension of how cells function. Structure is one of our best hints about RNA function, and high-resolution methods such as X-ray crystallography and nuclear magnetic resonance (nmr) spectroscopy are routinely applied to determine RNA three-dimensional (3-D) structures. However, these methods cannot be applied to all RNAs and in all cellular conditions. RNA 3-D structures can alternatively be predicted computationally or modelled interactively [44, 45, 46, 47, 48]. However, a major problem with almost all computational structure prediction methods is the generation of many different and energetically similar structures from which it is difficult to identify the native and active conformations. The problem is partially related to the facts that RNAs have a great potential to accommodate various reactions under several contextual conditions, for instance riboswitches, ribozymes, and microRNAs are subject to conformational induction in presence of their respective cofactors [216].

To help sort through many plausible models, much effort is being invested in using low-resolution experimental methods and data. These include variants of hydroxyl radicals footprinting [57], the ethylenediaminetetraacetic acid (EDTA) variants uridine-EDTA [265] and methidiumpropyl-EDTA (MPE) [58]; and, multiplexed hydroxyl radical cleavage analysis (MOHCA) [59]. Another emerging structure probing technique showing promises to accelerate the systematic resolution of RNA structures is small-angle X-ray scattering (SAXS) [60]. Here, we compare the above structure probing methods for their discriminative power to identify experimentally resolved structures from sets of theoretically generated all-atom models.

8.3 Results

8.3.1 Conformational sampling

We chose two well-studied RNAs: the *Eschericia coli* valine tRNA (tRNA), which has been solved in solution (PDB code 2K4C [261]), and the group I intron of the *Tetrahymena thermophila* P4-P6 domain (P4-P6), which structure has been determined by X-ray crystallography (PDB code 1GID [116]). These structures have been used as benchmarks for RNA secondary and tertiary structure prediction methods [48, 58, 59, 266]. We also chose these RNAs for the availability of structure probing data.

To challenge the RNA structure probing methods thoroughly, we needed an RNA structure generator capable of sampling the global shapes and producing atomic resolution models of the conformational space accessible to the tRNA and P4-P6 sequences. All-atoms were also desirable models to assess the reproducibility of the base pairs and stacks, and thus proper volume, electron density, and quality of the backbone paths. To the best of our knowledge, only the most recent version of the MC-Sym computer program [63, 215, 267] combines the desired attributes.

We built two sets of 3-D models for each RNA. The first set is termed “low”, where we only exploit local information, i.e. we input the right stems and their coaxial stacking. The second set is termed “high”, where we exploit some additional long-range distance base pairs. For the extra base pairs in the high set, we exploited structural data that have previously been inferred from sequence covariation analysis [139, 268, 269, 270]. Such long-range interactions are extremely valuable in the context of modelling since they provide strong constraints on the conformational search space by bringing together nucleotides that are distant in the sequence. The low and high sets simulate realistic situations of any RNA modeller in the first stage of his work, i.e. access to either a single or multiple sequences. Only in the latter scenario one can count on inferring a few covariations. Fig. 8.1 shows the distribution of the root-mean-square deviations (RMSD) (all-atoms) to their respective experimental structures of the models generated for both RNAs. For the tRNA, the RMSD range from about 6.1 up to 25.1 Å and 3.6 up to 12.4 for the tRNA low and tRNA high sets respectively. For P4-P6 domain which contains twice the number of nucleotides of the tRNA (158 vs. 76), the RMSD range from 13.1 to 49.4 Å and 6.9 to 14.2 Å for the low and high sets respectively. We note that the inclusion of a very few long-range tertiary contacts (see Methods) significantly affect the RMSD distributions of the generated models. For instance, the RMSD peaks shown on Fig. 8.1 go from 17 down to 5 Å and from 43 down to 9 Å for the tRNA and P4-P6 respectively.

8.3.2 Discriminative power of low-resolution experimental data

We use the correlation between the experimental data fitness and RMSD to assess the relative power of each experimental methods to discriminate the native fold among the sets of models (Figs 8.2 and 8.3). A visual inspection clearly shows that the best correlation is from using MOHCA on the P4-P6 low set (r^2 = 0.995). The next-best method is SAXS, again on the P4-P6 low set (r^2 = 0.347 in the RMSD region below 25 Å), as well as on the tRNA low set (r^2 = 0.234; in the RMSD region below 15 Å). The r^2 is below 0.2 for all other methods and model sets, indicating that for a any given fitness value many different models would qualify and any given RMSD distance to the solution structure a wide range in the fitness is observed.

The correlation of MPE on the tRNA low set is r^2 = 0.158. The problem of selecting one or a few models that represent the native fold is thus complicated in this case. One cannot simply pick the model that best fit the experimental data because the range in RMSD is quite large. Consider as another example the RMSD range of the OH fit above the 0.5 line in Fig. 8.2a).

Therefore, we devised a selection procedure (see Methods) that yields a handful of similar models that collectively best fit the experimental data. The selection procedure is repeated a hundred times, and then averaged on all selected models. We feel that this procedure is robust since the RMSD standard deviations between the selection cycles are small (around 1 Å on low sets, and around 0.5 Å on the high sets; Tables 8.1 and 8.2, values in parenthesis).

Results of applying the selection procedure on all sets are presented in Table 8.1 (tRNA) and Table 8.2 (P4-P6). For each individual and combination of experimental methods, the mean RMSD and Global Distance Test (GDT-TS [257, 258]) are shown. These measures show how close the selected models are compared to the experimentally resolved structure. Another interesting measure is the Q-value, which is the percentage of models in a set that have RMSD lower than the average RMSD. Lower Q values indicate that the selected structures have not been chosen by chance. On the low sets, SAXS seems the best method for both structures (Q=3.2 and 0.7 for the tRNA and P4-P6 respectively). MOHCA performs well but on the P4-P6 low set only (Q=0.4). MPE on the tRNA low set is less selective (Q=11.6) than SAXS.

Importantly, OH footprinting is of no help at the low modelling resolution. However, as the models gain in precision and get closer to the native fold, OH footprinting becomes increasingly discriminative. For instance, for the P4-P6 models selected from the high set and involving OH, we get Q-values of 2.0 and 11.5 for respectively OH+SAXS and OH+MOHCA. In the case of the tRNA high set, when only OH data are considered we get Q=36.9, and when OH+MPE are considered then we get Q=40.6. It is interesting to mention that OH data have been exploited in conjunction with MOHCA for the elucidation of the native fold of the P4-P6 domain [59]. It was found that the addition of OH data to that of MOHCA selects worst models than MOHCA data alone, most likely due to “kinetic traps”. The two RNA structure generators used in this experiment were FARNA [44] and NAST [48], which are energy-driven and hence susceptible to fall in local energy minima. In comparison, MC-Sym is geometrical constraint rather than energy based, and we found that incorporating OH data to that of MPE, MOHCA, and SAXS helps for identifying better models than without OH data. Here, we found that OH data are more useful to filter generated model sets (MC-Sym approach) than to guide the generation process itself (FARNA and NAST approaches).

8.4 Discussion

The observation of quite high Q-values for the high tRNA set (Table 8.1) indicates that either: i) our models in the high set are too precise for the resolution of the experimental methods tested here; or, ii) the tRNA is too small or compact to make a notable difference upon substantial conformational changes, i.e. the length of the EDTA probe is comparable to the radius of gyration of this tRNA.

Also noteworthy is the difference in the interpretation of the EDTA-based distance data between MPE and MOHCA. For MPE, the models are gratified for their fit to various distance constraints proportionally to the cleavage intensity at the measured sites (see Equation 1 in Methods), whereas in MOHCA, the cleavage intensities are not taken into account, and thus the models are penalised for distance violations from the estimated tether probe length. We reinterpreted the MPE data for the tRNA, MPE* and MPE**, in a MOHCA-like fashion, i.e. any site with a cleavage intensity ratio I/<I> above 1.0

should be within 30 Å of the probe's base, or else the excess distance $D = (d - 30)$ is scored as Δ^2 for MPE* (within 35 Å for MPE**). Fig. 8.4 shows how MPE* and MPE** vary with RMSD. On the high tRNA set, both MPE* and MPE** do not point to a model that is native-like. Given that the tRNA fold has the potential to be caught in many conformations [271], then it would be interesting to compare buffer solutions between the MPE-based probing and, say, that of SAXS, or to revisit how the EDTA-based data should be interpreted. That said, the radius of gyration of the tRNA has been measured to be about 23.5 Å [272], which is within the range of the tether probe.

For computational cost, computing the fitness to OH data is not efficient because it requires the ASA (see Methods). However, computing the fitness to MPE data is efficient since it provides distance constraints that can be used either in the model generation process or as a postprocessing step of it. Computing the fitness of MOHCA data is also efficient since it also provides distance constraints. Finally, evaluating the fitness to SAXS data is not efficient since it requires complex numerical computations and must consider the first hydration shells of the footmark on the scattering data, particularly in the case of RNAs. Progress toward fast SAXS profile evaluation is however under investigation by us in collaboration with Yang and Roux at the University of Chicago (unpublished).

As we can see from our results, there is plenty of space for improving RNA low-resolution structure determination methods, both at the RNA probing methods and quality of the models. It would be pleasant to select the best models using theoretically-derived force-fields, such as Amber and CHARMM, but unfortunately this is an even more difficult problem and for sometimes we rather focus our efforts on resolving low-resolution RNA structures experimentally.

8.5 Methods

8.5.1 Sets of 3-D models

We used the latest version of the MC-Sym computer program to generate the sets of models. MCSym uses nucleotide cyclic motifs (NCM), which are either single-stranded fragments closed by a single base pair (hairpin loops) and double-stranded fragments that are flanked by two base pairs (tandem of base pairs, bulge and interior loops) [63]. The fragments are either taken from the PDB when instances of the given sequence exist, or built by homology modelling otherwise. MC-Sym generates all-atom models. The use of NCMs reproduces accurately inter-nucleotide interactions such as the GNRA/receptor in the group I intron and the base triples in the tRNA structure. MC-Sym runs on desktop computers.

For the tRNA, we started with the secondary structure with the addition of in-stem non-Watson-Crick base pairs, such as the A14-A21, A26-G44, C32-A38, and U54-A58. The T-loop structure was imported from the PDB file 2K4C since it folds independently and in a well-defined motif [130, 131, 273, 274]. We also used the Fuller-Hodgson rule for the anticodon loop [275], i.e. we stacked nucleotides 34 to 38 as in an A-RNA helix. The coaxial stacking between the Acceptor and the T stems, as well as between the Anticodon and the D stems was used so that the four stems define two helices [135, 139]. The conformational sampling of the low tRNA set was made in three stages. First, we sampled and generated structures for the two helices: axis 1 formed by the T- and acceptor-stem; and, axis 2 by the D- and anticodon-stem. Second, the two helices are positioned relative to each other by sampling the dinucleotide 8-9 conformation. Finally, we completed the tRNA structure by building the D and variable loops by sampling trinucleotides using singlestranded fragments extracted from the PDB. The conformational sampling of the tRNA high set was also made in three stages. The first stage is identical to that of the tRNA low set, but we selected the conformations of the D-stem that formed the base triples 8-14-21, 9-12-23, and 13-22-4620. We also appended nucleotides 18 and 19 to form long-range base pairs with the T-loop nucleotides 55 and 56, respectively [139, 268]. Second, we positioned the two helices relative to each other using the dinucleotide 7-8. Finally, we added the 15-48 long-range base pair [139, 268], and completed the D and variable loops. Fig. 8.5 shows the information used for modelling the tRNA sets, and gives an idea of the conformational search space explored for each.

For the group I intron P4-P6 domain, we used a similar procedure as that of the tRNA. We divided the domain into two components: the P5-P4-P6 coaxial stems (axis 1), and the P5a and P5b coaxial stems (axis 2). We extracted the P5c stem and tetraloop receptor from PDB file 1GID. The tetraloop receptor adopts a specific structure that contains an adenosine platform [126] and a UA_handle [276]. In the first stage, we sampled and generated structures for each stem. For the P4-P6 low set, we then positioned the two components relative to each other by sampling the 122-126 and 196-199 strands using trinucleotide fragments extracted from the PDB. Hence, given no other long-range distance constraints, MC-Sym generated P4-P6 domain structures in the familiar U-shape, but also in the L- and flat shapes. For the P4-P6 high set, we added the long-range base pair 153-250 observed in the crystal structure [116]. This type of interaction was first postulated by Michel and Westhof [118]. Later, Murphy & Cech using protection data showed this interaction stabilises the P4-P6 domain tertiary structure [277]. Costa & Michel showed the interaction can be predicted by covariation analysis data [270], and it was further characterised by computer modelling [269], X-ray crystallography [116], and structural analysis [276]. For both sets, in the final stage we completed the domain by adding the A-rich loop (183-186) and the P6 tail (254-260). Fig. 8.6 shows the information used in the modelling of the P4-P6 sets and gives an idea of the conformational search space explored for each.

8.5.2 Hydroxyl radical footprinting

Cleavage of the RNA backbone by hydroxyl radicals (OH) provides a way to distinguish the inside and outside of a folded RNA molecule [278]. As the RNA folds in its native state, sections of the backbone may become protected from the solvent, which provides structural information. OH footprinting is gel-based [57] and amenable to robust and automated analysis [279]. OH attack the backbone chain to cleave it. Backbone atoms that are buried inside the RNA are less prone to these attacks. Hence, the OH footprint measures the accessibility of backbone atoms to the solvent at the nucleotide level, and the greater the cleavage activity the more accessible are the backbone atoms. To evaluate the fit of a 3-D model to the OH footprint, we first calculate the sum of the accessible surface area (ASA) of all backbone hydrogen atoms per nucleotide (NASA) in the model using the MSMS computer program [280]. We take the radius of a water probe (1.4 Å) as an approximation of the radius of the hydroxyl radical to evaluate the ASA. Then, we perform a linear least-squared best fit between the experimentally measured cleavage intensity and the NASA. The fit is quantified by the Pearson's correlation coefficient. For the tRNA, we used for the footprint profile the calculated NASA of the PDB solution structure (2K4C). For P4-P6 domain, we used published data [281].

8.5.3 Methidiumpropyl-EDTA

The methidiumpropyl (MPE) version of ethylenediaminetetraacetic acid (EDTA) is another gel-based method that provides distance constraints between different fragments of an RNA molecule [58], similarly to uridine-EDTA [265]. The idea behind these methods is the insertion of an Fe(II)-EDTA moiety at a specific position, and the consequent backbone cleavage pattern defining the accessibility and reach of the tether. EDTA combined with selective 2'-hydroxyl acylation and primer extension (SHAPE) [224], and a force-field based on a united residue representation [47], has proven to be sufficient to reproduce the tRNA fold with high accuracy (< 4 Å; P atoms in the secondary structure) [58].

Let us say that the Fe(II)-EDTA probe is attached at nucleotide position P. From the cleavage intensities, I at nucleotide S, a pseudo-potential energy, E, is assigned proportionally to I: E = -ln(I/<I>), where <I> is the profile's mean intensity. Then, for each pair of phosphate atoms (P, S) separated by a distance D, an energy, ε, is assigned in a stepwise manner:

$$\varepsilon = \begin{cases} -E, & \text{if D(P,S)} \leq 25\text{Å}, \\ -2E/3, & \text{if D(P,S)} \leq 30\text{Å}, \\ -E/3, & \text{if D(P,S)} \leq 35\text{Å}, \\ 0, & \text{otherwise} \end{cases} \tag{8.1}$$

We sum all ε for all probing experiments P and for all nucleotide positions S. The lower the energy, the better the fit with the pairwise distance constraints. In our study, we used the MPE data published by the Weeks's group on the tRNA-ASP structure [58]. We assumed that the MPE profiles for the tRNA-ASP would be appropriate for the tRNA-VAL by inserting an extra dummy nucleotide in the variable loop of the tRNA-ASP profile to match the sequence length of the tRNA-VAL.

8.5.4 Multiplex hydroxyl radical cleavage analysis

Another gel-based probing method that uses EDTA is the multiplexed hydroxyl radical cleavage analysis (MOHCA) [59]. Here, the EDTA probe cleavage agents are randomly inserted in the sequence. Then, the positions of radical hydroxyl cleavage is read in a two-dimensional gel against the positions of the cleavage agents, yielding a group of pairwise distance constraints corresponding to the reach of the probe's tether. The analysis of MOHCA data has recently been automated [282]. In the context of RNA modelling, the procedure generates pairwise distance constraints, which we assign between C1' backbone atoms. For each distance constraint, a squared penalty score, Δ^2, is given for any distance d beyond 30 Å; $\Delta = (d - 30)$. Hence, the fit of a model is the sum of its penalties. Since MC-Sym does not make use of internal energy, there is no need for a coupling constant between the MOHCA distance violations and the energy of a structure. For our study, we used the distance constraints for the native state of the P4-P6 domain [59].

8.5.5 Small-angle X-ray scattering

The idea is to bombard the RNA in solution with X-rays and to inspect how the scattering profile varies, averaged over all orientations of the RNA and according to small incident angular changes of the X-ray beam [283]. SAXS data can be used in two different ways: first, by comparing directly a theoretical scattering curve with that obtained experimentally. The Fourier transform of the scattering curve gives the pair-density distribution function, $P(r)$ (PDDF), i.e. the pairwise distance distribution between electrons of the molecule. Despite its usefulness in molecular modelling, as the PDDF of a set of structures can be computed quickly and compared easily to the experimental PDDF, the numerical methods used and the dependance of $P(r)$ on Dmax (the maximum pairwise distance observed) introduce large error bars plaguing the PDDF; second, by obtaining a beads model whose density agree with the scattering curve, and then building a set of structures that fit the beads model. We favour the first approach because of the degeneracy of the beads solutions and the inappropriateness of the beads model generators since they were developed in the context of globular protein structures.

