8

RNA Three-Dimensional Structure Determination Using Experimental Constraints

Feng Ding and Nikolay V. Dokholyan

CONTENTS

8.1 Introduction 159
8.2 Coarse-Grained RNA Modeling Using Discrete Molecule Dynamics 161
8.3 How to Evaluate an RNA 3D Structure Model 163
8.4 RNA Structure Determination Using Various Types of Structural Information 164
8.4.1 Base Pairing 164
8.4.2 Internucleotide Proximity Information 165
8.4.3 Solvent Accessibility 168
8.5 Conclusion 171
Acknowledgments 171
References 171

8.1 Introduction

RNAs function not only as bridges between the genetic information stored in DNA and the final protein products, as stated in the Central Dogma; recently, RNA has also been found to play diverse roles in almost every aspect of cell life (Cruz and Westhof, 2009; Nilsen, 2007; Sharp, 2009; Wan et al., 2011), from regulating transcription and translation (e.g., siRNA, miRNA, or riboswitch regulator motifs; Edwards et al., 2007) to catalyzing mRNA splicing (spliceosome RNA or self-splicing introns; Vicens and Cech, 2006) and protein synthesis (rRNA). These newly discovered RNA functions either are encoded in their primary sequences, through complementarity to target sequences, or originate from their ability to form complex secondary and high-order tertiary structures. The 3D RNA structures, formed by packing of base-paired helices, allow specific interactions with themselves or other biomolecules, including proteins, nucleic acids, and small-molecule ligands. The well-defined 3D structures of RNAs also determine the accessibility of specific sequences important for function. These novel functions of structural RNAs have been uncovered and characterized by studying a small fraction of the known RNA world. Whereas only 2% of a typical eukaryotic genome is translated into proteins, ~90% is transcribed into some kind of noncoding RNA, including antigene, long noncoding, small regulatory, and scaffolding RNAs (Janowski and Corey, 2010; Sharp, 2009; Wan et al., 2011; Wang et al., 2011a; Wang et al., 2011b). A large portion of these unknown RNAs form functional 3D structures, which remain to be characterized. The fact that RNAs adopt specific 3D structures in order to perform their functions also makes them potential drug targets

(Hermann and Westhof, 1998; Sucheck and Wong, 2000). Indeed, many well-known antibiotics bind to the RNA component of the bacterial ribosome. More recently, it was discovered that riboswitches could be targets for antibiotics (Kim et al., 2009; Lee et al., 2009; Mulhbacher et al., 2010; Ott et al., 2009). Therefore, the knowledge of the underlying RNA and RNA complex structures can not only enhance our understanding of RNA functions but also aid in design of novel drugs using structure-based rational drug design.

Traditional high-resolution structure determination methods such as x-ray crystallography and NMR spectroscopy offer crucial insight into the details of RNA structure–function relationships. However, as noted by many x-ray crystallography experts (Ke and Doudna, 2004), it is often difficult to grow RNA crystals due to the flexible nature of RNA molecules, many of which can either adopt multiple conformations or have significant unstructured components. On the other hand, RNAs amenable to NMR experiments are limited to small RNAs. For example, most RNAs in the Protein Databank (Berman et al., 2000) whose structures are determined by NMR are below 50 nucleotides (< 50 nts) in length. Therefore, there is a crucial need for novel methods of determining the 3D structures of RNAs. Computational modeling of RNA 3D structures offers the opportunity to incorporate the structural features of RNAs extracted from known RNA structures (Das and Baker, 2007; Jonikas et al., 2009; Jossinet and Westhof, 2005; Major et al., 1993; Major et al., 1991; Massire et al., 1998; Parisien and Major, 2008; Shapiro et al., 2007; Tsai et al., 2003), to integrate physical and chemical principles (Cao and Chen, 2011; Ding et al., 2008), and to include experimentally derived structural information in modeling (Jonikas et al., 2009). For instance, several recent RNA 3D structure modeling methods (Cao and Chen, 2011; Das and Baker, 2007; Ding et al., 2008; Parisien and Major, 2008) have yielded accurate structure predictions of small RNAs from sequence alone, highlighting the predictive power of RNA modeling approaches in general.

The ability to accurately predict RNA 3D structures is also important for the emerging field of RNA nanotechnology (Guo, 2010). In the bottom-up approaches of RNA nanotechnology, RNA building blocks are engineered to self-assemble into nanoscale materials with applications in nanomedicine and nanodevices (Guo, 2005). Computational modeling of RNA 3D structures, which accounts for noncanonical base–base pairs (Das et al., 2010), long-range tertiary interactions (Gherghe et al., 2009; Lavender et al., 2010), and ion-dependent folding (Draper et al., 2005), can help design the building blocks, predict the final structure, and characterize the assembly kinetics. The major challenges of computational RNA 3D structure modeling come from the vast conformational space of RNA and inaccuracy in the force field describing RNA folding. As RNA size increases, the available conformational space increases exponentially and the effects of force field inaccuracy accumulate. As a result, tertiary structure prediction for large RNAs with complex topologies is beyond the reach of the current *ab initio* approaches (Cao and Chen, 2011; Das and Baker, 2007; Ding et al., 2008; Parisien and Major, 2008). On the other hand, many biophysical and biochemical methods have been developed to probe RNA secondary and tertiary structure. For example, the selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) chemistry developed by Weeks and colleagues (Deigan et al., 2009; Weeks, 2010) characterizes the probability of base pairing for each nucleotide. Other experiments such as fluorescence resonance energy transfer (FRET) (Rueda et al., 2004), cross-linking (Harris et al., 1994; Pinard et al., 2001; Yu et al., 2008), and tethered hydroxyl radical probing (t-HRP) (Das et al., 2008; Gherghe et al., 2009) can probe internucleotide distances. The solvent accessibility of individual nucleotides can also be explored by solution hydroxyl radical probing (HRP) experiments (Cate et al., 1996; Pastor et al., 2000; Tullius and Greenbaum, 2005). Incorporation of experimentally derived structural information with computational

modeling can markedly reduce the allowed conformational space and thereby facilitate the computational prediction of native RNA ensembles (Das et al., 2008; Ding et al., 2012; Gherghe et al., 2009; Jonikas et al., 2009; Lavender et al., 2010; Yang et al., 2010; Yu et al., 2008).

Next, we first briefly introduce our computational RNA model. We will also discuss a novel approach to evaluate the statistical significance of an RNA structural model. We will then discuss our approaches to incorporate various pieces of experimentally derived structural information, including base pairs, long-range distance constraints, and solvent accessibilities, in RNA 3D structure refinement and prediction.

