HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2012 December 01.

Published in final edited form as:

Nat Methods. ; 9(6): 603–608. doi:10.1038/nmeth.1976.

Three-Dimensional RNA Structure Refinement by Hydroxyl Radical Probing

Feng Ding[1,2], **Christopher A. Lavender**[3], **Kevin M. Weeks**[3,*], and **Nikolay V. Dokholyan**[1,2,*]

[1]Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, North Carolina 27599

[2]Center for Computational and Systems Biology, University of North Carolina, Chapel Hill, North Carolina 27599

[3]Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599

Abstract

Molecular modeling guided by experimentally-derived structural information is an attractive approach for three-dimensional structure determination of complex RNAs that are not amenable to study by high-resolution methods. Hydroxyl radical probing (HRP), performed routinely in many laboratories, provides a measure of solvent accessibility at individual nucleotides. HRP measurements have, to date, only been used to evaluate RNA models qualitatively. Here, we report development of a quantitative structure refinement approach using HRP measurements to drive discrete molecular dynamics simulations for RNAs ranging in size from 80 to 230 nucleotides. HRP reactivities were first used to identify RNAs that form extensive helical packing interactions. For these RNAs, we achieved highly significant structure predictions, given inputs of RNA sequence and base pairing. This HRP-directed tertiary structure refinement approach generates robust structural hypotheses useful for guiding explorations of structure-function interrelationships in RNA.

Introduction

RNA molecules play central roles in gene expression, splicing, and translation[1]. Knowledge of the underlying three-dimensional structure is a fundamental prerequisite to a complete understanding of most RNA functions. High-resolution methods such as X-ray crystallography and NMR spectroscopy offer unparalleled atomic-level insight into RNA structure. However, many RNAs are not amenable to structural characterization by these methods because of their conformational flexibility or large size. Recent advances[2–5] in molecular modeling yield accurate structure predictions of small RNAs but, due to the vast

*To whom correspondence should be addressed. dokh@unc.edu (N.V.D.) and weeks@unc.edu (K.M.W.).

Author Contributions

F.D., K.M.W., and N.V.D conceived and designed the computational and experimental procedures. C.A.L. performed and analyzed the HRP measurements. F.D. developed the computational methodology and performed the computational analysis. F.D., C.A.L., K.M.W., and N.V.D. wrote the manuscript.

RNA conformational space and inaccuracies in available force fields describing atomic interactions, structure prediction for large RNA molecules with complex topologies is beyond the reach of current *ab initio* approaches. Incorporation of experimentally-derived structural information with computational modeling can dramatically reduce the allowed conformational space and thereby facilitate prediction of native RNA ensembles[6–11].

One can often establish the pattern of base pairing in an RNA, or secondary structure, with high accuracy by sequence covariation analysis[12,13] or by experimentally-constrained secondary structure prediction, especially with information obtained from selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) experiments[14,15]. Accurate knowledge of the RNA secondary structure greatly restrains the possible tertiary folds[16,17], but the size of the conformational space is still large[16]. Through-space distance constraints derived from biochemical experiments or bioinformatics analyses can provide information crucial for refining the fold of an RNA molecule. Critically, a small number of long-range, through-space distance constraints are often sufficient to limit the conformational space to allow accurate RNA structure prediction[10,12]. Experimental methods used to probe through-space distances, including site-directed hydroxyl radical footprinting, cross-linking, and fluorescence resonance energy transfer, can give high-quality distance information. However, these approaches often require synthesis of specialized RNA constructs, careful controls for unintended structural perturbations, and complex approaches for data interpretationr[15]. In contrast, HRP, which reports the approximate backbone solvent accessibility[18–20] (Fig. 1a), is relatively straightforward to implement. HRP measurements have been used to evaluate or filter RNA structural ensembles[9,18,21,22] but not to drive three-dimensional RNA structure determination in a quantitative and systematic way. Here we describe a framework for biasing discrete molecular dynamics (DMD)[23] simulations of RNA to generate structural ensembles consistent with experimental HRP measurements.

Results

We used a coarse-grained approach to model RNA molecules in which each RNA nucleotide is represented by three pseudo-atoms corresponding to the base, sugar, and phosphate groups. Three beads are sufficient to correctly recapitulate major features of RNA structure, including excluded volume, base pairing and stacking, and loop entropy, and sufficiently simple to allow efficient computational sampling[3]. This three-bead modeling approach has been used successfully to fold small RNAs with simple topologies from sequence alone[3] and to refine larger RNA structures using distance constraints[8,10]. This delineation into base, sugar, and phosphate groups is also compatible with HRP chemistry, where the hydroxyl radical reacts primarily at the ribose sugar[18].

We first optimized HRP-directed refinement with a training set of six structurally diverse RNAs ranging from 75 to 214 nucleotides in length (Table 1). Prediction accuracy was evaluated by comparison with available high-resolution structures. After optimization with the training set, HRP-directed refinement was applied to an independent set of four RNAs (from 152 to 412 nucleotides in length; Table 1). Structures in the test set were not used to optimize the method, and therefore the significance of the resulting models is expect to be indicative of the predictive capability of the HRP-directed structure refinement method.

We used a widely employed approach for the HRP experiment[24]. Our data are consistent with protection patterns described in previously reported HRP experiments (**Online Methods**). In order to incorporate experimentally measured HRP reactivities into DMD simulations, we needed to define a structural parameter reflective of the information obtained in an HRP measurement that could also be readily implemented as a constraint to drive the folding simulations. Hydroxyl radical reactivity is correlated with backbone solvent accessibility[19,25]; however, it is not straightforward to incorporate solvent accessibility as a constraint in a molecular dynamics simulation. We found that solvent accessibility is inversely proportional to the number of through-space neighbor atoms. In the example of the M-Box riboswitch, despite some outliers (Fig. 1a, asterisks), nucleotides with low HRP reactivities are generally buried and have many through-space contacts, whereas nucleotides with high reactivities have fewer through-space contacts and are more exposed (Fig. 1a). The number of through-space contacts can be readily incorporated as a constraint in DMD and other simulation methods, and we used it here to bias our simulations (**Online Methods**).

