



Computational Methods Toward Accurate RNA Structure Prediction Using Coarse-Grained and All-Atom Models

Andrey Krokhotin, Nikolay V. Dokholyan¹

Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina, Chapel Hill, North Carolina, USA

¹Corresponding author: e-mail address: dokh@unc.edu

Contents

1. Introduction	66
2. Discrete Molecular Dynamics	68
3. Three-Bead Model	69
3.1 Model geometry and interaction potential	69
3.2 Prediction of small RNA structure	73
4. Use of Hydroxyl-Radical Probing to Refine RNA Three-Dimensional Structure	75
4.1 Folding RNA with experimental constraints	75
4.2 Definition of through-space contacts	76
4.3 HRP bias potential	77
4.4 Assignment of interaction parameters	80
4.5 Simulation protocol	82
4.6 Results, significance of predictions, and the scope of applicability of the method	83
5. All-Atom Structure Reconstruction	85
6. iFoldRNA	86
7. Conclusions	87
References	87

Abstract

Computational methods can provide significant insights into RNA structure and dynamics, bridging the gap in our understanding of the relationship between structure and biological function. Simulations enrich and enhance our understanding of data derived on the bench, as well as provide feasible alternatives to costly or technically challenging experiments. Coarse-grained computational models of RNA are especially important in this regard, as they allow analysis of events occurring in timescales relevant to RNA biological function, which are inaccessible through experimental methods alone. We have developed a three-bead coarse-grained model of RNA for discrete molecular dynamics

simulations. This model is efficient in *de novo* prediction of short RNA tertiary structure, starting from RNA primary sequences of less than 50 nucleotides. To complement this model, we have incorporated additional base-pairing constraints and have developed a bias potential reliant on data obtained from hydroxyl probing experiments that guide RNA folding to its correct state. By introducing experimentally derived constraints to our computer simulations, we are able to make reliable predictions of RNA tertiary structures up to a few hundred nucleotides. Our refined model exemplifies a valuable benefit achieved through integration of computation and experimental methods.



1. INTRODUCTION

The central dogma of molecular biology casts ribonucleic acid (RNA) as the biological intermediary of genetic information transmission between deoxyribonucleic acid (DNA) and proteins. Over the past few decades, these essential biopolymers have been implicated in tasks reaching far beyond their canonical designation. Indeed, many noncoding RNAs bear important and diverse physiological functions: catalysis, regulation of cell activity, and cellular response to disease states. This diversity in function is attributed to the variety of three-dimensional (3D) structures accommodated by RNAs. Riboswitches, which are at the focus of this book, are a fundamental example of RNAs involved in cellular regulation through vast structural rearrangement. They are segments of mRNA located primarily in the 5' UTR that form unique 3D structures permissive to conformational change upon metabolite binding. Riboswitch conformational shifts regulate the downstream RNA coding sequence (Serganov & Nudler, 2013). This example emphasizes the importance of 3D structural knowledge for elucidation of RNA functioning mechanisms.

Traditional analytical methods to obtain 3D structure, such as X-ray crystallography or NMR, are limited in applicability to RNA, due to the high conformational flexibility of RNA polymers. X-ray crystallography is particularly sensitive in this regard, and special care must be taken to avoid RNA aggregation and/or misfolding prior to crystal formation (Reyes, Garst, & Batey, 2009). NMR experiments with RNA suffer from poor long-range correlations, due to the low proton density of RNA molecules (Addess & Feigon, 1996). As a result, the Protein Data Bank (PDB; Berman et al., 2000) currently contains 3032 solved RNA structures, out of approximately 6.6 million known noncoding sequences (RNA central, <http://rnacentral.org>). Hence, developing accurate computational methods for

RNA structure prediction is crucial to circumventing the bottleneck imposed by the constraints of current experimental methods.

Computational methods traditionally used for structure prediction include homology modeling, Monte Carlo-based methods and molecular dynamics (MD) simulations. Among these approaches, MD is the only approach that can provide a link between structure and dynamics of macromolecules. MD simulations rely on force fields to describe atomic interactions as a sum of pairwise interactions. Some of the most successful force fields include AMBER, CHARMM, GROMOS, and OPLS (Ponder & Case, 2003). Historically, greater emphasis has been placed on protein modeling, such that these force fields were extensively optimized to describe proteins. However, these force fields are not easily translatable to RNA modeling, due to the complex electrostatics introduced by the sugar and phosphate moieties of RNAs. The flexible conformations distinct to RNAs are not fully represented by the point charges adopted in common force fields for protein modeling. Additionally, the phosphodiester represents a highly polarizable anion, which is affected by solvation (McDowell, Špačková, Šponer, & Walter, 2007). The latter effect is typically neglected in most of the classical force fields, presenting further inaccuracy when extrapolating protein modeling to RNA. During the last two decades, an increasing number of known native RNA structures has improved the quality of current force fields to model RNA, such as variations of the Cornell force field implemented in the AMBER suite: parm94, parm98, parm99, or parmbsc0, and parmOL (Cornell et al., 1995; Pérez et al., 2007).

An extreme complexity of force fields and large number of atoms in all-atom simulations dictates the need to use extensive computational resources, to produce trajectories in the time frame relevant for biological functions. Current state of the art simulations, running in parallel on many processors, allow one to access millisecond range for relatively short proteins (Freddolino, Harrison, Liu, & Schulten, 2010). RNA simulations are even more resource consuming due to the higher number of degrees of freedom in RNA molecules versus proteins: in order to describe the RNA backbone one needs to employ six torsional angles as compared to only two torsional angles for proteins.

Coarse-grained models with a less exhaustive representation of RNA molecules can be computationally cheap and efficient alternatives to detailed all-atom simulations. The energy landscape of coarse-grained models is less rugged and can be sampled more efficiently, while retaining fundamental physicochemical properties of the system (Sim, Minary, & Levitt, 2012).

We have developed a coarse-grained three-bead model of RNA for an efficient implementation of MD, known as discrete molecular dynamics (DMD). Using this model, we are able to correctly predict 3D structures of short RNAs (less than 50 nucleotides) (Ding et al., 2008). With additional experimental constraints our model can be expanded to predict the structure of larger RNAs (up to a few hundred nucleotides) (Ding, Lavender, Weeks, & Dokholyan, 2012), entering the realm of functional RNAs, such as riboswitches, that are ~ 35 –200 nucleotides in length (Serganov & Patel, 2007). The constraints we use include information on base-pairing and solvent accessibilities of different nucleotides. We have also established the iFoldRNA Web server for prediction of RNA tertiary structure, which is freely available for researchers worldwide (Sharma, Ding, & Dokholyan, 2008).



2. DISCRETE MOLECULAR DYNAMICS

MD is routinely used for modeling the time-dependent motions (trajectories) of biological macromolecules (proteins and RNAs). In conventional implementations of MD, forces acting on the atoms (particles) are calculated as derivatives of potentials. These forces are substituted into Newton's equations of motion, which are iteratively solved for every particle in the system through femtosecond time steps. While precise, MD is also computationally costly. Through modern parallelization one is capable of modeling trajectories on the order of milliseconds. Unfortunately, these exhaustive simulations are still unable to address the majority of relevant biological processes, including folding of large macromolecules (i.e., proteins and RNAs) that occur on much longer timescales. To increase the time range available for simulations, we use a simplified description of forces. In DMD, the simplification is realized by approximating the potential force fields describing interactions between particles as stepwise potentials (Fig. 1). In this approximation, a particle changes its velocity only upon collision, while passing the region of a potential barrier. Between collisions, all velocities remain the same. The change of momentum as well as kinetic and potential energies of a particle is determined solely by conservation laws. If the height of a potential barrier is larger than the particle's kinetic energy, the particle bounces back retaining kinetic energy. The advantage of this tactic is that we do not explicitly calculate physical forces nor integrate equations of motion. DMD is a generalized approach, which has been successfully applied to biomolecules (Dokholyan, Buldyrev, Stanley, & Shakhnovich, 1998; Zhou, Karplus, Wichert, & Hall, 1997).