8.2 Coarse-Grained RNA Modeling Using Discrete Molecule Dynamics

We use DMD as the conformational sampling engine. A detailed description of the DMD algorithm can be found elsewhere (Dokholyan et al., 1998; Rapaport, 2004; Zhou and Karplus, 1997). The difference between discrete molecular dynamics and traditional molecular dynamics is in the interaction potential functions. Interatomic interactions in DMD are governed by stepwise potential functions (Figure 8.1a). Neighboring interactions (such as bonds, bond angles, and dihedrals) are modeled by infinitely high square well potentials (Figure 8.1b). By approximating the continuous potential functions with step functions of pairwise distances, DMD simulations are reduced to event-driven (collision) molecular dynamics simulation. In a DMD simulation, atoms move with constant velocity until they collide with another atom. As soon as the potential of interaction between the two atoms changes (i.e., the pairwise distance is at the step of the stepwise potential function) the velocities of the two interacting atoms change instantaneously (Figure 8.1a). These velocity changes are required to conform to the conservation laws of energy, momentum, and angular momentum. Each such collision is termed an "event."

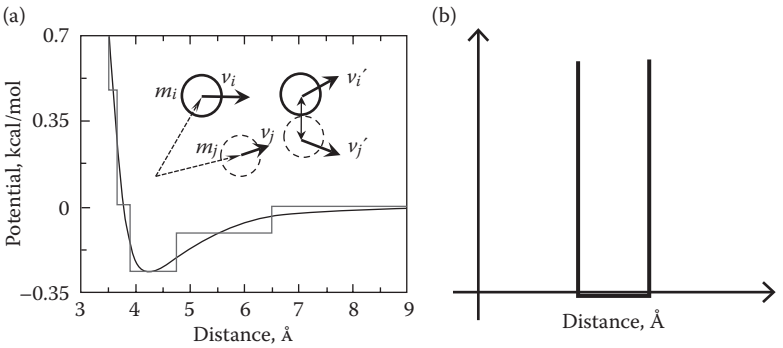

FIGURE 8.1
Discrete molecular dynamics simulations. (a) Schematic of the DMD potential. The stepwise function used in DMD is the approximation of the continuous function in traditional molecular dynamics. The insert depicts the collision of two atoms with masses of m_i and m_j at the initial position of r_i and r_j, respectively. The two atoms move with constant velocities (v) until they meet at distance of R_{ij}. (b) Schematic of the potential energy of bonds in DMD. The atom pairs remain within the distance range during the simulation.

The sampling efficiency of DMD over traditional MD is mainly due to rapid processing of collision events and localized updates of collisions (only colliding atoms are updated at each collision). In the limit of infinitesimally small steps, the discrete step function approaches the continuous potential function and DMD simulations become equivalent to traditional molecular dynamics.

We approximate the single-stranded RNA molecule as a coarse-grained 'beads-on-a-string' polymer with three beads representing each nucleotide, one for sugar (S), one for phosphate (P), and one for nucleotide base (B) (Figure 8.2b). The P and S beads are positioned at the centers of mass of the corresponding phosphate group and the five-atom ring of sugar group, respectively. For both purines (adenine and guanine) and pyrimidines (uracil and cytosine), we represent the base bead (B) as the center of the six-atom ring. The neighboring beads along the sequence, which may represent moieties that belong to the same or a neighboring nucleotide, are constrained to mimic the chain connectivity and local chain geometry (Figure 8.2b). Types of constraints include covalent bonds (solid lines), bond angles (dashed lines), and dihedral angles (dot-dashed lines). The parameters for bonded interactions mimic the folded RNA structure and are

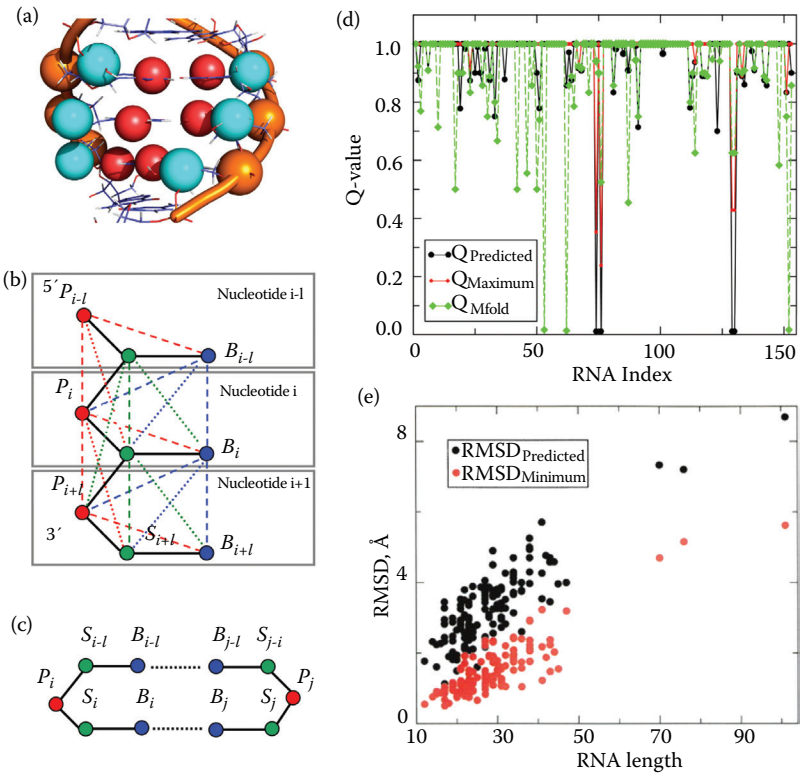

FIGURE 8.2

Ab initio RNA folding using the simplified RNA model. (a) Each nucleotide is represented by three coarse-grained beads. (b) The lines illustrate the bonded interactions, important for modeling RNA geometry. (c) The base pairing interactions are modeled by hydrogen bonding interactions (Ding et al., 2003). For 153 RNAs, the predicted structures by DMD simulations recapitulate not only the secondary structures (d, the fraction of native base pairs, Q-value) but also the tertiary structure (e, root mean square deviation from the native states, RMSD).

derived from a high resolution RNA structure database (Murray et al., 2003). Nonbonded interactions are crucial to model the folding dynamics of RNA molecules. In our model, we include base pairing (Watson-Crick pairs of A-U G-C, and *Wobble pair of* U-G), base stacking, short-range phosphate–phosphate repulsion, and hydrophobic interactions. The details of the interaction parameters can be found in Ref. (Ding and Dokholyan, 2012; Ding et al., 2008).

Using the simplified RNA model in DMD simulations, we were able to accurately fold a large set of 150 small RNAs (<50 nts) to their corresponding native states (Ding et al., 2008) (Figure 8.2e). The majority of the predicted structures are within 4 Å root mean square deviations (RMSD) from the native states. The average percentage of accurately predicted native base pairs for all 153 RNAs studied is 94% (Figure 8.2d). For comparison, we outperformed other secondary structure prediction methods, including the commonly used Mfold (Zuker, 2003), which yielded an average of 91% of native base pairs predicted. Given the high percentage of correctly predicted base pairs (94%), the average number of incorrectly predicted base pairs is less than one for the studied RNAs. These results highlight the robustness of our DMD-based RNA folding approach. Inspired by this result, we developed a web server, iFoldRNA (http://ifoldrna.dokhlab.org), which currently allows *ab initio* RNA 3D structure prediction of short RNAs for the RNA research community.