In RNA, because of the high residency of water molecules in both the minor and major grooves [91], a new and fast method employed here explicitly takes into account the scattering of the first hydration shells. This method has been developed in collaboration with Yang and Roux at the University of Chicago (unpublished). In a modelling context, a theoretical scattering curve is compared to an experimental one. This type of experience produces data that are not readily expressible as distance constraints. Alternatively, one could use $P(r)$ at the expense of a greater incertitude. In our study, we use the SAXS data provided by the Bax's lab for the tRNA [261], and the SAXS data from Doniach's lab for the P4-P6 domain [60].

8.5.6 Normalized Z-score

To express the fitness of a model to various experimental data types, we first compute a normalised Z-score, $Z_e(x)$, for the fitness of the model x in an experiment type

e. Then, the total fit, $Z(x)$, is simply given by $= \sum_e Z_e(x)$. The Z-score, $Z(x) = (x - \mu)/\sigma$, takes into account the mean, μ, and the standard deviation, σ, of the fits of all models for each experiment type. The normalised Z-score is $Z_e(x) = (Z_e(x) - Z_e^{\min})/(Z_e^{\max} - Z_e^{\min})$, so that each experimental data type contribution is between zero and one, where one signifies the best fit (care is taken to properly compute $Z_e(x)$ given that the highest or lowest experimental fit value is the best).

8.5.7 Model selection procedure

A question that naturally arises is how can we select a few representative structures that best fit the experimental data. Declaring the best-fit model as the native fold is dangerous because many models fit the data. Therefore, we sort the models from best to worst fit and we choose the top N models, where N is picked randomly between 100 and 500. We define this set as promising, and we partition it into 10 subsets using the K-clustering algorithm based on root-mean-square deviations (RMSD). We expect the K-clustering (K=10) to yield subsets of 10 to 50 members each. We choose the centroid centre of the subset, which on average fits best the experimental data. We define this model as the representative of the native fold. Because of the stochastic nature of the greedy K-clustering method, this selection procedure is repeated a hundred times, and we report the averaged RMSD and Global Distance Test (GDT-TS [257, 258]) between the centroids and corresponding reference structures.

8.6 Acknowledgements

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Institutes of Health Research (#MOP-93679) to FM. FM is a member of the Centre Robert-Cedergren of the Université de Montréal. MP is a recipient of PhD scholarships from the NSERC and the FQRNT (Fonds Québécois de Recherche sur la Nature et les Technologies).

Experiment			Resolution			
OH	MPE	SAXS	<RMSD>	<GDT-TS>	N	Q
tRNA_low ([6.1, 25.1] Å)						
■	□	□	15.92	0.07	2406	51.5
			(0.64)	(0.01)		
□	■	□	10.97	0.10	3277	11.6
			(0.57)	(0.02)		
□	□	■	9.11	0.20	2974	3.2
			(1.10)	(0.05)		
■	■	□	10.09	0.14	3997	7.0
			(0.98)	(0.03)		
■	□	■	9.16	0.18	4241	3.3
			(1.06)	(0.03)		
□	■	■	11.32	0.10	3276	14.3
			(0.65)	(0.01)		
■	■	■	9.46	0.17	4798	4.4
			(1.06)	(0.04)		
tRNA_high ([3.6, 12.4] Å)						
■	□	□	5.46	0.38	2535	36.9
			(0.81)	(0.05)		
□	■	□	5.71	0.35	2566	47.0
			(0.20)	(0.01)		
□	□	■	6.56	0.30	1649	74.4
			(0.19)	(0.02)		
■	■	□	5.55	0.38	2043	40.6
			(0.20)	(0.02)		
■	□	■	6.26	0.33	2890	65.8
			(0.49)	(0.03)		
□	■	■	7.51	0.25	1995	90.8
			(0.52)	(0.02)		
■	■	■	6.34	0.31	2760	68.8
			(0.56)	(0.03)		

Table 8.1: Performance of experimental data for the tRNA model sets. The first three columns are the experimental data sets: OH (hydroxyl radical footprinting); MPE (methidiumpropyl-EDTA); and, SAXS (smallangle X-ray scattering). A filled square in a column indicates that the experimental data from this method were considered in the selection; the empty square indicates that they were not. <RMSD> is the average RMSD in Å for 100 selection procedures (see Methods). <GDT-TS> is the average GDT-TS for 100 selection procedures. N indicates the total number of structures selected. Q is the percentage of structures in the set that have lower RMSD than <RMSD>. Values in parenthesis are the standard deviations.

Experiment			Resolution			
OH	MOHCA	SAXS	<RMSD>	<GDT-TS>	N	Q
P4-P6_low ([13.1, 49.4] Å)						
■	□	□	43.78	0.01	2770	84.0
			(1.79)	(0.00)		
□	■	□	16.57	0.06	2558	0.4
			(0.50)	(0.01)		
□	□	■	17.46	0.05	3029	0.7
			(1.17)	(0.01)		
■	■	□	19.76	0.04	3234	1.9
			(1.59)	(0.01)		
■	□	■	16.72	0.06	2382	0.4
			(1.18)	(0.01)		
□	■	■	17.35	0.06	2695	0.6
			(1.19)	(0.01)		
■	■	■	17.57	0.05	2799	0.7
			(1.27)	(0.01)		
P4-P6_high ([6.9, 14.2] Å)						
■	□	□	9.47	0.33	7556	52.5
			(0.09)	(0.00)		
□	■	□	8.76	0.27	2939	21.8
			(0.21)	(0.01)		
□	□	■	10.34	0.22	887	81.2
			(0.34)	(0.01)		
■	■	□	8.30	0.27	1735	11.5
			(0.16)	(0.02)		
■	□	■	7.48	0.35	6128	2.0
			(0.05)	(0.00)		
□	■	■	8.42	0.31	6765	14.4
			(0.09)	(0.00)		
■	■	■	7.48	0.35	6834	2.0
			(0.07)	(0.00)		

Table 8.2: Performance of experimental data for the P4-P6 model sets. The first three columns are the experimental data sets: OH (hydroxyl radical footprinting); MOHCA (multiplexed hydroxyl radical cleavage analysis); and, SAXS (small-angle X-ray scattering). A filled square in a column indicates that the experimental data from this method were considered in the selection; the empty square indicates that they were not. <RMSD> is the average RMSD in Å for 100 selection procedures (see Methods). <GDT-TS> is the average GDT-TS for 100 selection procedures. N indicates the total number of structures selected. Q is the percentage of structures in the set that have lower RMSD than <RMSD>. Values in parenthesis are the standard deviations.

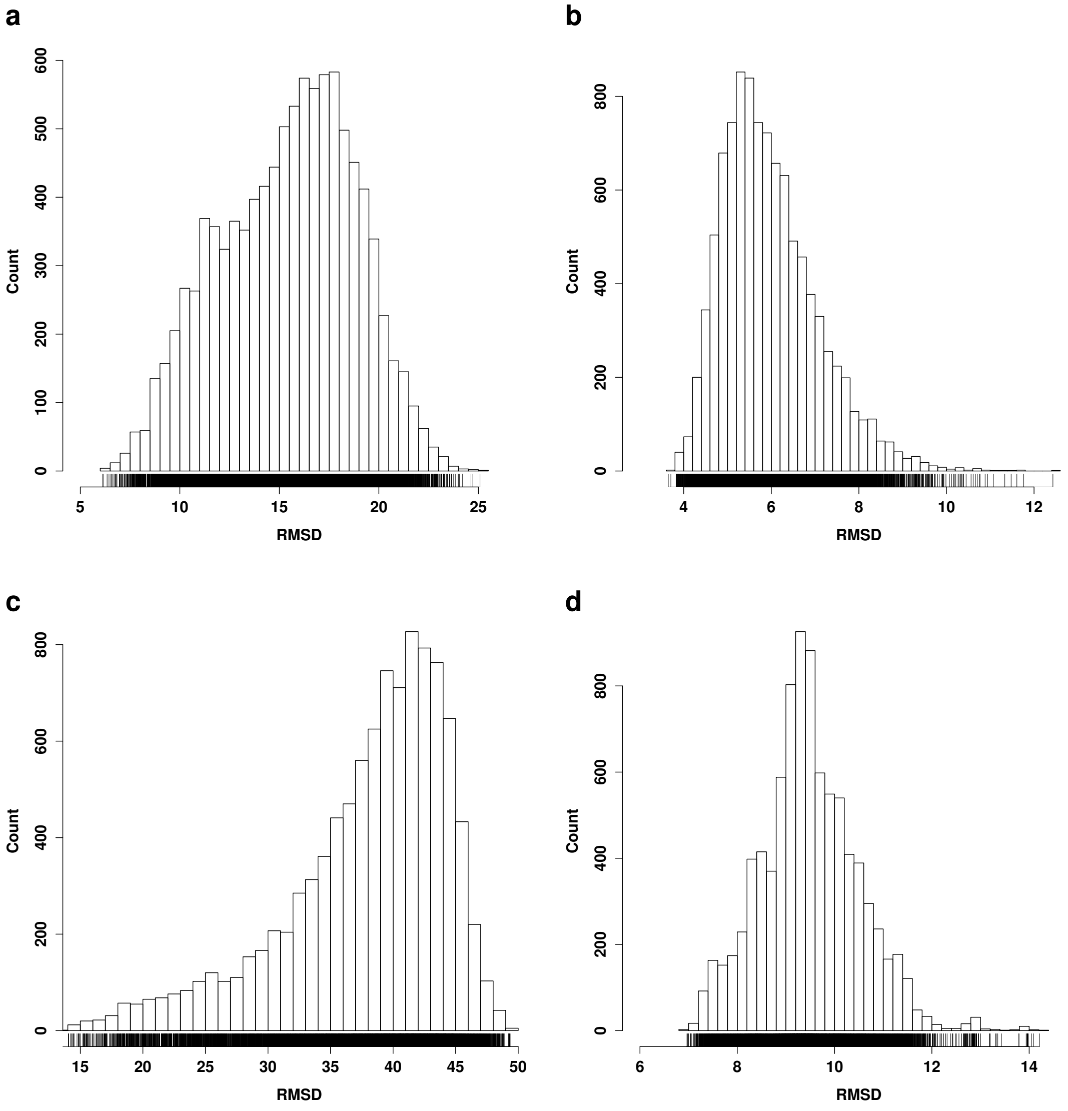


Figure 8.1: Conformational sampling RMSD ranges. All-atoms RMSD (Å) are computed against the tRNA solution structure (PDB code 2K4C) and P4-P6 crystal structure (PDB code 1GID). a) Low tRNA set. b) High tRNA set. c) Low tRNA set. d) High P4-P6 set.

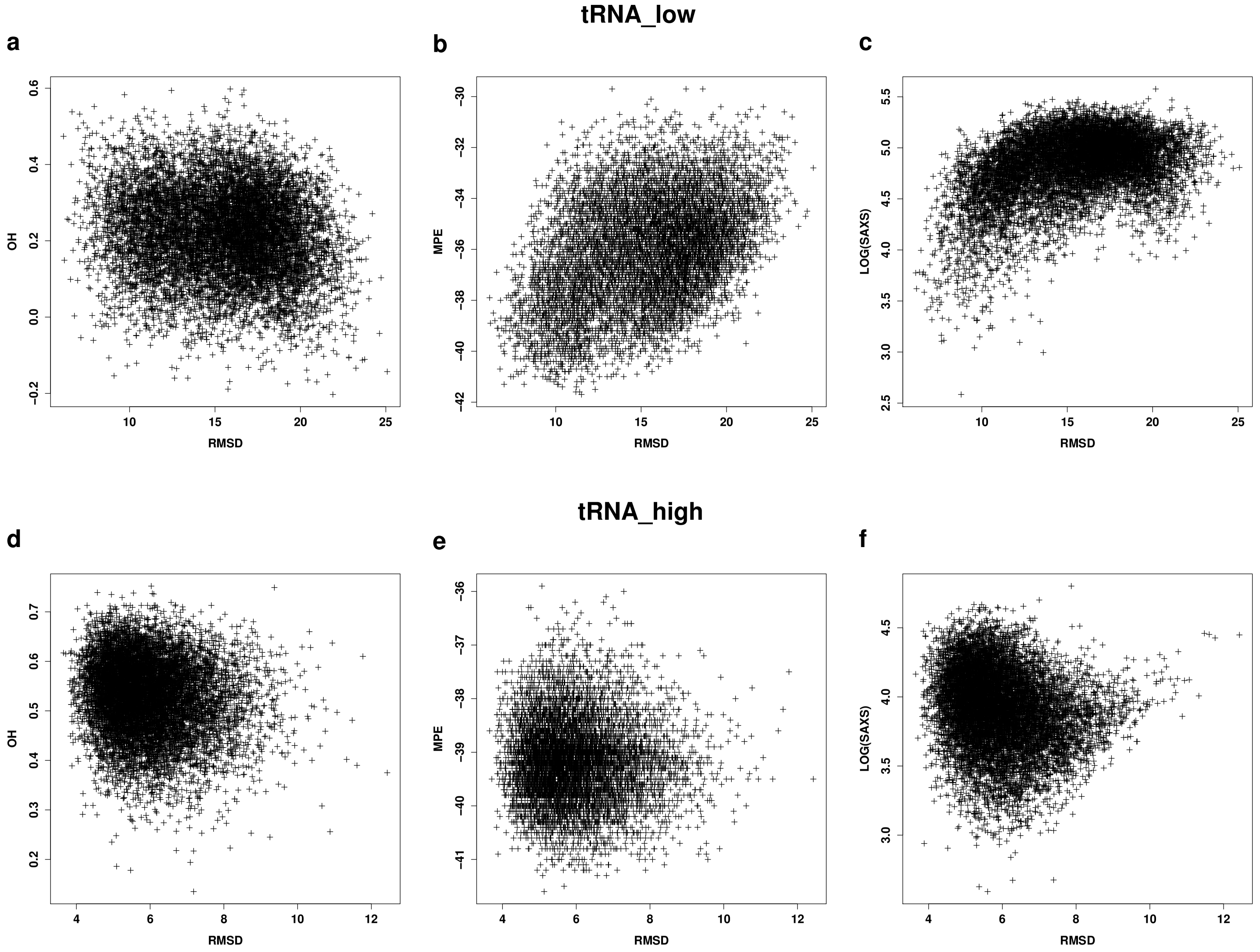


Figure 8.2: Performance of experimental data on tRNA. (Top) Low set. (Bottom) High set. ad) Hydroxyl radical footprinting (OH); high values better. be) Methidiumpropyl-EDTA (MPE); low values better. cf) Small-angle X-ray scattering (SAXS); low values better.

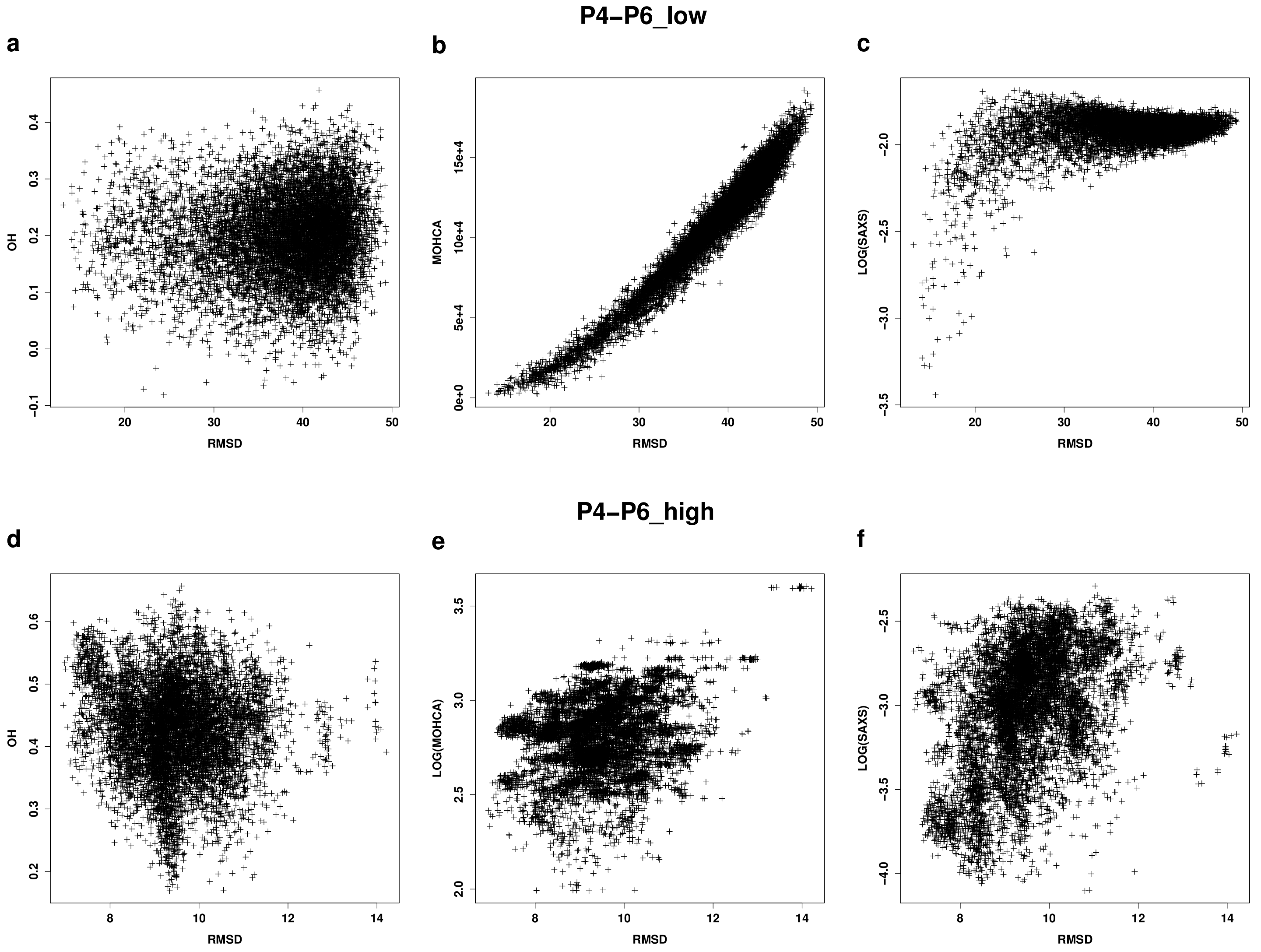


Figure 8.3: Performance of experimental data on P4-P6. (Top) Low set. (Bottom) High set. ad) Hydroxyl radical footprinting (OH); high values better. be) Multiplexed hydroxyl radical cleavage analysis (MOHCA); low values better. cf) Small-angle X-ray scattering (SAXS); low values better.