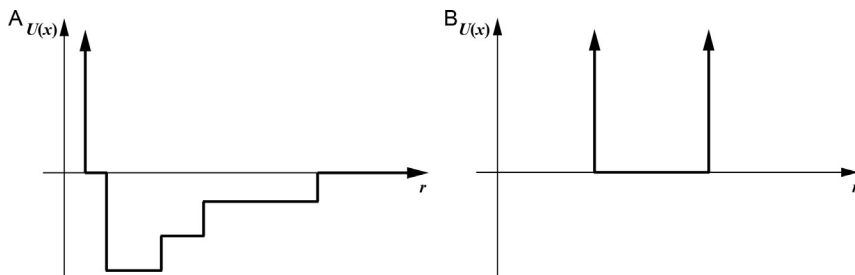


Figure 1 Example potentials used by discrete molecular dynamics (DMD). (A) Discretized van der Waals potential. (B) An infinite square-well potential used to describe covalent bonding.

Unlike MD, which is driven by physical forces, DMD is driven by collision events. This is a consequence of the discretization of interaction potentials characteristic of DMD. In the limit of very fine discretization of potentials, with step size approaching zero, DMD becomes identical to MD. Because DMD is event driven, the efficiency of the algorithm used to search for the next collision event is important for fast performance of DMD. We utilize a collision list approach, which restricts the search for the next collision to atoms located only in the same local area. We developed an efficient implementation of DMD (Proctor, Ding, & Dokholyan, 2011), which can be equally well applied for simulation of different systems starting from uncoupled molecules in an ideal gas system to large macromolecules, such as proteins and RNA. Recently, we also developed a parallel version of DMD (Shirvanyants, Ding, Tsao, Ramachandran, & Dokholyan, 2012). The parallelization was achieved using of a special algorithm, which predicts the possible collisions between atoms that are later accepted or rejected in accordance to the whole dynamics of a simulated system.



3. THREE-BEAD MODEL

3.1. Model geometry and interaction potential

We have developed a coarse-grained representation of RNA structure (Ding et al., 2008). Each nucleotide in our model consists of three beads, representing phosphate (P_i), sugar (S_i), and the nucleobase (B_i). The beads are placed in the center of mass of corresponding chemical moieties (Fig. 2A). Two types of interactions are considered: bonded and nonbonded. Bonded interactions are used to describe chain connectivity and local geometry and include constraints on the lengths of the bonds connecting adjacent beads

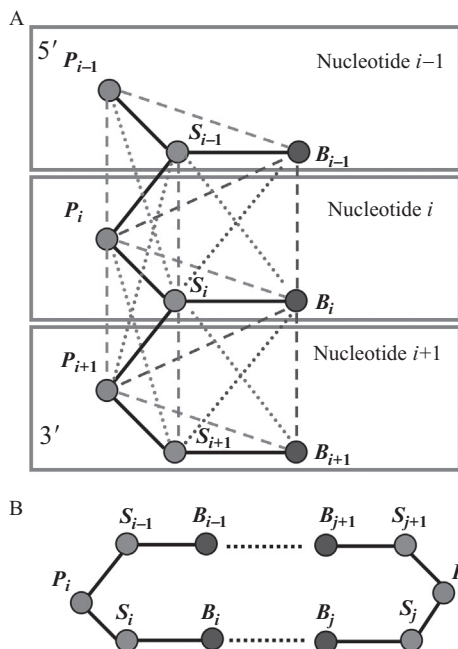


Figure 2 (A) RNA representation in three-bead model. P_i , S_i , and B_i represent phosphate, sugar, and nucleo-base correspondingly. Bonded interactions are shown with constraints on the bond length between nearest beads (solid lines), and on bond angles (dashed lines). (B) Base-pairing model. A hydrogen bond is formed if B_i and B_j come within interaction range, and distances between B_i and S_j , as well as between B_j and S_i provide correct orientation of the bond. Adapted from [Ding et al. \(2008\)](#) with permission.

(solid lines), on the bond angles (dashed lines), and on the dihedral angles (dot-dashed lines). The width of infinite square-well potentials for bonded interactions is determined empirically using structures deposited in the PDB. The assigned width is slightly larger than the standard deviation observed in PDB structures, which is a compromise between the desire to decrease the number of collisions (by increasing the well width), and to avoid large distortions in the structure (by minimizing the well width). Physical terms describing nonbonded interactions inside RNA include base-pairing (A–U, G–C, U–G), base-stacking, phosphate–phosphate repulsion, and hydrophobic interaction.

3.1.1 Base-pairing

Base-pairing is formed through hydrogen bonds between nucleo-bases of the corresponding nucleotides. Hydrogen bonds can be formed if B_i and

B_j fall within interaction range. To ensure that hydrogen binding is orientation dependent, we also check the distances between B_i and S_j as well as between B_j and S_i (Ding, Borreguero, Buldyrey, Stanley, & Dokholyan, 2003) (Fig. 2B).

3.1.2 Phosphate repulsion

The negative charges on phosphate atoms repel them from each other. However, in solution, the interaction potential between two phosphates departs from the ideal Coulomb law due to the screening effects of water on ions. To account for these effects, we use a Debye–Hückel model for phosphate–phosphate interactions where the potential is discretized with a step of 1 Å, and a cutoff distance of 10 Å.

3.1.3 Hydrophobic interactions

We account for the hydrophobicity of base pairs by including attractive interactions between all the bases. To avoid artificial overpacking, we introduce an additional energy term $E = dE \cdot \Theta(n - n_{\max})$, where dE is a repulsion coefficient, $\Theta(n)$ equals 0 when $n \leq 0$ and equals n otherwise; n is the number of nucleotides found in a sphere of 6.5 Å radius. Upon analyzing available RNA structures in the PDB, we found n_{\max} to be 4.2.

3.1.4 Base stacking

Base stacking is modeled by assuming every base can form no more than two stacking contacts, with stacked bases aligned linearly. After completing analysis of RNA structures from the PDB, we defined cutoff base–base distances for different base types: 4.65 Å between purines, 4.60 Å between pyrimidines, and 3.80 Å between purine and pyrimidine. To enforce the linearity of stacked bases, we additionally require that two bases forming stacking interactions to the third one should maintain a minimum distance of 6.5 Å.