8.3 How to Evaluate an RNA 3D Structure Model

To benchmark whether a given computational RNA modeling method is predictive, RMSD between predicted and accepted structures is commonly used. However, RMSD is not a straightforward measure of the significance of a given prediction. For example, a structural prediction of 10 Å RMSD for a short RNA stem is unlikely to be helpful in generating strong and testable biological hypotheses. However, a prediction of the same RMSD but for a large RNA as group I intron (Cate et al., 1996) is highly significant. Therefore, it is important to develop a quantity for structural evaluation that is length independent.

We performed extensive modeling of RNA structures with different lengths and generated decoy structures with alternative base pairing and helix packing (Hajdin et al., 2010). We found that for a given RNA, the pairwise RMSD between any two randomly generated RNA-like structures belongs to a Gaussian distribution. The average value of the Gaussian depends on the length of the RNA and on whether the native base pairs are included in generating the model structures (Figure 8.3). If the native base pairing information is used in modeling, the average RMSD between two random structures is significantly smaller since the available conformational space is reduced. We found that the average RMSD has a power-law dependence to the RNA length, with the exponent of 0.41. Similarly, similar behavior was observed for protein, except that the power-law exponent is ~1/3, as for a compact globular objects (Reva et al., 1998). This result suggests that RNA in general is less compact compared to proteins. Interestingly, the standard deviation of the Gaussian does not significantly depend on the RNA length, but rather is constant. Therefore, we developed an empirical relationship between average RMSD and RNA length, which can be used to compute the statistical significance of a prediction with given RMSD to the native state (Hajdin et al., 2010). The statistical significance calculation for a given prediction is available online at iFoldRNA.

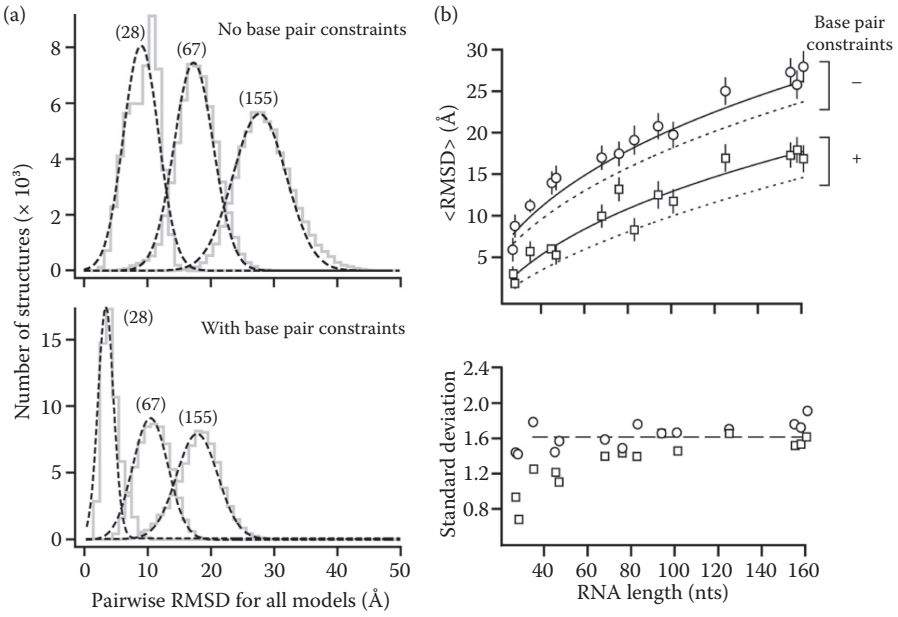

FIGURE 8.3
Statistical significance of a given RNA structure prediction. (a) The distribution of pairwise RMSD between two randomly generated RNA structures is computed as the function of RNA length. With incorporation of base pairs (low panel), the distribution is shifted toward lower average RMSD. (b) The average RMSD (upper) and standard deviation (lower) are plotted as the function of length.

8.4 RNA Structure Determination Using Various Types of Structural Information

Three out of 153 RNA molecules studied in the *ab initio* folding simulations are longer than 65 nucleotides, where the DMD-RNA method cannot be applied to predict the native secondary and tertiary structure from sequence alone (Figures 8.2d and 8.2e). The challenges to predict large RNA folding *ab initio* arise from the exponentially increasing size of the conformational space and inaccuracies in the force field. Incorporation of experimentally or bioinformatically derived RNA structural information as constraints in RNA modeling greatly reduces available the conformational space, and thus increases the prediction accuracy (Hajdin et al., 2010).

8.4.1 Base Pairing

RNA secondary structures can be obtained by evolutionary study of homologous sequences (Massire et al., 1998). The base-paired nucleotides tend to coevolve during evolution by maintaining the secondary structures. In order to obtain statistically significant predictions, the number of homologous sequences should be large. When a sufficient number of homologous sequences are not available, this approach is not applicable. On the other hand, chemical probing approaches have been widely used to probe RNA secondary structures (Fritz et al., 2002; Gopinath, 2009; Tijerina et al., 2007). Among them, SHAPE chemistry

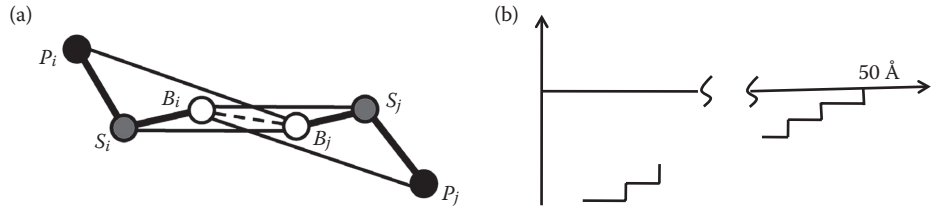

FIGURE 8.4
The base pair interaction. (a) The dot-dashed and thin solid lines between two nucleotides represent the multibody interactions, which capture the distance and angular dependence of base pairing. (b) The multistep potential between two bases is used to exert weak but long-range attraction that brings two nucleotides together.

(Deigan et al., 2009; Weeks, 2010) has been shown to be a powerful approach for analyzing secondary structure at single nucleotide resolution for RNAs of any length (Merino et al., 2005; Wilkinson et al., 2006). SHAPE exploits the discovery that the 2′-OH group in unconstrained or flexible nucleotides reacts preferentially with hydroxyl-selective electrophilic reagents. In contrast, nucleotides constrained by base pairing or tertiary interactions are unreactive. The resulting reactivity information can be used, in concert with a secondary structure prediction algorithm, to obtain accurate secondary structures (Deigan et al., 2009; Mathews et al., 2004; Mortimer and Weeks, 2007; Wang et al., 2008; Wilkinson et al., 2008).