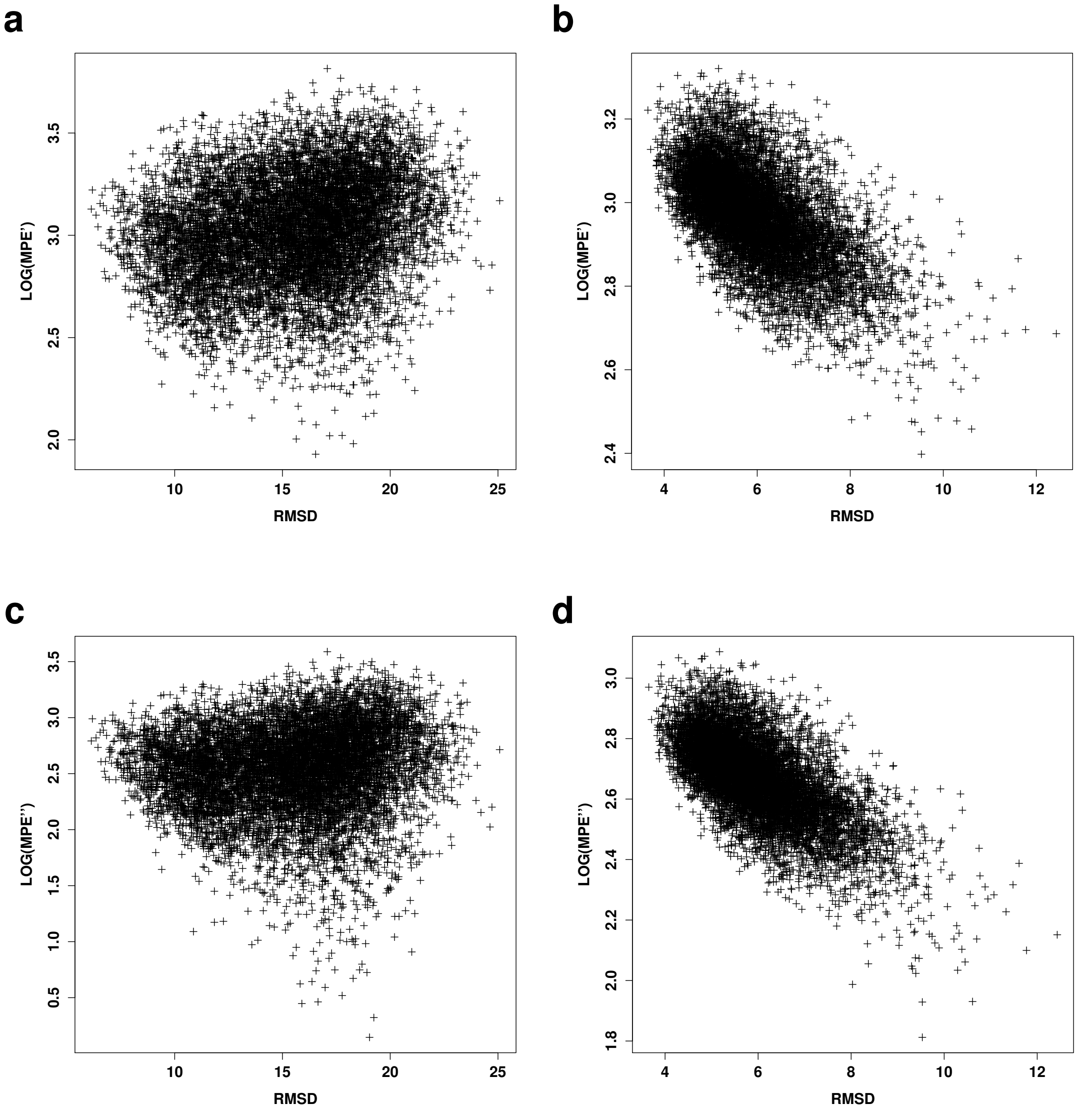


Figure 8.4: Reinterpretation of MPE data for tRNA. Distances that are: a) beyond 30 Å; and, b) beyond 35 Å. A quadratic penalty function is applied to: c) low set; and, d) high set.

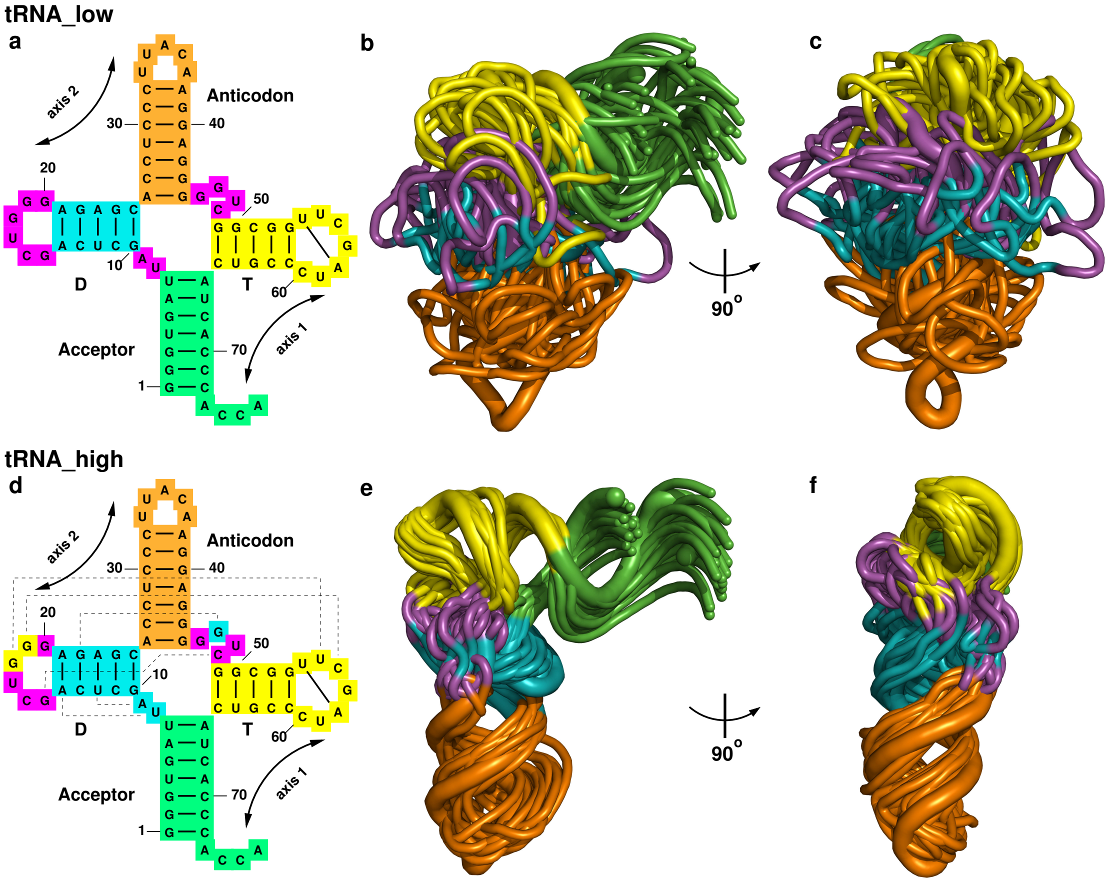


Figure 8.5: tRNA conformational sampling. (Top) Low set. a) Secondary structure coloured by stems: Acceptor (green); D (cyan); Anticodon (orange); and, T (yellow). All other nucleotides are in magenta. bc) Two views of twenty centroid centres (thin tubes) that are optimally aligned on the solution structure (PDB file 2K4C, thick tube). The colours are the same as in a. (Bottom) High set. d) The secondary structure and three base triples (8-14-21, 9-12-23, and 13-22-46) and three long-range base pairs (15-48, 18-55 and 19-56) shown using dashed lines. ef) Two views of twenty centroid centres (thin tubes) that are optimally aligned on the solution structure (PDB file 2K4C, thick tube).

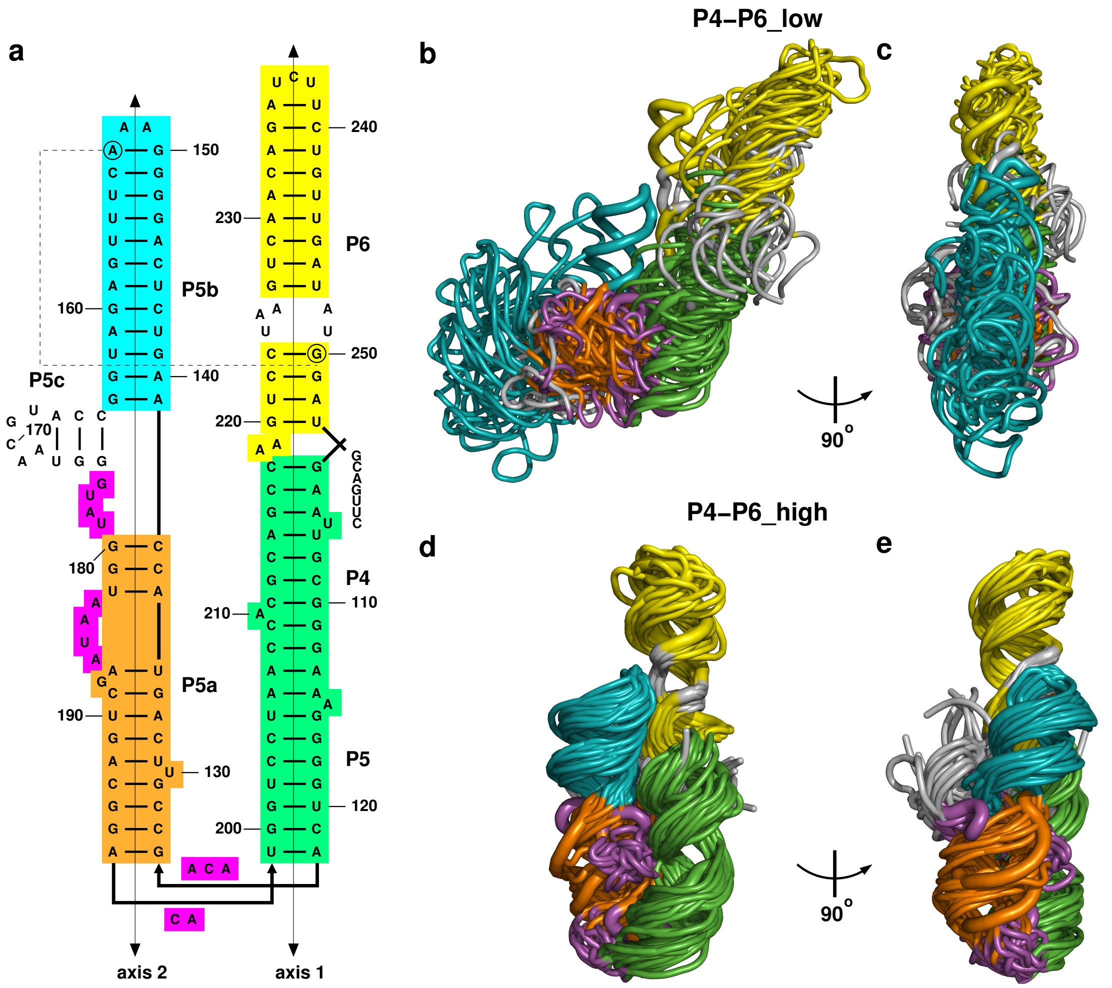


Figure 8.6: P4-P6 conformational sampling. (Top) Low set. a) Secondary structure coloured by stems: P4-P5 (green); P5a (orange); P5b (cyan); and, P6 (yellow). The other nucleotides are in magenta. The nucleotides that were not sampled are uncoloured. bc) Two views of twenty centroid centres (thin tubes) that are optimally aligned on the crystal structure (PDB file 1GID, thick tube). The colours are the same as in a. (Bottom) High set. d) The A-minor long-range base pair (A153-G250) is added in the constraints (circled nucleotides and dashed line). ef) Two views of twenty centroid centres (thin tubes) that are optimally aligned on the crystal structure (PDB file 1GID, thick tube).

CHAPTER 9

MODELING RNA

In this chapter we will discuss in details the modeling of two RNA molecules. It will highlight the use of the MC-Fold and MC-Sym pipeline. After all, it's pointless to develop tools if we don't use them afterward. Corollary to this exercise is the inherent difficulty of modeling RNA, despite its apparent low-complexity alphabet {A,C,G,U} and base pairing rules {A=U,C=G,G=U}.

As best resumed by late Dr. Cedergren and Dr. Major, modeling, "the process of combining, interpreting, and interpolating structural data, is the framework in which such an expression and synthesis can be made. Models serve to summarize and condense massive amounts of data into a visual, comprehensive format" [284]. The world leading expert on RNA modeling is Dr. Eric Westhof. His many modeling attempts are rich in hints on how to model RNA (and automate it). For recent RNA models made in our research lab we invite the reader to look at references [61] and [62].

9.1 Apical RNA transport

A notable feature of protein synthesis is the *in situ* sub-cellular localization of gene products, which explains why virtually all cells are polarized. The general belief is that proteins are transported once translated. However, messenger RNAs (mRNA) have also been found to be localized [285]. In *Drosophila*, it is estimated that about 70% of genes are subcellularly localized *before* translation [286]. Still in *Drosophila*, a conserved element in the *wingless* mRNA has recently been characterized [287]: WLE3. Here, we refine the sequence alignment and secondary structure, and propose three-dimensional models for a significant subset of known apically localized elements.

We first start the modeling process with an initial multiple-sequence alignment, shown in Figure 3 of [287]. The alignment provides sequences of similar lengths, suitable for the analysis by MC-Fold and MC-Cons. For each sequence, the top 100 sub-optimal structures were generated using MC-Fold. Then, using MC-Cons, a consensus structural assignment has been built. The sequences shown in Figure 3B of [287] served to build the consensus structure. Then, a total of nineteen sequences, including some from Figure 6 of [287], have been aligned on the consensus structure, to yield a 52-position multiple-sequence alignment shown in Figure 9.1. It is remarkable how complex is the alignment, already articulated in Dr. Krause's group [287], and its plasticity foretells an elaborate search scheme for linear scans of this element in whole genomes.

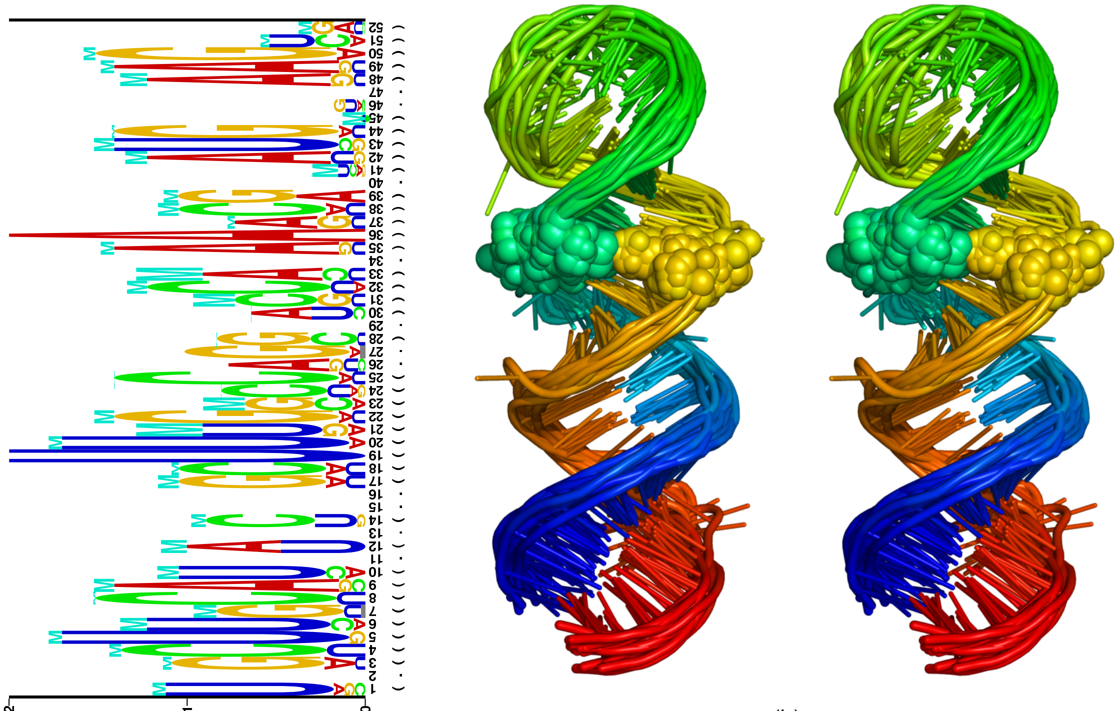
To further analyze the alignment, we used the RNA structure logo web server [288, 289] to produce Figure 9.2a. All nineteen sequences have been built in 3D using MC-Sym; Figure 9.2b. From these we can deduce certain features:

- The sequence is U-rich in 5' of the hairpin, while A-rich in 3'.
- Bulges don't seem to be tolerated between positions 3 to 10, and 17 to 25, inclusively. Bulges in the 3'-arm seem to be tolerated at only certain positions.
- Position 46 seem to provide for an unpaired nucleotide, but of any type.
- A perfectly conserved U=A base pair is found at positions 19=36, at a fixed distance from the hairpin's head.
- All hairpins seem to occupy the same volume.
- There seem to be no distinct electrostatic feature, except the disruption of the -1.5 kT/e isopotential surface at the location of bulged nucleotides, which could serve to lure the proteic complex (data not shown).