3.1.5 Parameterization of the hydrogen-bond, base-stacking, and hydrophobic interactions

Stacking and hydrophobic interaction parameters for all pairs of bases were determined by decomposition of the sequence-dependent free-energy parameters for individual nearest-neighbor hydrogen-bond model (INN-HB) (Mathews, Sabina, Zuker, & Turner, 1999). We assume that neighboring base pairs in INN-HB interact through hydrogen-bond, base-stacking, and hydrophobic interactions. Usually bases B_i and B_{i+1} stack on top of each

other, except when B_{i+1} and B_j are purines. In the latter case, B_{i+1} stacks on top of B_j , as opposed to B_{i+1} on B_i . The distance between B_i and B_{j-1} usually exceeds the cutoff value 6.5 Å. Therefore, we use the following equation to evaluate the pairwise interactions:

$$E \begin{pmatrix} 5' & B_i & B_{i+1} & 3' \\ 3' & B_j & B_{j-1} & 5' \end{pmatrix} = \begin{cases} \left(E_{B_i B_j}^{\text{HB}} + E_{B_{i+1} B_{j-1}}^{\text{HB}} \right) / 2 + E_{B_j B_{i+1}}^{\text{Stack}} \\ \quad + E_{B_i B_{i+1}}^{\text{Hydrophobic}} + E_{B_j B_{j-1}}^{\text{Hydrophobic}}, & B_{i+1}, B_j = \text{purines} \\ \left(E_{B_i B_j}^{\text{HB}} + E_{B_{i+1} B_{j-1}}^{\text{HB}} \right) / 2 + E_{B_i B_{i+1}}^{\text{Stack}} + E_{B_j B_{j-1}}^{\text{Stack}} \\ \quad + E_{B_{i+1} B_j}^{\text{Hydrophobic}}, & \text{otherwise} \end{cases}$$

Here, E^{Stack} , E^{HB} , and $E^{\text{Hydrophobic}}$ are interaction strengths of stacking, base-pair, and hydrophobic interactions. Their values were determined from experimentally tabulated values of interaction energies between neighboring base pairs (Mathews et al., 1999) using singular value decomposition.

3.1.6 Loop entropy

Correctly accounting for loop entropy is essential for predicting RNA folding kinetics (Tinoco & Bustamante, 1999). Due to the coarse-grained nature of our model, the entropy is often underestimated in DMD simulations, causing formation of long unnatural loops. To deal with this problem, we explicitly take entropy into account, using the experimentally determined values of free energy measured for different types of RNA loops (Mathews et al., 1999). We compute the free energy of loops in DMD simulations based on formation or breakage of base-pair contacts. The algorithm used to form base pairs in DMD includes the evaluation of the loop free-energy difference ΔG^{loop} due to base-pair formation followed by the formation of a base pair with probability calculated according to the Boltzmann distribution ($p = \exp(-\beta \Delta G^{\text{loop}})$). Kinetic energy of the nucleotides should be enough to overcome the potential barrier for base-pair formation. The breakage of a base pair is only governed by energy, momentum, and angular momentum conservation. We do not use stochastic procedures for breaking base pairs because this process is always entropically favorable.

The total potential energy E used for DMD simulation of RNA molecules is calculated as a sum of all interaction terms:

$$E = E_{\text{Bonded}} + E_{\text{Hbond}} + E_{\text{Stack}} + E_{\text{Hydrophobic}} + E_{\text{Overpacking}} + G_{\text{loop}}$$

3.2. Prediction of small RNA structure

The described method was tested on a set of 153 short RNAs, ranging in length from 10 to 100 nucleotides (Ding et al., 2008). All of these RNAs have experimentally obtained 3D structures deposited in the PDB. We used replica exchange DMD simulations in order to increase sampling efficiency (Sugita & Okamoto, 1999). In each simulation, we ran eight replicas at temperatures: 0.200, 0.208, 0.214, 0.220, 0.225, 0.230, 0.235, and 0.240. The temperature is given in abstract units of kcal/(mol k_B). The replicas were allowed to exchange every 2000 DMD time steps, with 2×10^6 total time steps. The quality of reconstructed structures was monitored using two observables: (1) root-mean-square deviation (RMSD), calculated between backbone phosphate atoms of the model and experimental structure, (2) the fraction of correctly predicted native base pairs (Q -value). We observed that for RNAs with length less than 50 nucleotides, the predicted RMSD fell below 6 Å (Fig. 3A). Models of longer RNA molecules demonstrated larger RMSD; however, 84% of all predicted structures have RMSD less than 4 Å. We also observed that predicted structures corresponding to the lowest free energy do not generally have the lowest RMSD. It can be attributed to the inaccuracy of the force field and coarse-graining nature of the RNA model. For the majority of the lowest free-energy RNA structures, the observed Q -values were close to unity (Fig. 3B), suggesting the correct formation of native base pairs. The average Q -value for all 153 RNA molecules was 94%. For comparison with existing secondary structure prediction methods, we calculated the Q -values using Mfold (Zuker, 2003). Its average Q -value was 91%.

Pseudoknot structure (PDB code: 1A60) demonstrates an example of the performance of our method. This pseudoknot represents the T-arm and acceptor stem of the turnip yellow mosaic virus (TYMV) and shares structural similarity with TYMV genomic tRNA (Kolk et al., 1998). We found the RMSD of the lowest free-energy structure to be 4.58 Å, while the RMSD of the structure with minimal deviation from the crystal structure was 2.03 Å (Fig. 4A). We studied the folding thermodynamics of this RNA using weighted histogram analysis method (Kumar, Rosenberg, Bouzida, Swendsen, & Kollman, 1992) as implemented in the MMTSB toolset (Feig, Karanicolas, & Brooks, 2004). The specific heat (Fig. 4B) has two peaks: one peak around $T^*=0.245$ and another one around $T^*=0.21$. This result suggests the presence of intermediate states in the folding pathway. To characterize folding intermediates, we computed the two-dimensional potential of mean force (2D-PMF) as a function of

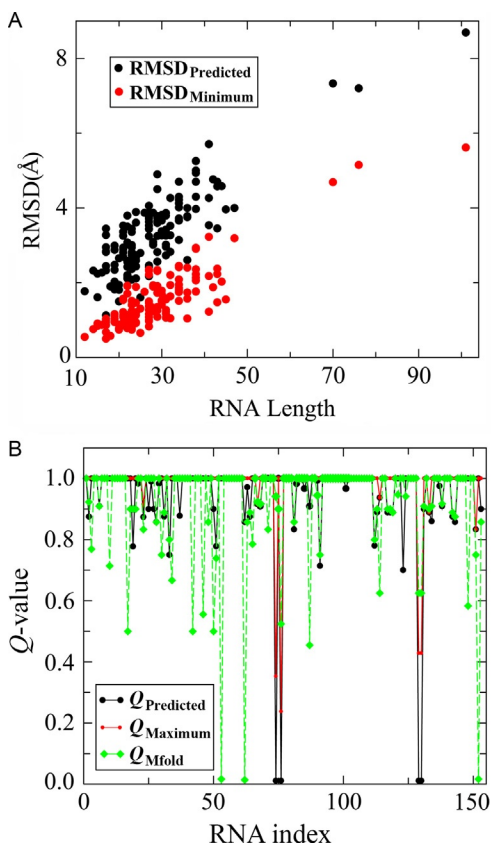


Figure 3 (A) RMSD calculated between RNA structure obtained in the course of DMD simulations and experimental structure from the PDB. Black points correspond to the predicted structures with the lowest free energy and red points correspond to structures with minimum RMSD. (B) Q-values for the predicted structures with the lowest free energy (black dots), the maximum Q-values (red dots) and structures obtained using Mfold (green dots). Adapted from [Ding et al. \(2008\)](#) with permission.

the total number of base pairs and the number of native base pairs at two peak temperatures (Fig. 4C and D). The 2D-PMF plots show two intermediate states with distinct free-energy basins. The first state corresponds to the folded 5' hairpin, while the second state corresponds to the formation of one of the helix stems in the 3' pseudoknot. Progression of the RNA folding toward the native state can be demonstrated at the contact frequencies of the folding intermediates (Fig. 4E and F) and the contact map of the native state (Fig. 4G).

