CHAPTER 6

Knowledge-Based Models

RNA

Raúl Méndez, Andrey Krokhotin, Marino Convertino, Jhuma Das, Arpit Tandon, and Nikolay V. Dokholyan

CONTENTS

6.1 INTRODUCTION

Nucleic acids were discovered early in 1869 by young Swiss physician Friederich Miescher, as a consequence of his studies on the molecules of life using leucocyte cultures (Dahm 2005). Miescher isolated a novel molecule from the cell nuclei with distinct biochemical properties from proteins. This was the first reported deoxyribonucleic acid (DNA) extraction. He termed the newly discovered molecule "nuclein" because it was isolated from cell nuclei. Several other experiments yielded the ensuing discovery of ribonucleic acid (RNA) in prokaryote organisms, which do not have nuclei. In the beginning of the twentieth century, the difference between DNA and RNA was still unclear. The two molecules were referred to as "pancreas nucleic acid" and "yeast nucleic acid," respectively, referring to the source for purification. Only in decade between 1930 and 1940, the chemical compositions of DNA and RNA were determined, resulting in the current distinction between sugar molecules (ribose and 2-deoxyribose for RNA and DNA, respectively), and one specific nucleobase (uracil and thymine in RNA and DNA, respectively) in the two nucleic acids (Worthington Allen 1941).

The next landmark discovery concerning RNA was its central role in protein biosynthesis. In 1956, Francis Crick stated his famous "central dogma" of molecular biology (Crick 1970), which claims that *DNA makes RNA and RNA makes proteins*. The "dogma" emphasized the role of RNA as messenger molecule. The discovery of the lactose (lac) operon, a cluster of metabolic genes transcribed together under the same promoter, in *Escherichia coli* allowed the study of the messenger RNA (mRNA) and the genetic code (i.e., the equivalence between nucleotide sequence in mRNA and actual amino acids in proteins). Later, radiolabeling experiments in the 1950s showed that amino acids bind ribosomes, those large complexes made of three long RNA strands and several proteins known to interact with mRNA. In the 1960s, the first ribosomes were isolated as they move along a single mRNA molecule during protein biosynthesis, catalyzing the new peptide bond formation (Geiduschek and Haselkorn 1969). During the same period of time, the transfer RNA (tRNA) was shown to covalently

bind amino acids at the 3′-terminus and to recognize a codon sequence that codes for a specific amino acid. The genetic code was fully elucidated using *in vitro* experiments, establishing a consistent equivalence between codons and amino acids.

The first tRNA molecule was sequenced in 1965 (Holley et al. 1965) and later in 1975, it was determined its three-dimensional structure (Ladner et al. 1975) revealing secondary structure conservation between homologous sequences. Also the first RNA virus, a bacteriophage, was sequenced, pointing the existence of the enzyme reverse transcriptase, which is able to synthesize DNA from RNA within the host organism (note that this observation introduces an important change in the "central dogma," i.e., *RNA can also make DNA*).

Beyond roles in protein synthesis as mRNA, tRNA, and ribosomal RNA (rRNA), most notably the 1980s yielded a new round of discoveries, pertaining to a new type of RNA molecule, namely the noncoding RNA (ncRNA). Many ncRNAs have catalytic activity (a property that besides the rRNA was previously believed to be the sole domain of proteins) such as Ribonuclease P, which catalyzes its tRNA preprocessing mechanism (Frank and Pace 1998). At the time, another breakthrough discovery was that eukaryotic genes are not continuous pieces of DNA, but discontinuous fragments that code for protein (exons) intercalated by noncoding fragments (introns) which are removed after transcription (pre-mRNA) by splicing (Green 1986). Moreover, intron removal is catalyzed by a complex of proteins and small nuclear ribonucleic particles (snRNPs) (Kramer 1996). In the last two decades, multitudes of ncRNAs have been found that are involved in several mechanisms such as DNA genome editing, gene regulation, interference or silencing, etc. For a wider review, please refer to Cech and Steitz (2014).

The increasing importance of RNA in biology raised the "RNA world" hypothesis, which advances the view that RNA is the precursor molecule to life because of its ability to self-replicate, catalyze reactions, and store genetic information (Copley, Smith, and Morowitz 2007). Interestingly, from the evolutionary point of view, RNA replication machinery (or "replicons") evolves quicker than DNA's, since reverse transcriptases and RNA replicases do not have exonuclease activity and there is no equivalent RNA repairing system as for damaged DNA molecules (Hansen, Long, and Schultz 1997).

To understand the dynamics, the physicochemical properties, and functions of RNA molecules, we need to determine the atoms' precise

spatial arrangement. Similarly to proteins, there are two main techniques to elucidate RNA three-dimensional (3D) structures: x-ray crystallography (Thomsen and Berger 2012) and nuclear magnetic resonance (NMR) (Scott and Hennig 2008). What makes RNA structure determination a more challenging task is the higher number of possible molecular conformations at physiological temperature. Despite the lower number of building blocks (i.e., only 4 nucleotides (nt), as compared to 20 amino acids in proteins), the number of conformations available to RNA polymers is much greater than what is accessible to proteins. In proteins, the conformations of amino acids' side chains impose serious steric hindrance to the backbone, so that in practice the angles of the rotatable bonds of the backbone are restricted to certain range of values (Ramachandran, Ramakrishnan, and Sasisekharan 1963) thus making the protein conformational space somewhat limited. RNA molecules, in contrast, have significantly large number of conformations accessible to the ribose subunits. Although some rotameric states can be found for RNA backbone conformations, they are less accurately defined than for proteins (Murray et al. 2003).

X-ray crystallography has allowed for the elucidation of a number of structures, notably the 3D structure of the ribosome was found in 2000 (Schluenzen et al. 2000; Ban et al. 2000; Wimberly et al. 2000). While the power of this technique yields exact structures at high resolution, care must be taken in sample preparation to avoid aggregates or misfolded structures (Reyes, Garst, and Batey 2009). NMR, while exceptionally useful and does not require crystallization, presents its own set of challenges. Solving an NMR spectrum for an RNA molecule is in general much harder than for proteins mostly because: (1) ^{1}H isotope proportion is much lower in RNA than in proteins; (2) chemical shifts from different active species such as ^{1}H, ^{15}N, or ^{13}C, are difficult to distinguish because they are in very similar chemical environments; (3) the 4 nt: adenosine monophosphate (A), uridine monophosphate (U), guanosine monophosphate (G), and cytosine monophosphate (C) are more chemically similar *per se* compared to the 20 possible amino acids; and (4) RNA molecules display helical structures that are rather linear, and not globular like proteins, making 2D correlation spectrums more difficult to resolve (Addess and Feigon 1996). In addition, backbone proton signal that is characteristic to polypeptides and can be easily identified, it is not as clear in RNA, because of the intercalated phosphate groups.

The limitations of experimental methodologies require that accurate computational techniques be applied to RNA structure prediction to gather knowledge about the different functional roles of RNA molecules. Any

computational modeling of a biomolecule consists of two important components: a description of the system, which involves an energy function that accounts for intra- and interatomic interactions, along with a sampling engine to explore the energy landscape derived from that energy function and, ideally, be able to identify the native conformation of the system as the absolute energy minimum. The description of the system can be as realistic as to include all atoms, typically coming with a molecular mechanics force field made of the sum of pairwise bonded and nonbonded interactions. There is always a tradeoff between the accuracy of the model description and the appropriate sampling of the energy landscape. Detailed atomistic models require efficient sampling of the conformational landscape (involving computationally extensive simulations); otherwise the optimal predicted structure may just be a conformation trapped on a local energy minimum that is not the absolute energy minimum. Coarse-grained methods simplify the system description and the interactions of their components. They provide a less rugged energy landscape that is easier to sample and consequently lightens the computational load as well. Coarse-grained methods are a common choice in RNA modeling, because they provide reasonably accurate results as compared to all-atom models, at a lower computational expense.

In this chapter, we summarize the most common coarse-grained models used to predict RNA 3D structure and provide a qualitative description of their energy functions. This classification is based on the study published by Xia and Ren (2013). This chapter is structured as follows. In Section 6.2, RNA molecule is revisited. Section 6.3 introduces the most common coarse-grained methods in RNA modeling. Section 6.4 summarizes the principles for the typical sampling algorithms used in combination with the previously described models. Section 6.5 illustrates how experimental information can be used to improve sampling and hence the final quality of the models. Section 6.6 is the comparison of the state-of-the-art RNA structure prediction methods as reported in the international blind contest *RNA Puzzles*. Not all the methods cited in that section are coarse-grained strictly speaking, but they best represent what it is possible to achieve using in the RNA structure-modeling field. Finally, Section 6.7 draws the conclusion.

6.2 RNA MOLECULE DESCRIPTION

Nucleic acids play a central role in almost all cellular processes. RNA molecules, like DNA, form polymeric chains, but unlike the latter, they

are single-stranded (except in special cases of RNA–RNA interaction) polymers. Each RNA monomer is made of a ribose sugar (containing a hydroxyl group at position 2′, see Figure 6.1), a phosphate group in position 5′, and like DNA, a purine (adenine or guanine) or pyrimidine (cytosine and uracil) nucleobase at position 1′ (Figure 6.1). At physiological conditions, phosphate groups are negatively charged, while nucleobases can

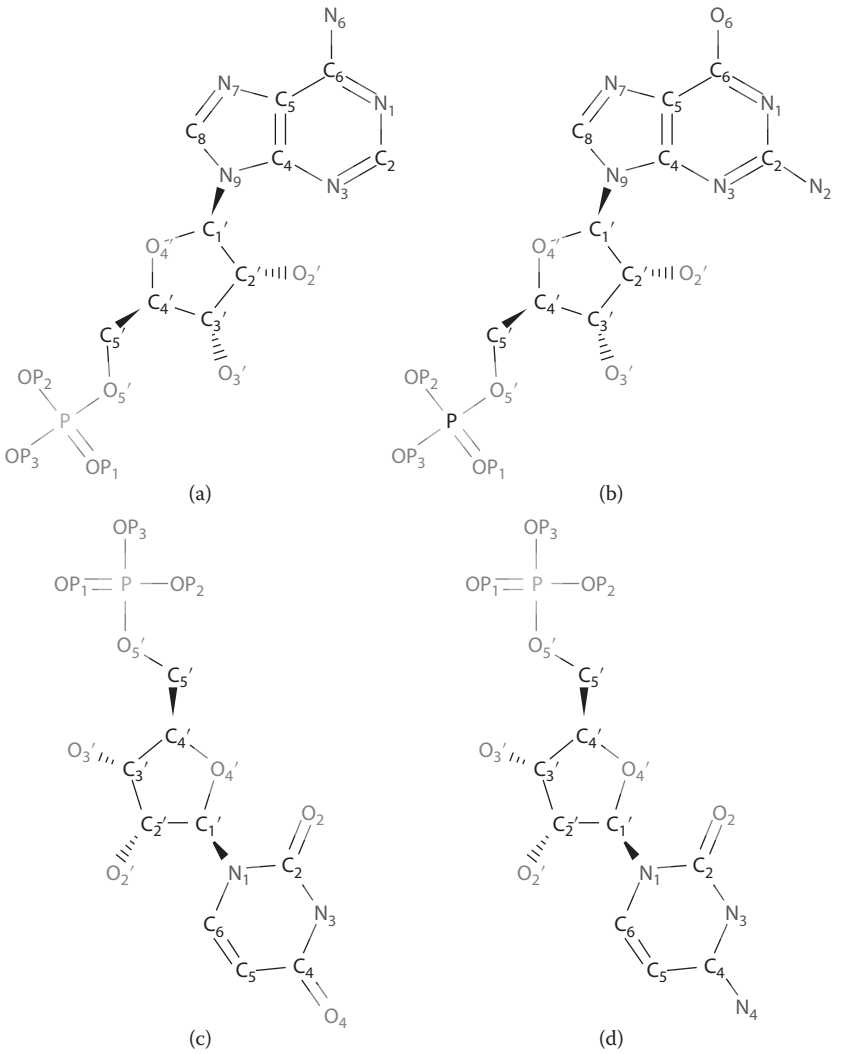


FIGURE 6.1 2D representations of the four nucleotide found in RNA. (a) Adenosine 5′-monophosphate, (b) Guanosine 5′-monophosphate, (c) Uridine 5′-monophosphate, and (d) Cytidine 5′-monophosphate.

make hydrogen bond interactions with other bases of the same strand or from different RNA molecules. Such base pairing is the major determinant of the 3D structure of RNA.

There are several possible base pairings in RNA molecules. In the canonical Watson–Crick pairing, adenine interacts via hydrogen bonds to uracil, while guanine interacts with cytosine (Figure 6.2). Hoogsteen base pairing also takes place between adenine (A) and uracil (U) and between guanine (G) and a positively charged cytosine (C^+). A–U pairing involves atom N_6 as hydrogen bond donor and N_7 as hydrogen bond acceptor, which bind the same N_3–O_4 Watson–Crick side of the pyrimidine base. G–C^+ pairing is made between O_6 (G, acceptor) and N_4 (C^+, donor), and N_7 (G, acceptor) and N_3^+ (C^+, donor). Note the base flipping in the pyrimidine base as compared to Watson–Crick pairing (see Figures 6.1 and 6.2 for details about atom nomenclature).