Hence, we have summarized in a 3D model all known apically localized sequences of the *wingless* 3'-UTR transcript. The model can now serve to propose new experiments to narrow down the specificity of binding to the localization proteic complex, and to unveil other genes which could potentially be transported apically. To be continued...

```
U-GCUUGCAU-A-CU-GCUUUGGCCAGG-ACCA-AAACG-UAUGCG-AAGUG D_melanogaster
1 23456789 0 12 345678901234 5678 90123 456789 01234
# #### ### # #  #########  # #### ##### ####   #####
(-((((((((-(-(.-(((((((((..)-))))-)))))-))))).-)))))
U-GCUUGCAC-A-CU-GCUUUGGCCAGG-ACCA-AAACG-UAUGCG-AAGUG D_teissieri
(-((((((((-(-(.-(((((((((..)-))))-)))))-))))).-)))))
1 23456789 0 12 345678901234 5678 90123 456789 01234
# #### ### # #  #########  # #### ##### ####   #####
U-GCUUGCAU-A-CU-GCUUUGGCCAGG-ACCA-AAACG-CAUGCG-AAGAA D_erecta
(-((((((((-(-(.-(((((((((..)-))))-)))))-))))).-)))))
1 23456789 0 12 345678901234 5678 90123 456789 01234
# #### ### # #  #########  # #### ##### ####   #####
U-GCUUGCAC-A-CU-GCUUUGGCCAGG-ACCA-AAACG-UGUGCG-AAGUG D_yakuba
(-((((((((-(-(.-(((((((((..)-))))-)))))-))))).-)))))
1 23456789 0 12 345678901234 5678 90123 456789 01234
# #### ### # #  #########  # #### ##### ####   #####
U-GCUUGCAU-A-CU-GCUUUGGUCAGG-ACCA-AAACG-UAUGUG-AAGUG D_takahashii
(-((((((((-(-(.-(((((((((..)-))))-)))))-))))).-)))))
1 23456789 0 12 345678901234 5678 90123 456789 01234
# #### ### # #  #########  # #### ##### ####   #####
U-GCUUGCAU-A-CU-GCUUUGGCCAGG-ACCA-AAACG-GAUGUG-AAGUG D_ficusphila
(-((((((((-(-(.-(((((((((..)-))))-)))))-))))).-)))))
1 23456789 0 12 345678901234 5678 90123 456789 01234
# #### ### # #  #########  # #### ##### ####   #####
U-GCUCGCAU-A-CU-GCUUUGGCCAGG-ACCA-AAACG-UAUGUG-GAGAU D_auraria
(-((((((((-(-(.-(((((((((..)-))))-)))))-))))).-)))))
1 23456789 0 12 345678901234 5678 90123 456789 01234
# #### ### # #  #########  # #### ##### ####   #####
U-GCUUGCAU-A-CU-GCUUUGGCCAGG-ACCA-AAACG-UAUGCA-AAGUG D_ananassae
(-((((((((-(-(.-(((((((((..)-))))-)))))-))))).-)))))
1 23456789 0 12 345678901234 5678 90123 456789 01234
# #### ### # #  #########  # #### ##### ####   #####
U-GCUUUCAUGU-U--GCUUUGGCCGGG-CUCA-AAGCA-GAUGGA-AAGCG D_pseudoobscura
(-((((((((.(-(--(((((((((..)-))))-)))))-))))).-)))))
1 2345678901 2  345678901234 5678 90123 456789 01234
# #### ### # #  #########  # #### ##### ####   #####
U-ACUUGCAU-U-UG-GCUUGGGCCUGG-ACCCUAAGCA-AAUGUC-AGGUU D_prosaltans
(-((((((((-(-(.-(((((((((..)-)))).)))))-))))).-)))))
1 23456789 0 12 345678901234 5678901234 567890 12345
# #### ### # #  #########  # #### ##### ####   #####
U-AUUUGCAU-U-UG-GCUUGGGCCUGGACCCC-AAGCA-AAUGUCCAAGUA D_willistoni
(-((((((((-(-(.-(((((((((..).))))-)))))-))))..))))))
1 23456789 0 12 34567890123456789 01234 567890123456
# #### ### # #  #########  # #### ##### ####   #####
U-GCUUGCAUUU-C--GCUUUGCCCAGC-UGCA-AAACG-CAUGUU-AAGCA D_robusta
(-((((((((.(-(--(((((((((..)-))))-)))))-))))).-)))))
1 2345678901 2  345678901234 5678 90123 456789 01234
# #### ### # #  #########  # #### ##### ####   #####
U-GCUUGCAU-UUCC-AAUUUGCCCAGC-UGCA-AAUCG-CAUGUU-AAGCA D_hydei
(-((((((((-(.(.-(((((((((..)-))))-)))))-))))).-)))))
1 23456789 0123 456789012345 6789 01234 567890 12345
# #### ### # #  #########  # #### ##### ####   #####
U-GCUUGCAU-U-CCCGCUUUGCCCAGC-UGCA-AAGUA-CAUGUU-AAGCA D_mojavensis
(-((((((((-(-(..(((((((((..)-))))-)))))-))))).-)))))
1 23456789 0 123456789012345 6789 01234 567890 12345
# #### ### # #  #########  # #### ##### ####   #####
U-GCUUGCAU-U-CC-GCUUUGCCCAGC-UGCA-AAACG-CAUGUU-AAGCA D_virilis
(-((((((((-(-(.-(((((((((..)-))))-)))))-))))).-)))))
1 23456789 0 12 345678901234 5678 90123 456789 01234
# #### ### # #  #########  # #### ##### ####   #####
U-GCUUGCAU-U-GC-GCUUUGCUCGGC-UGCA-AAGCA-AAUGUU-AAGCA Z_tuberculatus
(-((((((((-(-(.-(((((((((..)-))))-)))))-))))).-)))))
1 23456789 0 12 345678901234 5678 90123 456789 01234
# #### ### # #  #########  # #### ##### ####   #####
A-AUUU-CAA-U-U--UUUAAGAAAA-C-AUUU-UAAAA-AUUG-U-AAAUU TLS_orb
(-((((-(((-(-(--(((((((((.-)-))))-)))))-))))-.-)))))
1 2345 678 9 0  1234567890 1 2345 67890 1234 5 67890
# #### ### # #  #########  # #### ##### ####   #####
C-UUGAUUGU-A-U--UUUUAAAUUAAU-UCUU-AAAAACUACAAA-UUAAG TLS_K10
(-((((((((-(-(--(((((((((..)-))))-))))).))))).-)))))
1 23456789 0 1  234567890123 4567 890123456789 01234
# #### ### # #  #########  # #### ##### ####   #####
GUGCUC-UCA-A-C--AAUUGUCGCC-G-UCACAGAUUG-UUGU-UCGAGCC GLS_grk
(.((((-(((-(-(--(((((((((.-)-)))).)))))-))))-..)))))
123456 789 0 1  2345678901 2 3456 78901 2345 6789012
# #### ### # #  #########  # #### ##### ####   #####
|   |    |    |    |    |    |    |    |    |    |
1   5   10   15   20   25   30   35   40   45   50
```

Figure 9.1: Multiple-sequence alignment for the WLE3 element. A total of nineteen sequences are aligned. Each sequence is represented by a block of four rows. The first row features the sequence and its name tag. Insertions are marked using the '-' symbol. The second row keeps count of the sequential nucleotide number, starting from position one. The third row marks with '#' the positions which are base paired. A total of nineteen base pairs are highlighted and serves as anchor points to make the alignment. Finally, The fourth row shows the secondary structure adopted. Here also, insertions are marked using the '-' symbol.



(a)

(b)

Figure 9.2: RNA structure for the WLE3 element. **(a)** Structure logo. For each position in the alignment, the height of a column is proportional to the frequency of appearance of the symbol in that column. The height of M symbol is proportional to the mutual information contained in a co-varying base pairing scenario. **(b)** 3D structures. Cross-eye stereo view for the nineteen sequence studied. The nucleotides are colored from blue 5' to red 3'. The perfectly conserved U=A base pairs are rendered as atomic spheres.

9.2 Cap-independent translation element

Initiation of translation of mRNA into proteins relies on a sequence motif situated in the 5'-end of the messenger which recruits the ribosomal complex [290, 291]. However, certain viruses lack this cap motif, and lean instead on a cap-independent translation element (CITE) [292]. Recently, a first CITE has been identified in a pea enation mosaic virus RNA 2: PTE [293]. The structure of this CITE has been investigated using enzymatic (RNase 1) and chemical (SHAPE [224]) probing. A 3D modeling could now be sketched.

Starting from two sequences, PTE-PMV and PTE-PEMV, and the SHAPE data, a consensus structural assignment has been found using MC-Fold and MC-Cons, shown in Figure 9.3a. The SHAPE data provides clear evidence for the hairpin heads of SL1 and SL2. The pseudoknot has been put in evidence by sequence mutation; mutations in the region between H2 and H3 destabilized the structure, and rendered the nucleotides between SL1 and SL2 prone to SHAPE attacks (i.e. more flexible thus unpaired), indicating a connection between the two sequence parts.

To test the pseudoknot hypothesis we used a coarse-grained representation of RNA as in NAST [48]. Although extremely rare, parallel-running strands can form base pairs. Hence, the pseudoknot has been tested in the parallel and the anti-parallel cases. It turned out that the parallel case produced structures with real knots that cannot be unfolded, thus cannot be folded. Because NAST operates at a coarse-grained level, RNA single strands can go through one another, allowing folds featuring real knots (imagine pulling on both the 5' and 3' ends of the RNA once it is folded; can it be unfolded to a straigth line without the need to cross single-stranded regions?). The anti-parallel case did produce credible structures, at the least were unfoldable (by our previous imaginary thought experience). It also suggested a coaxial arrangement of H2, H3 and SL1 into an axis, and SL2 and PK into another axis.

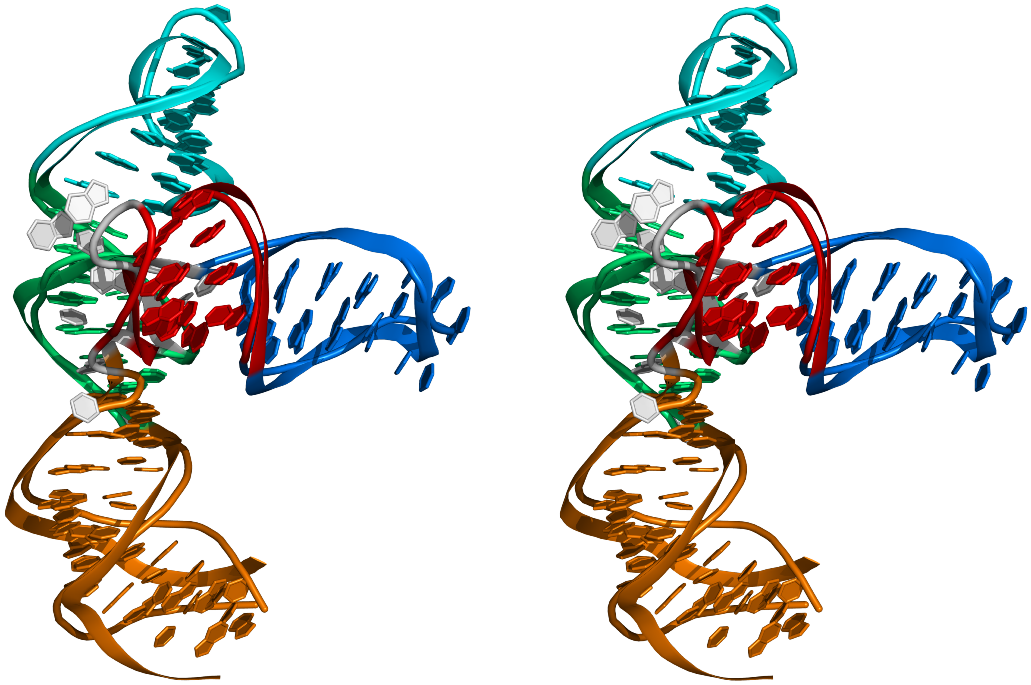
All-atoms 3D Modeling by MC-Sym proceeded with a divide-and-conquer approach: each of the five stems, H2, H3, SL1, SL2 and the PK, was generated independently. Next, we assembled the two main axis, by coaxially stacking the stems. Finally, complete models were assembled by bringing together the two main axes, and adding the few nucleotides left.

The PTE-PMV models exhibit a more constrained conformational space (angle between the two main axes) because of the few nucleotides flanking its pseudoknot. Therefore, all 3D models of PTE-PEMV have been aligned to all those of PTE-PMV, to identify the activated and common tertiary fold. The closest RMSD found between the two sequences is at 2.6 Å, and the optimal superposition of PMV and PEMV is shown if Figure 9.3b.

Additional structure probing could be performed to confirm the fold. In particular, hydroxyl radical footprinting [57] could be done to verify the protection of specific nucleotides in the pseudoknot core [278]. Furthermore, the footprint on the RNA following the binding of the eIF4E subunit of eIF4F protein could be unveiled [294]. Then the specific sequence or molecular requirements could be tested to custom-tailor a descriptor (see, for instance [232]), which can, in turn, be used to scan virus genomes for other such CITE elements.

```
>PEMV
AACACGUGGGAUAGGGGAUGACCUUGUCGACCGGUUAUCGGUCCCCUGCUCCUUCGAGCUGGCAAGGCGCUCACAGGUU
(((((((((((.[[[[.....(((((((((((((...))))))]]]](((((..))))).))))))).)))))))))))
               *                 ** **              * *
>PMV
CAGCAGAGACACGGGA--CG-CCACACCACCUUUG-CAGAGGUGCCCUUGGGAAACCAAUGGUGUGG-GGUCUCA-CUG
(((.(((((((.[[[[--..-(((((((((((((.-.))))))]]]](((((..))))).)))))))-)))))))-)))
|---H2----| |PK|     |-H3--||-----SL1-----||PK||---SL2----| |--H3-| |----H2---|
```

(a)



(b)

Figure 9.3: RNA structure for the PTE element. **(a)** Sequence and secondary structure for PTE-PMV and PTE-PEMV. Positions highly reactive to SHAPE are marked with a star. The pseudoknot base pairs are tagged with brackets []. Extent of helical regions are also delineated; H2, H3, SL1, SL2 and PK. **(b)** Stereo view of the optimal superposition of PTE-PEMV (tube) on PTE-PMV (ribbon). Stems are colored like; orange:H2, green:H3, cyan:SL1, marine:SL2 and red:PK. All-atoms RMSD is 2.6 Å.

CHAPTER 10

CONCLUSION

As the title of this thesis suggests, we have devised a new paradigm for the folding of ribonucleic acids. We will here summarize what has been done in the context of my Ph.D. studies, and possible research avenues to explore in the future.

10.1 On successes...

Recall that the principal goal of this thesis is to predict the structure of an RNA given its sequence. This goal is more known as the bio-polymer folding problem. An exact solution to this problem is within scope since 1) Anfinsen and colleagues demonstrated that the fold of a protein (hence, a bio-polymer) is encoded in its sequence, and that the adopted fold is the one of minimum free energy [34]; 2) RNA is believed to fold hierarchically [35], and its 3D structure is glass-like (low entropy) [256]; 3) the database of solved structures is ever-growing, providing more input data to new and powerful data-mining methods, which, altogether, has the potential to extract the folding rules from this database (shown herein); and, 4) Advances in biophysical studies of RNA refine our folding models (for instance, see [85, 93, 148]).

Because atomic-resolution structure determination is extremely expensive, we ask how precise can we be in predicting structures using computers only. The first paper published in this thesis addresses this question. The most prominent approach to RNA structure prediction is a physico-chemical formulation best known as the INN-HB (Individual Nearest Neighbor - Hydrogen Bond) model [172], which makes into play thermodynamics parameters for stacks of base pairs. This method predicts RNA secondary structure from sequence, mainly by using the dynamic programming approach. These parameters have been collected for more than 30 years now, and still have not pinned down the RNA folding problem. Furthermore, sketching an RNA tertiary structure from a secondary structure is a hefty task, considering the many degrees of freedom an unpaired nucleotide has.