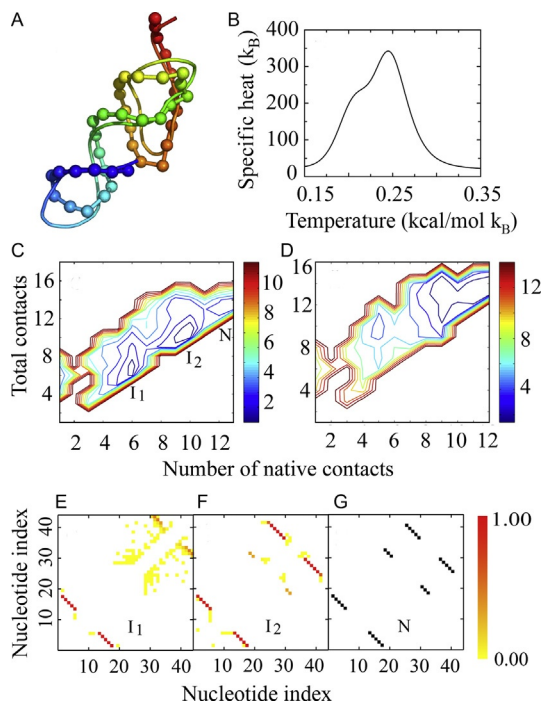


Figure 4 Folding kinetics of pseudoknot RNA. (A) Structure obtained as a result of DMD simulation (ribbon backbone trace with backbone spheres) superimposed over experimentally determined RNA structure (PDB code: [1A60](#)). Backbone ribbons are colored from blue (N terminus) to red (C-terminus). (B) Specific heat as a function of temperature. (C) Two-dimensional potential of mean force (2D-PMF) for pseudoknot folding at $T^* = 0.245$ (the major peak in the specific heat). I_1 , I_2 , and N denote two intermediate states and native state correspondingly. (D) 2D-PMF at $T^* = 0.21$. (E) Contact frequencies at the I_1 intermediate state, corresponding to the formation of 5' hairpin. (F) Contact frequencies at the I_2 intermediate state, corresponding to the formation of the helix stem in the 3' pseudoknot. (G) Contact map of the native state as observed in the experimental NMR structure (PDB code: [1A60](#)). Adapted from [Ding et al. \(2008\)](#) with permission.



4. USE OF HYDROXYL-RADICAL PROBING TO REFINE RNA THREE-DIMENSIONAL STRUCTURE

4.1. Folding RNA with experimental constraints

The quality of RNA tertiary structure prediction worsens with the increase of RNA length. There are two major reasons for this: (1) the inaccuracy of existing force fields, in particular, describing electrostatics by point charges

and ignoring polarizability effects, and (2) the need to sample larger conformational space in order to find the correct RNA fold. However, if any structural experimental data is available, it can be used to constrain RNA simulations and enhance sampling efficiency. For example, *a priori* knowledge of RNA secondary structure greatly restrains possible tertiary folds (Bailor, Mustoe, Brooks, & Al-Hashimi, 2011; Hajdin, Ding, Dokholyan, & Weeks, 2010). The secondary structure can be determined with high fidelity using comparative sequence analysis (Cannone et al., 2002; Gutell, Power, Hertz, Putz, & Stormo, 1992; Michel & Westhof, 1990) or SHAPE experiments (Deigan, Li, Mathews, & Weeks, 2008; Weeks, 2010). The conformational space can be constrained even further if the experimental information on the tertiary contacts, not involved in base-pairing, is available (Gherghe, Leonard, Ding, Dokholyan, & Weeks, 2009; Lavender, Ding, Dokholyan & Weeks, 2010). Here, we describe a method that relies on the information extracted from hydroxyl-radical probing (HRP) experiments to constrain RNA folds (Ding et al., 2012). This method shows high efficiency in predicting the structure of long RNAs. HRP experiments do not provide direct information about tertiary contacts between any particular pair of nucleotides. Instead they probe the solvent accessibility of nucleotides that can be related to the expected number of contacts that each particular nucleotide makes with its through-space neighbors. In general, nucleotides showing lower HRP reactivity are more buried and have more through-space contacts. This information can be translated into statistics-based potential biasing RNA toward correct 3D structure.

4.2. Definition of through-space contacts

The bias potential introduced here is based on the number of contacts that every nucleotide makes with its through-space neighbors. We define such contacts based on the distance between sugar pseudoatoms in our three-bead model. We exclude nucleotides located within four nucleotides in the sequence, and within four nucleotides from the nucleotide forming a base pair with the nucleotide in question, as these contacts reflect primary and secondary structure rather than higher order tertiary contacts. The number of through-space neighbor contacts depends on the cutoff distance d_{cutoff} used to define the contact. If this distance is too small, the number of contacts will be underestimated. On the other hand if the cutoff distance is too large, the number of contacts will be overestimated, and will not correlate with the

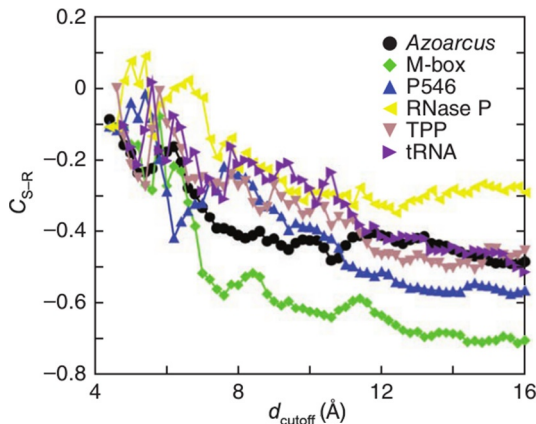


Figure 5 The structure–reactivity correlation, C_{S-R} as a function of the cutoff distance, used to define a contact. HRP reactivities were smoothed over a sliding window of three nucleotides to reduce noise (as described in the text). The data points are shown for six RNAs from the training set (Table 1). Adapted from Ding et al. (2012) with permission.

values of HRP reactivity. In order to optimize d_{cutoff} , we calculated how the structure–reactivity correlation C_{S-R} depends on d_{cutoff} , where C_{S-R} was defined as a Pearson correlation coefficient between nucleotide HRP reactivity and the number of contacts it makes within cutoff distance (Fig. 5). We used data collected from six RNAs as a training set to optimize model parameters (Table 1). The calculated C_{S-R} is negative because a lower HRP reactivity corresponds to a more buried nucleotide with a larger number of through-space contacts. We noticed that at $d_{\text{cutoff}} = 14 \text{ \AA}$, C_{S-R} reaches its minimum value. This optimized value of d_{cutoff} is further used to calculate the number of contacts.