For a given input list of base pairs, we assign energetic constraints between specific nucleotides to bias DMD simulations. We use the distance- and orientation-dependent base pairing interaction potential, as determined from statistical analysis of known RNA structures, to model base pair formation (Ding et al., 2008; Gherghe et al., 2009). The multibody interaction includes both the interaction between two bases B_i and B_j, and the auxiliary interactions between base B_i (B_j) of nucleotide i (j) and sugar S_j (S_i)/phosphate P_j (P_i) of nucleotide j (i) (Figure 8.4a). To efficiently form the base pairs in DMD simulations, we also assign a weak but long-range attractive interaction between the two bases with the interaction range of 50 Å (Figure 8.4b). The attractive force is about 0.15 kcal/(mol•Å), which is ~10 pN. We find that this weak attraction is able to effectively bring two nucleotides together to form a base pair in simulations. Using experimentally derived base pairing information, we predicted the 3D structure of the pseudoknot domain of the HCV IRES (Lavender et al., 2010). The recently solved crystal structure of the HCV IRES revealed (Berry et al., 2011) a structure that agreed closely with our model (Figure 8.5a). Consistent with this successful refinement, our structure prediction for the glycine riboswitch domain (~150 nts) in the first-ever blind RNA structure prediction competition, *RNAPuzzles*, was among the best submitted and included the recovery of atomistic features of the glycine binding pocket (Figure 8.5b) (Cruz, 2012).

8.4.2 Internucleotide Proximity Information

RNA tertiary structure is determined by long-range interactions between different secondary structure elements. Knowledge of proximity constraints between atoms or nucleotides can significantly reduce the conformational space and thus facilitate determination of the native structure. Experimentally determined atomic proximity information has been commonly used in structure refinement. For example, pairwise distances between protons detected by nuclear Overhauser effect (NOE) in NMR experiments can be used to determine protein and RNA structures at high resolution. However, RNA structure

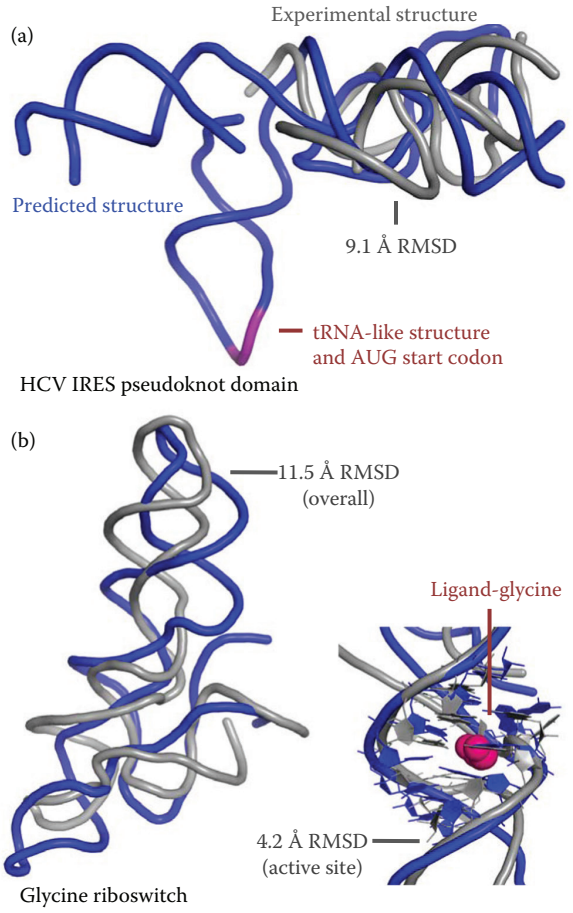

FIGURE 8.5
Blind RNA 3D structure prediction. The predicted structures are blue; crystallographic structures are gray. (a) The HCV IRES pseudoknot domain, which indicated that this domain likely functions by *tRNA mimicry* (Lavender et al., 2010), and (b) the glycine riboswitch, the second-best result in the *RNA Puzzles* community structure prediction exercise.

determination using NMR is often limited to small RNAs. On the other hand, long-range constraints for RNA modeling can be inferred from a variety of biochemical and bioinformatic techniques, ranging from t-HRP and cross-linking to sequence covariation (Gutell et al., 1992; Juzumiene et al., 2001; Michel and Westhof, 1990; Ziehler and Engelke, 2001). The derived structure information is low-resolution in nature, with inferred proximities between nucleotides rather than specific atoms. These low-resolution constraints are readily incorporated into our coarse-grained RNA modeling.

In collaboration with Weeks group, we developed a t-HRP approach to obtain tertiary proximity constraints (Gherghe et al., 2009). An Fe(II)-EDTA moiety was tethered specifically to RNA using the site-selective intercalation reagent methidiumpropyl-EDTA (MPE) (Hertzberg and Dervan, 1984). MPE preferentially intercalates at CpG steps in RNA at sites adjacent to a single nucleotide bulge (White and Draper, 1987, 1989), which can be introduced by mutations in helical regions. The nucleotides accessible to the Fe(II)-EDTA will have a high chance to be cleaved by the induced hydroxyl radical, while remote nucleotides

will not be cleaved. To apply the cleavage information to bias DMD simulations, we developed a generic approach to interpret the cleavage patterns as distance constraints (Figure 8.6a). The interaction potential features a "soft" energy wall at 25 Å, with smaller energy bonuses extending out to 35 Å (Figure 8.4a). The 25-Å barrier corresponds to the distance cutoff within which the nucleotides exhibit strong cleavage, and beyond which the nucleotides have weak cleavage. The cutoff value is also consistent with the length of the MPE-conjugated cleavage agent. The interaction strength is assigned according to the cleavage intensity [$E \sim \ln(I/\langle I \rangle)$], since the cleavage intensity is interpreted as the probability to be within the cutoff range cleavable by the reagent. This approach has two advantages: (1) no user input is required to decide whether a given cleavage is significant or not and (2) structure refinement is highly tolerant of measurement errors inherent in any hydroxyl radical footprinting experiment. By applying the experimentally derived constraints in DMD simulations, we were able to refine the structure of $tRNA^{Asp}$ to 6.4 Å RMSD with respect to the crystal structure (Gherghe et al., 2009).

In most high-profile RNA structure determination cases, secondary structures as well as some key tertiary structure information, such as internucleotide proximity information, were known before the high-resolution structure was solved using x-ray crystallography or NMR. For example, the T- and D-loop of $tRNA^{Asp}$ were known to be close to each other before the 3D structure was solved. To test the ability of DMD-based RNA structure refinement using a few long-range distance constraints, we used four RNAs: domain III of the cricket paralysis virus internal ribosome entry site (CrPV) (49 nts), a full-length hammerhead ribozyme from *S. mansoni* (HHR) (67 nts), *S. cerevisiae* $tRNA^{Asp}$ (75 nts), and the P546 domain of the *T. thermophilia* group I intron (P546) (158 nts). Each of these RNAs has a complex 3D fold, involving more than simple intrahelix interactions. Prior to publication of the high-resolution structures (Cate et al., 1996; Costantino et al., 2008; Martick and Scott, 2006; Westhof et al., 1988), significant biochemical or bioinformatic data describing tertiary interactions were available for each RNA. The secondary structure was also known to high accuracy in each case. Only this prior information of secondary and tertiary structures was used during DMD refinement. In all cases, we were able to generate a low-RMSD structure. The RMSDs between the predicted structures and the native states for the CrPV, HHR, $tRNA^{Asp}$, and P546 RNAs are 3.6, 5.4, 6.4, and 11.3 Å, respectively (Lavender et al., 2010) (Figure 8.7). This benchmark result highlights the efficiency of internucleotide proximity constraints in RNA structure determination.