Solution structure of the ribosomal complex by X-ray in early 2000 [159, 160] has opened a new era in RNA structure prediction. Indeed, an analysis of the structural content of the RNA portion of the ribosome shows that RNA is made of small and recurring blocks [156], often part of more elaborate motifs [121, 123, 295]. By using a small subset of these blocks, which we called Nucleotide Cyclic Motifs (NCM, Figure 3.2), we asked if this could be a viable alternative to the INN-HB model, by providing not only a hint on their relative stability for secondary structure prediction (already established here [205]), but also actual 3D instances for RNA tertiary structure prediction.

We took the NCM as a first order object to: 1) provide statistics on their relative thermodynamics stability given their sequences; 2) devise a model which predicts RNA secondary structure from sequence based on these statistics, MC-Fold; and, 3) use actual NCM instances from the PDB in MC-Sym. The welding of two consecutive NCMs at their common base pair limits the possibilities in base pair types and solves the spanning tree problem from earlier MC-Sym versions, in which some inter-nucleotide relations had to be left aside (see Dr. Lemieux's Ph.D. thesis [161]). A welcomed feature of NCMs is that they capture all base pair types, and not only the canonical ones (that are similar to those found in the DNA double-helix), and they provide a unified framework for RNA secondary and tertiary structure prediction (NCM fusion 1 and 2, Figure 5.1).

The results of our work have been published in *Nature* (reference [45] and chapter 4). The mathematical model and algorithms used are described in details in the accompanying text (chapter 5). We have shown that the new paradigm is not worse than the thermodynamics approach (Table 4.1 and 5.1) on hairpins extracted from the PDB, and allows for atomic precision RNA 3D models (Table 4.2 and Figure 4.1). Using the NCM paradigm, we also made 3D structure prediction for microRNA hairpins, that are processed by the Dicer protein complex [223], a step in RNA interference. We explain how the RNA double-helix can still be processed even though it contains non-canonical base pairs, bulges (insertions) and deletions (see Figure 4.2). We also proposed an alternative structure for the HIV-1 frame-shifting element (Figure 4.3). Although the whole

HIV-1 genome has been further characterized [296, 297], it is not clear from these recent studies if our model has been confirmed or invalidated.

To further extend the paradigm, we needed a new distance measure to quantify the success (or failure) of our (and future) RNA modeling attempts. Already, the root-mean-square-deviation (RMSD) was showing weaknesses in RNA structure comparison [56]. However, RMSD is the standard in structure comparison, so any new structure comparison method should have an equivalent distance measure. We took advantage of the fact that RNA side chains, the nucleobases, associate in specific manners through hydrogen bonding. This allows us to move away from the atomic positions to a simpler symbolic nomenclature, which describes base pairs through their interacting edges (the Leontis-Westhof (LW) nomenclature [67, 69]). Furthermore, nucleobases stack onto one another, also captured into a symbolic annotation [251, 298].

The base pairings and stackings found in a 3D structure define its Interaction Network (IN). By comparing the interaction annotations between a reference structure and a model, we can now measure how well the model reproduces the base pair and stack interactions, the Interaction Network Fidelity (INF), regardless of the minute atomic details inside the 3D structures. We then defined the Deformation Index (DI) as the RMSD divided by the INF. Hence, on one hand, if a 3D model cannot reproduce the interaction network of the reference structure (INF ≈ 0), then its RMSD to the reference structure is meaningless, and the modeling attempt should be qualified as a failure. On the other hand, the DI shifts attention to RMSD when the model reproduces the IN (INF ≈ 1). As a matter of fact, we demonstrated that there is no correlation between RMSD and INF, that is, we can build good RMSD models with bad INF, and vice-versa. We favor the INF over the RMSD, since an RNA molecule will most likely conserve its interaction network over its global shape (measured by the RMSD); a small angle change between the two arms in the tRNA fold has dramatic influence on the RMSD, even though the IN remains essentially the same. To spot the cause of a deviation between a model and a reference structure we introduced a Deformation Profile (DP), which highlights the regions that are poorly modeled. Taken together, these two new distance metrics should challenge the RNA modeling community to perfect RNA structure prediction. These results have been published in RNA (reference [63] and chapter 6).

Many low-resolution RNA structure probing methods have recently been developed, as an alternative to more costly but more precise techniques as NMR and X-ray. Our paradigm would be incomplete if we left aside these methods. Hence, we have implemented and evaluated the predictive power of the data issued from such methods. The results are not published at the time of writing, but are under peer review. It is the subject of chapter 8. We found that small-angle X-ray scattering (SAXS) offers the best structure discrimination power over all RNA sizes utilized. Hydroxyl radical footprints are even more discriminative of the native fold, but require RNA models that are very precise. EDTA-based approaches, that attach a probe in a particular spot on the 3D structure and then observe the consequent backbone cleavage pattern within the probe's reach and backbone accessibility, are also quite revealing of the fold, as long as the probe's length is smaller than the radius of gyration of the molecule under study (or else the probe will have access to a significant fraction of the nucleotides composing the RNA).

That said, we can conclude that we have made significant steps to:

1. better predict the RNA secondary and tertiary structures, by using the NCM paradigm, which naturally embraces all base pairings.
2. better compare RNA 3D structures, by introducing the Deformation Index and the Deformation Profile.
3. better determine RNA 3D structure, using data from the latest low-resolution experimental methods.
4. provide the RNA modeling community with a web site that implements all our concepts and findings.

10.2 On failures...

I have spent a significant portion of my time ($\approx$ 9 months) trying to find the Holly Grail of molecular biology, that is, a function f that takes as parameter an RNA atomic structure $\vec{R}$ and returns a score, say $E = f(\vec{R})$, with which one could sort many 3D models and obtain the native fold from the model that shows the best score. I have not succeeded yet to obtain such an f. Despite this, the history of my attempts is interesting (as this medium allows me to discuss about $\sim$negative$\sim$ results).

We first tackled this problem by computing a deformation matrix from which the coefficients represent "spring constants" [299], their magnitude indicating how easy or stiff it is to deform a base pair and a base pair stack in an RNA structure. The base pair and base pair stack parameters were those of Olson's group [300, 301]. Then, the energy of an RNA 3D structure is simply the sum of all its deformed base pairs and base pair steps. This approach did not work very well since it could not detect the proper base pairing types among many RNA decoys. Furthermore, base pairs and base pair steps may have multiple energy minima that our model would not be able to capture (because it is just a simple spring model with one rest position per parameter). This is a dead-end.

We then turned our attention to classical mechanical force-fields (FF), and in particular to the Amber '99 FF [302], because of its general acceptance for nucleic acid modeling (most tweaked for nucleic acids; [303, 304]). We already knew that it failed to identify the native fold in a set of decoys for a few RNA sequences. Since the two main non-bonded interactions in an RNA molecule are base stacking and pairing, we wondered if, by adding extra terms to the FF, it could now properly identify the native fold. The idea here is to decompose f in more basic and orthogonal weighted w_i phenomena f_i, like:

$$f(\vec{R}) = w_1 \times f_1(\vec{R}) + w_2 \times f_2(\vec{R}) + \cdots$$

We took the non-bonded terms of Amber '99 to describe the Coulomb and the van der Waals force-fields. For stacking, we added the π-π interaction, following the studies of Dr. Hunter's group in DNA [169, 170]. This interaction is best described as the influence of π-electron clouds above and under cyclic rings, such as nucleobases in nucleic acids [305]. Although this model has been invalidated for RNA [306], we still believe in the effects of π-π interactions on the structure, because: 1) the invalidation has been carried out on cytosine dimers only (not exhaustive); and, 2) these have been made in the context of "old" quantum-mechanical methodologies and parameters (Dr. Hunter's π-π parameters date back to 1971 [307]), and therefore should definitively be revisited. Stacking interaction is still subject to many research studies (see [308] for a recent review).

For pairing, mostly mediated through hydrogen bonds (Hbond), given the apparent directionality preferences between donor and acceptor groups (see, for instance, in the sibling protein prediction field [309]), we used an anisotropic formulation to score them, as prescribed here [310].

To see if we gained any discriminative power we used MC-Sym to generate ten thousand 3D models of a simple RNA duplex with four consecutive non-canonical base pairs. This duplex has been solved by NMR, but the structure is not disclosed in the PDB yet (hence, at the time of writing we cannot provide the modeled RNA sequence). The generated models have been partitioned by the base pair suite they feature; the goal here is to predict the correct base pair suite adopted by the mismatches. If we look at the score distributions for each suite of base pairs, the f function should, in theory, identify the native suite of base pairs. Figure 10.1 shows the distribution of energies or scores per suite. We compare side-by-side the Amber '99 FF, the Amber '99 FF with a Generalized-Born Solvent-Accessibility (GBSA) implicit solvation model [311], and our own f-score. We will let the reader draw its own conclusions from our results, given that the base pair suite adopted by the duplex in solution is #3.

10.3 On the future...

Our failure to devise an f function made us wonder if the numeric domain, i.e. the xyz position of each atom, is an appropriate vehicle to capture the various factors (electrostatics, van der Waals, etc) at play in RNA folding. This is, however, the ultimate goal, but the intensive quest for a physico-chemical formulation has given mixed results so far. The discrete states by which RNA nucleobases associate now fall into the symbolic domain, which is appropriate for machine-learning techniques. We are now investigating the predictive power of decision trees, such as ID3 [312] or C4.5 [313], to infer the base pair type of a base pair, given its nucleotide neighbors in an RNA duplex. The advantage of a decision tree over f is that it allows to predict the base pair type based on a few symbols only (like the type of nucleotides ACGU), rather than necessitating a 3D $\vec{R}$, a few hundred of atomic positions. Preliminary results are encouraging, but still need further investigations.

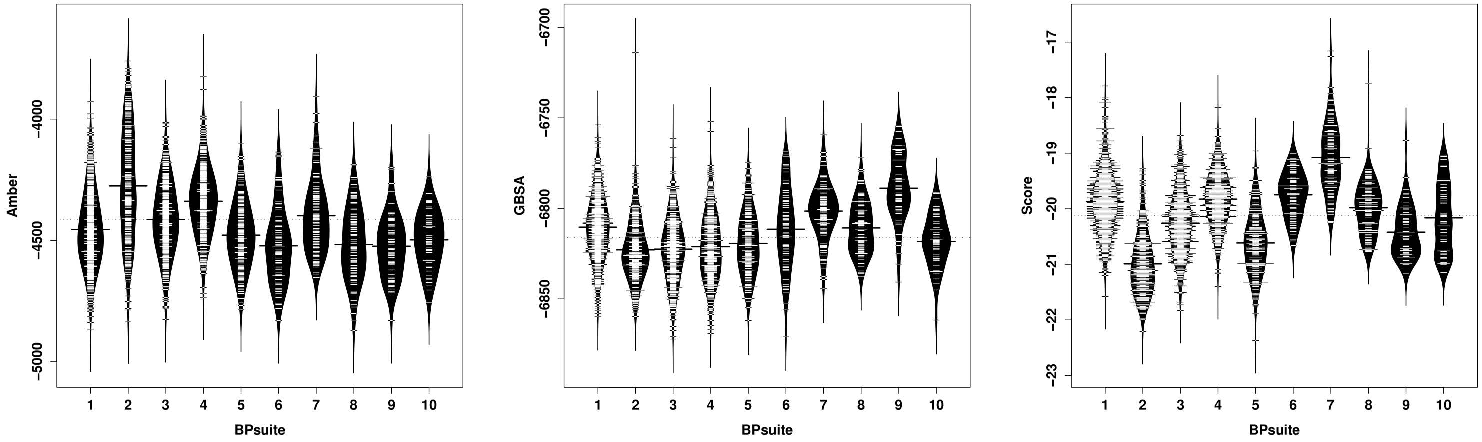
Failure to predict larger RNA secondary structures also raises eyebrows, following the tremendous effort to quantify the folding free energy of smaller RNA duplexes. We already discussed some possible explanations for this, like base triples, ionic cloud, kissing loops, etc, but one avenue left unexplored are the very basic principles of RNA folding; events that happen in a timely and locally fashion. The use of the dynamic programming algorithm, with its global optimization result, implies that the RNA chain, while collapsing on itself, has a certain sense of “omnipotence”. This is not in-line with recent results on RNA folding, where pausing during transcription may have its word to say on the final fold [102]. Furthermore, the complex RNA folding landscape combined to the 5’-to-3’ RNA synthesis can lead to viable intermediates [314]. And, the RNA chain can form early and local motifs, such as lonepair triloops [95].

One very interesting point of view is to look at the affinity of a small stretch of an RNA sequence against the rest of the sequence. The affinity can be best measured by MC-Fold’s force-field, which takes into account all base pair types (canonical as well as non-canonical). The affinity or density plot is computed in this fashion: consider all stems an RNA sequence can make. For each stem consider the energy density instead of the total energy. The density is obtained by dividing the total energy by the stem's length. Then, plot this energy density for all nucleotides in the stems. For the asparagine version of tRNA, this is shown in Figure 10.2. We can see that 1. the D- and T-loop hairpins are clearly identified (nucleotides 16 and 57 respectively) as the sequences near these hairpins find no attractive/complementary sub-sequences elsewhere in the sequence, leading to low energy densities, and 2. the first collapse of the most energy dense region would yield the T-stem, in accord with tRNA melting experiences [315], if we admit that, more less, unfolding is folding with the arrow of time flowing in the opposite direction (the T-stem/loop persists at high temperatures compared to the other stems in the tRNA). These two features of the density graph, that is, the absence and/or the best affinities, could be exploited to better RNA secondary structure predictions. The interest in this view comes from the fact that the secondary structure prediction, by MC-Fold and other thermodynamics-based approaches, for this particular RNA sequence is a three-way junction, and not the famous cloverleaf four-way junction! One could devise a folding algorithm in which the densest stem first form, then the density plot is recomputed (because pieces of the sequences are now no more available), and process applied recursively. This algorithm is local, timely and greedy. More research must be done to verify if this folding principle would apply to other RNA sequences. Variants with saddle points could also be envisaged (similar to [316] or [317]).

Allow me to finish on this most humorous note;

-“because, I'm Billy Mitchell”

Billy Mitchell, explaining why his Pac-Man can go through a monster without dying. He was the first to make a perfect Pac-Man gameplay, hence to obtain the highest possible score of 3,333,360.
ref: http://www.youtube.com/watch?v=Ok6kiLK0Idc



(a) (b) (c)

Figure 10.1: Various internal energy scoring schemes against base pair suites. The curved envelope highlights the density of the data, marked with horizontal ticks. The mean is shown using a thick line. **(a)** The Amber 99 force-field. The dashed horizontal line represents the mean of all data. **(b)** The Amber 99 force-field combined with a Generalized Born - Solvent Accessibility model. **(c)** Our own scoring function (unpublished).

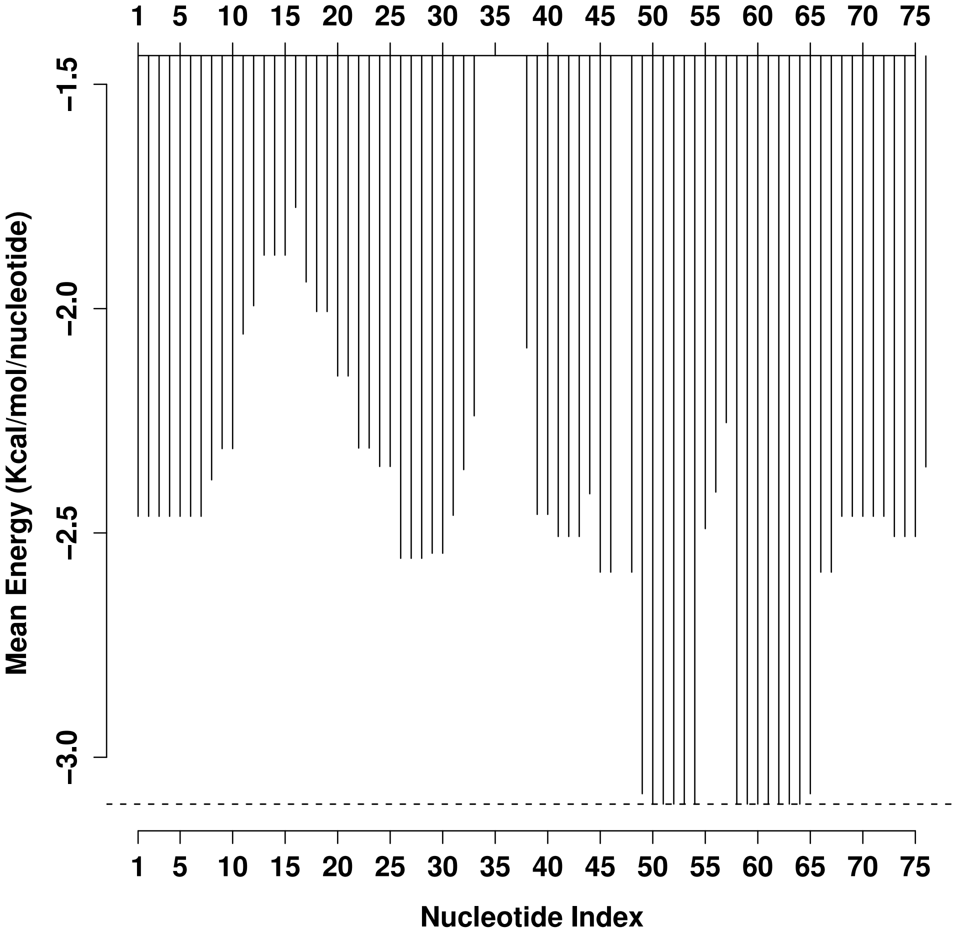


Figure 10.2: Mean energy density for tRNA-ASP. For all possible stems given the sequence, the best energy is shown for all nucleotides participating in that stem. The energy is converted into mean energy by dividing the total energy of the best stem by its length. Nucleotides that have zero mean energy are modified nucleotides known to remain unpaired. In tRNA-ASP, nucleotide 47 is absent to conform to the tRNA nucleotide numbering convention. The dotted horizontal line indicates the nucleotides involved in the lowest energy density stem.