We defined through-space contacts based on the sugar pseudo-atoms in our three-bead model for RNA[3]. We computed the number of contacts as the number of sugar beads within a cutoff distance, d_{cutoff}, of a given nucleotide sugar bead. We excluded immediate neighbors in the linear sequence and base-pairing partners in helical elements because these neighbors reflect primary and secondary structure rather than higher-order tertiary interactions. To find the optimal d_{cutoff}, we calculated the structure-reactivity correlation, $C_{S\text{-}R}$, the Pearson correlation coefficient between the number of contacts and the corresponding HRP reactivity for each nucleotide (Fig. 1b). Data obtained using the six RNAs from the training set were used in determining the optimal d_{cutoff} value. In these calculations, $C_{S\text{-}R}$ was negative because a lower HRP reactivity corresponds to a more buried nucleotide with a larger number of through-space neighbors. The absolute magnitude of $C_{S\text{-}R}$ was largest when the cutoff range was 13–15 Å (Fig. 1b). With an intermediate cutoff value of 14 Å, the correlation coefficients for the six training RNAs ranged from −0.5 to −0.7, with the exception of RNase P, for which $C_{S\text{-}R}$ was smaller (~−0.30).

To incorporate HRP data in DMD simulations, we assigned two bias interaction potentials (**Online Methods**)). The first included pair-wise attractive potentials for all nucleotides to encourage collapse of the RNA and general nucleotide packing. The second was an over-burial repulsion potential incurred when a given nucleotide exceeded the assigned threshold number of contacts (N_{max}) derived from its experimental HRP reactivity (Fig. 2a). To assign N_{max} values, we defined high and low HRP cutoff values corresponding to the highest and lowest mean HRP reactivities, respectively. Based on analysis of single chain RNAs in the RCSB database[26] and on exploratory simulations with the six training set RNAs, the largest and smallest N_{max} values were assigned as 11 and 0.5, respectively (Fig. 2b). Nucleotides with HRP values above and below HRP threshold values were assigned N_{max} values of 11 and 0.5, respectively. For nucleotides with intermediate HRP values, N_{max} was assigned by linear interpolation (**Online Methods**).

HRP experiments are intrinsically noisy (Fig. 1c and Supplementary Fig. 1), making assignment of interaction potentials challenging, especially in regions with moderate HRP

reactivities. To reduce the effects of noise on structure prediction, we incorporated stronger biasing interactions for RNA nucleotides that could be designated as exposed or buried with high confidence. We identified RNA segments (≥3 nts) with high or low HRP reactivities and selected the central nucleotide in each segment as the representative exposed or buried nucleotide, respectively (**Online Methods** and Fig. 1c, red and blue bars). These central representative nucleotides have a high probability of being buried or exposed in the native structure because the impact of the noise associated with HRP measurements is less significant when measured over several consecutive highly buried or solvent-exposed nucleotides. A strong pair-wise attraction was included between nucleotides identified as highly buried and the rest of the RNA molecule, while a strong over-burial repulsion was assigned for the nucleotides identified as either highly buried or highly exposed (**Online Methods**).

We used DMD simulations to obtain structural ensembles consistent with experimental HRP data, in three steps (Fig. 2c, **Online Methods**). First, we performed serial DMD simulations with inputs of RNA sequence and canonical base pairing taken from high-resolution structures. Following the formation of native secondary structures, we performed replica exchange DMD simulations with HRP-derived potentials. We then selected top 100 structures based on low energy and high $C_{S\text{-}R}$ values and identified representative structures through clustering analysis. Our goal was to define the RNA conformations that best represent clusters (sub-states) of low energy conformational ensembles that have strong correlations with experimental HRP reactivities.

The training set for the initial DMD refinements were the yeast $tRNA^{Asp}$ (75 nts), the TPP riboswitch (80 nts), the RNase P specificity domain (152 nts), the P546 domain (158 nts), the M-Box riboswitch (161 nts), and the *Azoarcus* group I intron (214 nts). These six RNAs have diverse folds and exhibit different levels of higher-order packing interactions. The *Azoarcus* group I intron, M-box riboswitch, and P546 domain RNAs have folds defined by close helical packing (Fig. 3); in contrast, folds for the RNase P domain, the TPP riboswitch, and $tRNA^{Asp}$ are characterized by local interactions between coaxially stacked helices (Supplementary Fig. 2). HRP is appropriate for *de novo* RNA structure refinement for the subset of RNAs with extensive helical packing. The extent of higher-order RNA packing can be estimated *a priori* from the fraction of nucleotides, $f(r)$, with HRP reactivities smaller than a given reactivity, r (Supplementary Fig. 3 and **Online Methods**). At $r = 0.25$ the RNAs with extensive helix packing interactions – the *Azoarcus* group I intron, M-box riboswitch, and P546 domain – have significantly larger $f(r)$ than the other RNAs (Table 1 and Supplementary Fig. 3).

We characterized the predicted structural ensembles for each RNA in terms of the number and population of clusters in the 100 final structures. For each cluster, we also computed the mean RMSD relative to the accepted structure and the prediction significance or *P*-value[16] (Table 1). The RMSD value corresponding to a significant prediction varies with RNA size, and it is not appropriate to apply a single cutoff for all RNAs[16]. For example, an RMSD of 7–10 Å is not significant for a small RNA but is highly significant for a 100-nt RNA[16]. The *P*-value quantifies the statistical significance of the RNA fold prediction as the probability of observing a given conformation in an unbiased simulation with a pre-constrained base

pairing arrangement. *P*-values less than 0.01 correspond to predictions with high statistical significance[16].

We obtained highly significant predictions for each of the three largest RNAs in the training set. For the *Azoarcus* group I intron and the M-Box riboswitch, all predicted structures fell into a single cluster with a low average RMSD and low *P*-value and were thus native-like (Fig. 3a,b). For the P546 domain, refined structures formed two highly populated clusters; both had low *P*-values and differed primarily in the location of a single helix (Fig. 3c). Simulations of the TPP riboswitch yielded three clusters of structures. *P*-values for two of these clusters were poor ($P > 0.01$), although the third cluster had a significant *P*-value (0.003) and correctly recapitulated the TPP ligand-binding pocket (Fig. 3d). For $tRNA^{Asp}$ and RNase P, HRP-directed structure refinement did not generate native-like structures ($P > 0.01$; Table 1 and Supplementary Fig. 2).

For the six training RNAs, we observed a strong correlation between the fraction of nucleotides protected from HRP cleavage, $f_{0.25}$, and the prediction significance (Table 1). $tRNA^{Asp}$, RNase P, and the TPP riboswitch had $f_{0.25}$ values less than 0.25 and yielded inaccurate predictions; whereas, we obtained statistically significant fold predictions for the three RNAs with higher $f_{0.25}$ values (Fig. 3). The $f_{0.25}$ values are calculated based on the HRP data alone, without reference to the accepted structure. Thus, we conclude that the HRP-directed structure refinement is appropriate for RNAs with extensive close packing of helices, corresponding to $f_{0.25} > 0.25$.