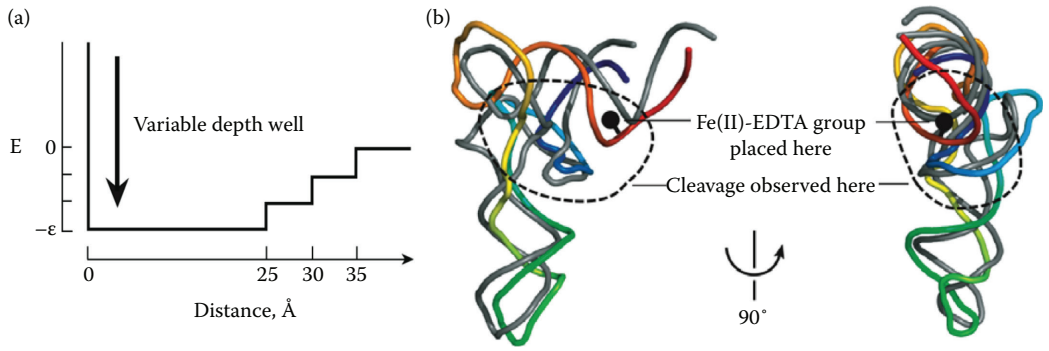

FIGURE 8.6
RNA structure refinement using t-HRP reactivity. (a) The interaction potential between the tethered nucleotide and the rest of RNA. The interaction strength ε depends on the intensity of cleavage reactivity. (b) Comparison between computational prediction (in rainbow color) and the experimental structure (in gray).

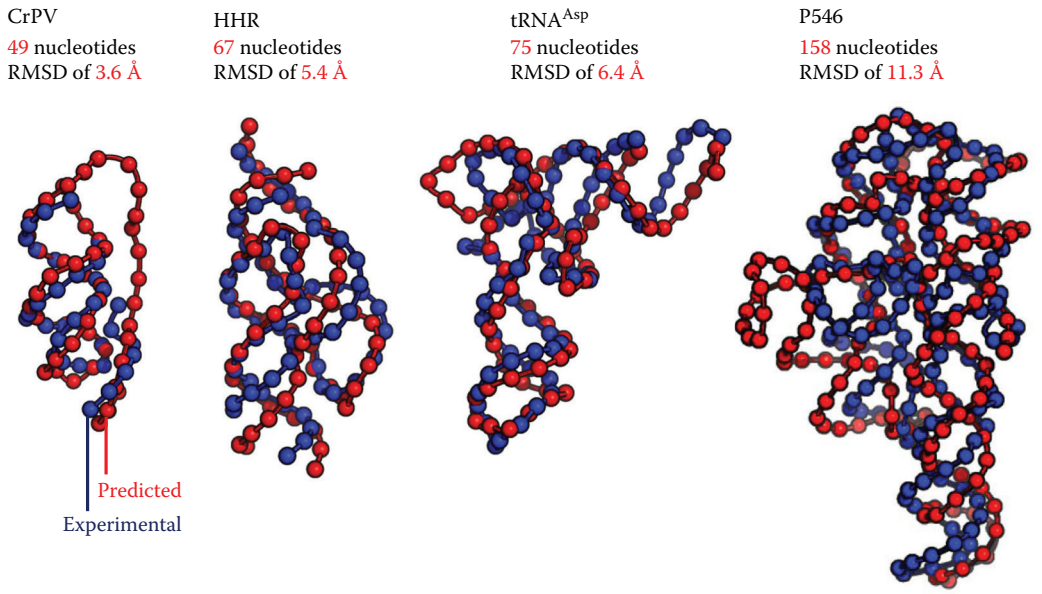

FIGURE 8.7
Benchmark of RNA structure refinement using information of base pairs and a small number of internucleotide tertiary contacts. The test sets are composed of CrPV, HHR, $tRNA^{Asp}$, and P546 domain of group I intron.

8.4.3 Solvent Accessibility

Experimental methods used to probe through-space distances, such as t-HRP (Das et al., 2008; Gherghe et al., 2009), cross-linking (Harris et al., 1994; Pinard et al., 2001; Yu et al., 2008), and FRET (Rueda et al., 2004), can give high-quality distance information. However, these techniques often require synthesis of specialized RNA constructs, careful controls for unintended structural perturbations, and complex approaches for data interpretation (Hajdin et al., 2010). In contrast, solution HRP, which reports the approximate backbone solvent accessibility for most nucleotides in an RNA molecule (Cate et al., 1996; Tullius and Greenbaum, 2005), is straightforward to implement. HRP measurements have been used to evaluate or filter RNA structural ensembles (Bergman et al., 2004; Jonikas et al., 2009; Rangan et al., 2003; Tullius and Greenbaum, 2005) but have not been used to drive RNA 3D structure determination in a quantitative and systematic way. We developed an approach that incorporates solvent accessibility information derived from HRP measurements to bias DMD (Dokholyan et al., 1998; Zhou and Karplus, 1997) simulations in order to generate structural ensembles consistent with experimental measurements.

In order to incorporate experimentally obtained HRP reactivities into DMD simulations, a structural parameter consistent with experimental measurements must be identified. Hydroxyl radical reactivity is correlated with backbone solvent accessibility (Balasubramanian et al., 1998; Cate et al., 1996); however, it is not straightforward to incorporate solvent accessibility as a constraint in a molecular dynamics simulation. We find that solvent accessibility is inversely proportional to the number of through-space neighbor atoms. Nucleotides in the M-Box riboswitch with low HRP reactivities are generally buried and have many through-space contacts, whereas nucleotides with high reactivities have fewer contacts and are more exposed (Figure 8.8a). The number of through-space

contacts can be readily incorporated as a constraint in DMD and other simulation methods (Vendruscolo et al., 2001), and we use it to bias our simulations. We assign a bias potential with two components. The first includes uniform pairwise attractive potentials for most of the nucleotides. This general attraction encourages collapse of the RNA and overall nucleotide packing. The second is an over-burial repulsion potential incurred when a given nucleotide exceeds an assigned threshold number of contacts (N_{max}) derived from its experimental HRP reactivity (Figure 8.8b). Taken together, we expect that each nucleotide forms its maximally allowed number of contacts, which drives conformational sampling of structures consistent with input HRP data.