BIBLIOGRAPHY

[1] W Gilbert. Origin of life: The RNA world. *Nature*, 319:618, 1986.

[2] RF Gesteland, TR Cech, and JF Atkins, editors. *The RNA World, 3rd Edn*. CSHL, Cold Spring Harbor, New York, NY, 2006.

[3] GF Joyce. The antiquity of RNA-based evolution. *Nature*, 418:214–221, 2002.

[4] K Kruger, PJ Grabowski, AJ Zaug, J Sands, DE Gottschling, and TR Cech. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell*, 31:147–157, 1982.

[5] C Guerrier-Takada, K Gardiner, T Marsh, N Pace, and S Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35:849–857, 1983.

[6] TJ Wilson and DM Lilley. Biochemistry. The evolution of ribozyme chemistry. *Science*, 323:1436–1438, 2009.

[7] YI Wolf and EV Koonin. On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. *Biol Direct*, 2:14, 2007.

[8] K Bokov and SV Steinberg. A hierarchical model for evolution of 23S ribosomal RNA. *Nature*, 457:977–980, 2009.

[9] C Briones, M Stich, and SC Manrubia. The dawn of the RNA world: toward functional complexity through ligation of random RNA oligomers. *RNA*, 15:743–749, 2009.

[10] TA Lincoln and GF Joyce. Self-sustained replication of an RNA enzyme. *Science*, 323:1229–1232, 2009.

[11] EA Schultes, A Spasic, U Mohanty, and DP Bartel. Compact and ordered collapse of randomly generated rna sequences. *Nat Struct Mol Biol*, 12:1130–1136, 2005.

[12] FH Crick. On protein synthesis. *Symp Soc Exp Biol*, 12:138–163, 1958.

[13] F Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970.

[14] M Selmer, CM Dunham, FV Murphy 4th, A Weixlbaumer, S Petry, AC Kelley, JR Weir, and V Ramakrishnan. Structure of the 70S ribosome complexed with mRNA and tRNA. *Science*, 313:1935–1942, 2006.

[15] A Korostelev, S Trakhanov, M Laurberg, and HF Noller. Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell*, 126:1065–1077, 2006.

[16] G Yusupova, L Jenner, B Rees, D Moras, and M Yusupov. Structural basis for messenger RNA movement on the ribosome. *Nature*, 444:391–394, 2006.

[17] A Fire, S Xu, MK Montgomery, SA Kostas, SE Driver, and CC Mello. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391:806–811, 1998.

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

[19] NC Lau, LP Lim, EG Weinstein, and DP Bartel. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, 294:858–862, 2001.

[20] RC Lee and V Ambros. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, 294:862–864, 2001.

[21] DP Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136:215–233, 2009.

[22] GA Soukup and RR Breaker. Nucleic acid molecular switches. *Trends Biotechnol*, 17:469–476, 1999.

[23] A Roth and RR Breaker. The structural and functional diversity of metabolite-binding riboswitches. *Annu Rev Biochem*, 78:305–334, 2009.

[24] SR Eddy. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet*, 2:919–929, 2001.

[25] JS Mattick. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep*, 2:986–991, 2001.

[26] G Storz. An expanding universe of noncoding RNAs. *Science*, 296:1260–1263, 2002.

[27] A Huttenhofer, P Schattner, and N Polacek. Non-coding RNAs: hope or hype? *Trends Genet*, 21:289–297, 2005.

[28] S Griffiths-Jones, A Bateman, M Marshall, A Khanna, and SR Eddy. Rfam: an RNA family database. *Nucleic Acids Res*, 31:439–441, 2003.

[29] S Griffiths-Jones, S Moxon, M Marshall, A Khanna, SR Eddy, and A Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, 33:D121–D124, 2005.

[30] PP Gardner, J Daub, JG Tate, EP Nawrocki, DL Kolbe, S Lindgreen, AC Wilkinson, RD Finn, S Griffiths-Jones, SR Eddy, and A Bateman. Rfam: updates to the RNA families database. *Nucleic Acids Res*, 37:D136–D140, 2009.

[31] KC Pang, S Stephen, ME Dinger, PG Engstrom, B Lenhard, and JS Mattick. RNAdb 2.0–an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res*, 35:D178–D182, 2007.

[32] V Bourdeau, G Ferbeyre, M Pageau, B Paquin, and R Cedergren. The distribution of RNA motifs in natural sequences. *Nucleic Acids Res*, 27:4457–4467, 1999.

[33] SA Tenenbaum, PJ Lager, CC Carson, and JD Keene. Ribonomics: identifying mRNA subsets in mRNP complexes using antibodies to RNA-binding proteins and genomic arrays. *Methods*, 26:191–198, 2002.

[34] CB Anfinsen, E Haber, M Sela, and FH White Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci (USA)*, 47:1309–1314, 1961.

[35] I Tinoco Jr and C Bustamante. How RNA folds. *J Mol Biol*, 293:271–281, 1999.

[36] BH Mooers. Crystallographic studies of dna and rna. *Methods*, 47:168–176, 2009.

[37] PJ Lukavsky and JD Puglisi. RNAPack: An integrated NMR approach to RNA structure determination. *Methods*, 25:316–332, 2001.

[38] N Hall. Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol*, 210:1518–1525, 2007.

[39] HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28:235–242, 2000.

[40] SR Holbrook. RNA structure: the long and the short of it. *Curr Opin Struct Biol*, 15:302–308, 2005.

[41] C Ehresmann, F Baudin, M Mougel, P Romby, JP Ebel, and B Ehresmann. Probing the structure of RNAs in solution. *Nucleic Acids Res*, 15:9109–9128, 1987.

[42] E Kierzek, R Kierzek, DH Turner, and IE Catrina. Facilitating RNA structure prediction with microarrays. *Biochemistry*, 45:581–593, 2006.

[43] RA Tinsley and NG Walter. Pyrrolo-C as a fluorescent probe for monitoring RNA secondary structure formation. *RNA*, 12:522–529, 2006.

[44] R Das and D Baker. Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci (USA)*, 104:14664–14669, 2007.

[45] M Parisien and F Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452:51–55, 2008.

[46] HM Martinez, JV Maizel Jr, and BA Shapiro. RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA. *J Biomol Struct Dyn*, 25:669–684, 2008.

[47] F Ding, S Sharma, P Chalasani, VV Demidov, NE Broude, and NV Dokholyan. Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, 14:1164–1173, 2008.

[48] MA Jonikas, RJ Radmer, A Laederach, R Das, S Pearlman, D Herschlag, and RB Altman. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, 15:189–199, 2009.

[49] J Frellsen, I Moltke, M Thiim, KV Mardia, J Ferkinghoff-Borg, and T Hamelryck. A probabilistic model of RNA conformational space. *PLoS Comput Biol*, 5:e1000406, 2009.

[50] W Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Cryst*, A32:827–828,

1976.

[51] W Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Cryst*, A34:827–828, 1978.

[52] PY Chou and GD Fasman. Prediction of protein conformation. *Biochemistry*, 13:222–245, 1974.

[53] AG Street and SL Mayo. Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc Natl Acad Sci (USA)*, 96:9074–9076, 1999.

[54] J DeBartolo, A Colubri, AK Jha, JE Fitzgerald, KF Freed, and TR Sosnick. Mimicking the folding pathway to improve homology-free protein structure prediction. *Proc Natl Acad Sci (USA)*, 106:3734–3739, 2009.

[55] G Faure, A Bornot, and AG de Brevern. Protein contacts, inter-residue interactions and side-chain modelling. *Biochimie*, 90:6614–6618, 2008.

[56] P Gendron, S Lemieux, and F Major. Quantitative analysis of nucleic acid three-dimensional structures. *J Mol Biol*, 308:919–936, 2001.

[57] TD Tullius and JA Greenbaum. Mapping nucleic acid structure by hydroxyl radical cleavage. *Curr Opin Chem Biol*, 9:127–134, 2005.

[58] CM Gherghe, CW Leonard, F Ding, NV Dokholyan, and KM Weeks. Native-like RNA tertiary structures using a sequence-encoded cleavage agent and refinement by discrete molecular dynamics. *J Am Chem Soc*, 131:2541–2546, 2009.

[59] R Das, M Kudaravalli, M Jonikas, A Laederach, R Fong, JP Schwans, D Baker, JA Piccirilli, RB Altman, and D Herschlag. Structural inference of native and partially folded RNA by high-throughput contact mapping. *Proc Natl Acad Sci (USA)*, 105:4144–4149, 2008.

[60] J Lipfert and S Doniach. Small-angle X-ray scattering from RNA, proteins, and protein complexes. *Annu Rev Biophys Biomol Struct*, 36:307–327, 2007.

[61] M Enstero, C Daniel, H Wahlstedt, F Major, and M Ohman. Recognition and coupling of A-to-I edited sites are determined by the tertiary structure of the RNA. *Nucleic Acids Res*, 2009. Epub ahead of print.

[62] AP McGraw, A Mokdad, F Major, PC Bevilacqua, and P Babitzke. Molecular basis of TRAP-5'SL RNA interaction in the *Bacillus subtilis* trp operon transcription attenuation mechanism. *RNA*, 15:55–66, 2009.

[63] M Parisien, JA Cruz, E Westhof, and F Major. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, 15:1875–1885, 2009.

[64] W Saenger. *Principles of Nucleic Acid Structure*. Springer-Verlag, New-York, 1984.

[65] ML Huggins. 50 years of hydrogen bond theory. *Angew Chem Int Ed Engl*, 10:147–152, 1971.

[66] T Steiner. The hydrogen bond in the solid state. *Angew Chem Int Ed Engl*, 41:49–76, 2002.

[67] NB Leontis and E Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7:499–512, 2001.

[68] JC Lee and RR Gutell. Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs. *J Mol Biol*, 344:1225–1249, 2004.

[69] S Lemieux and F Major. RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. *Nucleic Acids Res*, 30:4250–4263, 2002.

[70] JD Watson and FH Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[71] U Nagaswamy, N Voss, Z Zhang, and GE Fox. Database of non-canonical base pairs found in known RNA structures. *Nucleic Acids Res*, 28:375–376, 2000.

[72] U Nagaswamy, M Larios-Sanz, J Hury, S Collins, Z Zhang, Q Zhao, and GE Fox. NCIR: a database of non-canonical interactions in known RNA structures. *Nucleic Acids Res*, 30:395–397, 2002.

[73] Y Xin and WK Olson. BPS: a database of RNA base-pair structures. *Nucleic Acids Res*, 37:D83–D88, 2009.

[74] NB Leontis, J Stombaugh, and E Westhof. The non-Watson-Crick base pairs and their associated

isostericity matrices. *Nucleic Acids Res*, 30:3497–3531, 2002.

[75] J Stombaugh, CL Zirbel, E Westhof, and NB Leontis. Frequency and isostericity of RNA base pairs. *Nucleic Acids Res*, 37:2294–2312, 2009.

[76] D Gautheret and RR Gutell. Inferring the conformation of RNA base pairs and triples from patterns of sequence variation. *Nucleic Acids Res*, 25:1559–1564, 1997.

[77] A Lescoute, NB Leontis, C Massire, and E Westhof. Recurrent structural RNA motifs, isostericity matrices and sequence alignments. *Nucleic Acids Res*, 33:2395–2409, 2005.

[78] A Mokdad and AD Frankel AD. ISFOLD: Structure prediction of base pairs in non-helical RNA motifs from isostericity signatures in their sequence alignments. *J Biomol Struct Dyn*, 25:467–472, 2008.

[79] A Pérez, A Noy, F Lankas, FJ Luque, and M Orozco. The relative flexibility of B-DNA and A-RNA duplexes: database analysis. *Nucleic Acids Res*, 32:6144–6151, 2004.

[80] NJ Reiter, LJ Maher 3rd, and SE Butcher. DNA mimicry by a high-affinity anti-NF-kappaB RNA aptamer. *Nucleic Acids Res*, 36:1227–1236, 2008.

[81] TR Sosnick and T Pan. RNA folding: models and perspectives. *Curr Opin Struct Biol*, 119:309–316, 2003.

[82] R Das, LW Kwok, IS Millett, Y Bai, TT Mills, J Jacob, GS Maskel, S Seifert, SG Mochrie, P Thiyagarajan, S Doniach, L Pollack, and D Herschlag. The fastest global events in RNA folding: electrostatic relaxation and tertiary collapse of the *Tetrahymena* ribozyme. *J Mol Biol*, 332:311–319, 2003.

[83] PC Bevilacqua and R Russell. Editorial overview: exploring the vast dynamic range of RNA dynamics. *Curr Opin Chem Biol*, 12:601–603, 2008.

[84] R Russell, R Das, H Suh, KJ Travers, A Laederach, MA Engelhardt, and D Herschlag. The paradoxical behavior of a highly structured misfolded intermediate in RNA folding. *J Mol Biol*, 363:531–544, 2006.

[85] TR Sosnick. Kinetic barriers and the role of topology in protein and RNA folding. *Protein Sci*, 17:1308–1318, 2008.

[86] C Hyeon, G Morrison, and D Thirumalai. Force-dependent hopping rates of RNA hairpins can be estimated from accurate measurement of the folding landscapes. *Proc Natl Acad Sci (USA)*, 105:9604–9609, 2008.

[87] PT Li, J Vieregg, and I Tinoco Jr. How RNA unfolds and refolds. *Annu Rev Biochem*, 77:77–100, 2008.

[88] SA Woodson. RNA folding and ribosome assembly. *Curr Opin Chem Biol*, 12:667–673, 2008.

[89] B Honig and A Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268:1144–1149, 1995.

[90] K Chin, KA Sharp, B Honig, and AM Pyle. Calculating the electrostatic properties of RNA provides new insights into molecular interactions and function. *Nat Struct Biol*, 6:1055–1061, 1995.

[91] P Auffinger and Y Hashem. Nucleic acid solvation: from outside to insight. *Curr Opin Struct Biol*, 17:325–333, 2007.

[92] JH Cate, RL Hanna, and JA Doudna. A magnesium ion core at the heart of a ribozyme domain. *Nat Struct Biol*, 4:553–558, 1997.

[93] VB Chu, Y Bai, J Lipfert, D Herschlag, and S Doniach. A repulsive field: advances in the electrostatics of the ion atmosphere. *Curr Opin Chem Biol*, 12:619–625, 2008.

[94] SJ Chen and KA Dill. RNA folding energy landscapes. *Proc Natl Acad Sci (USA)*, 97:646–651, 2000.

[95] NJ Baird, E Westhof, H Qin, T Pan, and TR Sosnick. Structure of a folding intermediate reveals the interplay between core and peripheral elements in RNA folding. *J Mol Biol*, 352:712–722, 2005.

[96] I Shcherbakova, S Mitra, A Laederach, and M Brenowitz. Energy barriers, pathways, and dynamics during folding of large, multidomain RNAs. *Curr Opin Chem Biol*, 12:655–666, 2008.

[97] G Chen and DH Turner. Consecutive GA pairs stabilize medium-size RNA internal loops. *Biochemistry*, 45:4025–4043, 2006.

[98] G Chen, SD Kennedy, and DH Turner. A CA(+) pair adjacent to a sheared GA or AA pair stabilizes

size-symmetric RNA internal loops. *Biochemistry*, 48:5738–5752, 2009.

[99] AW Overhauser. Polarization of nuclei in metals. *Phys Rev*, 92:411–415, 1953.

[100] F Sanger, S Nicklen, and AR Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci (USA)*, 74:5463–5467, 1977.

[101] D Thirumalai. Native secondary structure formation in RNA may be a slave to tertiary folding. *Proc Natl Acad Sci (USA)*, 95:11506–11508, 1998.

[102] TN Wong, TR Sosnick, and T Pan. Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures. *Proc Natl Acad Sci (USA)*, 104:17995–18000, 2007.

[103] M Geis, C Flamm, MT Wolfinger, A Tanzer, IL Hofacker, M Middendorf, C Mandl, PF Stadler, and C Thurner. Folding kinetics of large RNAs. *J Mol Biol*, 379:160–173, 2008.

[104] V Moulton, M Zuker, M Steel, R Pointon, and D Penny. Metrics on RNA secondary structures. *J Comput Biol*, 7:272–292, 2000.

[105] M Zuker and P Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9:133–148, 1981.

[106] MS Waterman. Secondary structure of single-stranded nucleic acids. In *Studies on foundations and combinatorics, Advances in mathematics supplementary studies*, pages 167–212. Academic Press, New-York, 1978.

[107] WA Lorenz, Y Ponty, and P Clote. Asymptotics of RNA shapes. *J Comput Biol*, 15:31–63, 2008.

[108] JD Puglisi, JR Wyatt, and I Tinoco Jr. A pseudoknotted RNA oligonucleotide. *Nature*, 15:283–286, 1998.

[109] DP Aalberts and NO Hodas. Asymmetry in RNA pseudoknots: observation and theory. *Nucleic Acids Res*, 33:2210–2214, 2005.