Both Watson–Crick and Hoogsteen pairing yield hydrogen bond interactions that result in characteristic secondary structure elements, such as

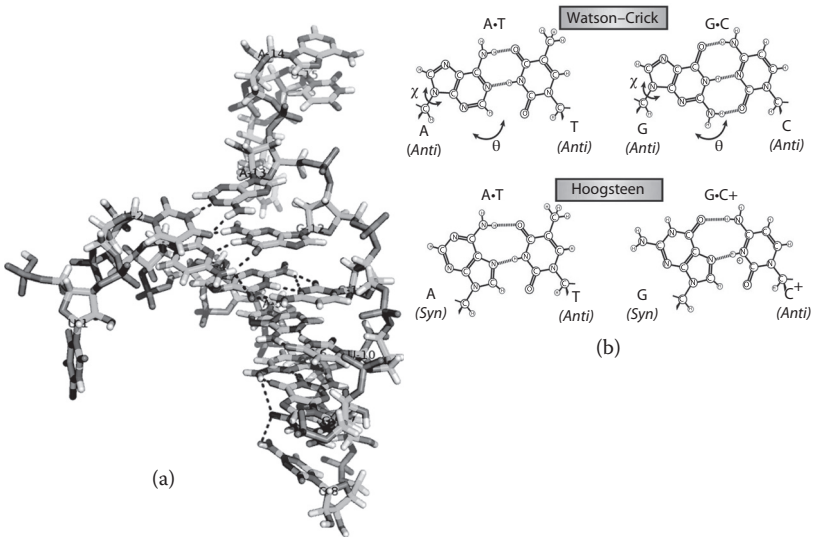


FIGURE 6.2 (a) Watson–Crick base pairing from RNA structure of a small interfering RNA (siRNA)-like, a noncoding RNA that interferes with the expression of certain genes, duplex from *Archaeoglobus fulgidus* (PDB-ID: 2BGG). Hydrogen bonds are highlighted as dashed black lines. (b) Hoogsteen base pairing scheme. Note that instead of the thymine as a pyrimidine, we should have uracil for the RNA, but the hydrogen bond pattern remains the same. (Reprinted by permission from MacMillan Publishers Ltd. *Nature,* Nikolova et al., 470: 498–502, copyright 2011.)

stem loops (Watson et al. 2013) (Figure 6.3a), tetraloops (Woese, Winker, and Gutell 1990) (Figure 6.3b), and pseudoknots (Staple and Butcher 2005) (Figure 6.3c and d). The tertiary structure of RNA can be in the form of a triplex, where either a third nucleotide is Watson–Crick paired to a Hoogsteen duplex in the major groove (Toor et al. 2008) (Figure 6.4a) or in the minor groove, involving the addition of ribose 2′-hydroxyl interactions. It can also form a quadruplex, made of four G nucleobases in a "Hoogsteen ring" (Figure 6.4b) or by G–C or A–U pairs using a combination of Watson–Crick and noncanonical base pairing in the minor groove (Su et al. 1999). Other tertiary structure motifs include coaxial stacking (Quigley and Rich 1976), in which two RNA duplex form a contiguous helix; an A-minor motif

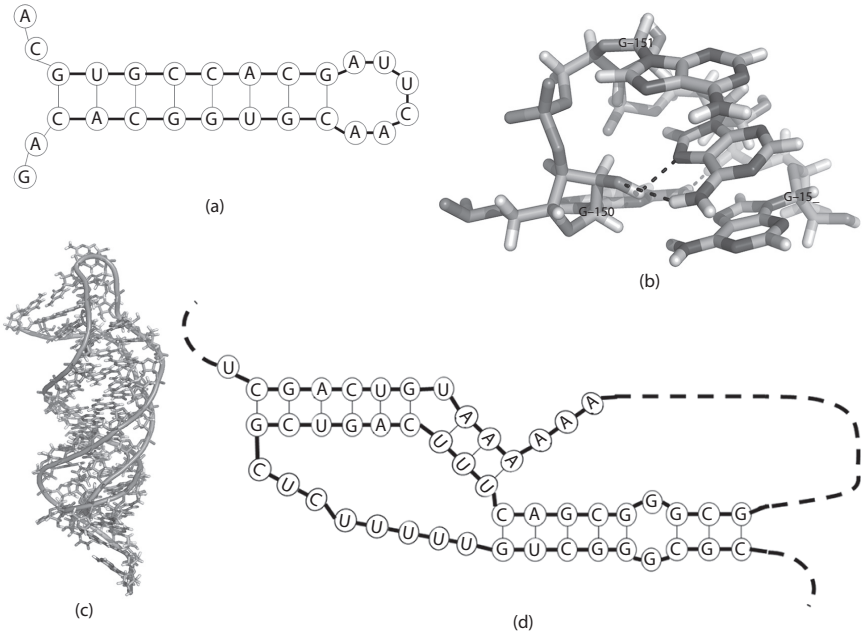
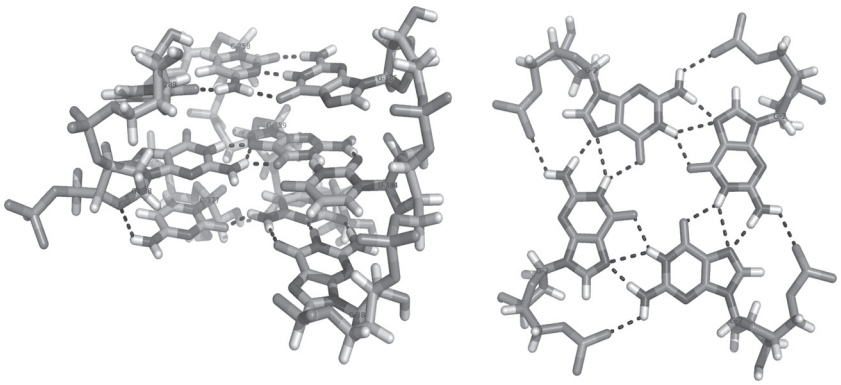


FIGURE 6.3 (a) Schematic view of a base pairing of a stem loop. (b) Tetraloop made of sequence GAAA, which is a hairpin loop of the P4–P6 domain of the *Tetrahymena thermophila* intron (PDB-ID: 1GID). This RNA molecule belongs to the class I self-splicing intron, being able to catalyze their own excision from an RNA precursor (Cate et al. 1996). (c) Stick representation of the 3D structure of an RNA pseudoknot, from the human telomerase (P2b-P3 fragment, PDB-ID: 1YMO). (d) Schematic view of a base pairing of pseudoknot from a human telomerase, showing the typical pattern made of two stem loops, in which half of one stem is intercalated between the two halves of the second stem. (From Chen, J-L and C W Greider, *Proceedings of the National Academy of Sciences of the United States of America*, 102, 8080–8085.)



(a) (b)

FIGURE 6.4 (a) Major groove triplex in the group II introns from *Oceanobacillus Iheyensis* (PDB-ID: 3BWP). Note the three layers of triplexes made of C377-C360-G383, C288-G359-U384, and C289-C358-G385. (b) Structure of a Hogsteen paired G-quadruplex from a synthetic tetraplex construct (UGGGGU, PDB-ID: 1RAU).

(Doherty et al. 2001), made of an unpaired nucleotide inserted into a minor groove of an RNA duplex; and ribose zipper that involves two RNA chains interacting through hydrogen bonds between the ribose 2′-hydroxyl atoms (Cate et al. 1996), which can act simultaneously as hydrogen bond acceptor and donor.

Due to the many combinations made possible through these structural motifs, RNA can serve many important biological functions. In a further challenge to the canonical role of RNA as the messenger between DNA and protein, high-throughput genomic analysis demonstrated that about 97% of the transcriptional output correspond to ncRNA sequences (Cech and Steitz 2014), which can either come from introns or RNA genes. The behavior and role of these molecules have not yet been fully elucidated indicating a significant role for computational modeling, which is the subject of the next section.

6.3 COARSE-GRAINED MODELING OF RNA MOLECULES

In an attempt to reduce the computational cost associated with multiple degrees of freedom in RNA modeling and to enhance sampling efficiency, several methods represent RNA molecules using different degree of detail/coarse-graining. Typically, most of the computational methods to predict RNA 3D structure can be divided into three families, according to the degree of "coarse-graining": (1) methods that focus on the coarse-grain

representation of the backbone, (2) those that focus on the nucleobase, and (3) others that consider both backbone and nucleobase. The selection of a particular type of coarse-grained method also requires choosing a potential energy model that describes interactions between particles (i.e., pseudoatoms or beads). In this section, we provide a qualitative review of the most important coarse-grained methods for RNA structure modeling along with their energy functions.

6.3.1 Backbone Models

When considering only the backbone to model RNA 3D structure, the simplest choice is to model each nucleotide with one pseudoatom (or bead) and pseudobonds connecting them. This yields the so-called one-bead model. Additionally, a second bead can be considered along the backbone in the two-bead models. One of the most well-known one-bead model is the Nucleic Acid Simulation Tool (NAST) (Jonikas et al. 2009), while Vfold applies two-bead model for RNA structure prediction (Cao and Chen 2005, 2009).

6.3.1.1 One-Bead Model: NAST

The NAST, developed by Altman and collaborators in 2009 (Jonikas et al. 2009), represents RNA nucleotides as single bead centered in the C_3' atom, with a Van der Waals radius of 4.0 Å (smaller radius would allow residues to intertwine within helical regions). A molecular dynamics (MD) engine (Alder and Wainwright 1959; Rahman 1964) is used to sample Cartesian degrees of freedom (see next section for more details). The potential energy function penalizes harmonic deviation from the ideal pseudobond length, angles, and tertiary contacts. In addition, there is a Fourier summation that accounts for the pseudorotation bonds and a repulsive Van der Waals term. Optimal parameters for the different energy terms are obtained in a two-step process. Probability distributions of distances, angles, and dihedrals, derived from a curated set of RNA structures, are fitted to ideal normal distributions (Figure 6.5). These ideal distributions are then converted to energies through the Boltzmann formula and fitted to the corresponding term in the potential energy, for which parameters are derived.

In addition to the primary nucleotide sequence as input, NAST requires the secondary structure information (base pairing) as well as available data about tertiary contacts (the latter increases performance significantly, since it reduces conformational space considerably). No information about hydrogen bonds is used, yet this method is able to find native-like structures

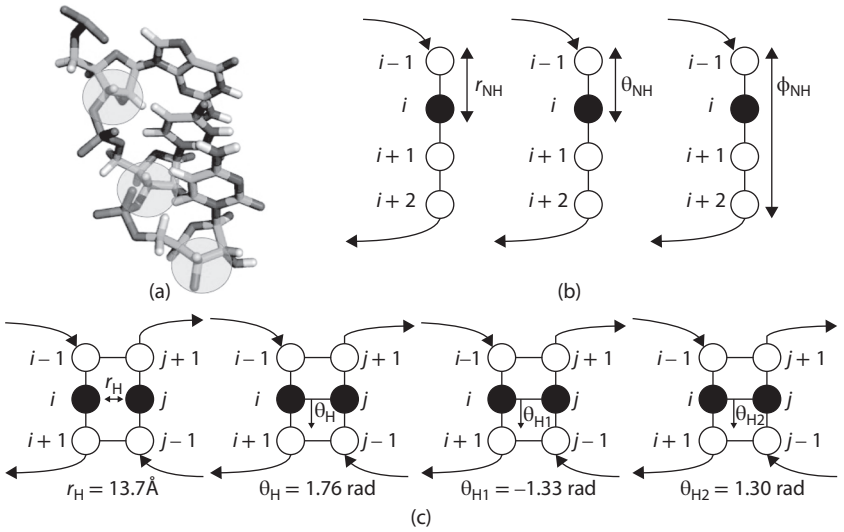


FIGURE 6.5 The NAST model and energy function. (a) Coarse-grained representation of RNA molecule, using one bead per nucleotide centered at C3′ atom. (b) Geometric restraints for nonhelical regions in the NAST energy function. (c) Geometric restraints used in helical regions.

that preserve global molecular shape. Typically, several clusters of structures are generated and rank ordered according to the experimental and predicted data on solvent accessible area, small-angle x-ray scattering (SAXS) data, and NAST energy (Jonikas et al. 2009).

The method is able to handle large RNA molecules, >100 nt. Predictions made for a 76-nt yeast phenylalanine tRNA and the 158-nt P4–P6 domain of *Tetrahymena thermophila* showed that NAST can attain structures as close as 8 Å to x-ray structure for the 76-nt tRNA, and up to 16 Å for the P4–P6 domain. This method can be used as a baseline or preliminary tool to produce a low-resolution model that can be further refined using more realistic energy functions.

6.3.1.2 Two-Bead Model

Still within the backbone coarse-grained representation, the resolution of the RNA model can be improved by using two beads to represent each nucleotide. The first two-bead mode, VFold, was developed by Cao and Chen (2005). VFold utilizes two pseudoatoms per nucleotide along the RNA backbone, placed on the phosphate and the C_4 sugar atoms, respectively (see Figure 6.6), since the bonds P–O_5–C_5–C_4 and the bonds C_4–C_3–C_3–P are approximately planar (Olson 1975; Olson and Flory 1972), they

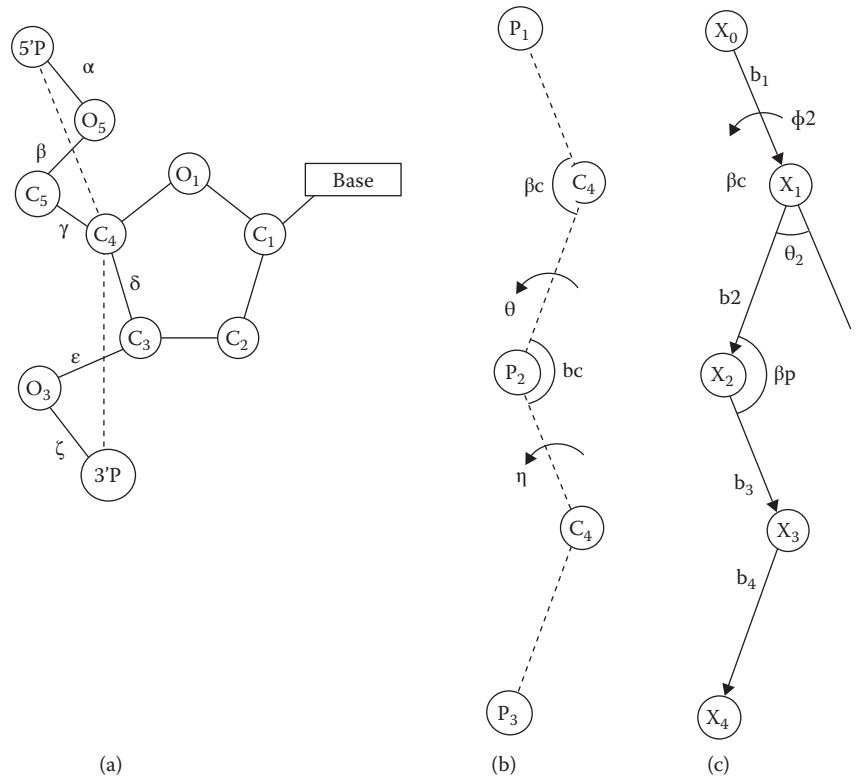


FIGURE 6.6 Coarse-grained model of the Vfold method. (a) Virtual bond scheme for nucleotide backbone. (b) The bond angles (βC, βP) and the torsional angle (η). (c) Vector model used to determine the atomic coordinates from the torsional angles.

can be considered rotameric (Murray et al. 2003). This allows us to use a limited number of degrees of freedom: two pseudobonds connecting P–C_4 of nucleotide *i* and C_4–P connecting nucleotide *i*+1, respectively, two pseudobond angles for the P–C–P and C_4–P–C_4 triads, and two dihedral angles through the P–C_4 and C_4–P bonds in nucleotide *i* and *i* + 1, respectively (Figure 6.6). An analysis of known RNA structure revealed a distance 3.9 Å for pseudobonds (Rich et al. 1961). The modeling of the different secondary structure elements utilizes distinct methods for each type of characteristic. Within helices, the pseudobond angles can be considered of $105 \pm 5°$ and $95 \pm 5°$ (RNA–A helix (Biswas, Mitra, and Sundaralingam 1998), while the torsion angles, θ, η, Duarte et al. (Duarte and Pyle 1998; Duarte, Wadley, and Pyle 2003) reported values close to 170°, 210°).

The Vfold method performs an exhaustive enumeration of all the possible conformations (torsion angles θ, η) compatible with a given nucleotide

sequence using the rotameric property of RNA. In contrast, the more flexible loops are modeled by self-avoiding walks on a diamond lattice, in which the bond length of the lattice equals that of the pseudobond (Flory 1969; Mattice and Suter 1994; Rapold and Mattice 1995). Mapping helix atoms onto the closest diamond lattice site makes joining between loops and off lattice helix nodes. In both cases, helices and loops, a translation from the previous pseudobond vector and composite rotation on the first pseudo rotation and bond angle generate the different conformations (Cao and Chen 2005). Some initial coordinates from reference RNA structure are used for the P–C_4–P and C_4–P–C_4 pseudobonds.

By employing this pseudobond/diamond lattice, this method can enumerate all the possible conformations of a given sequence to compute the minimized partition function of the system and obtain the lowest free energy conformation. Only base-stacking interactions are considered. They can be canonical base stack if both base pairs are any of A–U, G–C, or G–U, or base mismatch, if only one of the base pairs is one from the previous list. The base stack has effect on the enthalpy (including base pair mismatches) and entropy, while the nonstacked nucleotide (at loops) contributes only to the entropy.