We next applied HRP-directed structure refinement to the test set of four additional RNA molecules: the glmS ribozyme (152 nts), the lysine riboswitch (174 nts), the catalytic domain of RNase P (231 nts), and a group II intron (412 nts). Based on $f_{0.25}$ values (Table 1), three of these RNAs – the glmS ribozyme, the lysine riboswitch, and the catalytic domain of RNase P – were appropriate candidates for structure refinement. In contrast, with an $f_{0.25}$ value of 0.21, the group II intron was not a suitable candidate for refinement using HRP-derived constraints. The three RNAs with appropriate $f_{0.25}$ values all refined to native-like folds with significant *P*-values (Table 1 and Supplementary Fig. 4). The major structural variations between different clusters for a given RNA corresponded to regions without well-defined HRP data (for example, the 3′ end of RNase P catalytic domain) (Supplementary Fig. 4 and Supplementary Dataset 1).

HRP-directed structure predictions often resulted in multiple clusters with distinct structures (Fig. 3 and Supplementary Fig. 4), suggesting that all experimental constraints cannot be satisfied simultaneously. Predictions that yielded multiple clusters reflect either the intrinsic structural heterogeneity of an RNA molecule or non-ideal experimental data. To explore the relationship between prediction accuracy and experimental HRP data quality, we generated idealized datasets by assuming perfect structure-reactivity correlations ($C_{S\text{-}R} = 1$) for the M-Box, P546 domain, and TPP riboswitch RNAs (**Online Methods** and Supplementary Table 1). Additional simulated datasets with decreasing $C_{S\text{-}R}$ values were generated by introducing random noise into the idealized data. Larger $C_{S\text{-}R}$ values generally yielded significant increases in the RNA prediction significance (Supplementary Table 1).

Discussion

HRP-directed structure refinement is unique among RNA structure refinement methods as prediction quality is highest for larger and more complex RNA folds with extensive helical packing and a significant fraction of nucleotides occluded from solvent. (Table 1, Fig. 3 and Supplementary Fig. 4). The HRP-directed fold prediction is also highly tolerant of the noise intrinsic to the RNA HRP experiment. Although the overall correlations between structure and HRP reactivity, as illustrated by $C_{S\text{-}R}$, were modest (Fig. 1b), highly significant refinements were obtained because our algorithm reduces the impact of noise by identifying subsets of nucleotides with high probability of being buried or exposed (Fig. 1c) and imposes strong energy terms on these nucleotides to drive RNA folding.

Previous RNA tertiary structure prediction studies have shown that a relatively small number of long-range constraints are often sufficient to reduce allowable conformational space and to make possible prediction of diverse native-like structures[8,10]. In three of the RNAs studied here, the *Azoarcus* group I intron, the lysine riboswitch, and the glmS ribozyme, long-range pseudoknot base-pairing constraints were included in structure prediction. Even for these partially pre-constrained RNAs, the HRP-directed structural refinement improved prediction beyond what is possible by including the pseudoknot base-pairing constraints alone (see **Methods**). One can thus use HRP-directed structural refinement in conjunction with other classes of information. Moreover, the correlation between the structure prediction accuracy and the quality of the input HRP data (Supplementary Table 1) suggests that if it becomes possible to improve the HRP approach or if experiments that better measure the solvent accessibility of RNA molecules are developed, it will be possible to refine RNA folds with an even higher level of accuracy.

The goal of the method is to reconstruct structural models for challenging RNA molecules not amendable to high-resolution methods. These structural models are especially useful for developing experimentally-testable hypotheses and for guiding the exploration of structure-function relationships for RNA.

Methods and any associated references are available in the online version. All software packages developed in this work for analyzing hydroxyl radical data and for predicting RNA structural models are available at http://troll.med.unc.edu/ifoldrna/HRP-1.0-openmpi.tgz.

Online Methods

Hydroxyl radical probing (HRP) measurements

RNA Preparation—RNAs were synthesized by T7 RNA polymerase-mediated *in vitro* transcription[35] using double-stranded PCR-generated templates. Sequences were transcribed in the context of 5′ and 3′ structure cassette sequences to facilitate analysis by primer extension[36]. Transcribed RNAs were purified on 10% denaturing polyacrylamide gels (7 M urea, 1× TBE). Bands containing full-length product were excised; RNA was recovered by passive elution in 1× TE (10 mM Tris, pH 8.0; 1 mM EDTA) and precipitation with ethanol. Samples were resuspended in 1× TE and quantified by absorbance measurements at 260 nm.

Hydroxyl Radical Cleavage—HRP datasets for the *Azoarcus* group I intron and the RNase P specificity domain were taken from a previous study, which used essentially the same approach as outlined below[37]. Hydroxyl radical cleavage experiments for the other RNAs were performed as described[24]. RNAs were first refolded by heat denaturation, snap-cooling on ice, and incubation at 37 °C. The HRP data reported here are consistent with previously reported experiments.[24,29,38,39]

For the ligand-binding RNAs, 1 μL of a 5 μM RNA solution of RNA was combined with 2μL sterile water and 3 μL folding mix (333 mM HEPES, pH 8.0, 333 mM NaCl, 33 mM $MgCl_2$ for the TPP riboswitch; 333 mM HEPES, pH 8.0, 333 mM KCl, 33 mM $MgCl_2$ for the lysine riboswitch; 167 mM HEPES, pH 7.5, 6.7 mM $MgCl_2$ for the glmS ribozyme). RNAs were heated at 95 °C for 2 min, cooled on ice, and then incubated at 37 °C for 10 min. 1 μL of ligand solution (10 μM thiamine pyrophosphate, 50 μM lysine, or 1 mM glucoasamine-6-phosphate for the TPP riboswitch, the lysine riboswitch, and the glmS ribozyme, respectively) was added, and the RNA was incubated in the presence of ligand at 37 °C for 20 min.