[110] R Oliva, L Cavallo, and A Tramontano. Accurate energies of hydrogen bonded nucleic acid base pairs and triplets in tRNA tertiary interactions. *Nucleic Acids Res*, 34:865–879, 2006.

[111] M Levitt. Detailed molecular model for transfer ribonucleic acid. *Nature*, 172:759–763, 1969.

[112] JE Tabaska, RB Cary, HN Gabow, and GD Stormo. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14:691–699, 1998.

[113] M Tamura and SR Holbrook. Sequence and structural conservation in RNA ribose zippers. *J Mol Biol*, 320:455–474, 2002.

[114] P Nissen, JA Ippolito, N Ban, PB Moore, and TA Steitz. RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. *Proc Natl Acad Sci (USA)*, 98:4899–4903, 2001.

[115] HW Pley, KM Flaherty, and DB McKay. Three-dimensional structure of a hammerhead ribozyme. *Nature*, 372:68–74, 1994.

[116] JH Cate, AR Gooding, E Podell, K Zhou, BL Golden, CE Kundrot, TR Cech, and JA Doudna. Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science*, 273:99–109, 1996.

[117] RR Gutell, HF Noller, and CR Woese. Higher order structure in ribosomal RNA. *EMBO J*, 5:1111–1113, 1986.

[118] F Michel and E Westhof. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J Mol Biol*, 216:585–610, 1990.

[119] A Lescoute and E Westhof. The interaction networks of structured RNAs. *Nucleic Acids Res*, 34:6587–6604, 2006.

[120] Y Xin, C Laing, NB Leontis, and T Schlick. Annotation of tertiary interactions in RNA structures reveals variations and correlations. *RNA*, 14:2465–2477, 2008.

[121] PB Moore. Structural motifs in RNA. *Annu Rev Biochem*, 68:287–300, 1999.

[122] NB Leontis and E Westhof. Analysis of RNA motifs. *Curr Opin Struct Biol*, 13:300–308, 2003.

[123] NB Leontis, A Lescoute, and E Westhof. The building blocks and motifs of RNA architecture. *Curr Opin Struct Biol*, 16:279–287, 2006.

[124] DJ Klein, TM Schmeing, PB Moore, and TA Steitz. The kink-turn: a new RNA secondary structure

motif. *EMBO J*, 20:4214–4221, 2001.

[125] GJ Quigley and A Rich. Structural domains of transfer RNA molecules. *Science*, 194:796–806, 1976.

[126] JH Cate, AR Gooding, R Podell, Z Zhou, BL Golden, AA Szewczak, CE Kundrot, TR Cech, and JA Doudna. RNA tertiary structure mediation by adenosine platforms. *Science*, 273:1696–1699, 1996.

[127] C Tuerk, P Gauss, C Thermes, DR Groebe, M Gayle, N Guild, G Stormo, Y d'Aubenton Carafa, OC Uhlenbeck, I Tinoco Jr, EN Brody, and L Gold. CUUCGG hairpins: extraordinarily stable RNA secondary structures associated with various biochemical processes. *Proc Natl Acad Sci (USA)*, 85:1364–1368, 1988.

[128] CR Woese, S Winker, and RR Gutell. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc Natl Acad Sci (USA)*, 87:8467–8471, 1990.

[129] U Nagaswamy and GE Fox. Frequent occurrence of the T-loop RNA folding motif in ribosomal RNAs. *RNA*, 8:1112–1119, 2002.

[130] AS Krasilnikov and A Mondragón. On the occurrence of the T-loop RNA folding motif in large RNA molecules. *RNA*, 9:640–643, 2003.

[131] JC Lee, JJ Cannone, and RR Gutell. The lonepair triloop: a new motif in RNA structure. *J Mol Biol*, 325:65–83, 2003.

[132] SH Kim, JL Sussman, FL Suddath, GJ Quigley, A McPherson, AH Wang, NC Seeman, and A Rich. The general structure of transfer RNA molecules. *Proc Natl Acad Sci (USA)*, 8:4970–4974, 1974.

[133] AE Walter, DH Turner, J Kim, MH Lyttle, P Muller, DH Mathews, and M Zuker. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc Natl Acad Sci (USA)*, 91:9218–9222, 1994.

[134] RR Gutell, JJ Cannone, D Konings, and D Gautheret. Predicting U-turns in ribosomal RNA with comparative sequence analysis. *J Mol Biol*, 300:791–803, 2000.

[135] R Tyagi and DH Mathews. Predicting helical coaxial stacking in RNA multibranch loops. *RNA*, 13:939–951, 2007.

[136] DE Draper. The RNA-folding problem. *Accounts Chem Res*, 25:201–207, 1992.

[137] PB Moore. The RNA folding problem. In RF Gesteland, TR Cech, and JF Atkins, editors, *The RNA World, 2nd Edn*, pages 381–401. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, NY, 1999.

[138] KD Gibson and HA Scheraga. Minimization of polypeptide energy. I. Preliminary structures of bovine pancreatic ribonuclease S-peptide. *Proc Natl Acad Sci (USA)*, 58:420–427, 1967.

[139] M Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol*, 104:59–107, 1976.

[140] MR Shirts and VS Pande. Screen savers of the world, unite! *Science*, 290:1903–1904, 2000.

[141] C Levinthal. Are there pathways for protein folding? *J Chim Phys*, 65:44–45, 1968.

[142] R Zwanzig, A Szabo, and B Bagchi. Levinthal's paradox. *Proc Natl Acad Sci (USA)*, 89:20–22, 1992.

[143] J Lee, A Liwo, and HA Scheraga. Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: application to the 10-55 fragment of staphylococcal protein A and to apo calbindin D9K. *Proc Natl Acad Sci (USA)*, 96:2025–2030, 1999.

[144] S Cao and SJ Chen. Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA*, 11:1884–1897, 2005.

[145] CN Pace, BA Shirley, M McNutt, and K Gajiwala. Forces contributing to the conformational stability of proteins. *FASEB J*, 10:75–83, 1996.

[146] MT Sykes and M Levitt. Simulations of RNA base pairs in a nanodroplet reveal solvation-dependent stability. *Proc Natl Acad Sci (USA)*, 104:12336–12340, 2007.

[147] SJ Chen. RNA folding: conformational statistics, folding kinetics, and ion electrostatics. *Annu Rev Biophys*, 37:197–214, 2008.

[148] S Moghaddam, G Caliskan, S Chauhan, C Hyeon, RM Briber, D Thirumalai, and SA Woodson. Metal

ion dependence of cooperative collapse transitions in RNA. *J Mol Biol*, 2009. Epub ahead of print.

[149] Y Zhang. Protein structure prediction: when is it useful? *Curr Opin Struct Biol*, 19:145–155, 2009.

[150] BA Shapiro, YG Yingling, W Kasprzak, and E Bindewald. Bridging the gap in RNA structure prediction. *Curr Opin Struct Biol*, 17:157–165, 2007.

[151] N Spackova and J Sponer. Molecular dynamics simulations of sarcin-ricin rRNA motif. *Nucleic Acids Res*, 34:697–708, 2006.

[152] P Sharma, A Mitra, S Sharma, and H Singh. Base pairing in RNA structures: a computational analysis of structural aspects and interaction energies. *J Chem Sci*, 119:525–531, 2007.

[153] ME Christiansen and BM Znosko. Thermodynamic characterization of tandem mismatches found in naturally occurring RNA. *Nucleic Acids Res*, 37:4696–4706, 2009.

[154] CR Woese and RR Gutell. Higher order structural elements in ribosomal RNAs: pseudo-knots and the use of noncanonical pairs. *Proc Natl Acad Sci (USA)*, 87:663–667, 1990.

[155] CM Duarte, LM Wadley, and AM Pyle. RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Res*, 31:4755–4761, 1999.

[156] S Lemieux and F Major. Automated extraction and selection of RNA tertiary structure cyclic motifs. *Nucleic Acids Res*, 34:2340–2346, 2006.

[157] H Yang, F Jossinet, N Leontis, L Chen, J Westbrook, H Berman, and E Westhof. Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res*, 31:3450–3460, 2003.

[158] JD Hornton. A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM J Comp*, 16:358–366, 1987.

[159] BT Wimberly, DE Brodersen, WM Clemons Jr, RJ Morgan-Warren, AP Carter, C Vonrhein, T Hartsch, and V Ramakrishnan. Structure of the 30S ribosomal subunit. *Nature*, 407:327–339, 2000.

[160] N Ban, P Nissen P, J Hansen, PB Moore, and TA Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 å resolution. *Science*, 289:905–920, 2000.

[161] S Lemieux. *Modélisation automatisée de la structure 3-D des ARN*. PhD thesis, University of Montreal, 2001.

[162] F Major, S Lemieux, M Larose, and P Thibault. Modelling RNA three-dimensional structure by combining short nucleotide interaction cycles. *Eur Biophys J*, 34:560, 2005.

[163] K St-Onge, P Thibault, S Hamel, and F Major. Modeling RNA tertiary structure motifs by graph-grammars. *Nucleic Acids Res*, 35:1726–1736, 2007.

[164] KT Simons, C Kooperberg, E Huang, and D Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*, 268:209–225, 1997.

[165] KT Simons, I Ruczinski, C Kooperberg, BA Fox, C Bystroff, and D Baker. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, 34:82–95, 1999.

[166] KT Simons, R Bonneau, I Ruczinski, and D Baker. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins*, Suppl 3:171–176, 1999.

[167] CA Rohl, CE Strauss, KM Misura, and D Baker. Protein structure prediction using Rosetta. *Methods Enzymol*, 383:66–93, 2004.

[168] P Bradley, KM Misura, and D Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309:1868–1871, 2005.

[169] CA Hunter. Sequence-dependent DNA structure. The role of base stacking interactions. *J Mol Biol*, 230:1025–1054, 1993.

[170] CA Hunter and XJ Lu. DNA base-stacking interactions: a comparison of theoretical calculations with oligonucleotide x-ray crystal structures. *J Mol Biol*, 265:603–619, 1997.

[171] WK Olson, M Esguerra, Y Xin, and XJ Lu. New information content in RNA base pairing deduced from quantitative analysis of high-resolution structures. *Methods*, 47:177–186, 2009.

[172] T Xia, J SantaLucia Jr, ME Burkard, R Kierzek, SJ Schroeder, X Jiao, C Cox, and DH Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37:14719–14735, 1998.

[173] M Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31:3406–3415, 2003.

[174] Y Ding and CE Lawrence. A bayesian statistical algorithm for RNA secondary structure prediction. *Comput Chem*, 23:387–400, 1999.

[175] Y Ding and CE Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res*, 31:7280–7301, 2003.

[176] Y Ding, CY Chan, and CE Lawrence. Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res*, 32:135–141, 2004.

[177] IL Hofacker, W Fontana, PF Stadler, S Bonhoeffer, M Tacker, and P Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*, 125:167–188, 1994.

[178] IL Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res*, 31:3429–3431, 2003.

[179] S Wuchty, W Fontana, IL Hofacker, and P Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49:145–165, 1999.

[180] DH Mathews, MD Disney, JL Childs, SJ Schroeder, M Zuker, and DH Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci (USA)*, 101:7287–7292, 2004.

[181] W Dawson, K Suzuki, and K Yamamoto. A physical origin for functional domain structure in nucleic acids as evidenced by cross-linking entropy: I. *J Theor Biol*, 213:359–386, 2001.

[182] W Dawson, K Suzuki, and K Yamamoto. A physical origin for functional domain structure in nucleic acids as evidenced by cross-linking entropy: II. *J Theor Biol*, 213:387–412, 2001.

[183] I Tinoco Jr, OC Uhlenbeck, and MD Levine. Estimation of secondary structure in ribonucleic acids. *Nature*, 230:362–367, 1971.

[184] KJ Doshi, JJ Cannone, CW Cobaugh, and RR Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, 5:105, 2004.

[185] DM Layton and R Bundschuh. A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation. *Nucleic Acids Res*, 33:519–524, 2005.

[186] T Dale, R Smith, and MJ Serra. A test of the model to predict unusually stable RNA hairpin loop stability. *RNA*, 6:608–615, 2000.

[187] JM Diamond, DH Turner, and DH Mathews. Thermodynamics of three-way multibranch loops in RNA. *Biochemistry*, 40:6971–6981, 2001.

[188] A Lescoute and E Westhof. Topology of three-way junctions in folded RNAs. *RNA*, 12:83–93, 2006.

[189] JS McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers*, 29:1105–1119, 1990.

[190] M Andronescu, A Condon, HH Hoos, DH Mathews, and KP Murphy. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, 23:i19–i28, 2007.

[191] RB Lyngsø and CNS Pedersen. Pseudoknots in RNA secondary structures. In *RECOMB*, pages 201–209, 2000.

[192] R Nussinov and AB Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci (USA)*, 77:6309–6313, 1980.

[193] E Rivas and SR Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol*, 285:2053–2068, 1999.

[194] J Ruan, GD Stormo, and W Zhang. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 20:58–66, 2004.

[195] Y Wexler, CBZ Zilberstein, and M Ziv-Ukelson. A study of accessible motifs and RNA folding complexity. In *RECOMB*, pages 473–487, 2006.

[196] RB Lyngso, M Zuker, and CN Pedersen. Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics*, 15:440–445, 1999.

[197] AY Ogurtsov, SA Shabalina, AS Kondrashov, and MA Roytberg. Analysis of internal loops within the RNA secondary structure in almost quadratic time. *Bioinformatics*, 22:1317–1324, 2006.

[198] M Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52, 1989.

[199] MS Waterman. Sequence alignments in the neighborhood of the optimum with general application to dynamic programming. *Proc Natl Acad Sci (USA)*, 80:3123–3124, 1983.

[200] MS Waterman and TH Byers. A dynamic programming algorithm to find all solutions in the neighborhood of the optimum. *Math Biosci*, 77:179–188, 1985.

[201] M Zuker. Calculating nucleic acid secondary structure. *Curr Opin Struct Biol*, 10:303–310, 2000.

[202] DH Mathews. Revolutions in RNA secondary structure prediction. *J Mol Biol*, 359:526–532, 2006.

[203] PP Gardner and R Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5:140, 2004.

[204] CB Do, DA Woods, and S Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22:e90–e98, 2006.

[205] RI Dima, C Hyeon, and D Thirumalai. Extracting stacking interaction parameters for RNA from the data set of native structures. *J Mol Biol*, 347:53–69, 2005.

[206] JC Wu, DP Gardner, S Ozer, RR Gutell, and P Ren. Correlation of RNA secondary structure statistics with thermodynamic stability and applications to folding. *J Mol Biol*, in press, 2009.

[207] GM Clore and J Kuszewski. Improving the accuracy of NMR structures of RNA by means of conformational database potentials of mean force as assessed by complete dipolar coupling cross-validation. *J Am Chem Soc*, 125:1518–1525, 2003.

[208] GR Bowman, X Huang, Y Yao, J Sun, G Carlsson, LJ Guibas, and VS Pande. Structural insight into RNA hairpin folding intermediates. *J Am Chem Soc*, 130:9676–9678, 2008.

[209] C Hyeon and D Thirumalai. Mechanical unfolding of RNA: from hairpins to structures with internal multiloops. *Biophys J*, 92:731–743, 2007.

[210] C Hyeon and D Thirumalai. Multiple probes are required to explore and control the rugged energy landscape of RNA hairpins. *J Am Chem Soc*, 130:1538–1539, 2008.

[211] P Kapranov, J Cheng, S Dike, DA Nix, R Duttagupta, AT Willingham, PF Stadler, J Hertel, J Hackermüller, IL Hofacker, I Bell, E Cheung, J Drenkow, E Dumais, S Patel, G Helt, M Ganesh, S Ghosh, A Piccolboni, V Sementchenko, H Tammana, and TR Gingeras. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316:1484–1488, 2007.

[212] DA Benson, I Karsch-Mizrachi, DJ Lipman, J Ostell, and DL Wheeler. Genbank. *Nucleic Acids Res*, 35:D21–D35, 2007.

[213] DH Mathews and DH Turner. Prediction of RNA secondary structure by free energy minimization. *Curr Opin Struct Biol*, 16:270–278, 2006.

[214] RR Gutell, JC Lee, and JJ Cannone. The accuracy of ribosomal RNA comparative structure models. *Curr Opin Struct Biol*, 12:310–318, 2002.

[215] F Major, M Turcotte, D Gautheret, G Lapalme, E Fillion, and R Cedergren. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science*, 253:1255–1260, 1991.

[216] JR Williamson. Induced fit in RNA-protein recognition. *Nat Struct Biol*, 7:834–837, 2000.

[217] N Shankar, SD Kennedy, G Chen, TR Krugh, and DH Turner. The NMR structure of an internal loop from 23S ribosomal RNA differs from its structure in crystals of 50s ribosomal subunits. *Biochemistry*, 45:11776–11789, 2006.

[218] J Kondo, A Urzhumtsev, and E Westhof. Two conformational states in the crystal structure of the *Homo sapiens* cytoplasmic ribosomal decoding A site. *Nucleic Acids Res*, 34:676–685, 2006.

[219] BM Lee, J Xu, BK Clarkson, MA Martinez-Yamout, HJ Dyson, DA Case, JM Gottesfeld, and PE Wright. Induced fit and "lock and key" recognition of 5S RNA by zinc fingers of transcription factor IIIA. *J Mol Biol*, 357:275–291, 2006.

[220] DP Giedroc, CA Theimer, and PL Nixon. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J Mol Biol*, 298:167–185, 2000.

[221] S Griffiths-Jones, RJ Grocock, S van Dongen, A Bateman, and AJ Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34:D140–D144, 2006.