The method can accurately predict thermodynamic properties such as heat capacity for different 58–59 nt RNA fragments (Cao and Chen 2005). In general, the VFold method is suited to predict secondary structure rather than tertiary structure, although the formalism in which it is based would allow the latter as well. VFold was later modified to accept an extra pseudobond (and an extra pseudoatom) between C_4 and N_1 (pyrimidines) or N_9 (purines) (Figure 6.7). This extra pseudobond accounts for the base orientation, so that more sophisticated secondary structure element such as pseudoknots up to 65 nt can be predicted accurately (Cao and Chen 2009).

6.3.2 Nucleobase Models (Inspired in Rosetta)

The Baker group has adapted their protocol to model protein structure *ab initio* (Rohl et al. 2004) and applied it to RNA. The fragment assembly RNA (FARNA) uses a library of 3-nt fragments and a low-resolution energy function to bias the Monte Carlo (MC) simulation process toward native RNA structures (Das and Baker 2007).

To avoid inclusion of fragments belonging to evolutionary related molecules, the library of fragments is taken from a single crystal structure, the large ribosomal subunit from *Haloarcula marismortui*, Protein Data

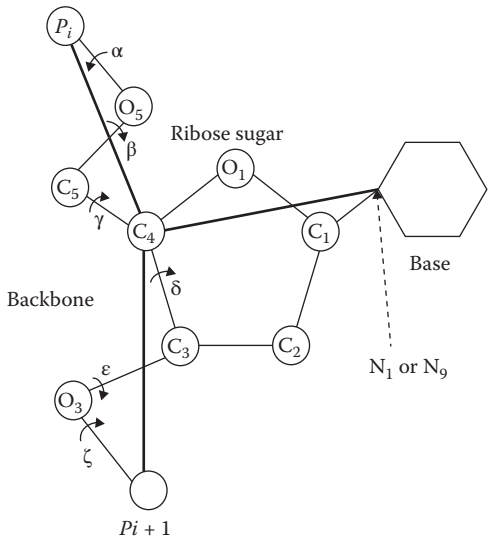


FIGURE 6.7 In addition to the two pseudobonds between P–C_4–P atoms, a third pseudobond is added to the Vfold model to better describe the base orientation, between the C_4–N_1 (pyrimidine) and C_4–N_9 (purine) atoms.

Bank identifier (PDB-ID) (Bernstein et al. 1977), 1FFK. Considering a two letter alphabet (purines/pyrimidines) yields more than 300 3-nt fragments.

The energy function includes several terms: (1) a term to ensure RNA compactness proportional to the radius of gyration, (2) a term to avoid steric clashes, (3) a term to reward Watson–Crick base pairing as a knowledge-based potential, (4) a term to enforce coplanarity between base pairs also as a knowledge-based potential of an orientation between two base pair planes, and a distance among the latter, and (5) an additional term to reward stacking of $-1k_BT$, where k_B is the Boltzmann constant. This energy function is derived from and applied to a coarse-grained representation of the RNA molecule, in which there is a Cartesian reference system placed on the geometric center of a given base pair. The x-axis goes through the N1 atom for the purines or the N3 atom for the pyrimidines; the z-axis is perpendicular to the base plane (Figure 6.8). Noncanonical base pairing is not included in the energy function, however, it is found to be recapitulated (Das and Baker 2007).

When applied to a set of 20 RNA structures up to 46 nt, most of them had Watson–Crick interactions predicted, and 11 of them showed root-mean-square distance (RMSD) <4.0 Å with respect to the native structure. In general, poor predictions in FARNA are the results of inaccuracies in the

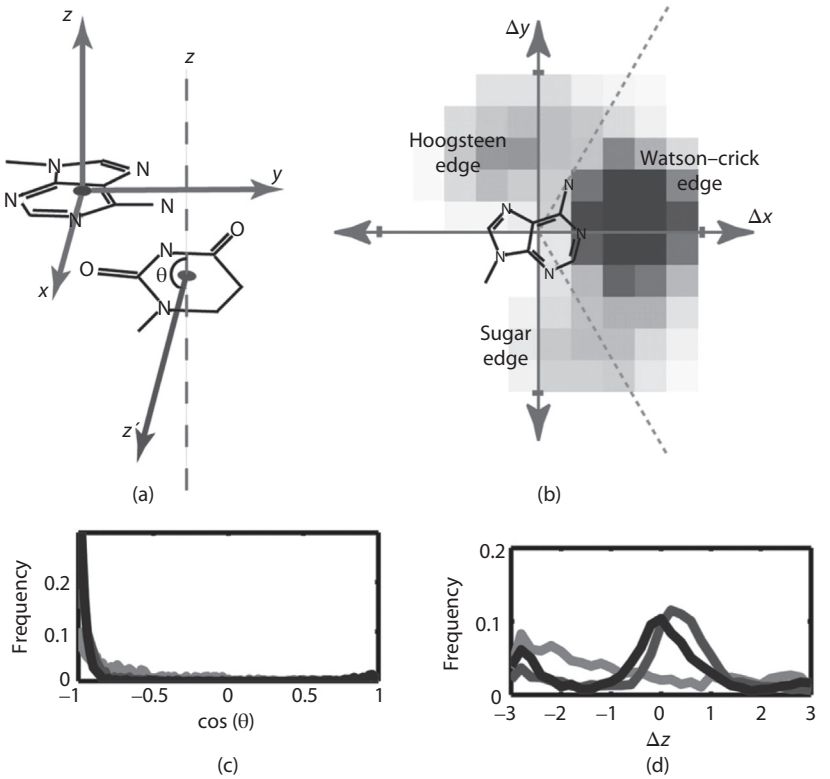


FIGURE 6.8 (a) The FARNA coordinate reference system. (b) Distribution of the Δx, Δy coordinates for a uridine residue in the vicinity of adenosine residues in the ribosome crystal structure. The logarithm of this distribution provides the knowledge-based potential. (c) Distribution of angles between base planes. (d) Distribution of relative stagger of base planes. Black line corresponds to the observed values, light gray, from the models built by FARNA without coplanarity, and dark gray, values considering coplanarity. (Reprinted from Das, R and D Baker. 2007. Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences of the United States of America* 104: 14664–14669. With permission from PNAS, Copyright [2007] National Academy of Sciences, U.S.A.)

energy function rather than a lack of sampling. In order to improve RNA models, the Baker group introduced a second refinement step using traditional all-atom Rosetta energy function (Rohl et al. 2004), which includes pairwise terms to account for van der Waals forces, hydrogen bonds, packing of hydrophobic groups interactions, and a penalty for burying hydrophilic groups. The entire protocol was named FARFAR (Full Atom Refinement FARNA). In addition to the previous terms, the latest updates from the Rosetta development community included a term to account for

the carbon–hydrogen bonds, an alternative orientation-dependent model for desolvation, and a term to describe the screened electrostatics interactions between phosphate groups (Das, Karanicolas, and Baker 2010).

The two steps procedure is able to model RNA structure up to an RMSD < 2.0 Å from native for 14 out of 32 RNA fragments of short lengths (6–23 residues). Failures in RNA structure prediction may be due to lack of proper sampling (Das, Karanicolas, and Baker 2010). Moreover, the same authors tested FARFAR on the reverse design problem, by redesigning the sequence (nucleobases) on 15 x-ray high-resolution RNA template structures with good recapitulation of Watson–Crick and non-Watson–Crick base pairing at 40% and 62%, respectively. Some of the divergent designed sequences were more stable than native; however, the fact that they were not selected by natural evolution suggests functional restrictions (interaction with other macromolecules, like proteins) that cannot be properly modeled by force fields.

6.3.3 Backbone—Nucleobase Hybrid Models

These models use several beads to describe atoms in both backbone and nucleobases, as to improve the accuracy of the results. Naturally, a rise in computational expense is the cost of using more sophisticated models. Here, we describe two of the most common hybrid models: the five- and seven-bead models.

6.3.3.1 Three-Bead Model

In the three-bead model, introduced by Dokholyan and collaborators (Ding et al. 2008), geometry of RNA molecule is described by three beads representing phosphate, sugar, and nucleobase (Figure 6.9). The beads are placed in the center of mass of the corresponding chemical moieties. To describe forces acting between atoms of RNA, two types of interactions are introduced: bonded and nonbonded. Bonded interactions are used to preserve connectivity and local geometry of RNA chain. The nonbonded interactions include base-pairing (A–U, G–C, and U–G), base-stacking, short-range phosphate–phosphate repulsion, and hydrophobic interactions. All the interactions are approximated by square-well potentials in order to make the model compatible with discrete molecular dynamics (DMD) sampling engine, as described in the next section. By choosing appropriate discretization potential step, the computer simulations can be accelerated. Due to the specific nature of DMD, all the potentials are expressed as a function of the interatomic distances. For example, to place a restraint

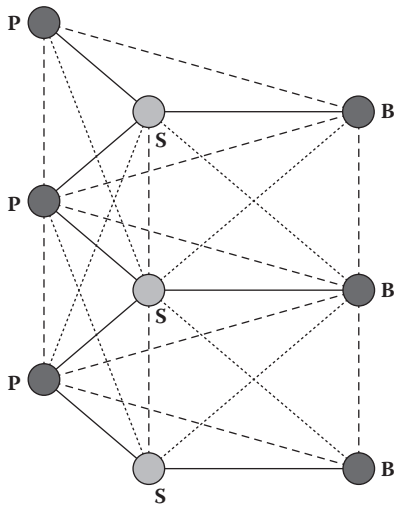


FIGURE 6.9 Three-bead model of RNA. The beads represent phosphate (P), sugar (S), and nucleobase (B). Covalent interactions are shown by thick lines, dashed lines show bond angular restraints, and dashed-dotted lines show dihedral restraints.

on a bond angle in the RNA backbone, three pairwise terms are required to restrict the atoms related by a certain bond angle. For bonded potentials, infinite square wells are used to preserve connectivity of a chain. The parameters of bonded interactions were derived from an RNA database with high spatial resolution (Murray et al. 2003). Base pairing interactions are modeled using a "reaction" algorithm (Ding et al. 2003), which allows formation of direction-dependent hydrogen bonds. Phosphate–phosphate electrostatic repulsion is modeled by Debye–Hückel potential (Debye and Hückel 1923). Parameters for the stacking and hydrophobic interactions were determined by sequence-dependent free energy decomposition for individual nearest-neighbor hydrogen bond (INN-HB) model (Mathews et al. 1999). An essential part of the simulation algorithm is explicit modeling of loop entropy. Due to the reduction of the degrees of freedom as a consequence of a coarse-grained nature of the model, the entropy of a system tends to be underestimated in the simulations, which leads to formation of long unnatural loops that trap RNA in nonnative conformations. To solve this problem, the authors use experimentally determined free energies of the loops, which were tabulated for loops of different sizes and types (i.e., hairpin, bulge, or internal loops) (Mathews et al. 1999). These values are used to guide formation of base pairs during the course of a simulation.

The model was validated on the set of 153 RNAs, ranging in length from 10 to 100 nt. The accuracy of the predictions was evaluated using two criteria: (1) RMSD between predicted and experimentally determined structure from the PDB and (2) percentage of native contacts formed in predicted structure. It was shown that for RNA molecules with length less than 50 nt, predicted RMSD is less than 6 Å, and for 85% of all predicted structures, the RMSD is less than 4 Å. The average percentage of native contacts predicted in the simulations was 94%. The performance of this model can be greatly enhanced with the use of experimental restraints, as described in Section 6.5.

6.3.3.2 Five-Bead Models

In the model developed by Ren and collaborators (Xia et al. 2010), every nucleotide is represented by five pseudoatoms: two pseudoatoms for the phosphate and the sugar and three pseudoatoms for the nucleobase (Figure 6.10). Three pseudoatoms describe well every nucleobase, each

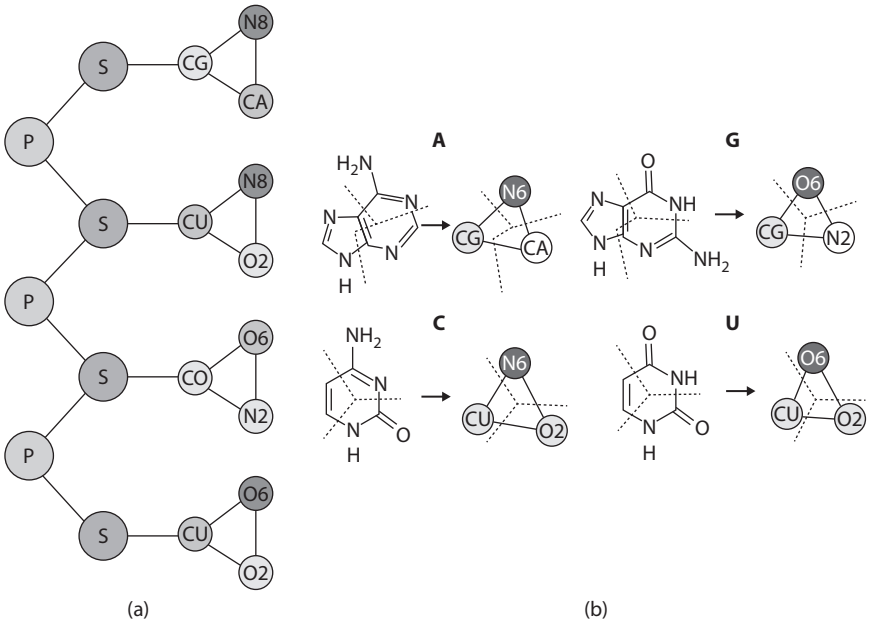


FIGURE 6.10 (a) Coarse-grained five-bead model of RNA. Phosphate (P) and sugar (S) are represented by one bead. Three beads are used to represent nucleobase. (b) Composition of the beads. The atoms of nucleobase are grouped into different pseudoatoms by lines to coarse-grained pseudoatoms. (Reprinted with permission from Xia et al., 13497–13506. Copyright 2010 American Chemical Society.)

representing a certain group of real atoms in the nucleobase. It was found that five different types of pseudoatoms are enough to describe all four nucleobases, since the same groups of atoms are present in different nucleobases. Potential energy of the coarse-grained field consists of two terms, corresponding to bonded and nonbonded interactions. Bonded interactions are described by the sum of bond stretching, bond angle bending, and dihedral angle rotation terms:

$$E_{\text{bonded}} = K_{\text{bond}}\left(b - b_0\right)^2 + K_a\left(\theta - \theta_0\right)^2 + \sum_{n=1}^{3} K_n\left[1 + \cos\left(n\phi - \delta_n\right)\right]$$

Here, b is a bond length, θ is a bond angle, and ϕ is a dihedral angle. In total, there are 14 different types of bonds, 25 types of bond angles, and 28 types of dihedral angles, each corresponding to their own set of parameters. These parameters were extracted from statistical distributions, fitted to a normal distribution, of bond lengths and bond and dihedral angles obtained from a set of 668 RNA structures from the PDB. The Buckingham potential describes the nonbonded interactions as

$$E_{\text{nonbonded}} = \varepsilon_{ij}\left[-2.25\left(\sigma_{ij}/r_{ij}\right)^6 + 1.84 \times 10^5 e^{-12.00\left(r_{ij}/\sigma_{ij}\right)}\right]$$

Here, ε_{ij} is the depth of potential well, σ_{ij} is the radius, and r_{ij} is the distance between the atoms. There are 19 pairs of (σ_{ij}, r_{ij}) parameters describing nonbonded interactions between different types of pseudoatoms. The values of these parameters were derived upon least square fit of the Buckingham potential to the following potential of mean force:

$$E_{\text{nonbonded}}(r) = -k_B T \ln g(r)$$

Here, $g(r)$ is a radial distribution function (RDF) defined as

$$g_{ij}(r) = \frac{1}{N_i d_j}\frac{n_{ij}(r)}{4\pi r^2 \delta r}$$

Here, n_{ij} is a number of pairs at the distance from r to $r + \mathrm{d}r$, N_i is the total number of type i particles in the system, and d_j is the mean density of type j particles. All the 19 RDFs were calculated using the same data set of RNA structures as was used to determine the parameters for bonded interactions. The nonbonded interactions were further refined on the set of seven structurally diverse RNA by minimizing the difference between

predicted and experimental structures. This method was tested on a set of 15 RNA molecules with the length of 12–27nt. MD simulations with simulated annealing, led 75% of tested RNA molecules to native-like structures at least once during the course of the simulation. With imposed restraints on secondary structure content, all 15 RNAs were folded to within 6.5 Å relatively to native structures.