To fold the other RNAs, 1 μL of a 5 μM RNA solution was combined with 3 μL sterile water and 3 μL folding mix (46.6 mM HEPES, pH 7.5, 23.3 mM $MgCl_2$ for $tRNA^{Asp}$; 333 mM HEPES, pH 7.5, 333 mM NaCl, 33 mM $MgCl_2$ for the P546 domain; 46.6 mM HEPES, pH 7.5, 23.3 mM $MgCl_2$ for the M-Box riboswitch; 33 mM HEPES, pH 7.5, 333 mM NaCl, 33 mM $MgCl_2$ for the RNase P catalytic domain; 333 mM HEPES, pH 8.0, 333 mM KCl, 416 mM $MgCl_2$ for the group I intron; 300 mM HEPES, pH 8.0, 300 mM KCl, 375 mM $MgCl_2$ for the group II intron). These RNAs were then heated at 95 °C for 2 min, cooled on ice, and then incubated at 37 °C for 20 min.

The glmS ribozyme construct contained a point mutation (G40A) to prevent autolytic RNA cleavage during the HRP experiment; this mutant induces minimal structural disruption to the RNA[40].

Hydroxyl radical cleavage was initiated by addition of Fe(II)-EDTA, sodium ascorbate, and hydrogen peroxide to the folded RNA. Fresh Fe(II)-EDTA [10 mM ammonium iron (II) sulfate and 20 mM EDTA, pH 8.0] and 50 mM sodium ascorbate solutions were made prior to each experiment. The Fe(II)-EDTA and ascorbate solutions were combined in a 1:1 ratio, and 2 μL of this 1:1 solution and 1 μL of 0.03% hydrogen peroxide were spotted in separate areas of the reaction lid. Hydroxyl radical cleavage was initiated by brief centrifugation. After incubation at 37 °C for 2 min, reactions were quenched by addition of a solution containing 169 μL water, 20 μL 3 M sodium acetate (pH 5.5), and 1 μL 20 μg/μL glycogen, followed by addition of 500 μL ethanol. Modified RNA was recovered by precipitation with ethanol and washed with 70% ethanol. For each reaction, a no-reaction control without Fe(II)-EDTA and ascorbate was performed in parallel.

Primer Extension—Sites of hydroxyl radical-mediated cleavages were analyzed by primer extension using fluorescently labeled primers[37,41], labeled with fluorophores from the Applied Biosystems G5 dye set: (+) reaction, FAM; (−) reaction, VIC; sequencing ladder, NED. For each primer extension reaction, 3 μL 0.3 μM fluorescently-labeled DNA

primer was added to 1 pmol RNA in 10 μL 0.5× TE. This solution was incubated at 65 °C for 5 min, then cooled on ice for 1 min. To this solution, 6 μL Superscript III enzyme mix (250 mM KCl, 167 mM Tris, pH 8.3, 1.67 mM each dNTP, 17 mM DTT, 10 mM $MgCl_2$) and 1 μL Superscript III (Invitrogen) were added. For sequencing reactions, 1.67 mM of a ddNTP was included in the Superscript III enzyme mix. The solution was incubated at 45 °C for 1 min, 52 °C for 25 min, and 65 °C for five min. The (+) and (−) reagent and sequencing reactions were then combined in 1 mL ethanol to quench extension and to precipitate the cDNA. Recovered cDNAs were washed with 70% ethanol and resuspended in 10 μL dry formamide (Applied Biosystems).

cDNAs were resolved on Applied Biosystems 3130 or 3500 Genetic Analyzer capillary electrophoresis instruments. Signal processing, sequencing alignment, and peak integration of raw traces were performed using ShapeFinder[42] and custom signal processing software. A representative processed electropherogram is provided in Supplementary Fig. 5. Net reactivity at each nucleotide was defined as the area of the (+) reaction peak after subtracting the area of the corresponding (−) reaction peak. Nucleotides with high (−) signal were excluded from further analysis as high-background regions; the number of these high-background regions was small in the analyzed RNAs. Net reactivities were normalized by dividing reactivities by the average reactivity of the top 10% of nucleotides, excluding the top 2%. HRP reactivities for each of the ten RNAs are provided in Supplementary Dataset 1.

Computational modeling using HRP reactivities

Overview of the DMD Refinement Approach—We used a coarse-grained RNA model for DMD simulations[3] in which each nucleotide is represented by three pseudo-atoms, representing the phosphate, sugar, and base groups. Bonded terms, including covalent bond lengths, angles, and dihedrals, were used to model local RNA geometry. Non-bonded interactions included base pairing, base stacking, phosphate-phosphate repulsion, and hydrophobic interactions. We explicitly modeled the entropy loss for loop formation. To bias the DMD simulation toward the structural ensemble consistent with experimental measurements, we added additional potential terms based on the experimental hydroxyl radical probing data.

DMD simulations and analysis were performed in three steps. First, serial DMD simulations were performed with inputs of RNA sequence and canonical base pairing, including pseudoknotted pairs, as obtained from high-resolution structures. Although the base-pairing arrangements were taken from X-ray crystallographic analyses, this information can be obtained with high accuracy from sequence covariation analysis[12,13] or SHAPE-directed secondary structure prediction[14,15]. The result of these simulations was the formation of native secondary structures. Second, replica exchange DMD simulations with the HRP-derived potentials were applied to enrich for conformations consistent with the experimental HRP data. Third, top 100 structures were selected with the lowest energies and highest correlations between HRP reactivities and numbers of contacts ($C_{S\text{-}R}$). Clustering analysis based on pairwise root-mean-square deviation (RMSD) was then performed to identify representative structures consistent with the predicted structural ensemble.

HRP Bias Potential—For each nucleotide, we assigned a favorable energy increment, $E_{attr}(i)$ for forming a contact; a threshold number of contacts, $N_{max}(i)$; and a repulsive energy, $dE_{rep}(i)$, for exceeding the threshold. The HRP E_{bias} potential equals:

$$E_{bias}=\sum_{i<j}E_{ij}^{attr}+\sum_{i}E_{i}^{over-bury}. \quad (1)$$

The first term is the pairwise attraction, $E_{ij}^{attr}(r_{ij})=\min\{E_{attr}(i),E_{attr}(j)\}F(r_{ij})$, where $F(x)$ is a step function,

$$F(x)=\begin{cases}0 & IR<r\\ 1 & R_{hc}<r\leq IR\\ \infty & r\leq R_{hc}\end{cases}. \quad (2)$$

IR is the interaction range of 14 Å, and R_{hc} is the hard core diameter, 3.0 Å. The second term prevents over-burying by exceeding the threshold number of contacts:

$$E^{over-bury}(i)=dE_{rep}(i)(n_c(i)-N_{\max}(i))\Theta(n_c(i)-N_{\max}(i)), \quad (3)$$

where $n_c(i)$ is the number of contacts for each nucleotide *i*, $dE_{rep}(i)$ is the penalty energy for over-burying, and $\Theta(x)$ is the unit step function, which equals 1 if *x* is positive and zero otherwise. The number of contacts for each nucleotide was computed as the number of non-local sugar beads within the 14 Å cutoff. For each nucleotide *i*, we excluded contacts with nucleotides that were adjacent (within 4 nucleotides) to *i* or were adjacent to a nucleotide to which *i* base pairs (for *i* pairing with *I*, these nucleotides are $|i-j|>4$, or $|I-j|>4$).