[222] J Han, Y Lee, KH Yeom, JW Nam, I Heo, JK Rhee, SY Sohn, Y Cho, BT Zhang, and VN Kim. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell*, 125:887–901, 2006.

[223] IJ Macrae, K Zhou, F Li, A Repic, AN Brooks, WZ Cande, PD Adams, and JA Doudna. Structural basis for double-stranded RNA processing by Dicer. *Science*, 311:195–198, 2006.

[224] EJ Merino, KA Wilkinson, JL Coughlan, and KM Weeks. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J Am Chem Soc*, 127:4223–4231, 2005.

[225] V Perret, A Garcia, J Puglisi, H Grosjean, JP Ebel, C Florentz, and R Giegé. Conformation in solution of yeast tRNA(Asp) transcripts deprived of modified nucleotides. *Biochimie*, 72:735–743, 1990.

[226] C Brunel, P Romby, E Westhof, C Ehresmann, and B Ehresmann. Three-dimensional model of *Escherichia coli* ribosomal 5S RNA as deduced from structure probing in solution and computer modeling. *J Mol Biol*, 221:293–308, 1991.

[227] NB Leontis and PB Moore. NMR evidence for dynamic secondary structure in helices II and III of the RNA of *Escherichia coli*. *Biochemistry*, 25:3916–3925, 1986.

[228] MW Hentze and LC Kuhn. Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc Natl Acad Sci (USA)*, 93:8175–8182, 1996.

[229] SR Jaffrey, DJ Haile, RD Klausner, and JB Harford. The interaction between the iron-responsive element binding protein and its cognate RNA is highly dependent upon both RNA sequence and structure. *Nucleic Acids Res*, 21:4627–4631, 1993.

[230] H Sierzputowska-Gracz, RA McKenzie, and EC Theil. The importance of a single G in the hairpin loop of the iron responsive element (IRE) in ferritin mRNA for structure: an NMR spectroscopy study. *Nucleic Acids Res*, 23:146–153, 1995.

[231] R Leipuviene and EC Theil. The family of iron responsive RNA structures regulated by changes in cellular iron and oxygen. *Cell Mol Life Sci*, 64:2945–2955, 2007.

[232] A Cléry, V Bourguignon-Igel, C Allmang, A Krol, and C Branlant. An improved definition of the RNA-binding specificity of SECIS-binding protein 2, an essential component of the selenocysteine incorporation machinery. *Nucleic Acids Res*, 35:1868–1884, 2007.

[233] T Jacks, MD Power, FR Masiarz, PA Luciw, PJ Barr, and HE Varmus. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature*, 331:280–283, 1988.

[234] C Gaudin, MH Mazauric, M Traikia, E Guittet, S Yoshizawa, and D Fourmy. Structure of the RNA signal essential for translational frameshifting in HIV-1. *J Mol Biol*, 349:1024–1035, 2005.

[235] DW Staple and SE Butcher. Solution structure and thermodynamic investigation of the HIV-1 frameshift inducing element. *J Mol Biol*, 349:1011–1023, 2005.

[236] R Haralick and G Elliott. Increasing tree search efficiency for constraint satisfaction problems. *Artificial Intelligence*, 14:263–313, 1980.

[237] J Reeder and R Giegerich. Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics*, 21:3516–3523, 2005.

[238] MS Bazaraa, HD Sherali, and CM Shetty. *Nonlinear pogramming theory and algorithms, 3rd Edn*. John Wiley & Sons, Inc, Hoboken, NJ, 2006.

[239] T Elgavish, JJ Cannone, JC Lee, SC Harvey, and RR Gutell. AA.AG@helix.ends: A=A and A=G base-pairs at the ends of 16S and 23S rRNA helices. *J Mol Biol*, 310:735–753, 2001.

[240] D Dulude, M Baril, and L Brakier-Gingras. Characterization of the frameshift stimulatory signal controlling a programmed -1 ribosomal frameshift in the human immunodeficiency virus type 1. *Nucleic Acids Res*, 30:5094–5102, 2002.

[241] M Baril, D Dulude, K Gendron, G Lemay, and L Brakier-Gingras. Efficiency of a programmed -1

ribosomal frameshift in the different subtypes of the human immunodeficiency virus type 1 group M. *RNA*, 9:1246–1253, 2003.

[242] R Girnary, L King, L Robinson, R Elston, and I Brierley. Structure-function analysis of the ribosomal frameshifting signal of two human immunodeficiency virus type 1 isolates with increased resistance to viral protease inhibitors. *J Gen Virol*, 88:226–235, 2007.

[243] C Massire and E Westhof. MANIP: an interactive tool for modeling RNA. *J Mol Graphics Modell*, 16:197–205, 1998.

[244] M Djelloul and A Denise. Automated motif extraction and classification in RNA tertiary structures. *RNA*, 14:2489–2497, 2008.

[245] HC Huang, U Nagaswamy, and GE Fox. The application of cluster analysis in the intercomparison of loop structures in RNA. *RNA*, 11:412–423, 2005.

[246] V Lisi and F Major. A comparative analysis of the triloops in all high-resolution RNA structures reveals sequence structure relationships. *RNA*, 13:1537–1545, 2007.

[247] M Abraham, O Dror, R Nussinov, and HJ Wolfson. Analysis and classification of RNA tertiary structures. *RNA*, 14:2274–2289, 2008.

[248] RD Dowell and SR Eddy. Evaluation of several lightweight stochastic context-free grammers for RNA secondary structure prediction. *BMC Bioinformatics*, 5:71–85, 2004.

[249] AS Yang and B Honig. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J Mol Biol*, 301:665–678, 2000.

[250] M Shatsky, R Nussinov, and HJ Wolfson. Flexible protein alignment and hinge detection. *Proteins*, 48:242–256, 2002.

[251] HA Gabb, SR Sanghani, CH Robert, and C Prevost. Finding and visualizing nucleic acid base stacking. *J Mol Graph*, 14:6–11, 1996.

[252] F Major and P Thibault. RNA tertiary structure prediction. In T Lengauer, editor, *Bioinformatics: from genomes to therapies*, pages 491–539. Wiley-VCH, Weinheim, Germany, 2007.

[253] J Gorodkin, SL Stricklin, and GD Stormo. Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res*, 29:2135–2144, 2001.

[254] CC Correll, J Beneken, MJ Plantinga, M Lubbers, and YL Chan. The common and the distinctive features of the bulged-G motif based on a 1.04 å resolution RNA structure. *Nucleic Acids Res*, 31:6806–6818, 2003.

[255] CM Dunham, JB Murray, and WG Scott. A helical twist-induced conformational switch activates cleavage in the hammerhead ribozyme. *J Mol Biol*, 332:327–336, 2003.

[256] A Laederach, JM Chan, A Schwartzman, E Willgohs, and RB Altman. Coplanar and coaxial orientations of RNA bases and helices. *RNA*, 13:643–650, 2007.

[257] A Zemla, C Venclovas, J Moult, and K Fidelis. Processing and analysis of CASP3 protein structure predictions. *Proteins*, Suppl 3:22–29, 1999.

[258] K Ginalski, NV Grishin, A Godzik, and L Rychlewski. Practical lessons from protein structure prediction. *Nucleic Acids Res*, 33:1874–1891, 2005.

[259] Y Zhang. Progress and challenges in protein structure prediction. *Curr Opin Struct Biol*, 18:342–348, 2008.

[260] RL Dunbrack Jr and FE Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci*, 6:1661–1681, 1997.

[261] A Grishaev, J Ying, MD Canny, A Pardi, and A Bax. Solution structure of tRNAVal from refinement of homology model against residual dipolar coupling and SAXS data. *J Biomol NMR*, 42:99–109, 2008.

[262] WL DeLano. *The PyMOL molecular graphics system*. DeLano Scientific, Palo Alto, CA, USA, 2002.

[263] MH Hao, S Rackovsky, A Liwo, MR Pincus, and HA Scheraga. Effects of compact volume and chain stiffness on the conformations of native proteins. *Proc Natl Acad Sci (USA)*, 89:6614–6618, 1992.

[264] PA Sharp. The centrality of rna. *Cell*, 136:577–580, 2009.

[265] H Han and PB Dervan. Visualization of RNA tertiary structure by RNA-EDTA.Fe(II) autocleavage: analysis of tRNA(Phe) with uridine-EDTA.Fe(II) at position 47. *Proc Natl Acad Sci (USA)*, 91:4955–4959, 1994.

[266] F Major, D Gautheret, and R Cedergren. Reproducing the three-dimensional structure of a tRNA molecule from structural constraints. *Proc Natl Acad Sci (USA)*, 90:9408–9412, 1993.

[267] F Major. Building three-dimensional ribonucleic acid structures. *Comp Sci Eng*, 5:44–53, 2003.

[268] TM Klingler and DL Brutlag. Detection of correlations in tRNA sequences with structural implications. *Proc Int Conf Intell Syst Mol Biol*, 1:225–233, 1993.

[269] L Jaeger, F Michel, and E Westhof. Involvement of a GNRA tetraloop in long-range RNA tertiary interactions. *J Mol Biol*, 236:1271–1276, 1994.

[270] M Costa and F Michel. Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J*, 14:1276–1285, 1995.

[271] L Nilsson, R Rigler, and P Laggner. Structural variability of tRNA: small-angle x-ray scattering of the yeast tRNA^{Phe}-*Escherichia coli* tRNA$_2{}^{Glu}$ complex. *Proc Natl Acad Sci (USA)*, 79:5891–5895, 1982.

[272] X Fang, K Littrell, XJ Yang, SJ Henderson, S Siefert, P Thiyagarajan, T Pan, and TR Sosnick. Mg2+-dependent compaction and folding of yeast tRNA-Phe and the catalytic domain of the *B. subtilis* RNase P RNA determined by small-angle X-ray scattering. *Biochemistry*, 39:11107–11113, 2000.

[273] Z Zhuang, L Jaeger, and JE Shea. Probing the structural hierarchy and energy landscape of an RNA T-loop hairpin. *Nucleic Acids Res*, 35:6995–7002, 2007.

[274] P Bouchard, J Lacroix-Labonté, G Desjardins, P Lampron, V Lisi, SLemieux, F Major, and P Legault. Role of SLV in SLI substrate recognition by the neurospora VS ribozyme. *RNA*, 14:736–748, 2008.

[275] W Fuller and A Hodgson. Conformation of the anticodon loop in tRNA. *Nature*, 215:817–821, 1967.

[276] L Jaeger, EJ Verzemnieks, and C Geary. The UA_handle: a versatile submotif in stable RNA architectures. *Nucleic Acids Res*, 37:215–230, 2009.

[277] FL Murphy and TR Cech. GAAA tetraloop and conserved bulge stabilize tertiary structure of a group I intron domain. *J Mol Biol*, 236:49–63, 1994.

[278] JA Latham and TR Cech. Defining the inside and outside of a catalytic RNA molecule. *Science*, 245:276–282, 1989.

[279] S Mitra, IV Shcherbakova, RB Altman, M Brenowitz, and A Laederach. High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. *Nucleic Acids Res*, 36:e63, 2008.

[280] MF Sanner, AJ Olson, and JC Spehner. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, 38:305–320, 1996.

[281] K Takamoto, R Das, Q He, S Doniach, M Brenowitz, D Herschlag, and MR Chance. Principles of RNA compaction: insights from the equilibrium folding pathway of the P4-P6 RNA domain in monovalent cations. *J Mol Biol*, 343:1195–1206, 2004.

[282] J Kim, S Yu, B Shim, H Kim, H Min, EY Chung, R Das, and S Yoon. A robust peak detection method for RNA structure inference by high-throughput contact mapping. *Bioinformatics*, 25:1137–1144, 2009.

[283] MH Koch, P Vachette, and DI Svergun. Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Q Rev Biophys*, 36:147–227, 2003.

[284] R Cedergren and F Major. Modeling the tertiary structure of RNA. In RW Simons and M Grunberg-Manago, editors, *RNA structure and function*, pages 37–75. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, NY, 1998.

[285] E Lécuyer, H Yoshida, and HM Krause. Global implications of mRNA localization pathways in cellular organization. *Curr Opin Cell Biol*, 21:409–415, 2009.

[286] E Lécuyer, H Yoshida, N Parthasarathy, C Alm, T Babak, T Cerovina, TR Hughes, P Tomancak, and HM Krause. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, 131:174–187, 2007.

[287] G dos Santos, AJ Simmonds, and HM Krause. A stem-loop structure in the wingless transcript

defines a consensus motif for apical RNA transport. *Development*, 135:133–143, 2008.

[288] J Gorodkin, LJ Heyer, S Brunak, and GD Stormo. Displaying the information contents of structural RNA alignments: the structure logos. *Comput Appl Biosci*, 13:583–586, 1997.

[289] TD Schneider and RM Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18:6097–6100, 1990.

[290] TV Pestova, VG Kolupaeva, IB Lomakin, EV Pilipenko, IN Shatsky, VI Agol, and CU Hellen. Molecular mechanisms of translation initiation in eukaryotes. *Proc Natl Acad Sci (USA)*, 98:7029–7036, 2001.

[291] N Sonenberg and AG Hinnebusch. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, 136:731–745, 2009.

[292] WA Miller, Z Wang, and K Treder. The amazing diversity of cap-independent translation elements in the 3'-untranslated regions of plant viral RNAs. *Biochem Soc Trans*, 35:1629–1633, 2007.

[293] Z Wang, K Treder, and WA Miller. Structure of a viral cap-independent translation element that functions via high affinity binding to the eIF4E subunit of eIF4F. *J Biol Chem*, 284:14189–14202, 2009.

[294] XD Wang and RA Padgett. Hydroxyl radical footprinting of RNA: application to pre-mRNA splicing complexes. *Proc Natl Acad Sci (USA)*, 86:7795–7799, 1989.

[295] DK Hendrix, SE Brenner, and SR Holbrook. RNA structural motifs: building blocks of a modular biomolecule. *Q Rev Biophys*, 38:221–243, 2005.

[296] KA Wilkinson, RJ Gorelick, SM Vasa, N Guex, A Rein, DH Mathews, MC Giddings, and KM Weeks. High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol*, 6:e96, 2008.

[297] JM Watts, KK Dang, RJ Gorelick, CW Leonard, JW Bess Jr, R Swanstrom, CL Burch, and KM Weeks. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, 460:711–716, 2009.

[298] F Major and P Thibault. Computer modeling of RNA 3D structure. In RA Meyers, editor, *Encyclopedia of molecular biology and molecular medicine*, pages 605–636. VCH Publishers, Inc., New York, 2005.

[299] WK Olson, AA Gorin, XJ Lu, LM Hock, and VB Zhurkin. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci (USA)*, 95:11163–11168, 1998.

[300] MS Babcock, EP Pednault, and WK Olson. Nucleic acid structure analysis. Mathematics for local cartesian and helical structure parameters that are truly comparable between structures. *J Mol Biol*, 237:125–156, 1994.

[301] XJ Lu and WK Olson. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*, 31:5108–5121, 2003.

[302] J Wang, P Cieplak, and PA Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem*, 21:1049–1074, 2000.

[303] A Perez, I Marchan, D Svozil, J Sponer, TE Cheatham III, CA Laughton, and M Orozco. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J*, 92:3817–3829, 2007.

[304] E Fadrna, N Spackova, J Sarzynska, J Koca, M Orozco, TE Cheatham III, T Kulinski, and J Sponer. Single stranded loops of quadruplex DNA as key benchmark for testing nucleic acids force fields. *J Chem Theory Comput*, 2009. Epub ahead of print.

[305] CA Hunter and JKM Sanders. The nature of pi-pi interactions. *J Am Chem Soc*, 112:5525–5534, 1990.

[306] J Sponer, J Leszczynski, and P Hobza. Base stacking in cytosine dimer. A comparison of correlated ab initio calculations with three empirical potential models and density functional theory calculations. *J Comp Chem*, 17:841–850, 1996.

[307] V Renugopalakrishnan, AV Lakshminarayanan, and V Sasisekharan. Stereochemistry of nucleic acids and polynucleotides III. Electronic charge distribution. *Biopolymers*, 10:1159–1167, 1971.

[308] J Sponer, KE Riley, and P Hobza. Nature and magnitude of aromatic stacking of nucleic acid bases. *Phys Chem Chem Phys*, 10:2595–2610, 2008.

[309] AV Morozov and T Kortemme. Potential functions for hydrogen bonds in protein structure prediction and design. *Adv Protein Chem*, 72:1–38, 2005.

[310] DN Boobbyer, PJ Goodford, PM McWhinnie, and RC Wade. New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. *J Med Chem*, 32:1083–1094, 1989.

[311] W Hasel, TF Hendrickson, and WC Still. A rapid approximation to the solvent accessible surface areas of atoms. *Tetrahedron Comput Method*, 1:103–116, 1988.

[312] JR Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[313] JR Quinlan. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann, 1993.

[314] D van Meerten, G Girard, and J van Duin. Translational control by delayed RNA folding: identification of the kinetic trap. *RNA*, 7:483–94, 2001.

[315] KA Wilkinson, EJ Merino, and KM Weeks. RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNA(Asp) transcripts. *J Am Chem Soc*, 127:4659–4667, 2005.

[316] C Flamm, W Fontana, IL Hofacker, and P Schuster. RNA folding at elementary step resolution. *RNA*, 6:325–338, 2000.

[317] SR Morgan and PG Higgs. Barrier heights between ground states in a model of RNA secondary structure. *J Phys A Math Gen*, 31:3153–3170, 1998.