In an alternative five-bead model by Levitt (Bernauer et al. 2011), the beads are placed at the positions of P and C_4' backbone atoms, and at the positions of C_2, C_4, and C_6 nucleobase atoms. The authors define statistical potential as

$$E = -kT \sum_{ij} \ln \left(\frac{P_{\text{obs}}\left(d_{ij}\right)}{P_{\text{ref}}\left(d_{ij}\right)} \right)$$

Here, $P_{\text{obs}}\left(d_{ij}\right)$ is a probability to find pseudoatoms of types i and j at the distance d, calculated from the set of RNA structures from the PDB. On total, 77 RNA structures were used, chosen to have high experimental resolution, and to be nonredundant. $P_{\text{ref}}\left(d_{ij}\right)$ is a reference probability calculated from overall mole fractions of atom types i and j (Lu and Skolnick 2001). The authors used their model to identify native-like RNA structure from a set of near-native models (decoys). Three sets of decoys were generated using position-restrained MD, normal modes, and assembly from RNA-like fragments, which had no native base-pairing enforced. The performance of the model was evaluated by counting the number of decoys with energy lower than the energy of the native structure. Out of five RNAs used to generate decoys with position-restrained MD. For one of the RNA structures, only one decoy was found to have lower energy than native structure. Out of 15 RNAs, used to produce decoys with normal modes, the native structures had the lowest energy for 14 RNAs, and only for one RNA 11 decoys were scored better than the native structure. The most difficult task was to distinguish native structure from the set of decoys produced from RNA fragments. Out of 19 tested RNAs, only in nine of them the native structure was scored as having the lowest energy. In each of other 10 RNAs, the number of decoys scored better than native structure was close to 500.

6.3.3.3 Seven-Bead Models

HiRE-RNA (Pasquali and Derreumaux 2010) is a coarse-grained model, which relies on seven-bead representation of a single nucleotide: one bead for phosphate, four beads for sugar, O_5', C_5', C_4', C_1', one bead for the pyrim-

idine bases (C, U), and two beads for the purine bases (G, A). The beads representing pyrimidine and purine bases are placed in the centers of mass of nonhydrogen atoms of the corresponding aromatic rings (Figure 6.11). The HiRE-RNA force field is represented by a sum of local, nonbonded, and hydrogen-bonding terms. The local interaction term is expressed as

$$K_b\left(r - r_{\text{eq}}\right)^2 + \sum_{\alpha} K_\alpha\left(\alpha - \alpha_{\text{eq}}\right)^2 + \sum_{d} K_d\left[1 + \cos(\tau - \gamma)\right]$$

$$E_{\text{local}} = \sum_{b}$$

Nonbonded and hydrogen bond interactions are expressed as

$$E_{\text{nonbonded}}(r_{ij}) = \varepsilon_{ij}\left[\left(\frac{G\left(\sigma_{ij}\right)}{r_{ij}}\right)^6 e^{-2r_{ij}} + \frac{3}{5}\tanh\left[2\left(r_{ij} - \sigma_{ij} - \frac{1}{2}\right) - 1\right]\right]$$

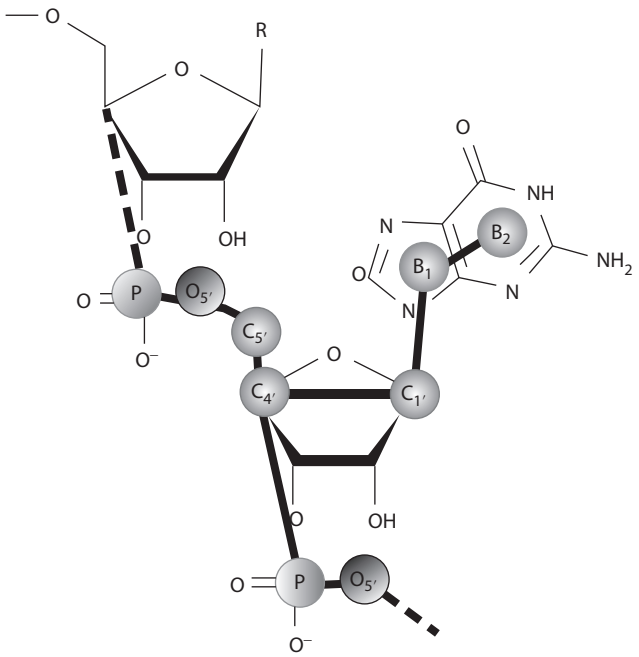


FIGURE 6.11 Coarse-grained representation of seven-bead model. Phosphate (P), O5′, C5′, C4′, and C1′ are represented by single beads. Beads for purine (two beads) and pyrimidine (one bead) are placed in the centers of corresponding aromatic rings. (Reprinted with permission from Pasquali, Samuela, and Philippe, 11957–11966. Copyright 2010 American Chemical Society.)

Here, r_{ij} and σ_{ij} are instantaneous and equilibrium distances between atoms, ε_{ij} is a coupling constant, and $G(\sigma_{ij})$ is a numerical parameter depending on equilibrium distance between i and j. Hydrogen-bonding interactions are described by a sum of two-body, three-body, and four-body terms. Recently, the authors of HiRE-RNA revised their force field (Cragnolini et al. 2014), introducing new features, in particular noncanonical base pairing and the possibility for a base to form multiple contacts. The performance of the improved force field was evaluated on a set of seven topologically different RNAs ranging in length between 22 and 79 nt. For three modeled structures, including two hairpins (PDB-IDs 1F9L and 1N8X) and a duplex (PDB-ID 433D), the reached precision was 2–3 Å as compared to the x-ray structure. For the triple helix (PDB-ID: 2K96) of 49 nt, the precision was 4.3 Å. For two pseudoknots (PDB-IDs: 2G1W and 1L2X) of 22 and 28 nt, respectively, the experimental structure was predicted as the most stable structure, however, the authors have not reported obtained RMSD. The last predicted structure was the riboswitch (PDB-ID: 1Y26) of 79 nt. This prediction was done using secondary structure restraints, resulting in a precision of 7–8 Å for the best models.

6.4 SAMPLING METHODS

Simulating RNA, in order to predict its native conformation, ideally involves the exploration of the full conformational landscape. This is almost unreachable using the current state-of-the-art computers. Most common sampling algorithms aim at generating a representative number of conformations to allow a reasonable estimation of thermodynamic and structural properties. In this section, we summarize the basic working principles behind those sampling algorithms used for modeling RNA structures, i.e., MD, MC, and DMD simulations.

Note that, to enhance the sampling efficiency and rapid exploration of the conformational space overcoming big energy barriers, these sampling methodologies often use a variety of different enhancement strategies, such as umbrella sampling (Torrie and Valleau 1977) and steered MD (Patel et al. 2014).

6.4.1 MD Simulations

MD simulations intend to compute the real dynamics of a given system. Regardless of the model used to account for interatomic particle interactions of the system, whether all-atom or coarse-grained, the positions

and velocities of each particle in the system are calculated by integrating Newton's equations of motion numerically (discretized into small time steps of the order of 1 fs). At each time step, the forces acting on each particle are computed from the gradient of the potential energy. The corresponding accelerations derived from those forces are then used to update positions and velocities of each particle at current step. New forces are then computed and the process iterates for the entire simulation trajectory. Forces are considered constant within an integration step. At the beginning of the simulation, the velocities are randomly assigned to each particle according to a Maxwell–Boltzmann distribution that depends on the simulation temperature. The potential energy between two interacting atoms typically contains bonded and nonbonded terms (Laurendeau 2005; McCammon, Gelin, and Karplus 1977; Gelin and Karplus 1975). The bonded terms usually include three harmonic potential terms: bond, angle, and torsional angles. The harmonic potential is applied to each atom with respect to the equilibrium values of bond distances and bend angles, and a Fourier summation for the torsional angles. Nonbonded interactions in all-atom simulations are typically the sum of the electrostatic and van der Waals interaction functions. Notably, analogous interactions are developed in case of more coarse-grained models. Several methods reviewed in Section 6.3 adopted an MD-based sampling method combined with a coarse-grained potential for modeling RNA 3D structure: NAST (Jonikas et al. 2009), five-bead model (Xia et al. 2010), and HiRe-RNA (Pasquali and Derreumaux 2010).

6.4.2 MC Simulations

In MC simulations, a new conformation is generated by randomly changing one or several degrees of freedom of the system. At this point, the potential energy is calculated and compared with the energy of the conformation at the previous time step. If the energy of the new conformation is lower than the previous one, the current conformation is accepted. If the new energy is higher than its predecessor, then a random number between 0 and 1 is generated and compared to the computed Boltzmann factor $\exp\left[-\left(E_{i+1} - E_i\right)/k_B T\right]$. If the random number is lower than the Boltzmann factor, the new conformation is accepted otherwise it is rejected (Metropolis acceptance criterion, Metropolis and Ulam 1949). Generating system trajectories using an MC method is computationally less demanding than using MD, requiring only the energies of the system and not the forces.

However, MD provides time dependence of the variables of the system, while in MC, each move attempt depends exclusively on its predecessor.

Das and collaborators (Das and Baker 2007; Das, Karanicolas, and Baker 2010) developed a MC-based sampling approach, within the Rosetta family of methods to sample the conformational RNA conformational space and find the native conformation as described in Section 6.3.2.

6.4.3 DMD Simulations

DMD simulation is a special case of a generalized MD simulation, in which the particles move with constant velocity upon no interaction, and change velocities upon interaction and/or elastic collision so that total energy and momenta are preserved (Alder and Wainwright 1959). Elastic collisions between two interacting atoms take place when the increase in the potential energy cannot be compensated by a decrease in the kinetic energy (Figure 6.12). After a collision, the velocities and positions are updated for the colliding atoms, keeping the potential energy constant and generating a new list of putative collisions with the neighbor atoms is generated. The consecutive repetition of these events yields a complete trajectory.

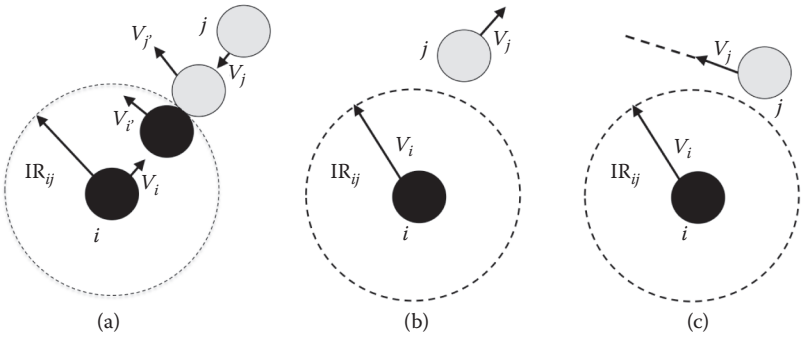


FIGURE 6.12 DMD interactions and collisions. (a) When two particles approach each other within the interaction range, IR_{ij}, higher than the minimum distance for hard-core repulsion and lower than the maximum interacting distance, they interact. Whenever within the interaction range, if two atoms are approaching, they will experience hard-core repulsion (collision) if the decrease in the kinetic energy does not compensate for the increase in the potential energy. Upon collision, two particles with divergent velocities will not interact (b). Another example of noninteraction is the case in which the particles are approaching but not intersecting the interaction range (c). (With kind permission from **Springer Science+Business Media**: *Computational Modeling of Biological Systems: From Molecules to Pathways*, Discrete molecular dynamics simulation of biomolecules, 2012, 55–73, Ding, F and N V Dokholyan. Copyright 2012.)

Besides the "event-driven" nature of DMD, it has been successfully parallelized so that it can be used in combination with replica exchange algorithm (Earl and Deem 2005), where different simulations are run in parallel at different temperatures. In order to overcome potential energy barriers, a given pair of replicas can exchange temperature according to a Metropolis criterion. The Dokholyan group has developed a tool to model RNA native structures that integrates replica exchange DMD sampling with a three-bead model representation of the RNA molecule and a discrete stepwise interaction potential. This tool has proven to be computationally efficient and generates reliable native-like structures of RNA (see previous section for details) (Ding et al. 2008; Sharma, Ding, and Dokholyan 2008).

What makes DMD particularly interesting is the fact that it does not require the computation of the gradients of the potentials to obtain forces. Thus, it does not have to integrate any differential equation, allowing longer time steps and simulation runs as compared to conventional MD, at lower computational cost. Also the discrete nature of the potential energy function makes calculations much faster than a traditional MD simulation of the same time length.

6.5 GUIDING CONFORMATIONAL SAMPLING BY MEANS OF EXPERIMENTAL RESTRAINTS

Even with efficient sampling algorithms, the conformational space is so large that it becomes almost impossible to exhaustively sample it all. An efficient way to improve the quality of the predictions is by imposing restraints (typically distance restraints) between certain atom pairs, according to experimental data. By doing so, simulations become more reliable in overcoming force field limitations and are more likely to converge to the native state.

There are several experimental techniques that can provide restraints between atom pairs in the form of a pairwise potential to improve RNA structure prediction methods. In this section, we describe the most common techniques. It is not the purpose to deeply review them all here, but rather provide an introduction and appropriate reference to the original works to the interested reader.

6.5.1 Most Common Experimental Methods to Derive Distance Restraints

As for protein structure determination, the two most well-developed techniques to determine RNA 3D structures are x-ray crystallography

and NMR. Since x-ray crystallography builds the entire 3D model for the RNA molecule under consideration upon analysis of the diffraction pattern (Kendrew et al. 1958), it will not be considered here as a mean to provide experimental restraints. However, one can derive knowledge-based potentials on the deposited structures in the PDB or use a particular solved structure as a template in homology modeling to build the structure of a homologous RNA piece (Flores et al. 2010). Despite the intrinsic limitations to solve NMR spectra (see Section 6.1), computer simulation is needed to build up the final model with the torsion angles derived from J-coupling constants and the distances from the Nuclear Overhauser Effect SpectroscopY (NOESY) cross peaks (Addess and Feigon 1996), taking into account that the data to parameter ratio is much poorer than in x-ray crystallography. In particular, it is not necessary to resolve the entire NOESY spectrum in order to model RNA structure, using few distance restraints is enough for DMD to attain conformations close to the native (Tandon et al. submitted 2014).

AQ1: Tandon et al. (2014) is not listed in the References list. Please provide complete reference details or remove text citation.

In the recent years, a set of chemical probing methods have been developed to extract structural features on RNA by exploiting the chemical properties of its nucleotides. The most used chemical reagents comprise a mixture of H_2O_2, ascorbic acid, and Fe(II)-ethylenediaminetetraacetic acid (EDTA) to generate hydroxyl radicals; dimethyl sulfate (DMS), 1-cyclohexyl-(2-morpholinoethyl) carbodiimide metho-*p*-toluenesulfonate (CMCT), kethoxal; and the **s**elective 2′-**h**ydroxyl **a**cylation analyzed by **p**rimer **e**xtension (SHAPE) method. We will treat more deeply the hydroxyl radical probing (HRP) and SHAPE methods since they have been able to produce reliable results in conjunction with DMD sampling and three-bead model combined method (Ding et al. 2012; Gherghe et al. 2009).