To obtain structural ensembles consistent with HRP data, we perform simulations and analysis in three steps (Figure 8.8c). First, we perform serial DMD simulations with inputs of RNA sequence and canonical base pairs. We use the approach described in Aim 1 to bias formation of base pairs. The result of these simulations is the formation of native

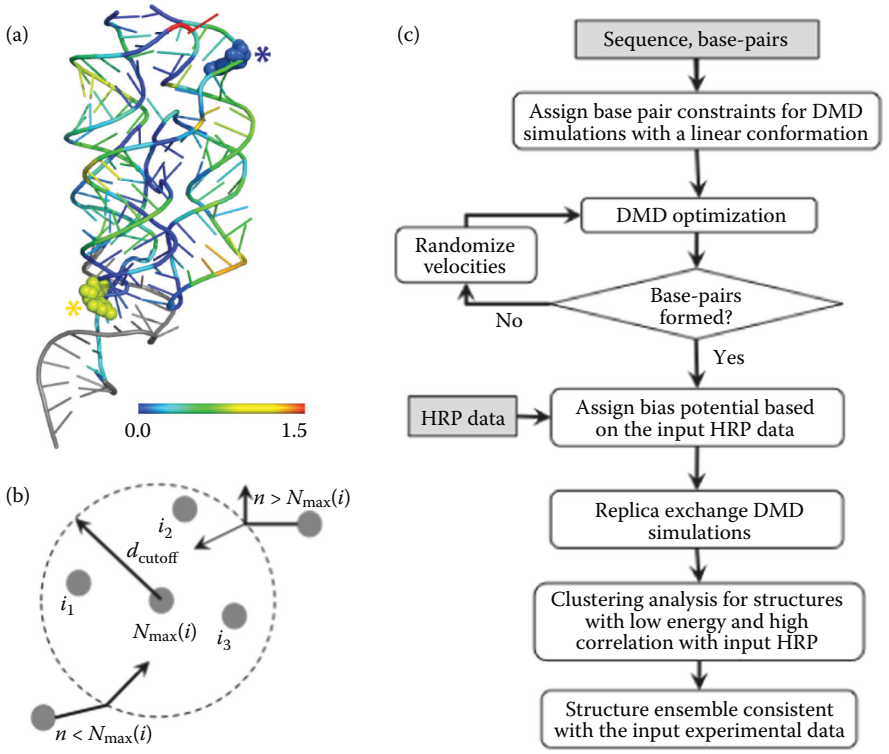

FIGURE 8.8
HRP-driven RNA structure refinement. (a) The structure of the M-box riboswitch is illustrated. Nucleotides are colored according to HRP reactivity (blue to red); nucleotides without HRP data are shown in gray. A solvent-exposed nucleotide with low HRP reactivity (blue) and a buried nucleotide with high HRP reactivity (red) are emphasized with all-atom representations (asterisks). (b) The assignment of potentials for incorporating HRP reactivities into DMD simulations. Each nucleotide is assigned a threshold number of contacts (N_{max}) within the cutoff distance (d_{cutoff} = 14 Å). For a given nucleotide *i*, its *n* through-space neighbors are denoted as i_1, i_2, i_3.... An approaching nucleotide can form a new contact (indicated by the inward arrow) if the number of total contacts is smaller than N_{max}. If *n* is larger than N_{max}, the approaching nucleotide can form a contact only if the total DMD kinetic energy is sufficient to overcome the energy penalty for overpacking. Otherwise, the nucleotide reflects back without forming a new contact (denoted by the outward arrow). (c) The HRP-directed DMD simulation algorithm. (Adapted from Ding, F. et al. *Nat Methods*, 9, 603–608, 2012. With permission.)

secondary structures. Second, we perform replica exchange DMD simulations and impose the HRP-derived bias potentials to enrich conformations consistent with the experimental HRP data. Replica exchange simulations have been shown to be efficient in RNA conformational sampling. Third, we select 100 structures with lowest energies and highest correlations between HRP reactivities and numbers of contacts and perform RMSD-based clustering analysis to identify representative structures of the predicted structural ensemble. The resulting model features well-defined RNA structure and agrees with the input experimental data.

We tested our HRP-based RNA modeling approach on nine structurally diverse RNAs, with length ranging from 80 to 230 nts. In all cases of compact RNAs, we obtained RNA 3D structures with high statistical significance (Figures 8.9a through 8.9d). It is interesting that the performance of our prediction is independent of RNA lengths. However, our methods failed to reproduce the structures of less-compact RNAs (Figures 8.9e and 8.9f). Therefore, it is necessary to determine whether we can know *a priori* whether a given RNA is compact or not, and thus whether HRP reactivity can be used to refine the 3D structure. A compact RNA has a high fraction of buried nucleotides. Since HRP measures the extent of nucleotide burial (number of through-space neighbors), we can use the fraction of nucleotides with low HRP reactivity to measure the compactness. For a given RNA with HRP data, we compute the fraction of nucleotides, $f(r)$, with HRP reactivities below a given value, r

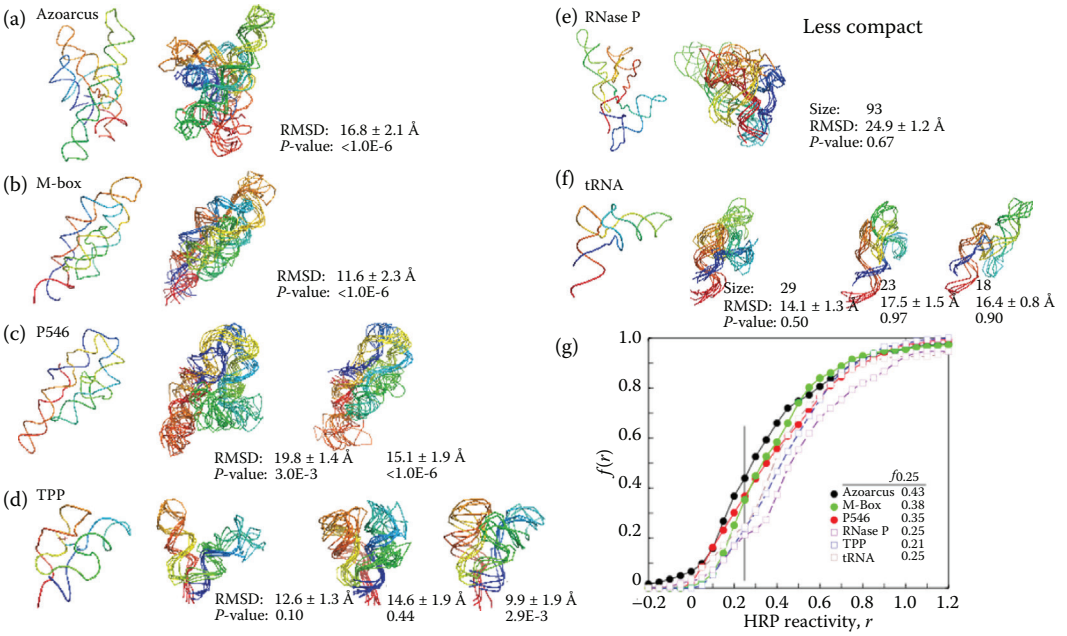

FIGURE 8.9

HRP-directed RNA refinement. (a through f) RNAs are shown with backbone traces. The leftmost panel shows the accepted structure for each RNA. Right-hand panels show representative structures for each highly populated cluster. Backbones are colored from blue to red in the 5′ to 3′ direction. For each cluster, the number of structures, mean RMSD, and *P*-value are shown. Significant *P*-values are emphasized in bold. Panels (a) through (d) correspond to four examples of compact RNAs that have been studied. Panels (e) and (f) are two noncompact RNAs, where our predictions failed to recapitulate the native structures (Ding et al., 2012). (g) Fraction of nucleotides, $f(r)$, with HRP reactivities below a given value r. Normalized HRP reactivities are shown; the vertical line indicates the $f_{0.25}$ cutoff. (Adapted from Ding, F. et al. *Nat Methods*, 9, 603–608, 2012. With permission.)