4.3. HRP bias potential

Our HRP bias potential consists of two terms. The first term is the sum of pairwise attractive potentials assigned to every pair of nucleotides. The second term is an overburial repulsion potential that penalizes the energy function when a given nucleotide exceeds its maximum number of allowed contacts (N_{max}) (Fig. 6A). The strength of attraction and repulsion potentials as well as N_{max} is assigned to every nucleotide depending on its HRP reactivity. In general, the bias interaction potential can be written as:

$$E_{\text{bias}} = \sum_{i < j} E_{ij}^{\text{attr}} + \sum_i E_i^{\text{overbury}}$$

Table 1 The first six RNAs comprise the training set used for algorithm optimization and determination of its applicable range and the last four RNAs comprise the verification set used to rate the performance of the algorithm

RNA	Length (nt)	<i>f</i> (0.25)	<i>C</i> _{S-R}	Large clusters			
				Number of clusters	<i>n</i> (out of 100)	RMSD (Å)	<i>P</i> -value
<i>Azoarcus</i> group I intron	214	0.43	−0.45	1	100	16.8±2.1	<10 ^{−6}
M-box riboswitch	161	0.35	−0.68	1	100	11.6±2.3	<10 ^{−6}
P546 domain	158	0.37	−0.57	3	66	19.8±1.4	3.0×10 ^{−3}
					32	15.1±1.9	<10 ^{−6}
RNase P specificity domain	152	0.25	−0.30	3	93	24.9±1.2	0.67
					4	24.1±3.2	0.50
					3	22.7±0.5	0.22
TPP riboswitch	80	0.21	−0.50	4	44	12.6±1.3	0.10
					42	14.6±1.9	0.44
					12	9.9±1.8	2.9×10 ^{−3}
tRNA ^{Asp}	75	0.25	−0.45	8	29	14.1±1.3	0.50
					23	17.5±1.5	0.97
					18	16.4±0.8	0.90
					8	18.9±0.8	0.99
					6	12.3±1.5	0.17
					6	15.7±0.4	0.82
					5	18.9±0.8	0.99
<i>O. iheyensis</i> group II intron	412	0.21	−0.30	−	−	−	−
RNase P catalytic domain	231	0.28	−0.50	4	46	19.2±1.6	<10 ^{−6}
					41	21.6±2.1	<10 ^{−6}
					8	25.0±0.7	1.3×10 ^{−4}
					5	24.4±2.0	3.5×10 ^{−5}

Table 1 The first six RNAs comprise the training set used for algorithm optimization and determination of its applicable range and the last four RNAs comprise the verification set used to rate the performance of the algorithm—cont'd

RNA	Length (nt)	$f(0.25)$	C_{S-R}	Large clusters			
				Number of clusters	n (out of 100)	RMSD (Å)	P -value
Lysine riboswitch	174	0.36	−0.57	3	57	12.0 ± 1.6	$<10^{-6}$
					42	18.1 ± 1.0	1.6×10^{-6}
glmS ribozyme	152	0.35	−0.55	2	74	16.6 ± 1.7	1.5×10^{-5}
					26	8.5 ± 1.3	$<10^{-6}$

The structure–reactivity correlation, C_{S-R} , was calculated with reference to the experimental structure. For each RNA, one hundred of the lowest free-energy structures were selected and clustered with the cluster centroid chosen as a representative structure. Low populated clusters with less than two structures were excluded. RMSD was calculated between the centroid of a cluster and the corresponding experimental structure. P -values were calculated based on the average RMSD with respect to the accepted experimental structure. Structure refinement was not performed for *O. ihayensis* group II intron RNA in the test set because of the low $f(0.25)$ value. Adapted from Ding et al. (2012) with permission.

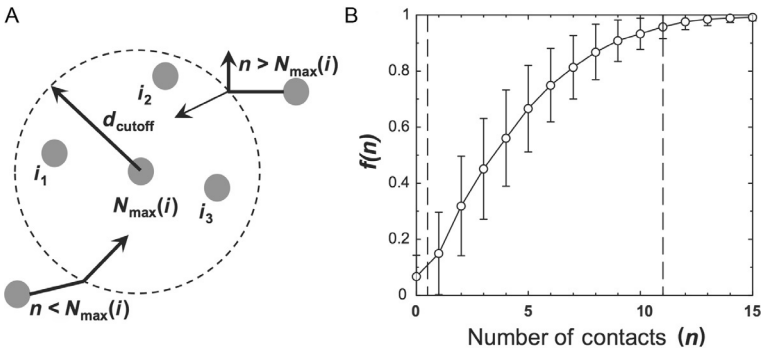


Figure 6 (A) Two nucleotides make a contact if they fall within cutoff distance, d_{cutoff} . Every nucleotide is assigned a threshold number of contacts N_{max} . If the number of already existing contacts is less than N_{max} , a new contact is energetically favorable (shown by an inward arrow). However, if the number of contacts exceeds N_{max} , every new approaching nucleotide can form a contact only if the total DMD kinetic energy is sufficient to overcome the energy penalty for overpacking; otherwise it bounces back (shown by outward arrow). (B) The cumulative distribution of the fraction of nucleotides, $f(n)$, as a function of the number of contacts, n , as follows from statistical analysis of all single chain RNAs in the PDB. The values of NC_{min} and NC_{max} are shown by vertical dashed lines. Adapted from Ding et al. (2012) with permission.

Here, $E_{ij}^{\text{attr}} = \min\{E_{\text{attr}}(i), E_{\text{attr}}(j)\}F(r_{ij})$. $E_{\text{attr}}(i)$ defines the strength of attractive interaction for every particular nucleotide based on its reactivity value, and

$$F(x) = \begin{cases} 0 & d_{\text{cutoff}} < r \\ 1 & R_{\text{hc}} < r < d_{\text{cutoff}} \\ \infty & r \leq R_{\text{hc}} \end{cases}$$

Here, $d_{\text{cutoff}} = 14 \text{ \AA}$ (as defined above), and R_{hc} is a hardcore diameter, 3.0 \AA .

$$E^{\text{overbury}}(i) = dE_{\text{rep}}(i)(n_{\text{C}}(i) - N_{\text{max}}(i))\Theta(n_{\text{C}}(i) - N_{\text{max}}(i))$$

Here, $n_{\text{C}}(i)$ is the number of contacts formed by nucleotide i , $dE_{\text{rep}}(i)$ is the penalty energy for overburying, which depends on HRP-reactivity value of nucleotide i , and Θ is a unit step function:

$$\Theta(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

4.4. Assignment of interaction parameters

The maximally and minimally buried nucleotides make the largest (NC_{max}) and smallest (NC_{min}) possible number of through-space contacts, respectively. To determine these numbers, we performed a statistical analysis of contacts between nucleotides in all single chain RNAs from the PDB (Berman et al., 2000). Based on the cumulative distribution of the fraction of nucleotides $f(n)$ as a function of the number of contacts n (Fig. 6B), we derived that $NC_{\text{min}} = 0.5$ and $NC_{\text{max}} = 11$. By choosing these values, we avoided the tails of the $f(n)$ distribution. We want to assign NC_{min} and NC_{max} to nucleotides exhibiting maximal and minimal reactivity values, correspondingly. HRP experiments are intrinsically noisy, so to account for the noise we average reactivities over a sliding window of three nucleotides. We also define threshold values of HRP reactivity R_{min} and R_{max} . Nucleotides with reactivity R , lower than R_{min} are maximally buried and assigned $N_{\text{max}} = NC_{\text{max}}$, while nucleotides exceeding R_{max} are maximally exposed and assigned $N_{\text{max}} = NC_{\text{min}}$. The R_{min} and R_{max} values are calculated as averages of subsets of rank-ordered R values from 2% to 20% and from 80% to 98% correspondingly. The top and bottom 2% of R values are not taken into account to reduce the effect of extreme R values observed