DMS modifies nucleobases by methylating N_1 atom in the adenosine and N_3 at cytosine. Those chemical modifications can only take place in single-stranded RNA, base paired at the end of a helix or a base pair next to a G–U. The modified RNA is then analyzed through reverse transcriptase polymerase chain reaction (RT-PCR) (Freeman, Walker, and Vrana 1999), and run on a gel. Since modified sites cannot be base paired, they produce different size bands, as the polymerase fails to add a nucleotide in this particular position. Protected positions can be due to base pairing, tertiary contacts, or interactions with other macromolecules. The band pattern generated by DMS treated can thus be used to infer RNA secondary structure (Tijerina, Mohr, and Russell 2007).

Similarly, CMT also reacts to exposed N_3 atoms in uridine or N_1 atoms in guanine, preventing base pairing, so that the unprotected sites correspond to single-stranded RNA, and the protected to double stranded. The RT-PCR and gel allow also to assign secondary structure (Fritz et al. 2002). The kethoxal (1,1-dihydroxy-3-ethoxy-2-butanone) reagent forms a covalent adduct with adenosine upon reaction on N_1 and N_2 atoms that prevents base pairing, so that when the reverse transcriptase reaches the modified position, it falls off yielding a band in the gel (Gopinath 2009). By combining the three chemical probing, DMS, CMT, and kethoxal, it is possible to gather data about base pairing and/or tertiary contacts.

AQ2: Please provide the expansion of CMT.

Haptoglobin-related protein (HPR) consists of the reaction of hydroxyl radicals with RNA, by nucleophilic attack on the ribose, breaking the phosphate backbone. Hydroxyl radical are chemically unstable, but can be generated using a mixture of H_2O_2, ascorbic acid, and Fe(II)-EDTA. Upon digestion, the reaction to those radicals RNA molecules is cleaved and then amplified by RT-PCR. Since the last nucleotide remains intact upon cleavage, the recovered band length on a gel is informative about the last nucleotide cleaved. The different protection strength against HRP does not depend on base pairing (secondary structure) since the backbone is not involved, but rather on tertiary contacts (Karaduman et al. 2006). In Section 6.5.3, we explain how to incorporate the data on HRP accessibility to bias RNA structure modeling, in the context of DMD.

The SHAPE method uses some reactants such as N-methylisotoic anhydride (NMIA) and 1-methyl-7-nitroisatoic anhydride (1M7) that make covalent adducts with RNA upon reaction with flexible 2′ hydroxyl group. Adduct formation at 2′OH position is quantified by extension of a complimentary DNA using reverse transcriptase and two primers flanking the RNA fragment upon consideration. The 2′-O modified nucleotide stops reverse transcriptase one nucleotide before. By comparing the band profile upon reaction to either NMIA or 1M7 with respect to a nonreactive control and to a sequencing control for nucleotide identification in the same gel, one can get precise data about secondary structure (Merino et al. 2005). In Section 6.5.2, we discuss the use of SHAPE data as restraints in DMD simulation to model 3D RNA structure.

6.5.2 Secondary Structure from SHAPE in Addition to Long-Range Restraints

SHAPE chemistry can be used to determine RNA base pairing very accurately, since 2′-OH reactivity depends whether bases are paired and/or

making tertiary contacts. Secondary structure is typically assigned using SHAPE reactivity data in combination to secondary structure prediction software. In brief, the reactivity data (Figure 6.13) can be converted into a free energy profile (Deigan et al. 2008) that can be input to dynamic programming algorithms to predict RNA secondary structure (Mathews et al. 2004). In the case of the $^{t}RNA^{Asp}$, SHAPE data in combination to dynamic programing produced the correct secondary structure, as displayed in Figure 6.14. Base pairing is implemented as restraints in a multibody energy potential that accounts for both distance and orientation restraints between base–base, base–sugar, and base–phosphate pseudoatoms (Figure 6.15), in addition to a soft attractive up to 50 Å term between bases. Distances are taken from statistics on known RNA structures.

A particular derivation of the HRP chemistry, named tethered hydroxyl radical probing (t-HRP) can be used to infer long-range contact information and derive pairwise distance restraint potential energy (Gherghe et al. 2009). Pairwise interactions are generated between a Fe(II)-EDTA moiety covalently bound between GpC steps adjacent to a single nucleotide bulge upon reaction to methydiumpropyl-EDTA (MPE) (Figure 6.16). In the context of the DMD, a pairwise step potential can be set up with a minimum energy in the range of 25 Å to the tethered position, whose depth is assigned from the normalized t-HRP reactivity relative intensity (as compared to nontethered RNA molecule), interpreted as the probability of cleavage within the cutoff distance. Some soft penalties can extend the interaction range up to 35 Å (Figure 6.17). Long-range restraints are incorporated into DMD simulations in a second step after base pairing (see Figure 6.18 for the DMD protocol). The last stage involves clustering and analysis of the most populated cluster. The method has been applied to the refinement of $^{t}RNA^{Asp}$ structure up to 4 Å in average of the crystal structure (from the most populated cluster), when one or several sets of MPE (differing in the nucleotide tethered) are used (Gherghe et al. 2009). However, in the absence of long-range restraints, the closest model attained is at 11 Å from the crystal structure.

In a more general context, the combined use of DMD with restraints from tertiary contact (between 2 and 7) data available from different experiments plus secondary structure (base pairing) data yielded high-resolution models for four RNAs: domain III of the cricket paralysis virus internal ribosome entry site (CrPV) (49 nt), a full-length hammerhead ribozyme from *Schistosoma mansoni* (HHR) (67 nt), *Saccharomyces cerevisiae* $^{t}RNA^{Asp}$ (75 nt), and the P546 domain of the *T. thermophila* group

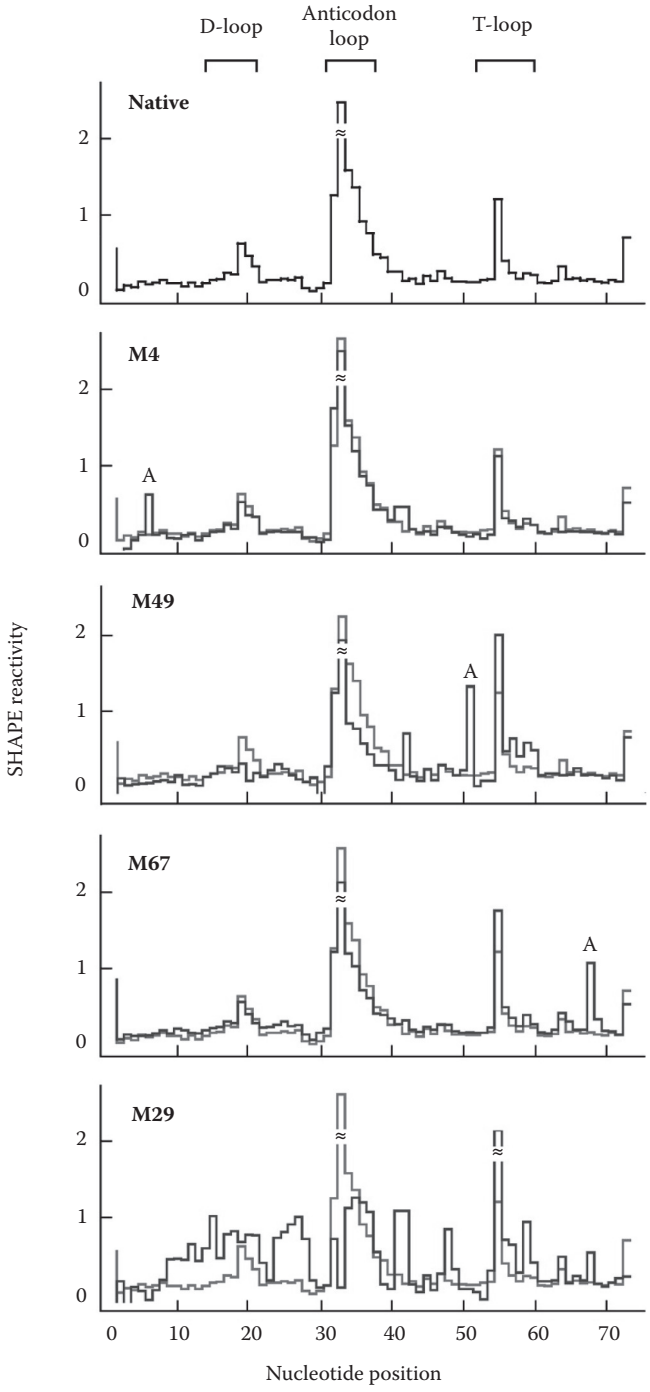


FIGURE 6.13 SHAPE reactivity profile for the $^{t}RNA^{Asp}$. The SHAPE profile tells us the compatibility of the MPE binding sites with $^{t}RNA^{Asp}$ tertiary structure. Here the profile for each mutant is shown as dark bar plot, superimposed to the profile on the wild type in gray. The position where the bulged adenosine is inserted is indicated. (Reprinted with permission from Gherghe et al., 2541–2546. Copyright 2009 American Chemical Society.)

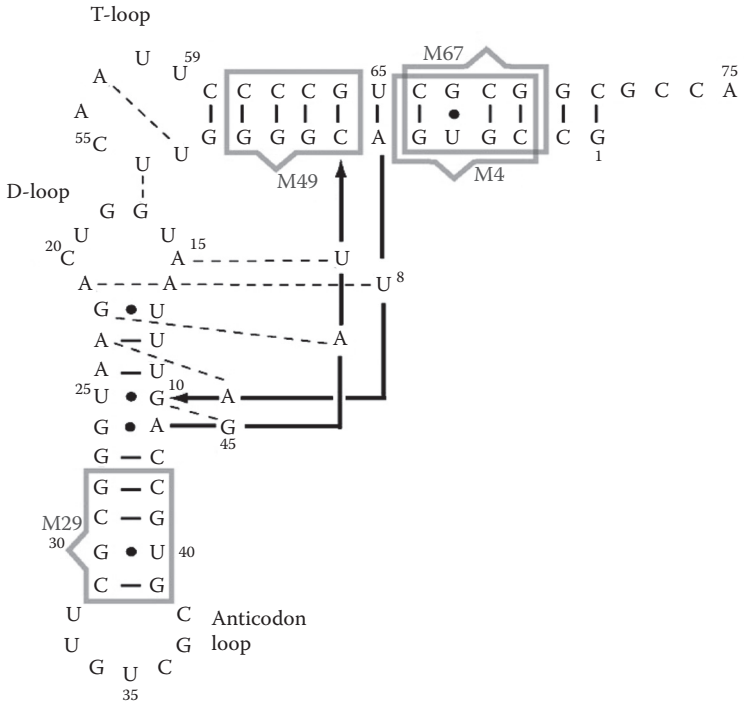


FIGURE 6.14 Secondary structure for the $^{t}RNA^{Asp}$, showing the different mutants, which are numbered according to the closest residue to the tethered Fe(II)-EDTA group. (Reprinted with permission from Gherghe et al., 2541–2546. Copyright 2009 American Chemical Society.)

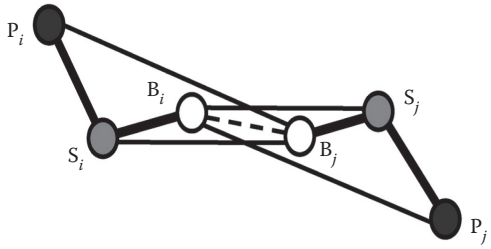


FIGURE 6.15 Three-bead model indicating the different pairwise interactions between: Bases, B_i–B_j or base pairing; nucleotide orientation, defined by the base–sugar interactions, B_i–S_j and B_j–S_i; and the base–phosphate interaction B_i–P_j and B_j–P_i, which define RNA rigidity. (Reprinted with permission from Gherghe et al., 2541–2546. Copyright 2009 American Chemical Society.)

I intron (P546) (158 nt), at 3.6, 5.4, 6.4, and 11.3 Å, respectively, from the crystal structure (Lavender et al. 2010). There the restraint potentials were set generically as 2 kcal/mol well depth and 15 Å width.

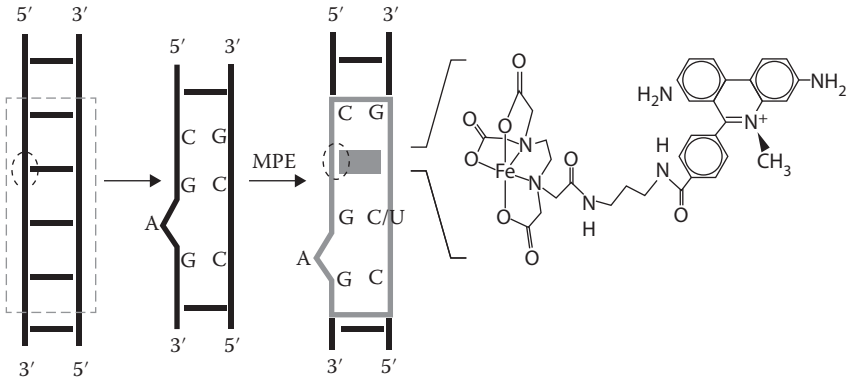


FIGURE 6.16 Intercalation of PME in a bulged helix to replace four canonical base pair. MPE has its Fe(II)-EDTA moiety oriented toward the bulged adenosine nucleotide. (Reprinted with permission from Gherghe et al., 2541–2546. Copyright 2009 American Chemical Society.)

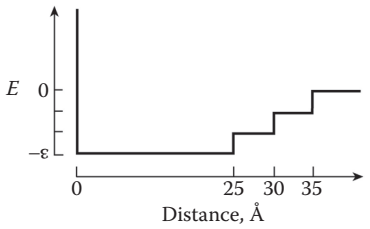


FIGURE 6.17 Discrete energy potential applied to the long-range interaction derived from the site-directed hydroxyl chemistry. (Reprinted with permission from Gherghe et al., 2541–2546. Copyright 2009 American Chemical Society.)