Assignment of Interaction Parameters—The energy parameters, $N_{max}(i)$, $E_{attr}(i)$, and $dE_{rep}(i)$ were assigned using HRP reactivities for each nucleotide in the following three steps:

1. Assignment of the threshold number of contacts: Threshold number of contacts, N_{max}, were assigned according to reactivities, *R*, smoothed over a sliding window of three nucleotides. Smoothing reduced the influence of the noise intrinsic to HRP experiments performed with RNA and increased the correlations to the accepted structure, $C_{S\text{-}R}$. We defined two threshold values, R_{min} and R_{max}, corresponding to the maximally buried and exposed nucleotides. R_{min} and R_{max} were the average of the subsets from 2% to 20% and from 80% to 98% of the rank-ordered *R*. The top and bottom 2% *R* values were discarded to reduce the influence of extreme *R* values observed in typical HRP experiments. For nucleotides with *R* smaller than R_{min} or higher than R_{max}, the threshold number of contacts was defined as NC_{max} = 11 and NC_{min} = 0.5, respectively (Fig. 2b). For a nucleotide *i* with intermediate reactivity, the threshold number of contacts was assigned by linear interpolation,

$$N_{max}(i)=NC_{max}+(NC_{min}-NC_{max})(R(i)-\langle R_{min}\rangle)/(\langle R_{max}\rangle-\langle R_{min}\rangle). \quad (4)$$

2. Assignment of representative buried and exposed nucleotides: We first identified segments of strongly buried and exposed nucleotides. We defined three values, R^{EXP}, R^{INT}, and R^{BUR}, corresponding to the threshold values of exposed, intermediate, and buried residues (Fig. 1c). The buried threshold R^{BUR} and exposed threshold R^{EXP} correspond to lowest 20% and highest 80% of the rank-ordered reactivities R, respectively. There are two types of intermediate R values, the average value of all the reactivities $\langle R\rangle$ and the median value of the rank-ordered reactivities R_{50}. For simplicity, we chose the mean of these two values as R^{INT}.

We defined buried segments as those with more than three consecutive nucleotides having R smaller than R^{INT} and at least one with R smaller than R^{BUR}. For each buried segment, we selected the one with the lowest reactivity as the buried representative, excluding the first and last residues in the segment. Similarly, we defined exposed segments as those with more than three consecutive nucleotides having R larger than R^{INT} and at least one nucleotide having R larger than R^{EXP} and, for two-nucleotide segments, with both nucleotides having R values larger than R^{EXP}. For each exposed segment, we defined the nucleotide with largest R value as the exposed representative.

3. Assignment of attractions and repulsions: Two attractive energy scales were used, E_{low} = −0.10 kcal/mol and E_{high} = −0.05 kcal/ml, based on the simulation temperature (see below). We assigned a strong attractive energy, E_{low}, to the buried representative nucleotides identified in Step 2 and the median value of $(E_{high}+E_{low})/2$ to their nearest neighbors. For all remaining nucleotides, we assigned the weak attractive energy of E_{high}. We defined a strong repulsive over-burial energy, $dE_{rep}(i)$ = 0.3 kcal/mol, for both the buried and exposed representative positions. We assigned the repulsive energy $dE_{rep}(i)$ = $-E_{attr}(i)$ to all other nucleotides, where $E_{attr}(i)$ equals E_{low} or E_{high}. By making over-burial repulsion potentials equal to those for attractions, these nucleotides were allowed to make additional contacts ($>N_{max}$) without a net energy penalty. This approach reduced the impact of noise in HRP experiments on RNA structure refinement by promoting compaction while imposing strong energy terms correlated with solvent accessibility for the subset of nucleotides identified as having a high probability of being buried or exposed. The HRP-derived values – threshold number of contacts (N_{max}), attractive (E_{attr}), and repulsive ($dE_{rep}(i)$) energies – are listed for all tested RNAs in Supplementary Dataset 1.

Replica Exchange DMD Simulations—Because the HRP-directed potential is non-specific with respect to any two nucleotides (in contrast to the distance and bonded constraints between specific nucleotides[8,10]) we performed replica exchange DMD simulations to obtain sufficient sampling of conformational space. We used eight replicas with temperatures of 0.200, 0.225, 0.250, 0.270, 0.300, 0.333, 0.367, and 0.400 kcal/(mol·k_B). Every 1000 DMD time units, we exchanged replicas with neighboring temperatures according to a Metropolis-based Monte Carlo algorithm using instantaneous potential energies[3]. For each replica, we performed simulations over 5×10^5 DMD time

units. Replica exchange DMD simulations were performed in parallel on 2.27 GHz Intel Xeon computing nodes. Representative running times for the TPP riboswitch (80 nts), M-Box (160 nts), and *Azoarcus* RNAs (214 nts) were 60, 170 and 264 CPU hours, respectively. The wall-clock time is one-eighth of the total CPU time.

Identifying Structure Ensembles Consistent with Experiments—To identify structural ensembles consistent with the experimentally measured HRP data, we computed structure-reactivity correlation coefficients, $C_{S\text{-}R}$, for snapshot structures, computed every 100 time units, yielding 4×10^4 snapshots for each refinement. We rank-ordered these snapshots by $C_{S\text{-}R}$ and selected the 2000 structures with the lowest (negative) correlation coefficients. From these, we selected 100 structures with the lowest energies. We also selected structures applying these rules in the reverse order: from the 4×10^4 structures, we selected 2000 structures by energy from which we then selected the 100 structures with the lowest $C_{S\text{-}R}$.