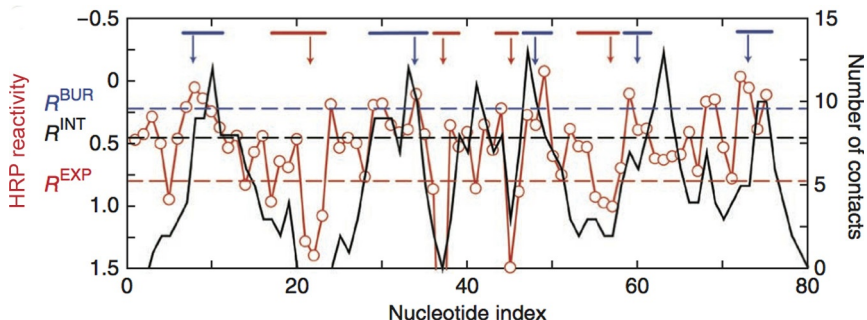


Figure 7 Experimentally measured HRP reactivities (red (light gray in the print version)) and the number of through-space contacts (black) for TPP riboswitch. Buried/exposed nucleotide segments are highlighted in blue (dark gray in the print version)/red (light gray in the print version) bars at the top of the figure. Arrows pointing downward show nucleotides, which can be identified as buried or exposed with high confidence. The dashed horizontal lines represent threshold values used to define exposed (R^{EXP}), intermediate (R^{INT}), or buried (R^{BUR}) nucleotides. Adapted from [Ding et al. \(2012\)](#) with permission.

in a typical HRP experiment. For nucleotides with intermediate reactivities ($R_{\min} < R < R_{\max}$), we define N_{\max} as

$$N_{\max}(i) = NC_{\max} + (NC_{\min} - NC_{\max})(R(i) - \langle R_{\min} \rangle) / (\langle R_{\max} \rangle - \langle R_{\min} \rangle)$$

The intrinsic noise inherent to HRP experiments ([Fig. 7](#)) render estimations of nucleotide solvent exposure less reliable. However, we can identify nucleotides as buried or exposed with high confidence if those nucleotides are found in buried or exposed RNA segments. To define buried or exposed segments, we first define three values, R^{EXP} , R^{INT} , and R^{BUR} , corresponding to the threshold values of exposed, intermediate, and buried residues. Intermediate R values can be defined in two different ways: as an average value of the reactivities, $\langle R \rangle$, or as the median value of the rank-ordered reactivities. To avoid ambiguity we use the mean of these two values, denoted as R^{INT} . We define buried segments as segments with more than three consecutive nucleotides having $R < R^{\text{INT}}$ and at least one nucleotide having $R < R^{\text{BUR}}$. Inside each buried segment, a nucleotide with the lowest R (excluding the first and the last nucleotide in the segment) is considered as buried with high confidence. Exposed segments are defined as segments with more than three

consecutive nucleotides having $R > R^{\text{INT}}$, and at least one nucleotide having $R > R^{\text{EXP}}$, and for two-nucleotide segments with both nucleotides having $R > R^{\text{EXP}}$. Inside each exposed segment, a nucleotide having the largest reactivity is considered as exposed with high confidence.

Two attractive energy scales were used $E_{\text{attr}}^{\text{low}} = -10 \text{ kcal/mol}$ and $E_{\text{attr}}^{\text{high}} = -0.05 \text{ kcal/mol}$. The strong attractive energy $E_{\text{attr}}^{\text{low}}$ was assigned to the buried nucleotides identified with high confidence, and intermediate values $(E_{\text{attr}}^{\text{low}} + E_{\text{attr}}^{\text{high}})/2$ were assigned to their immediate neighbors. For all remaining nucleotides $E_{\text{attr}}^{\text{high}}$ was assigned. The strong repulsive overburial energy was assigned to all high-confidence buried and exposed nucleotides. The repulsive energy $dE_{\text{rep}}(i) = -E_{\text{attr}}(i)$ was assigned to all other nucleotides, where $E_{\text{attr}}(i)$ is equal to $E_{\text{attr}}^{\text{low}}$ or $E_{\text{attr}}^{\text{high}}$. By making overburial repulsive potential equal to attractive potential, we allow nucleotides to make additional contacts without a net energy penalty.

4.5. Simulation protocol

The simulation protocol, which we apply along with HRP bias potential, is an essential part of this method. The aim of this protocol is to increase the sampling efficiency of conformational space and hence increase the precision of our method. The protocol consists of three steps (Fig. 8). At the first step, we perform DMD simulations with the input RNA sequence and base-pairing constraints. This allows us to form an initial 3D structure of modeled RNA. For all the RNAs shown in Table 1, high-resolution structures are available in PDB (Berman et al., 2000). However in the course of *de novo* prediction, reliable base-pairing information can be obtained from sequence comparative analysis (Cannone et al., 2002; Gutell et al., 1992; Michel & Westhof, 1990) or from SHAPE data (Deigan et al., 2008; Weeks, 2010). At the second step, after the formation of native secondary structure, we perform a second round of simulations with HRP bias potential (Fig. 2C). After each round of simulations, we check if the experimental constraints are satisfied and if they are not, we repeat the simulation. In order to sample conformational space of RNAs more efficiently, we use DMD with replica exchange. At the third step, we select one hundred structures based on low energy and high $C_{\text{S-R}}$ values. We cluster these structures and take centroids of every cluster as a representative structure.

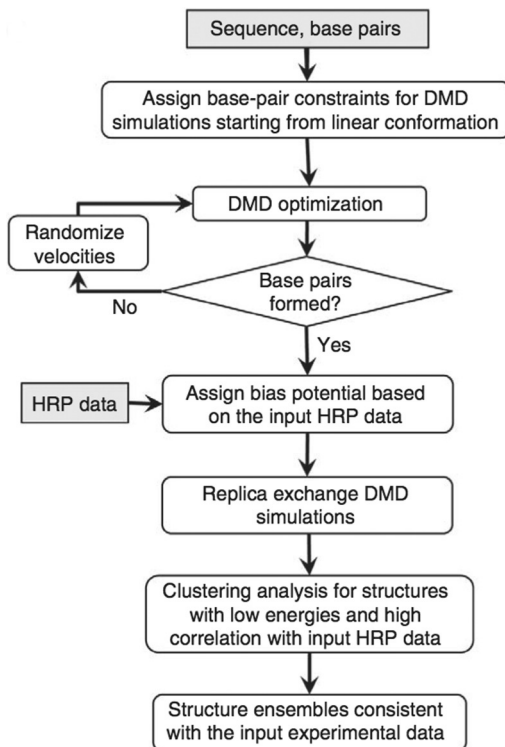


Figure 8 RNA simulation algorithm with the use of base-pairing constraints and constraints from hydroxyl-radical probing experiments. Adapted from [Ding et al. \(2012\)](#) with permission.

4.6. Results, significance of predictions, and the scope of applicability of the method

We optimized the modeling procedure on a training set of six RNA: the *Azoarcus* group I intron, the M-box riboswitch, the P546 domain of the *Tetrahymena thermophila* group I intron, the RNase P specificity domain, the thiamine pyrophosphate (TPP) riboswitch, and the yeast aspartic acid transfer (tRNA^{Asp}) RNA (Table 1). These are structurally diverse RNAs, with lengths ranging from 75 to 214 nucleotides. We used replica exchange DMD with eight replicas at the temperatures: 0.200, 0.225, 0.250, 0.270, 0.300, 0.333, 0.367, and 0.400. Replicas were exchanged every 1000