6.5.3 Incorporating HPR Data as a Restraints into DMD Simulations

Traditional HRP chemistry reactivity profile, on the other hand, provides information about solvent accessibility for the backbone atoms, meaning that low accessible backbone atoms are making high number of interatomic contacts (Figure 6.19a). For a particular nucleotide, there is thus a negative correlation between the number of atoms in contact at a given distance cutoff and its HRP reactivity (Figure 6.19b). The Dokholyan group implemented a procedure, based on the three-bead model, which maximizes the number of contacts between sugar beads according to the data derived from HRP, in DMD simulations. The biasing potential is made of two terms, one generically attractive to most of the nucleotide pairs, ensuring RNA collapse and packing, and a second term that is repulsive once a

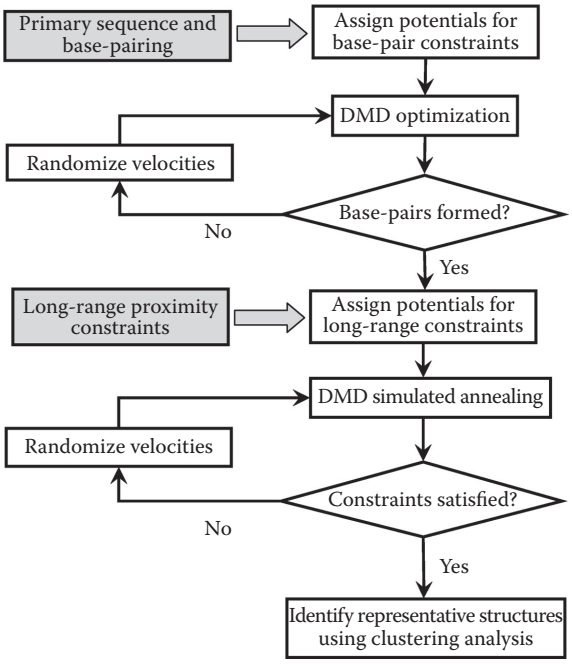


FIGURE 6.18 DMD protocol for RNA structure refinement, including of the secondary structure data from SHAPE and long-range restraints from site-directed hydroxyl radical cleavage. It starts from the RNA sequences and the base pairing data. Once the base pairing is satisfied, the weighted potentials to account for the long-range interactions are added. The system is then cooled down. Upon fulfillment of the latter restraints, there is a final equilibration run. (Reprinted with permission from Gherghe et al., 2541–2546. Copyright 2009 American Chemical Society.)

given nucleotide has reached its maximum number of assigned contacts as derived from the HRP reactivity profile. If a given nucleotide achieved the maximum number of contacts, it will only accept a new one if the kinetic energy during DMD simulation overcomes the penalty energy barrier of overpacking (Ding et al. 2012) (Figure 6.20a).

The entire DMD protocol is made of three steps. Initially, conventional DMD simulations are performed to ensure base pairing. A second round, involves replica exchange DMD simulations and HRP restraints as explained above. The second step expects to have an enriched population of structures in agreement with the experimental data (Figure 6.20b). At the end, the limited number of trajectory snapshots is selected based on their energy and correlation between number of contacts and HRP reactivities. These structures are clustered according to the RMSD. Representative

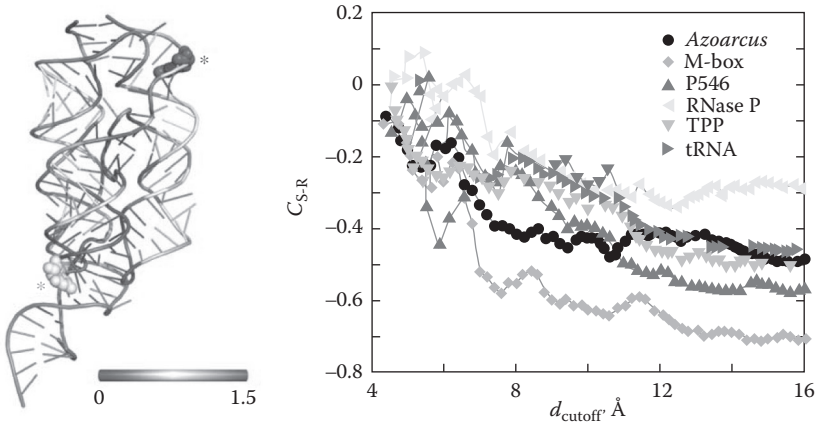


FIGURE 6.19 (a) Structure of the M-box riboswitch shaded according to HRP reactivity. Dark zones correspond to low HRP reactivity and are buried, while light areas are rather solvent exposed. However, two residues are displayed as spheres that correspond to a solvent accessible residue with high HRP reactivity (light spheres) and a buried residue with low HRP reactivity (dark spheres). (b) Pearson correlation coefficient between the HRP reactivity and the number of atomic contacts $C_{S\text{-}R}$ as a function of the distance cutoff. Best correlation in absolute values appears to be in the range of 14–16 Å. (Reprinted by permission from MacMillan Publishers Ltd. *Nature Methods*, Ding et al., 9 (6): 603–8, copyright 2012.)

structures of the most populated clusters tend to reproduce native RNA structures, even for large RNA fragments up to 230 nucleotide (Ding et al. 2012).

6.6 COMPARISON OF THE PERFORMANCE OF THE DIFFERENT METHODS: RNA PUZZLES

In the recent years, several computational techniques have been developed for the *de novo* 3D structure prediction of middle-sized RNA sequences (i.e., up to a few hundreds). Therefore, it has become essential to benchmark the quality and the performance of the current methods and tools for structural predictions. In order to meet this need of the scientific community, Dr. Westhof and colleagues have established RNA Puzzles (http://paradise-ibmc.u-strasbg.fr/rnapuzzles/index.html), a collective experiment for blind RNA 3D structure prediction. The main aim of RNA Puzzles is the comparison of strengths and limits of the existent methodologies for modeling RNA structures in order to assess the status of research in the field and promote its further development (Cruz et al. 2012).

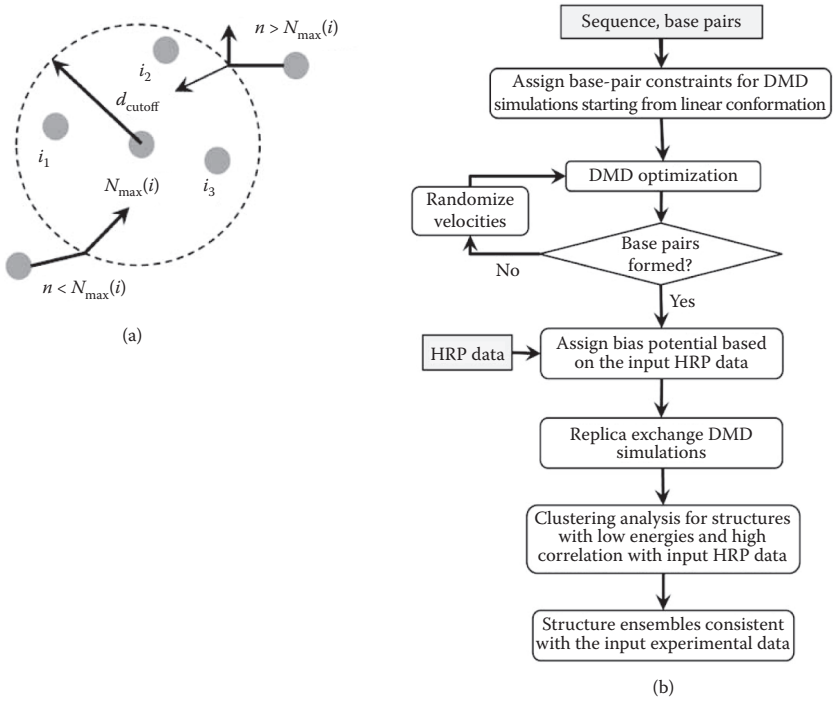


FIGURE 6.20 Incorporating HRP reactivities into energy potentials in DMD simulations. (a) A number of maximum (N_{max}) contacts are assigned to each nucleotide according to the cutoff distance. A nucleotide approaching a given one i can form new contact if the number of neighbors (denoted by $i_1, i_2, \ldots$) is smaller than N_{max}. If it is bigger, a new contact can form only if the total kinetic energy is sufficient to overcome the overpacking energy penalty. (b) Schematic representation of the flow chart for the HRP-directed DMD simulation protocol. (Reprinted by permission from MacMillan Publishers Ltd. *Nature Methods*, Ding et al., 9 (6): 603–8, copyright 2012.)

In RNA Puzzles, crystallographic RNA structures are provided to the structural biology community in order to challenge several research groups to generate accurate 3D structural models. Starting from 2011, eight exercises (i.e., puzzles) have been proposed (Figure 6.21), two of which are currently open.

The quality of each submitted model is evaluated using the RMSD and the deformation index (DI). RMSD accounts for the average distance between the model and the target structure (from x-ray crystallography), upon optimal superposition of the former onto the latter:

$$\mathrm{RMSD}(A,B) = \sqrt{\frac{\sum_{i=1}^{N} d_i^{\,2}}{N}}$$

FIGURE 6.21 RNA-Puzzles crystallographic molecular structures and challenges. (a) Challenge 1: "*What is the structure of the following sequence: 5′-CCGCCGCGCCAUGC CUGUGGCGG-3′ knowing that the crystal structure shows a homodimer that contains two strands of the sequence? The strands hybridize with blunt ends (C–G closing base pairs).*" Crystal structure provided by T. Hermann (From Dibrov, S et al., *Acta Crystallographica Section D*, 67, 97–104, 2011), 46 nt. (b) Challenge 2: "*The crystal structure shows a 100 nucleotide square that assembles from four inner and four outer strands. 3D coordinates of the nucleotides in the inner strands were provided. What are the structures of the outer strands?*" Crystal structure provided by T. Hermann (From Dibrov, S et al., *Acta Crystallographica Section D*, 67, 97–104, 2011). (c) Challenge: 3 *"A domain of a riboswitch was crystallized. The sequence is the following:* 5′*-CUCUGGAGAGAACCGUUUAAUCGGUCGCCGAA GGAGCAAGCUCUGCGCAUAUGCAGAGUGAAACUCUCAGGCAAAAGGACAGAG-3′. The crystallized sequence was slightly different (an apical loop was replaced by a GAAA loop) but it was not mentioned to protect the work of crystallographers.*" Crystal structure provided by D. Patel (From Huang, L et al., *Molecular Cell*, 40, 774–786, 2010), 84 nt. (d) Challenge 4: "*What is the structure of the following sequence:* 5′*-GGCUUAUCAAGAGAGGUGGAGGGAC UGGCCCGAUGAAACCCGGCAACCACUAGUCUAGCGUCAGCUUCGGCUGACGCUAG GCUAGUGGUGCCAAUUCCUGCAGCGGAAACGUUGAAAGAUGAGCCA-3′?*" Crystal structure provided by A. Ferré-D'amaré (From Baird, N J et al., *RNA*, 18, 759–770, 2012), 126 nt. (e) Challenge 6: "*This problem consists in predicting the structure of an RNA molecule with the following sequence:* 5′*-CGGCAGGUGCUCCCGACCCUGCGGUCGGGAGUUAAA AGGGAAGCCGGUGCAAGUCCGGCACGGUCCCGCCACUGUGACGGGGAGUCGCCC CUCGGGAUGUGCCACUGGCCCGAAGGCCGGGAAGGCGGAGGGGCGGCGAGGAUC CGGAGUCAGGAAACCUGCCUGCCG-3′.*" Crystal structure provided by A. Serganov (From Peselis, A and A Serganov, *Nature Structural & Molecular Biology*, 19, 1182–1184, 2012), 168 nt. (f) Challenge 10: "*This problem consists in predicting the structure of a T-box riboswitch and tRNA 1:1 complex with the following sequence:* 5′*UGCGAUGAGAA GAAGAGUAUUAAGGAUUUACUAUGAUUAGCGACUCUAGGAUAGUGAAAGCUAGAG GAUAGUAACCUUAAGAAGGCACUUCGAGCA-3′ (T-box).* 5′*-GCGGAAGUAGUUCAGU GGUAGAACACCACCUUGCCAAGGUGGGGGUCGCGGGUUCGAAUCCCGUCUUCCG CUCCA-3′ (tRNA).*" Crystal structure provided by A. Ferré-D'Amaré (From Zhang, J and A R Ferré-D'Amaré, *Nature*, 500, 363–366, 2013), 171 nt. 3D structural models and challenges are reported from RNA-Puzzles website (http://paradise-ibmc.u-strasbg.fr/rnapuzzles). (*Continued*)

FIGURE 6.21(*Continued*) RNA-Puzzles crystallographic molecular structures and challenges. (a) Challenge 1: "*What is the structure of the following sequence: 5′-CCGC CGCGCCAUGCCUGUGGCGG-3′ knowing that the crystal structure shows a homodimer that contains two strands of the sequence? The strands hybridize with blunt ends (C–G closing base pairs).*" Crystal structure provided by T. Hermann (From Dibrov, S et al., *Acta Crystallographica Section D*, 67, 97–104, 2011), 46 nt. (b) Challenge 2: "*The crystal structure shows a 100 nucleotide square that assembles from four inner and four outer strands. 3D coordinates of the nucleotides in the inner strands were provided. What are the structures of the outer strands?*" Crystal structure provided by T. Hermann (From Dibrov, S et al., *Acta Crystallographica Section D*, 67, 97–104, 2011). (c) Challenge: 3 *"A domain of a riboswitch was crystallized. The sequence is the following: 5′-CUCUGGAGAGAACCGUUUAAUCGG UCGCCGAAGGAGCAAGCUCUGCGCAUAUGCAGAGUGAAACUCUCAGGCAAAAGG ACAGAG-3′. The crystallized sequence was slightly different (an apical loop was replaced by a GAAA loop) but it was not mentioned to protect the work of crystallographers.*" Crystal structure provided by D. Patel (From Huang, L et al., *Molecular Cell*, 40, 774–786, 2010), 84 nt. (d) Challenge 4: *"What is the structure of the following sequence: 5′-GGCUUAUCAA GAGAGGUGGAGGGACUGGCCCGAUGAAACCCGGCAACCACUAGUCUAGCGUCAG CUUCGGCUGACGCUAGGCUAGUGGUGCCAAUUCCUGCAGCGGAAACGUUGAAA GAUGAGCCA-3′?*" Crystal structure provided by A. Ferré-D'amaré (From Baird, N J et al., *RNA*, 18, 759–770, 2012), 126 nt. (e) Challenge 6: *"This problem consists in predicting the structure of an RNA molecule with the following sequence: 5′-CGGCAGGUGCUCCCG ACCCUGCGGUCGGGAGUUAAAAGGGAAGCCGGUGCAAGUCCGGCACGGUCCCG CCACUGUGACGGGGAGUCGCCCCUCGGGAUGUGCCACUGGCCCGAAGGCCGGGA AGGCGGAGGGGCGGCGAGGAUCCGGAGUCAGGAAACCUGCCUGCCG-3′."* Crystal structure provided by A. Serganov (From Peselis, A and A Serganov, *Nature Structural & Molecular Biology*, 19, 1182–1184, 2012), 168 nt. (f) Challenge 10: *"This problem consists in predicting the structure of a T-box riboswitch and tRNA 1:1 complex with the following sequence: 5′UGCGAUGAGAAGAAGAGUAUUAAGGAUUUACUAUGAUUAGCGACUC UAGGAUAGUGAAAGCUAGAGGAUAGUAACCUUAAGAAGGCACUUCGAGCA-3′ (T-box). 5′-GCGGAAGUAGUUCAGUGGUAGAACACCACCUUGCCAAGGUGGGGGUC GCGGGUUCGAAUCCCGUCUUCCGCUCCA-3′ (tRNA).*" Crystal structure provided by A. Ferré-D'Amaré (From Zhang, J and A R Ferré-D'Amaré, *Nature*, 500, 363–366, 2013), 171 nt. 3D structural models and challenges are reported from RNA-Puzzles website (http://paradise-ibmc.u-strasbg.fr/rnapuzzles).

where d_i is the distance between the i atom in the modeled (A) and experimental (B) structures, and N is the number of considered atoms. *DI* corrects the RMSD between a model and the corresponding target structure by considering the interaction network fidelity (INF) between the two, i.e., how similar the base pairing and base stacking are

$$\mathrm{DI}(A,B) = \frac{\mathrm{RMSD}(A,B)}{\mathrm{MCC}(A,B)}$$

$$\mathrm{INF}(A,B) = \mathrm{MCC}(A,B) = \sqrt{\mathrm{PPV} \times \mathrm{STY}}$$
$$\mathrm{PPV}\left(\text{specificity}\right) = \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FP}|}$$
$$\mathrm{STY}\left(\text{sensitivity}\right) = \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FN}|}$$

where $\mathrm{MCC}(A,B)$ is the Matthews correlation coefficient calculated as the geometric mean of the specificity and sensitivity for the base–base interaction predictions (Matthews 1975; Parisien et al. 2009). Stereochemistry accuracy of RNA structural models, however, is estimated with MolProbity (Davis et al. 2007), a web service that performs a number of structure quality checks on macromolecules, such as torsional angles and atomic clashes. The *clash score* is chosen as a measure of stereochemical correctness. Since RMSD metric depends largely on the size of the RNA molecules considered, it requires an assessment of the statistical significance. The groups of Weeks and Dokholyan (Hajdin et al. 2010) have shown that all-against-all RMSD between pairs of computationally random generated RNA conformations (decoys) is normally distributed, and that the RMSD expected by chance correlates with the length of the molecule via a power law:

$$\langle \mathrm{RMSD} \rangle = a \times N^{0.41} - b$$

where N is the number of nucleotides and a and b are constants whose values depend on the use of secondary structure information as restraints when sampling the RNA conformational space. Hence, for a given comparison RMSD, its associated p-value can be determined from the cumulative density function of the normal distribution (assuming a standard deviation of 1.8):

$$p = \left[1 + \mathrm{erf}\left(\frac{(\mathrm{RMSD} - \langle \mathrm{RMSD} \rangle)/1.8}{\sqrt{2}}\right)\right]/2$$

Generally, RNA structural models with *p*-value lower than .01 represent a statistical significant prediction of a global RNA fold.