For the combined 200 structures, we removed duplicates and selected top 100 structures to represent the predicted structural ensemble. The structures were ranked according to the combined rank-order using both energy and $C_{S\text{-}R}$. We clustered these 100 structures according to pairwise RMSD using a hierarchical clustering algorithm and grouped similar structures into clusters using a cutoff RMSD. For simplicity, we used a cutoff value of 4 Å (roughly two standard deviations) below the average RMSD for a given RNA length[16] (see below), or three quarters of the average RMSD, whichever is smaller:

$$\begin{cases} R(n)-4, & R(n)-4<0.75R(n) \\ 0.75R(n), & 0.75R(n) \le R(n)-4 \end{cases} . \quad (5)$$

Here, n is the RNA length, and $R(n)$ is the average RMSD as the function of RNA length.

P-Value Calculation

A recent study of a large set of RNA decoy structures derived from both simulations and from threading suggests that the RMSD between two random RNA structures of the same length follows a Gaussian distribution with a length-dependent average RMSD and a length-independent standard deviation (~1.8 Å)[16]. For an RNA with known secondary structure, the average RMSD between two random decoy structures is smaller relative to a decoy set generated without knowledge of the secondary structure. We computed the statistical significance, or *P*-value, corresponding to the probability that an HRP-constrained structure prediction, evaluated by its RMSD from the accepted structure, is significantly better than that expected by chance. The *P*-value calculation is available online at http://ifoldrna.dokhlab.org[16].

The question of how to interpret the significance of a structure model with a given RMSD value has been a major challenge in the RNA folding field. Some groups have suggested that RMSDs should correspond in some qualitative way with the physical dimensions of RNA. For example, the RMSDs should be less than 7 Å (the average distance between two nucleotides) or within with width of an RNA helix (~20 Å). In fact, the average RMSD

between any two structural models is strongly dependent on RNA length and on whether the secondary structure is used as a constraint[16]. Thus, we argue that an appropriate way to understand the significance of a structure prediction is in terms of a *P*-value. Prior work using the 7 or 20 Å heuristic rules tended to overestimate the quality of predictions for short RNAs and to underestimate the significance of predictions for large RNAs. For large RNAs, seemingly large RMSD values with low P-values correspond to native-like folds with high significance (see Fig. 3).

Generation of Ideal and Randomized Reactivity Profiles—We generated idealized HRP reactivities based on the number of contacts in the native structure, $R^{ideal}(i) = 1 - N_c(i)/N_{max}$. We added noise to these idealized reactivities to generate randomized reactivities, $R^{rand}(i) = R^{ideal}(i)(1+ \sigma x)$, where x is a random number from −1 to 1, and σ is the amplitude of the noise, determined by the relative error:

$$1/N\sum_{i=1}^{N}|(R^{Rand}(i)-R^{ideal}(i))/R^{ideal}(i)|=1/N\sum_{i=1}^{N}\sigma|x|\sim\sigma/2 \quad (6)$$

where the sum is over all nucleotides in a RNA. By varying σ, we generated randomized reactivity profiles with different levels of noise and, thus, different structure-reactivity correlations (Supplementary Fig. 6a). Notably, the M-Box riboswitch had the least noise-induced decrease in the structure-reactivity correlation $C_{S\text{-}R}$, while $tRNA^{Asp}$ had the greatest decrease in $C_{S\text{-}R}$, which correlates with their respective prediction significances (Supplemental Fig. 6b).

For the M-box, the P546 domain, the TPP riboswitch, and $tRNA^{Asp}$ RNAs, we selected seven sets of computationally generated HRP data with $C_{S\text{-}R}$ values ranging from −0.4 to −1.0 (Supplementary Table 1). Using the generated HRP reactivities as the input, we applied our structure refinement protocol to generate structural ensembles (Supplementary Table 1). For all tested RNAs, except $tRNA^{Asp}$, we found that HRP reactivities with high $C_{S\text{-}R}$ resulted in low RMSDs and highly significant predictions. As the $C_{S\text{-}R}$ of input HRP reactivities decreased, the RMSDs of the predicted structures and the corresponding *P*-values increased, indicating less accurate predictions.

There are two critical implications of this analysis. First, the high *P*-value predictions for $tRNA^{Asp}$ using both experimental and computationally generated HRP reactivities suggest that RNAs like tRNA, with few buried nucleotides, are not good candidates for HRP-directedrefinement. Importantly, these RNAs can be identified (and excluded) in advance using the $f_{0.25}$ metric (Supplementary Fig. 3). Second, our simulations indicate that the level of noise and resulting structure-reactivity correlation for the input HRP data play a determining role in the accuracy of HRP-directed structure prediction. If a better experimental method with reduced noise in HRP (or solvent accessibility) reactivities were developed, our approach would immediately lead to significantly more accurate RNA structure refinements.

Structural refinement for RNAs with pseudoknot base pairs—In our study, we assumed that all base pairs, including pseudoknots, were known. A relatively small number

of constraints based on long-range contacts, such as pseudoknots, are sufficient to direct the prediction of highly significant RNA structures[10]. The *Azoarcus* group I intron, the lysine riboswitch, and the glmS ribozyme RNA contain long-range pseudoknots that likely reduce available conformational space and may themselves lead to significant structure predictions. To examine the effects of incorporating HRP data for RNA refinements in which long-range pseudoknot constraints were included, we compared the results of RNA structure prediction with and without HRP data.

First, we evaluated whether incorporation of HRP data as constraints drives the conformational sampling toward native states during the course of simulations for the pseudoknot-containing RNAs. We calculated the RMSD for all RNA conformations sampled during DMD simulations both with and without HRP data as constraints. For both the lysine riboswitch and glmS ribozyme, incorporation of HRP data in the DMD simulations significantly enhanced sampling of native-like conformations (Supplementary Fig. 7). Second, we applied the structure selection approach to reconstruct conformational ensembles for simulations that did not incorporate HRP data. Critically, for these large RNAs, if the HRP data were not used to drive refinement, the resulting structural ensembles fell into multiple small clusters with a wide range of RMSD values (Supplementary Table 2); in contrast, using the HRP data to drive refinement yielded only a few clusters, each with well-defined structures and highly significant RMSD values (Table 1). Therefore, although the pseudoknotted base pairs reduced the available conformational space, the HRP-directed structural refinement drove RNA folding to native-like states.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank E.A. Proctor, R. Redler, and S. Ramachandran for critical readings of the manuscript. This work was supported by grants from the US National Institutes of Health to K.M.W. (GM064803) and N.V.D. (GM080742 and CA084480), by an NIH ARRA supplement (to K.M.W.), and by the UNC Research Council (to F.D.).