(Figure 8.9g). Interestingly, we find that compact and noncompact RNAs have clear differences in $f(r)$ values around $r = 0.25$. Compact RNAs usually have $f(0.25)$ larger than 0.25, while less-compact RNAs have smaller $f(0.25)$ values. Our results (Figure 8.9g) suggest that $f(0.25)$ is indeed a good predictor for the applicability of our HRP-driven RNA refinement method. The programs for HRP-driven RNA refinement are also available online at iFoldRNA (http://troll.med.unc.edu/ifoldrna/HRP-1.0-openmpi.tgz).

8.5 Conclusion

With advances in high-throughput sequencing, RNAs with novel functions are being discovered at a rapid pace. Knowledge of the underlying 3D structures of these RNAs is a fundamental prerequisite to complete understanding and further manipulation of their functions. Due to experimental challenges in RNA structure determination, our knowledge of structure–function relationships for RNAs lags behind that attained for proteins. Many biochemical and biophysical methods have been proposed to probe the RNA 2D and 3D structures. Although the structural information derived from these approaches is often low resolution in nature, incorporation of these pieces of information in computational RNA modeling can greatly increase the prediction accuracy. With the rapid development of experimental characterizations and computational modeling, we expect the convergence of two fronts for the emergence of hybrid approaches that can rapidly and accurately generate RNA 3D structures.

Acknowledgments

This work is supported by NIH grants GM080742 and CA084480 (to NVD) and by startup funds from Clemson University (to FD).

References

Balasubramanian B, Pogozelski WK, and Tullius TD (1998). DNA strand breaking by the hydroxyl radical is governed by the accessible surface areas of the hydrogen atoms of the DNA backbone. *Proc Natl Acad Sci U S A 95*, 9738–9743.

Bergman NH, Lau NC, Lehnert V, Westhof E, and Bartel DP (2004). The three-dimensional architecture of the class I ligase ribozyme. *RNA 10*, 176–184.

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

Berry KE, Waghray S, Mortimer SA, Bai Y, and Doudna JA (2011). Crystal structure of the HCV IRES central domain reveals strategy for start-codon positioning. *Structure 19*, 1456–1466.

Cao S, and Chen SJ (2011). Physics-based de novo prediction of RNA 3D structures. *J Phys Chem B 115*, 4216–4226.

Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR, and Doudna JA (1996). Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science 273*, 1678–1685.

Costantino DA, Pfingsten JS, Rambo RP, and Kieft JS (2008). tRNA-mRNA mimicry drives translation initiation from a viral IRES. *Nat Struct Mol Biol 15*, 57–64.

Cruz JA, and Westhof E (2009). The dynamic landscapes of RNA architecture. *Cell 136*, 604–609.

Cruz JA et al. (2012). RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. *RNA 18*, 610–625.

Das R, and Baker D (2007). Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci U S A 104*, 14664–14669.

Das R, Karanicolas J, and Baker D (2010). Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat Methods 7*, 291–294.

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

Deigan KE, Li TW, Mathews DH, and Weeks KM (2009). Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci U S A 106*, 97–102.

Ding F, and Dokholyan NV (2012). Multiscale modeling of RNA Structure and Dynamics. In: *RNA 3D Structure Analysis and Prediction*, Leontis NB, and Westhof E, eds. Verlag-Berlin-Heidelberg: Springer.

Ding F, Borreguero JM, Buldyrey SV, Stanley HE, and Dokholyan NV (2003). Mechanism for the alpha-helix to beta-hairpin transition. *Proteins 53*, 220–228.

Ding F, Sharma S, Chalasani V, Demidov V, Broude NE, and Dokholyan NV (2008). Ab initio RNA folding by discrete molecular dynamics: From structure prediction to folding mechanisms. *RNA 14*, 1164–1173.

Ding F, Lavender CA, Weeks KM, and Dokholyan NV (2012). Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nat Methods 9*, 603–608.

Dokholyan NV, Buldyrev SV, Stanley HE, and Shakhnovich EI (1998). Discrete molecular dynamics studies of the folding of a protein-like model. *Fold Des 3*, 577–587.

Draper DE, Grilley D, and Soto AM (2005). Ions and RNA folding. *Annu Rev Biophys Biomol Struct 34*, 221–243.

Edwards TE, Klein DJ, and Ferre-D'Amare AR (2007). Riboswitches: Small-molecule recognition by gene regulatory RNAs. *Curr Opin Chem Biol 17*, 273–279.

Fritz JJ, Lewin A, Hauswirth W, Agarwal A, Grant M, and Shaw L (2002). Development of hammerhead ribozymes to modulate endogenous gene expression for functional studies. *Methods 28*, 276–285.

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

Gopinath SC (2009). Mapping of RNA–protein interactions. *Anal Chim Acta 636*, 117–128.

Guo P (2005). RNA nanotechnology: Engineering, assembly and applications in detection, gene delivery and therapy. *J Nanosci Nanotechnol 5*, 1964–1982.

Guo P (2010). The emerging field of RNA nanotechnology. *Nat Nanotechnol 5*, 833–842.

Gutell RR, Power A, Hertz GZ, Putz EJ, and Stormo GD (1992). Identifying constraints on the higher-order structure of RNA: Continued development and application of comparative sequence analysis methods. *Nucleic Acids Res 20*, 5785–5795.

Hajdin CE, Ding F, Dokholyan NV, and Weeks KM (2010). On the significance of an RNA tertiary structure prediction. *RNA 16*, 1340–1349.

Harris ME, Nolan JM, Malhotra A, Brown JW, Harvey SC, and Pace NR (1994). Use of photoaffinity crosslinking and molecular modeling to analyze the global architecture of ribonuclease P RNA. *EMBO J 13*, 3953–3963.

Hermann T, and Westhof E (1998). RNA as a drug target: Chemical, modelling, and evolutionary tools. *Curr Opin Biotechnol 9*, 66–73.

Hertzberg RP, and Dervan PB (1984). Cleavage of DNA with methidiumpropyl-EDTA-iron(II): Reaction conditions and product analyses. *Biochemistry 23*, 3934–3945.

Janowski BA, and Corey DR (2010). Minireview: Switching on progesterone receptor expression with duplex RNA. *Mol Endocrinol 24*, 2243–2252.