RNA-Puzzles participants adopt a variety of sampling methods, and computer-assisted strategies to model the RNA 3D structure provided in each single puzzle (Hajdin et al. 2010; Kosinski, Cymerman, and Feder 2003; Chen 2008; Cao and Chen 2011; Sripakdeevong, Kladwang, and Das 2011; Das and Baker 2008; Das, Karanicolas, and Baker 2010; Ding and Dokholyan 2012; Parisien and Major 2008; Rother et al. 2011, 2012; Cao and Chen 2005, 2006a,b; Gherghe et al. 2009; Lavender et al. 2010; Cruz et al. 2012; Sijenyi et al. 2012). For example, the Bujnicki group used a hybrid strategy, bases on homology modeling, previously developed for protein modeling in the course of the Critical Assessment of protein Structure Prediction (CASP) experiment (Kosinski, Cymerman, and Feder 2003). The Chen lab used Vfold, a multiscale, free energy landscape-based RNA folding model (Chen 2008; Cao and Chen 2011). The Das group used the stepwise assembly method (Sripakdeevong, Kladwang, and Das 2011) as implemented in Rosetta (Das and Baker 2008; Das, Karanicolas, and Baker 2010). The Dokholyan group adopted a combined three-bead model RNA molecule representation with a DMD sampling approach (Ding and Dokholyan 2012; Ding et al. 2008). The Major group applied their in-house developed MC-Fold and MC-Sym pipeline (Parisien and Major 2008). The results of their contributions to the past six challenges are summarized in Table 6.1. Generally, all submitted models correctly reproduced the base-pair interactions and accurately predicted the structure of helical regions in RNA. Loops remained the most difficult elements to model. Similarly, non-Watson–Crick base pairs were not consistently reproduced, clearly indicating the need of a better description of alternative interactions between RNA base pairs in the current computational techniques. Furthermore, the high values of *clash scores* in the submitted models can be improved by a more precise parameterization of geometric features like atomic bond, distances, and angles. The human and computer times required to produce the submitted models vary according to the approach adopted by individual research groups, and the computational resources that they employed. Therefore, it is difficult to have a direct comparison of the time-efficiency of each specific approach. Overall, the results of this first blind exercise for *de novo* RNA 3D structure prediction reflect the state of the art of the field.

TABLE 6.1 Comparison of RNA 3D Structural Models Submitted for RNA-Puzzles Challenges

Challenge 1 (November 2011)					
Lab	**RMSD**	**DI**	**Clash score**	***p*-Value**	**References**
Das	3.413	3.657	0	2.220e−16	(Sripakdeevong, Kladwang, and Das 2011; Das and Baker 2008; Das, Karanicolas, and Baker 2010)
Major	4.064	4.567	66.4	3.830e−15	(Parisien and Major 2008)
Chen	4.113	5.008	0.68	4.774e−15	(Chen 2008; Cao and Chen 2005, 2006a,b, 2011)
Bujnicki	4.60	5.754	54.73	4.985e−14	(Kosinski, Cymerman, and Feder 2003; Rother et al. 2011, 2012; Boniecki et al. 2003; Surles et al. 1994)
Santalucia	5.688	6.750	28.40	3.203e−12	(Sijenyi et al. 2012)
Dokholyan	6.939	8.552	31.74	3.306e−10	(Hajdin et al. 2010; Ding and Dokholyan 2012; Ding et al. 2008; Gherghe et al. 2009; Lavender et al. 2010)
Challenge 2 (November 2011)					
Lab	**RMSD**	**DI**	**Clash score**	***p*-Value**	**References**
Bujnicki	2.3	2.827	14.54	0.000e+00	(Kosinski, Cymerman, and Feder 2003; Rother et al. 2011, 2012; Boniecki et al. 2003; Surles et al. 1994)
Das	2.496	2.899	19.25	0.000e+00	(Sripakdeevong, Kladwang, and Das 2011; Das and Baker 2008; Das, Karanicolas, and Baker 2010)
Dokholyan	2.543	3.089	9.77	0.000e+00	(Hajdin et al. 2010; Ding and Dokholyan 2012; Ding et al. 2008; Gherghe et al. 2009; Lavender et al. 2010)
Chen	2.83	3.739	18.66	1.110e−16	(Chen 2008; Cao and Chen 2005, 2006a,b, 2011)
Major	2.98	3.819	134.26	2.220e−16	(Parisien and Major 2008)
Wildauer	3.479	4.403	165.57	1.988e−15	NA
Santalucia	3.650	4.537	26.4	4.274e−15	(Sijenyi et al. 2012)

(*Continued*)

TABLE 6.1 (*Continued*) Comparison of RNA 3D Structural Models Submitted for RNA-Puzzles Challenges

Challenge 3 (November 2011)					
Lab	**RMSD**	**DI**	**Clash score**	***p*-Value**	**References**
Chen	7.241	9.842	1.1	2.010e−05	(Chen 2008; Cao and Chen 2005, 2006a,b, 2011)
Dokholyan	11.460	16.104	41.21	3.902e−02	(Hajdin et al. 2010; Ding and Dokholyan 2012; Ding et al. 2008; Gherghe et al. 2009; Lavender et al. 2010)
Das	11.965	16.419	1.1	6.932e−02	(Sripakdeevong, Kladwang, and Das 2011; Das and Baker 2008; Das, Karanicolas, and Baker 2010)
Bujnicki	12.186	17.494	14.72	8.711e−02	(Kosinski, Cymerman, and Feder 2003; Rother et al. 2011, 2012; Boniecki et al. 2003; Surles et al. 1994)
Major	13.701	23.331	93.52	3.026e−02	(Parisien and Major 2008)
Lab	**RMSD**	**DI**	**Clash score**	***p*-Value**	**References**
Chen	3.347	5.178	4.43	0.000e+00	(Chen 2008; Cao and Chen 2005, 2006a,b, 2011)
Santalucia	4.117	6.651	52.07	0.000e+00	(Sijenyi et al. 2012)
Bujnicki	4.191	6.803	34.92	0.000e+00	(Kosinski, Cymerman, and Feder 2003; Rother et al. 2011, 2012; Boniecki et al. 2003; Surles et al. 1994)
Adamiak	4.342	7.54	29.27	0.000e+00	NA
Das	4.498	7.183	14.28	0.000e+00	(Sripakdeevong, Kladwang, and Das 2011; Das and Baker 2008; Das, Karanicolas, and Baker 2010)
Major	4.515	7.229	1.23	0.000e+00	(Parisien and Major 2008)
Dokholyan	5.366	8.516	19.42	0.000e+00	(Hajdin et al. 2010; Ding and Dokholyan 2012; Ding et al. 2008; Gherghe et al. 2009; Lavender et al. 2010)
Mikolajczak	12.801	25.341	40.11	1.362e−06	NA

(*Continued*)

TABLE 6.1 (*Continued*) Comparison of RNA 3D Structural Models Submitted for RNA-Puzzles Challenges

Lab	RMSD	DI	Clash score	*p*-Value	References
Das	11.699	16.426	32.84	9.015e−14	(Sripakdeevong, Kladwang, and Das 2011; Das and Baker 2008; Das, Karanicolas, and Baker 2010)
Blanchet	21.379	30.813	2.2	2.359e−02	NA
Chen	22.152	34.442	3.49	5.996e−02	(Chen 2008; Cao and Chen 2005, 2006a,b, 2011)
Dokholyan	22.769	34.744	15.23	1.126e−01	(Hajdin et al. 2010; Ding and Dokholyan 2012; Ding et al. 2008; Gherghe et al. 2009, Lavender et al. 2010)
Bujnicki	30.968	46.451	2.75	9.996e−01	(Kosinski, Cymerman, and Feder 2003; Rother et al. 2011, 2012; Boniecki et al. 2003; Surles et al. 1994)
Lab	**RMSD**	**DI**	**Clash score**	***p*-Value**	**References**
Das	6.803	12.687	11.09	0.000e+00	(Sripakdeevong, Kladwang, and Das 2011; Das and Baker 2008; Das, Karanicolas, and Baker 2010)
Bujnicki	9.339	18.149	0.91	0.000e+00	(Kosinski, Cymerman, and Feder 2003; Rother et al. 2011, 2012; Boniecki et al. 2003; Surles et al. 1994)
Chen	11.34	22.981	7.81	4.852e−14	(Chen 2008; Cao and Chen 2005, 2006a,b, 2011)
Ding	12.496	26.208	7.8	5.143e−12	(Lavender et al. 2010; Ding and Dokholyan 2012)

Note: Data are retrieved from the RNA-puzzles website (http://paradise-ibmc.u-strasbg.fr/rnapuzzles/). Among the several models submitted by each research group per single challenge, only the ones with the lowest RMSD with respect to the experimental structure are reported herein. RMSD and DI values are given in Å.

6.7 CONCLUSIONS

In this chapter, we have reviewed some of the most employed methodologies to model the 3D structure of RNA molecules, in particular those ones based on knowledge-based energy potentials. It is very difficult to establish

a quantitative comparison about their performance, even if they are contrasted to the same problem, like in the RNA-Puzzles initiative (Cruz et al. 2012), because neither the amount of human expertise versus computer power utilized, nor the way they handle the experimental information is the same for each method. However, at this point there are relevant questions we should address. Does the degree of coarse-graining have a clear impact in the results? What is the relevance of using experimental data to restraint the sampling? To answer those questions we can have a look to the results of the RNA Puzzles. Interestingly, the degree of coarse-graining does not seem to always be critical. The group of Chen used Vfold, a two-bead model combined with MD sampling (Cao and Chen 2005, 2009), still being able to produce significant good models (RMSD with respect the x-ray structure close to 4 Å and low *clash score* values) in three out the six Puzzles experiments reviewed in Table 6.1, producing the lowest RMSD model in two occasions (on the first below 4 Å and second at 7.2 Å). Notwithstanding, the Das group using a stepwise assembly procedure and an all-atom model (Sripakdeevong, Kladwang, and Das 2011) produced models lower than 4 Å in just two experiments, although it produced the absolute lowest RMSD model in three occasions. The hybrid approach between the two is the one employed by Bujnicki, which combines homology modeling with posterior MC refinement on a three-bead model to incorporate secondary structure restraints (Rother et al. 2011; Magnus et al. 2014) is able to generate good models in two of the analyzed Puzzles rounds, being the absolute lowest RMSD prediction in one of them. Similar approach by Santalucia group yielded good models also for two rounds. The use of experimental information in the past two Puzzles experiments makes it possible to obtain models below 10 Å, for large RNA fragments of 171 and 187 nt, respectively (Challenges 10 and 5, data not shown).

REFERENCES

Addess, K J and J Feigon. 1996. Introduction to 1H NMR spectroscopy of DNA. In *Bioorganic Chemistry: Nucleic Acids*, edited by S M Hecht, 500. New York, NY: Oxford University Press.

Alder, B J and T E Wainwright. 1959. Studies in molecular dynamics. I. General method. *Journal of Chemical Physics* 31 (2): 459.

Baird, N J, J Zhang, T Hamma, and A R Ferré-D'Amaré. 2012. YbxF and YlxQ are bacterial homologs of L7Ae and bind K-turns but not K-loops. *RNA* 18 (4): 759–70.

Ban, N, P Nissen, J Hanssen, P B Moore, and T A Steitz. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* 289 (5481): 905–20.

Bernauer, J, X Huang, A Y L Sim, and M Levitt. 2011. Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA* 17 (6): 1066–75.

Bernstein, F C, T F Koetzle, G J Williams, E F Meyer, M D Brice, J R Rodgers, O Kennard, T Shimanouchi, and M Tasumi. 1977. The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology* 112 (3): 535–42.

Biswas, R, S N Mitra, and M Sundaralingam. 1998. 1.76 Å structure of a pyrimidine start alternating A-RNA hexamer r(CGUAC)dG. *Acta Crystallographica. Section D, Biological Crystallography* 54 (Pt 4): 570–6.

Cao, S and S-J Chen. 2005. Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA* 11 (12): 1884–97.

Cao, S and S-J Chen. 2006a. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Research* 34 (9): 2634–52.

Cao, S and S-J Chen. 2006b. Free energy landscapes of RNA/RNA complexes: With applications to snRNA complexes in spliceosomes. *Journal of Molecular Biology* 357 (1): 292–312.

Cao, S and S-J Chen. 2009. Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA* 15 (4): 696–706.

Cao, S and S-J Chen. 2011. Physics-based de novo prediction of RNA 3D structures. *Journal of Physical Chemistry B* 115 (14): 4216–26.

Cate, J H, A R Gooding, E Podell, K Zhou, B L Golden, C E Kundrot, T R Cech, and J A Doudna. 1996. Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science* 273 (5282): 1678–85.

Cech, T R, and J A Steitz. 2014. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* 157 (1): 77–94.

Chen, J-L and C W Greider. 2005. Functional analysis of the pseudoknot structure in human telomerase RNA. *Proceedings of the National Academy of Sciences of the United States of America* 102 (23): 8080–5; discussion 8077–9.

Chen, S-J. 2008. RNA folding: Conformational statistics, folding kinetics, and ion electrostatics. *Annual Review of Biophysics* 37: 197–214.

Copley, S D, E Smith, and H J Morowitz. 2007. The origin of the RNA world: Co-evolution of genes and metabolism. *Bioorganic Chemistry* 35 (6): 430–43.

Cragnolini, T, Y Laurin, P Derreumaux, and S Pasquali. 2014. Predicting complex 3D RNA structures via a high resolution coarse-grained model. *Biomolecules* (1): 1–5.

Crick, F. 1970. Central dogma of molecular biology. *Nature* 227: 561–3.

Cruz, J A, M-F Blanchet, M Boniecki, J M Bujnicki, S-J Chen, S Cao, R Das et al. 2012. RNA-puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. *RNA* 18 (4): 610–25.

Dahm, R. 2005. Friedrich Miescher and the discovery of DNA. *Developmental Biology* 278 (2): 274–88.

AQ3: Please provide volume number for Cragnolini et al. (2014).

Das, R and D Baker. 2007. Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences of the United States of America* 104 (37): 14664–9.