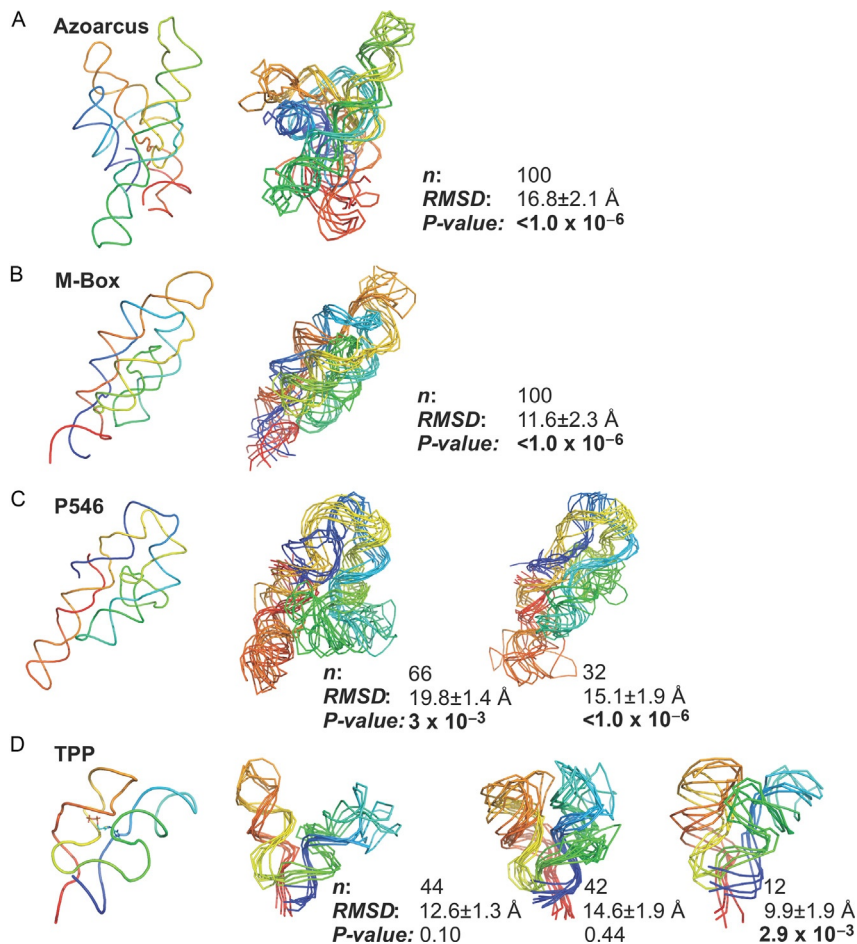


Figure 9 Predicted structures of RNAs from the training set. RNAs are shown with backbone traces. The accepted structures are shown on the left. Representative structures for each of the highly populated cluster are shown on the right. Backbones are colored from blue to red in the direction from 5' to 3'. The mean RMSD and *P*-values are shown, with significant *P*-values emphasized in bold. Adapted from [Ding et al. \(2012\)](#) with permission.

DMD time steps. The simulations were run for 5×10^5 DMD time steps. The temperature is measured in abstract units of kcal/(mol k_B). To evaluate the accuracy of our method we compared predicted structures with available structures in the PDB ([Table 1](#); [Fig. 9](#)). In particular, for each representative predicted structure we calculated RMSD relative to the structure from the PDB, and a *P*-value showing the significance of the prediction. *P*-values

lower than 0.01 correspond to highly statistically significant predictions (Hajdin et al., 2010). It is interesting to note that observed RMSD values corresponding to significant predictions are different for RNAs of different length. For example, a prediction with an RMSD of 7–10 Å is not significant for small RNAs, but is highly significant for RNAs beyond 100 nucleotides. The structures of the three largest RNAs in the training set were predicted with high significance values (Table 1). In addition, all of the predicted structures for the *Azoarcus* group I intron and for M-box riboswitch fell into a single cluster and the P546 domain showed two resulting clusters. The structures for TPP riboswitch fell into three clusters, and only one of them was highly significant. For tRNA^{Asp} and RNase P specificity domain no predictions were statistically significant (Table 1). We noticed that the quality of prediction is related to the packing density of RNA. Those RNAs with higher order RNA packing returned better predictions. The extent of RNA packing can be estimated *a priori* as a fraction of nucleotides $f(r)$ falling below a given reactivity r . At $r=0.25$, the RNAs with extensive packing (the *Azoarcus* group I intron, M-box riboswitch, and P546 domain) have significantly larger $f(r)$ values than other RNAs (Table 1). We determined that our HRP directed refinement method generally predicts correct structure for RNAs with $f(0.25) > 0.25$. After the optimization, we tested our methods on a set of four RNAs: *Oceanobacillus iheyensis* group II intron, RNase P catalytic domain, lysine riboswitch, and the glmS ribozyme (ranged from 152 to 412 nucleotides in length). We observed that three RNAs from the training set yielded highly significant predicted structures grouped in a few clusters. For the last RNA (*O. iheyensis* group II intron), we did not find any reasonable structures at all, which confirms the empirically derived rule of $f(0.25)$.



5. ALL-ATOM STRUCTURE RECONSTRUCTION

We developed an algorithm to translate RNA from a three-bead representation to an all-atom structure.

The algorithm consists of two steps. In the first step, we build an initial structure from template A, G, C, and U nucleotides, which are randomly chosen from the PDB (Berman et al., 2000). To accomplish this step, we define a three-bead representation to every template nucleotide and determine an optimal translation and rotation that superimposes the three-bead template nucleotide over the three-bead nucleotide from the RNA. The resulting transformation of coordinate system applied to all atoms of

template nucleotides provides an initial all-atom RNA structure, with many nonideal bonds, angles, and dihedral angles between adjacent nucleotides. In the second step, we further refine the all-atom structure through all-atom DMD simulations to remove clashes and connect bonds. We also impose base-pairing constraints derived from the coarse-grained simulations (or experimental input). Our all-atom DMD relies on the Medusa force field (Yin, Biedermannova, Vondrasek, & Dokholyan, 2008). Severe clashes in the initial all-atom structure can lead to instantaneous heating of the RNA and cause large distortions compared to the initial three-bead model before the system cools down to the thermostat temperature. We overcome this obstacle using a high heat-exchange coefficient, which allows the system to cool quickly thus preventing it from destruction.



6. iFoldRNA

We created and support the iFoldRNA Web server (Sharma et al., 2008) for the prediction of RNA structure and the analysis of various thermodynamic properties of RNA folding. iFoldRNA operates in two different prediction modes. In the first mode, RNA structures can be predicted *de novo* without any additional constraints. This approach works well for short RNAs with a length of RNA limited to 50 nucleotides at most (Ding et al., 2008). In the second mode, a user can provide base-pairing constraints and an optional HRP-reactivity profile. If an HRP-reactivity profile is provided, the additional biasing potential is used to guide RNA folding (Ding et al., 2012). To sample RNA conformational space more efficiently, the server runs eight replicas at different temperatures. The default values of replica temperatures, replica exchange time and number of steps are the same as described in the text above for the unconstrained and constrained simulations correspondingly. Constant temperature throughout the course of the simulation is supported by the Andersen thermostat (Andersen, 1980). To perform simulations, we use a Linux-based computing cluster with more than 1840 computing cores from the University of North Carolina. During the simulation, the state of the system in each of the eight trajectories is saved every 100 DMD steps. After the DMD simulation is finished, saved states are subjected to automated clustering. The centroids of these clusters are selected as representative structures. The default value for the number of output clusters is 1. The Web interface of iFoldRNA server allows one to rewrite the default values for the maximum number of simulation DMD time steps, replica temperatures, and the number of output clusters.

For each of the representative structures, we reconstruct an all-atom geometry based on the three-bead model obtained from the previous steps of the simulation. Simulation time depends on the size of RNA molecule and on the constraints applied during the simulation. We observed that simulation timescales linearly with RNA size. As an example, the simulation of M-box riboswitch (160 nucleotides) takes 170 CPU hours. The wall-clock time is one-eighth of the total CPU time.