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

Jossinet F, and Westhof E (2005). Sequence to Structure (S2S): Display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics 21*, 3320–3321.

Juzumiene D, Shapkina T, Kirillov S, and Wollenzien P (2001). Short-range RNA–RNA crosslinking methods to determine rRNA structure and interactions. *Methods 25*, 333–343.

Ke A, and Doudna JA (2004). Crystallization of RNA and RNA–protein complexes. *Methods 34*, 408–414.

Kim JN, Blount KF, Puskarz I, Lim J, Link KH, and Breaker RR (2009). Design and antimicrobial action of purine analogues that bind Guanine riboswitches. *ACS Chem Biol 4*, 915–927.

Lavender CA, Ding F, Dokholyan NV, and Weeks KM (2010). Robust and generic RNA modeling using inferred constraints: A structure for the hepatitis C virus IRES pseudoknot domain. *Biochemistry 49*, 4931–4933.

Lee ER, Blount KF, and Breaker RR (2009). Roseoflavin is a natural antibacterial compound that binds to FMN riboswitches and regulates gene expression. *RNA Biol 6*, 187–194.

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

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

Martick M, and Scott WG (2006). Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell 126*, 309–320.

Massire C, Jaeger L, and Westhof E (1998). Derivation of the three-dimensional architecture of bacterial ribonuclease P RNAs from comparative sequence analysis. *J Mol Biol 279*, 773–793.

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

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

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

Mortimer SA, and Weeks KM (2007). A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J Am Chem Soc 129*, 4144–4145.

Mulhbacher J, Brouillette E, Allard M, Fortier LC, Malouin F, and Lafontaine DA (2010). Novel riboswitch ligand analogs as selective inhibitors of guanine-related metabolic pathways. *PLoS Pathog 6*, e1000865.

Murray LJ, Arendall WB, 3rd, Richardson DC, and Richardson JS (2003). RNA backbone is rotameric. *Proc Natl Acad Sci U S A 100*, 13904–13909.

Nilsen TW (2007). RNA 1997–2007: A remarkable decade of discovery. *Mol Cell 28*, 715–720.

Ott E, Stolz J, Lehmann M, and Mack M (2009). The RFN riboswitch of *Bacillus subtilis* is a target for the antibiotic roseoflavin produced by *Streptomyces davawensis*. *RNA Biol 6*, 276–280.

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

Pastor N, Weinstein H, Jamison E, and Brenowitz M (2000). A detailed interpretation of OH radical footprints in a TBP–DNA complex reveals the role of dynamics in the mechanism of sequence-specific binding. *J Mol Biol 304*, 55–68.

Pinard R, Lambert D, Heckman JE, Esteban JA, Gundlach CWt, Hampel KJ, Glick GD, Walter NG, Major F, and Burke JM (2001). The hairpin ribozyme substrate binding-domain: a highly constrained D-shaped conformation. *J Mol Biol 307*, 51–65.

Rangan P, Masquida B, Westhof E, and Woodson SA (2003). Assembly of core helices and rapid tertiary folding of a small bacterial group I ribozyme. *Proc Natl Acad Sci U S A 100*, 1574–1579.

Rapaport DC (2004). *The Art of Molecular Dynamics Simulation*. Cambridge, UK: Cambridge University Press.

Reva BA, Finkelstein AV, and Skolnick J (1998). What is the probability of a chance prediction of a protein structure with an rmsd of 6 A? *Fold Des 3*, 141–147.

Rueda D, Bokinsky G, Rhodes MM, Rust MJ, Zhuang X, and Walter NG (2004). Single-molecule enzymology of RNA: Essential functional groups impact catalysis from a distance. *Proc Natl Acad Sci U S A 101*, 10066–10071.

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

Sharp PA (2009). The centrality of RNA. *Cell 136*, 577–580.

Sucheck SJ, and Wong CH (2000). RNA as a target for small molecules. *Curr Opin Chem Biol 4*, 678–686.

Tijerina P, Mohr S, and Russell R (2007). DMS footprinting of structured RNAs and RNA–protein complexes. *Nat Protoc 2*, 2608–2623.

Tsai HY, Masquida B, Biswas R, Westhof E, and Gopalan V (2003). Molecular modeling of the three-dimensional structure of the bacterial RNase P holoenzyme. *J Mol Biol 325*, 661–675.

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

Vendruscolo M, Paci E, Dobson CM, and Karplus M (2001). Three key residues form a critical contact network in a protein folding transition state. *Nature 409*, 641–645.

Vicens Q, and Cech TR (2006). Atomic level architecture of group I introns revealed. *Trends Biochem Sci 31*, 41–51.

Wan Y, Kertesz M, Spitale RC, Segal E, and Chang HY (2011). Understanding the transcriptome through RNA structure. *Nat Rev Genet 12*, 641–655.

Wang B, Wilkinson KA, and Weeks KM (2008). Complex ligand-induced conformational changes in $tRNA^{Asp}$ revealed by single nucleotide resolution SHAPE chemistry. *Biochemistry 47*, 3454–3461.

Wang X, Song X, Glass CK, and Rosenfeld MG (2011a). The long arm of long noncoding RNAs: Roles as sensors regulating gene transcriptional programs. *Cold Spring Harb Perspect Biol 3*, a003756.

Wang XQ, Crutchley JL, and Dostie J (2011b). Shaping the genome with non-coding RNAs. *Curr Genomics 12*, 307–321.

Weeks KM (2010). Advances in RNA structure analysis by chemical probing. *Curr Opin Struct Biol 20*, 295–304.

Westhof E, Dumas P, and Moras D (1988). Restrained refinement of 2 crystalline forms of yeast aspartic-acid and phenylalanine transfer-RNA crystals. *Acta Crystallogr A 44*, 112–123.

White SA, and Draper DE (1987). Single base bulges in small RNA hairpins enhance ethidium binding and promote an allosteric transition. *Nucleic Acids Res 15*, 4049–4064.

White SA, and Draper DE (1989). Effects of single-base bulges on intercalator binding to small RNA and DNA hairpins and a ribosomal RNA fragment. *Biochemistry 28*, 1892–1897.

Wilkinson KA, Merino EJ, and Weeks KM (2006). Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): Quantitative RNA structure analysis at single nucleotide resolution. *Nat Protoc 1*, 1610–1616.

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

Yang S, Parisien M, Major F, and Roux B (2010). RNA structure determination using SAXS data. *J Phys Chem B 114*, 10039–10048.

Yu ET, Hawkins A, Eaton J, and Fabris D (2008). MS3D structural elucidation of the HIV-1 packaging signal. *Proc Natl Acad Sci U S A 105*, 12248–12253.

Zhou Y, and Karplus M (1997). Folding thermodynamics of a model three-helix-bundle protein. *Proc Natl Acad Sci U S A 94*, 14429–14432.

Ziehler WA, and Engelke DR (2001). Probing RNA structure with chemical reagents and enzymes. *Curr Protoc Nucleic Acid Chem*. Unit 6.1.

Zuker M (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res 31*, 3406–3415.