Das, R and D Baker. 2008. Macromolecular modeling with rosetta. *Annual Review of Biochemistry* 77: 363–82.

Das, R, J Karanicolas, and D Baker. 2010. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature Methods* 7 (4): 291–4.

Davis, I W, A Leaver-Fay, V B Chen, J N Block, G J Kapral, X Wang, L W Murray et al. 2007. MolProbity: All-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research* 35 (Web Server issue): W375–83.

Debye, P and E Hückel. 1923. The theory of electrolytes. I. Lowering of freezing point and related phenomena. *Physikalische Zeitschrift* 24: 185–206.

Deigan, K E, T W Li, D H Mathews, and K M Weeks. 2008. Accurate SHAPE-directed RNA structure determination a SHAPE experiment can be interpreted as a pseudo-free energy high accuracy. Free energy minimization, by using SHAPE pseudo.

Dibrov, S, J McLean, and T Hermann. 2011. Structure of an RNA dimer of a regulatory element from human thymidylate synthase mRNA. *Acta Crystallographica Section D* 67 (Pt 2): 97–104.

Dibrov, S M, J McLean, J Parsons, and T Hermann. 2011. Self-assembling RNA square. *Proceedings of the National Academy of Sciences of the United States of America* 108 (16): 6405–8.

Ding, F, C A Lavender, K M Weeks, and N V Dokholyan. 2012. Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nature Methods* 9 (6): 603–8.

Ding, F, J M Borreguero, S V Buldyrey, H E Stanley, and N V Dokholyan. 2003. Mechanism for the alpha-helix to beta-hairpin transition. *Proteins: Structure, Function and Genetics* 53 (2): 220–8.

Ding, F and N V Dokholyan. 2012. Multiscale modeling of RNA structure and dynamics. *RNA 3D Structure Analysis and Prediction*.

Ding, F and N V Dokholyan. 2012. Discrete molecular dynamics simulation of biomolecules. In *Computational Modeling of Biological Systems: From Molecules to Pathways*, 55–73. Springer.

Ding, F, S Sharma, P Chalasani, V V Demidov, N E Broude, and N V Dokholyan. 2008. Ab initio RNA folding by discrete molecular dynamics: From structure prediction to folding mechanisms. *RNA* 14 (6): 1164–73.

Doherty, E A, R T Batey, B Masquida, and J A Doudna. 2001. A universal mode of helix packing in RNA. *Nature Structural Biology* 8 (4): 339–43.

Duarte, C M and A M Pyle. 1998. Stepping through an RNA structure: A novel approach to conformational analysis. *Journal of Molecular Biology* 284 (5): 1465–78.

Duarte, C M, L M Wadley, and A M Pyle. 2003. RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Research* 31 (16): 4755–61.

AQ4: Please provide complete details for Deigan et al. (2008).

AQ5: Please provide Editor details, Publisher location for Ding and Dokholyan (2012).

AQ6: Please provide volume number and page range for Ding and Dokholyan (2012).

AQ7: Ding and Dokholyan (2012) has been added from Figure 6.12 source. Please confirm whether it is appropriate.

AQ8: Please provide volume number and page range for Ding and Dokholyan (2012).

Earl, D J and M W Deem. 2005. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics* 7 (23): 3910.

Flores, S C, Y Wan, R Russell, and R B Altman. 2010. Predicting RNA structure by multiple template homology modeling. *Pacific Symposium on Biocomputing* (10): 216–27.

Flory, P J. 1969. *Statistical Mechanics of Chain Molecules*. Interscience Publishers.

Frank, D N and N R Pace. 1998. Ribonuclease P: Unity and diversity in a tRNA processing ribozyme. *Annual Review of Biochemistry* 67: 153–80.

Freeman, W M, S J Walker, and K E Vrana. 1999. Quantitative RT-PCR: Pitfalls and potential. *BioTechniques* 26 (1): 112–25.

Fritz, J J, A Lewin, W Hauswirth, A Agarwal, M Grant, and L Shaw. 2002. Development of hammerhead ribozymes to modulate endogenous gene expression for functional studies. *Methods* 28 (2): 276–85.

Geiduschek, B E P and R Haselkorn. 1969. Messenger RNA. *Annual Review of Biochemistry* 38 (1): 647–76.

Gelin, B R and M Karplus. 1975. Sidechain torsional potentials and motion of amino acids in proteins: Bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences of the United States of America* 72 (6): 2002–6.

Gherghe, C M, C W Leonard, F Ding, N V Dokholyan, and K M Weeks. 2009. Native-like RNA tertiary structures using a sequence-encoded cleavage agent and refinement by discrete molecular dynamics. *Journal of the American Chemical Society* 131 (7): 2541–6.

Gopinath, S C B. 2009. Mapping of RNA-protein interactions. *Analytica Chimica Acta* 636 (2): 117–28.

Green, M R. 1986. Pre-mRNA splicing. *Annual Review of Genetics* 20: 671–708.

Hajdin, C E, F Ding, N V Dokholyan, and K M Weeks. 2010. On the significance of an RNA tertiary structure prediction. *RNA* 16 (7): 1340–9.

Hansen, J L, A M Long, and S C Schultz. 1997. Structure of the RNA-dependent RNA polymerase of poliovirus. *Structure* 5 (8): 1109–22.

Holley, R W, J Apgar, G A Everett, J T Madison, M Marquisee, S H Merrill, J R Penswick, and A Zamir. 1965. Structure of a ribonucleic acid. *Science* 147 (3664): 1462–5.

Huang, L, A Serganov, and D J Patel. 2010. Structural insights into ligand recognition by a sensing domain of the cooperative glycine riboswitch. *Molecular Cell* 40 (5): 774–86.

Jonikas, M A, R J Radmer, A Laederach, R Das, S Pearlman, D Herschlag, and R B Altman. 2009. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA* 15 (2): 189–99.

Karaduman, R, P Fabrizio, K Hartmuth, H Urlaub, and R Lührmann. 2006. RNA structure and RNA-protein interactions in purified yeast U6 snRNPs. *Journal of Molecular Biology* 356 (5): 1248–62.

Kendrew, J C, G Bodo, H M Dintzis, G Parrish, H Wickoff, and D C Phillips. 1958. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* 181: 662–6.

AQ9: please provide volume number for Flores et al. (2010).

AQ10: Please provide publisher location for Flory (1969).

Kosinski, J, I A Cymerman, and M Feder. 2003. A 'FRankenstein's monster' approach to comparative modeling: Merging the finest fragments of fold-recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins: Structure*.

Kramer, A. 1996. The structure and function of proteins pre-mRNA splicing. *Annual Review of Biochemistry* 65: 367–409.

Ladner, J E, A Jack, J D Robertus, R S Brown, D Rhodes, B F Clark, and A Klug. 1975. Structure of yeast phenylalanine transfer RNA at 2.5 A resolution. *Proceedings of the National Academy of Sciences* 72 (11): 4414–8.

Laurendeau, N M. 2005. *Statistical Thermodynamics: Fundamentals and Applications*. Cambridge University Press.

Lavender, C A, F Ding, N V Dokholyan, and K M Weeks. 2010. Robust and generic RNA modeling using inferred constraints: A structure for the hepatitis C virus IRES pseudoknot domain. *Biochemistry* 49 (24): 4931–3.

Lu, H and J Skolnick. 2001. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins* 44 (3): 223–32.

Magnus, M, D Matelska, G Lach, G Chojnowski, M J Boniecki, E Purta, W Dawson, S Dunin-Horkawicz, and J M Bujnicki. 2014. Computational modeling of RNA 3D structures, with the aid of experimental restraints. *RNA Biology* 11 (5): 1–15.

Mathews, D H, J Sabina, M Zuker, and D H Turner. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology* 288 (5): 911–40.

Mathews, D H, M D Disney, J L Childs, S J Schroeder, M Zuker, and D H Turner. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America* 101 (19): 7287–92.

Matthews, B W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta* 405 (2): 442–51.

Mattice, W L and U W Suter. 1994. *Conformational Theory of Large Molecules: The Rotational Isomeric State Model in Macromolecular Systems.*

McCammon, J A, B R Gelin, and M Karplus. 1977. Dynamics of folded proteins. *Nature* 267 (5612): 585–90.

Merino, E J, K A Wilkinson, J L Coughlan, and K M Weeks. 2005. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *Journal of the American Chemical Society* 127 (12): 4223–31.

Metropolis, N and S Ulam. 1949. The Monte Carlo method. *Journal of the American Statistical Association* 44 (247): 335–41.

Murray, L J W, W B Arendall, D C Richardson, and J S Richardson. 2003. RNA backbone is rotameric. *Proceedings of the National Academy of Sciences of the United States of America* 100 (24): 13904–9.

Nikolova, E N et al. 2011. Transient Hoogsteen base pairs in canonical duplex DNA. *Nature*. 470: 498–502.

AQ11: Please provide volume number and page range for Kosinski et al. (2003).

AQ12: Please provide publisher location for Laurendeau (2005).

AQ13: Please provide publisher and location for Mattice and Suter (1994).

AQ14: Nikolova et al. (2011) has been added from Figure 6.2 source. Please confirm whether it is appropriate.

Olson, W K. 1975. Configurational statistics of polynucleotide chains. A single virtual bond treatment. *Macromolecules* 8 (3): 272–5.

Olson, W K and P J Flory. 1972. Spatial configurations of polynucleotide chains. I. Steric interactions in polyribonucleotides: A virtual bond model. *Biopolymers* 11 (1): 1–23.

Parisien, M and F Major. 2008. The MC-fold and MC-sym pipeline infers RNA structure from sequence data. *Nature* 452 (7183): 51–5.

Parisien, M, J A Cruz, E Westhof, and F Major. 2009. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA* 15 (10): 1875–85.

Pasquali, S and P Derreumaux. 2010. HiRE-RNA: A high resolution coarse-grained energy model for RNA. *Journal of Physical Chemistry B* 114 (37): 11957–66.

Patel, J S, A Berteotti, S Ronsisvalle, W Rocchia, and A Cavalli. 2014. Steered molecular dynamics simulations for studying protein-ligand interaction in cyclin-dependent kinase 5. *Journal of Chemical Information and Modeling* 54 (2): 470–80.

Peselis, A and A Serganov. 2012. Structural insights into ligand binding and gene expression control by an adenosylcobalamin riboswitch. *Nature Structural & Molecular Biology* 19 (11): 1182–4.

Quigley, G J and A Rich. 1976. Structural domains of transfer RNA molecules. *Science* 194 (4267): 796–806.

Rahman, A. 1964. Correlations in the motion of atoms in liquid argon. *Physical Review* 136 (2A): A405–A411.

Ramachandran, G N, C Ramakrishnan, and V Sasisekharan. 1963. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* 7 (1): 95–9.

Rapold, R F and W L Mattice. 1995. New high-coordination lattice model for rotational isomeric state polymer chains. *Journal of the Chemical Society, Faraday Transactions* 91 (16): 2435.

Reyes, F E, A D Garst, and R T Batey. 2009. *Strategies in RNA Crystallography. Methods in Enzymology*. 1st ed, Vol. 469. Elsevier.

Rich, A, D R Davies, F H C Crick, and J D Watson. 1961. The molecular structure of polyadenylic acid. *Journal of Molecular Biology* 3 (1): 71–86.

Rohl, C A, C E M Strauss, K M S Misura, and D Baker. 2004. Protein structure prediction using rosetta. *Methods in Enzymology* 383: 66–93.

Rother, K, M Rother, M Boniecki, and T Puton. 2012. Template-based and template-free modeling of RNA 3D structure: Inspirations from protein structure modeling. *RNA 3D Structure*.

Rother, M, K Rother, T Puton, and J M Bujnicki. 2011. ModeRNA: A tool for comparative modeling of RNA 3D structure. *Nucleic Acids Research* 39 (10): 4007–22.

Schluenzen, F, A Tocilj, R Zarivach, J Harms, M Gluehmann, D Janell, A Bashan et al. 2000. Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell* 102: 615–23.

Scott, L G and M Hennig. 2008. RNA structure determination by NMR. In *Bioinformatics. Volume I: Structure, Sequence Analysis and Evolution*, edited by J M

AQ15: Please provide publisher location for Reyes et al. (2009).

AQ16: Please provide volume number and page range for Rother et al. (2012).

Keith, Vol. 452, 29–61. *Methods in Molecular Biology*. Totowa, NJ: Humana Press.

Sharma, S, F Ding, and N V Dokholyan. 2008. iFoldRNA: Three-dimensional RNA structure prediction and folding. *Bioinformatics* 24 (17): 1951–2.

Sijenyi, F, P Saro, Z Ouyang, K Damm-Ganamet, M Wood, J Jiang, and J Santalucia. 2012. The RNA folding problems: Different levels of sRNA structure prediction. In *RNA 3D Structure Analysis and Prediction*, Vol. 27, 91–117. Berlin, Heidelberg: Springer Berlin Heidelberg.

Sripakdeevong, P, W Kladwang, and R Das. 2011. An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling. *PNAS* 108 (51): 20573–8.

Staple, D W and S E Butcher. 2005. Pseudoknots: RNA structures with diverse functions. *PLoS Biology* 3 (6): e213.

Su, L, L Chen, M Egli, J M Berger, and A Rich. 1999. Minor groove RNA triplex in the crystal structure of a ribosomal frameshifting viral pseudoknot. *Nature Structural Biology* 6 (3): 285–92.

Thomsen, N D and J M Berger. 2012. *Crystallization and X-Ray Structure Determination of an RNA-Dependent Hexameric Helicase. Methods in Enzymology*. 1st ed, Vol. 511. Elsevier.

AQ17: Please provide publisher location for Thomsen and Berger (2012).

Tijerina, P, S Mohr, and R Russell. 2007. DMS footprinting of structured RNAs and RNA-protein complexes. *Nature Protocols* 2 (10): 2608–23.

Toor, N, K S Keating, S D Taylor, and A M Pyle. 2008. Crystal structure of a self-spliced group II intron. *Science* 320 (5872): 77–82.

Torrie, G M and J P Valleau. 1977. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* 23 (2): 187–99.

Watson, J D, T A Baker, S P Bell, A Gann, M Levine, and R Losick. 2013. *Molecular Biology of the Gene*. 7th ed. Benjamin Cummings.

AQ18: Please provide publisher location for Watson et al. (2013).

Wimberly, B T, D E Brodersen, W M Clemons Jr, R J Morgan-warren, A P Carter, C Vonrhein, T Hartsch, and V Ramakrishnan. 2000. Structure of the 30S ribosomal subunit. *Nature* 407: 327–39.

Woese, C R, S Winker, and R R Gutell. 1990. Architecture of ribosomal RNA: Constraints on the sequence of 'tetra-loops'. *Proceedings of the National Academy of Sciences* 87 (21): 8467–71.

Worthington Allen, F. 1941. The biochemistry of the nucleic acids, purines and pyrimidines. *Annual Review of Biochemistry* 10: 221–45.

Xia, Z, D P Gardner, R R Gutell, and P Ren. 2010. Coarse-grained model for simulation of RNA three-dimensional structures. *Journal of Physical Chemistry. B* 114 (42): 13497–506.

Xia, Z and P Ren. 2013. Prediction and coarse-grained modeling of RNA structures. In *Biophysics of RNA Folding, Biophysics for the Life Sciences*, edited by R Russell, Vol. 3, 53–68. New York, NY: Springer.

Zhang, J and A R Ferré-D'Amaré. 2013. Co-crystal structure of a T-box riboswitch stem I domain in complex with its cognate tRNA. *Nature* 500 (7462): 363–6.