References

1. Gesteland, RF.; Cech, TR.; Atkins, JF., editors. The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World. Cold Spring Harbor Lab Press; Plainview, NY: 2006.
2. Das R, Baker D. Automated *de novo* prediction of native-like RNA tertiary structures. Proc Natl Acad Sci USA. 2007; 104 (37):14664–14669. [PubMed: 17726102]
3. Ding F, Sharma S, Chalasani V, et al. Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. RNA. 2008; 14:1164–1173. [PubMed: 18456842]
4. Parisien M, Major F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature. 2008; 452:51–55. [PubMed: 18322526]
5. Cao S, Chen SJ. Physics-based de novo prediction of RNA 3D structures. J Phys Chem B. 2011; 115 (14):4216–4226. [PubMed: 21413701]
6. Das R, Kudaravalli M, Jonikas M, et al. Structural inference of native and partially folded RNA by high-throughput contact mapping. Proc Natl Acad Sci USA. 2008; 105 (11):4144–4149. [PubMed: 18322008]
7. Yu ET, Hawkins A, Eaton J, Fabris D. MS3D structural elucidation of the HIV-1 packaging signal. Proc Natl Acad Sci USA. 2008; 105 (34):12248–12253. [PubMed: 18713870]

8. Gherghe CM, Leonard CW, Ding F, et al. Native-like RNA tertiary structures using a sequence-encoded cleavage agent and refinement by discrete molecular dynamics. J Am Chem Soc. 2009; 131 (7):2541–2546. [PubMed: 19193004]
9. Jonikas MA, Radmer RJ, Laederach A, et al. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA. 2009; 15 (2):189–199. [PubMed: 19144906]
10. Lavender CA, Ding F, Dokholyan NV, Weeks KM. Robust and generic RNA modeling using inferred constraints: a structure for the hepatitis C virus IRES pseudoknot domain. Biochemistry. 2010; 49 (24):4931–4933. [PubMed: 20545364]
11. Yang S, Parisien M, Major F, Roux B. RNA structure determination using SAXS data. J Phys Chem B. 2010; 114 (31):10039–10048. [PubMed: 20684627]
12. Michel F, Westhof E. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J Mol Biol. 1990; 216 (3):585–610. [PubMed: 2258934]
13. Gutell RR, Power A, Hertz GZ, et al. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. Nucleic Acids Res. 1992; 20 (21):5785–5795. [PubMed: 1454539]
14. Deigan KE, Li TW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci USA. 2009; 106:97–102. [PubMed: 19109441]
15. Weeks KM. Advances in RNA structure analysis by chemical probing. Curr Opin Struct Biol. 2010; 20 (3):295–304. [PubMed: 20447823]
16. Hajdin CE, Ding F, Dokholyan NV, Weeks KM. On the significance of an RNA tertiary structure prediction. RNA. 2010; 16 (7):1340–1349. [PubMed: 20498460]
17. Bailor MH, Mustoe AM, Brooks CL 3rd, Al-Hashimi HM. Topological constraints: using RNA secondary structure to model 3D conformation, folding pathways, and dynamic adaptation. Curr Opin Struct Biol. 2011; 21 (3):296–305. [PubMed: 21497083]
18. Tullius TD, Greenbaum JA. Mapping nucleic acid structure by hydroxyl radical cleavage. Curr Opin Chem Biol. 2005; 9 (2):127–134. [PubMed: 15811796]
19. Cate JH, Gooding AR, Podell E, et al. Crystal structure of a group I ribozyme domain: principles of RNA packing. Science. 1996; 273 (5282):1678–1685. [PubMed: 8781224]
20. Pastor N, Weinstein H, Jamison E, Brenowitz M. 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. 2000; 304 (1):55–68. [PubMed: 11071810]
21. Bergman NH, Lau NC, Lehnert V, et al. The three-dimensional architecture of the class I ligase ribozyme. RNA. 2004; 10 (2):176–184. [PubMed: 14730016]
22. Rangan P, Masquida B, Westhof E, Woodson SA. Assembly of core helices and rapid tertiary folding of a small bacterial group I ribozyme. Proc Natl Acad Sci U S A. 2003; 100 (4):1574–1579. [PubMed: 12574513]
23. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Discrete molecular dynamics studies of the folding of a protein-like model. Fold Des. 1998; 3 (6):577–587. [PubMed: 9889167]
24. Dann CE 3rd, Wakeman CA, Sieling CL, et al. Structure and mechanism of a metal-sensing regulatory RNA. Cell. 2007; 130(5):878–892. [PubMed: 17803910]
25. Balasubramanian B, Pogozelski WK, Tullius TD. 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 USA. 1998; 95 (17):9738–9743. [PubMed: 9707545]
26. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28 (1): 235–242. [PubMed: 10592235]
27. Westhof E, Dumas P, Moras D. Restrained Refinement of 2 Crystalline Forms of Yeast Aspartic-Acid and Phenylalanine Transfer-Rna Crystals. Acta Cryst A. 1988; 44:112–123. [PubMed: 3272146]
28. Serganov A, Polonskaia A, Phan AT, et al. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. Nature. 2006; 441 (7097):1167–1171. [PubMed: 16728979]
29. Krasilnikov AS, Yang X, Pan T, Mondragon A. Crystal structure of the specificity domain of ribonuclease P. Nature. 2003; 421 (6924):760–764. [PubMed: 12610630]

30. Adams PL, Stahley MR, Kosek AB, et al. Crystal structure of a self-splicing group I intron with both exons. Nature. 2004; 430 (6995):45–50. [PubMed: 15175762]

31. Cochrane JC, Lipchock SV, Smith KD, Strobel SA. Structural and chemical basis for glucosamine 6-phosphate binding and activation of the glmS ribozyme. Biochemistry. 2009; 48 (15):3239–3246. [PubMed: 19228039]

32. Serganov A, Huang L, Patel DJ. Structural insights into amino acid binding and gene control by a lysine riboswitch. Nature. 2008; 455 (7217):1263–1267. [PubMed: 18784651]

33. Kazantsev AV, Krivenko AA, Pace NR. Mapping metal-binding sites in the catalytic domain of bacterial RNase P RNA. RNA. 2009; 15 (2):266–276. [PubMed: 19095619]

34. Toor N, Keating KS, Fedorova O, et al. Tertiary architecture of the Oceanobacillus iheyensis group II intron. RNA. 2009; 16 (1):57–69. [PubMed: 19952115]