7. CONCLUSIONS

Computational modeling can be used to obtain information about RNA 3D structure and functions that are otherwise difficult and expensive to obtain experimentally. Methods like DMD and geometry coarse graining are especially important in this respect because they allow access to biologically relevant timescales, which is beyond the reach of ordinary MD in the extreme complexity of RNA modeling. We have developed an RNA simulation package based on our coarse-grained three-bead RNA model coupled to an efficient implementation of MD, known as DMD. Our approach efficiently predicts the structure of short RNAs (less than 50 nucleotides) *de novo*, from only RNA sequence. Additionally, experimental constraints, such as base-pairing information or reactivities derived from HRP experiments, allow us to get precise predictions for RNAs in the range of a few hundred nucleotides. Our method consistently shows good performance at the RNA-Puzzle competition (Cruz et al., 2012). RNA modeling is a rapidly developing field and in order to keep pace we continue to develop our method. In particular, we are currently optimizing the Medusa force field used to run all-atom RNA simulations. We are also working to develop better ways to incorporate experimental constraints into DMD simulations.

REFERENCES

- Adress, K. J., & Feigon, J. (1996). Introduction to ^1H NMR spectroscopy of DNA. In S. M. Hecht (Ed.), *Bioorganic chemistry: Nucleic acids* (pp. 163–185). New York: Oxford University Press.
- Andersen, H. C. (1980). Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4), 2384–2393.
- Bailor, M. H., Mustoe, A. M., Brooks, C. L., III, & Al-Hashimi, H. M. (2011). Topological constraints: Using RNA secondary structure to model 3D conformation, folding pathways, and dynamic adaptation. *Current Opinion in Structural Biology*, 21(3), 296–305.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Research*, 28(1), 235–242.

- Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D'Souza, L. M., Du, Y., et al. (2002). The comparative RNA web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, 3(1), 2.
- Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., et al. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19), 5179–5197.
- Cruz, J. A., Blanchet, M. F., Boniecki, M., Bujnicki, J. M., Chen, S. J., Cao, S., et al. (2012). RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, 18(4), 610–625.
- Deigan, K. E., Li, T. W., Mathews, D. H., & Weeks, K. M. (2008). Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences*, 106(1), 97–102, pnas-0806929106.
- Ding, F., Borreguero, J. M., Buldyrev, S. V., Stanley, H. E., & Dokholyan, N. V. (2003). Mechanism for the α -helix to β -hairpin transition. *Proteins: Structure, Function, and Genetics*, 53, 220–228.
- Ding, F., Lavender, C. A., Weeks, K. M., & Dokholyan, N. V. (2012). Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nature Methods*, 9(6), 603–608.
- Ding, F., Sharma, S., Chalasani, P., Demidov, V. V., Broude, N. E., & Dokholyan, N. V. (2008). Ab initio RNA folding by discrete molecular dynamics: From structure prediction to folding mechanisms. *RNA*, 14(6), 1164–1173.
- Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E., & Shakhnovich, E. I. (1998). Discrete molecular dynamics studies of the folding of a protein-like model. *Folding and Design*, 3(6), 577–587.
- Feig, M., Karanicolas, J., & Brooks, C. L., III. (2004). MMTSB Tool Set: Enhanced sampling and multiscale modeling methods for applications in structural biology. *Journal of Molecular Graphics and Modelling*, 22(5), 377–395.
- Freddolino, P. L., Harrison, C. B., Liu, Y., & Schulten, K. (2010). Challenges in protein-folding simulations. *Nature Physics*, 6(10), 751–758.
- Gherghe, C. M., Leonard, C. W., Ding, F., Dokholyan, N. V., & Weeks, K. M. (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), 2541–2546.
- Gutell, R. R., Power, A., Hertz, G. Z., Putz, E. J., & Stormo, G. D. (1992). Identifying constraints on the higher-order structure of RNA: Continued development and application of comparative sequence analysis methods. *Nucleic Acids Research*, 20(21), 5785–5795.
- Hajdin, C. E., Ding, F., Dokholyan, N. V., & Weeks, K. M. (2010). On the significance of an RNA tertiary structure prediction. *RNA*, 16(7), 1340–1349.
- Kolk, M. H., van der Graaf, M., Wijmenga, S. S., Pleij, C. W., Heus, H. A., & Hilbers, C. W. (1998). NMR structure of a classical pseudoknot: Interplay of single- and double-stranded RNA. *Science*, 280(5362), 434–438.
- Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., & Kollman, P. A. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8), 1011–1021.
- Lavender, C. A., Ding, F., Dokholyan, N. V., & Weeks, K. M. (2010). Robust and generic RNA modeling using inferred constraints: A structure for the hepatitis C virus IRES pseudoknot domain. *Biochemistry*, 49(24), 4931–4933.
- Mathews, D. H., Sabina, J., Zuker, M., & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5), 911–940.

- McDowell, S. E., Špačková, N., Šponer, J., & Walter, N. G. (2007). Molecular dynamics simulations of RNA: An in silico single molecule approach. *Biopolymers*, 85(2), 169–184.
- Michel, F., & Westhof, E. (1990). Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *Journal of Molecular Biology*, 216(3), 585–610.
- Pérez, A., Marchán, I., Svozil, D., Šponer, J., Cheatham, T. E., III, Laughton, C. A., et al. (2007). Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers. *Biophysical Journal*, 92(11), 3817–3829.
- Ponder, J. W., & Case, D. A. (2003). Force fields for protein simulations. *Advances in Protein Chemistry*, 66, 27–86.
- Proctor, E. A., Ding, F., & Dokholyan, N. V. (2011). Discrete molecular dynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(1), 80–92.
- Reyes, F. E., Garst, A. D., & Batey, R. T. (2009). Strategies in RNA crystallography. *Methods in Enzymology*, 469, 119–139.
- Serganov, A., & Nudler, E. (2013). A decade of riboswitches. *Cell*, 152(1), 17–24.
- Serganov, A., & Patel, D. J. (2007). Ribozymes, riboswitches and beyond: Regulation of gene expression without proteins. *Nature Reviews Genetics*, 8(10), 776–790.
- Sharma, S., Ding, F., & Dokholyan, N. V. (2008). iFoldRNA: Three-dimensional RNA structure prediction and folding. *Bioinformatics*, 24(17), 1951–1952. <http://troll.med.unc.edu/ifoldrna.v2/index.php>.
- Shirvanyants, D., Ding, F., Tsao, D., Ramachandran, S., & Dokholyan, N. V. (2012). Discrete molecular dynamics: An efficient and versatile simulation method for fine protein characterization. *The Journal of Physical Chemistry B*, 116(29), 8375–8382.
- Sim, A. Y., Minary, P., & Levitt, M. (2012). Modeling nucleic acids. *Current Opinion in Structural Biology*, 22(3), 273–278.
- Sugita, Y., & Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1), 141–151.
- Tinoco, I., Jr., & Bustamante, C. (1999). How RNA folds. *Journal of Molecular Biology*, 293(2), 271–281.
- Weeks, K. M. (2010). Advances in RNA structure analysis by chemical probing. *Current Opinion in Structural Biology*, 20(3), 295–304.
- Yin, S., Biedermannova, L., Vondrasek, J., & Dokholyan, N. V. (2008). MedusaScore: An accurate force field-based scoring function for virtual drug screening. *Journal of Chemical Information and Modeling*, 48(8), 1656–1662.
- Zhou, Y., Karplus, M., Wichert, J. M., & Hall, C. K. (1997). Equilibrium thermodynamics of homopolymers and clusters: Molecular dynamics and Monte Carlo simulations of systems with square-well interactions. *The Journal of Chemical Physics*, 107(24), 10691–10708.
- Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13), 3406–3415.