35. Milligan JF, Groebe DR, Witherell GW, Uhlenbeck OC. Oligoribonucleotide synthesis using T7 RNA polymerase synthetic DNA templates. Nucleic Acids Res. 1987; 15 (21):8783–8798. [PubMed: 3684574]

36. Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM. RNA structure analysis at single nucleotide resolution by Selective 2′-Hydroxyl Acylation Primer Extension (SHAPE). J Am Chem Soc. 2005; 127 (12):4223–4231. [PubMed: 15783204]

37. Duncan CDS, Weeks KM. The Mrs1 Splicing Factor Binds the bI3 Group I Intron at Each of Two Tetraloop-Receptor Motifs. Plos One. 2010; 5(2)

38. Murphy FL, Cech TR. An independently folding domain of RNA tertiary structure within the Tetrahymena ribozyme. Biochemistry. 1993; 32 (20):5291–5300. [PubMed: 7684607]

39. Latham JA, Cech TR. Defining the inside and outside of a catalytic RNA molecule. Science. 1989; 245 (4915):276–282. [PubMed: 2501870]

40. Klein DJ, Been MD, Ferre-D'Amare AR. Essential role of an active-site guanine in glmS ribozyme catalysis. J Am Chem Soc. 2007; 129 (48):14858–14859. [PubMed: 17990888]

41. McGinnis JL, Duncan CD, Weeks KM. High-throughput SHAPE and hydroxyl radical analysis of RNA structure and ribonucleoprotein assembly. Methods Enzymol. 2009; 468:67–89. [PubMed: 20946765]

42. Vasa SM, Guex N, Wilkinson KA, et al. ShapeFinder: A software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. RNA. 2008; 14 (10):1979–1990. [PubMed: 18772246]

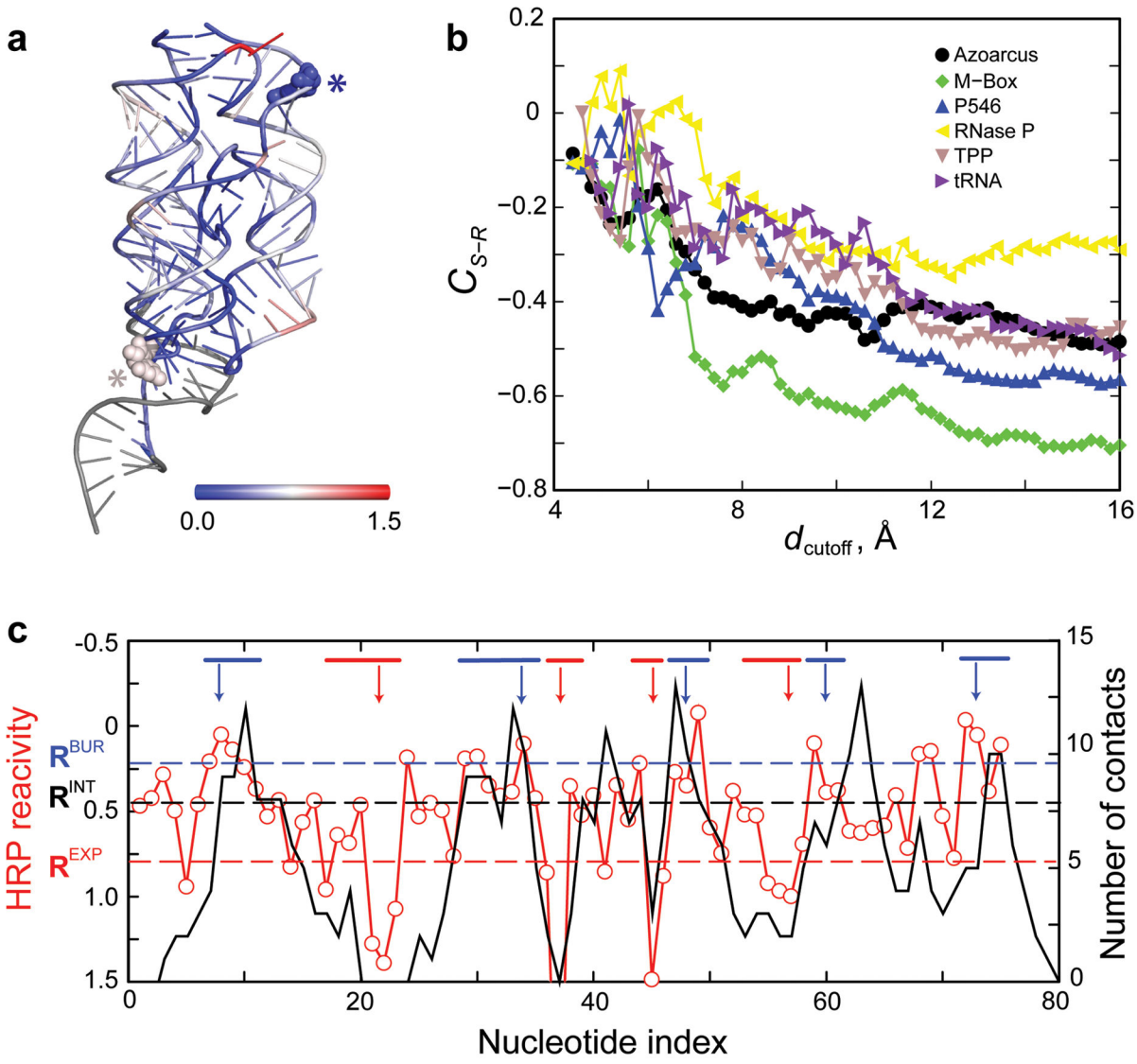

Figure 1. Relationship between RNA structure and HRP reactivity

(a) Structure of the M-Box riboswitch shown in cartoon representation. Nucleotides are colored according to HRP reactivity (blue to red); nucleotides without HRP data are 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) Structure-reactivity correlation coefficient, $C_{S\text{-}R}$, as a function of d_{cutoff} for the six training RNAs using HRP data smoothed over a three-nucleotide window (**Online Methods**). (c) Comparison of experimentally measured HRP reactivities (red) with the number of through-space contacts (black) for the TPP riboswitch RNA using a d_{cutoff} of 14.0 Å. Buried and exposed nucleotide segments are denoted with blue and red lines, respectively (top); arrows indicate the representative nucleotides characteristic of each nucleotide segment. Dashed horizontal lines represent the exposed (R^{EXP}), buried (R^{BUR}), and intermediate (R^{INT}) threshold values.

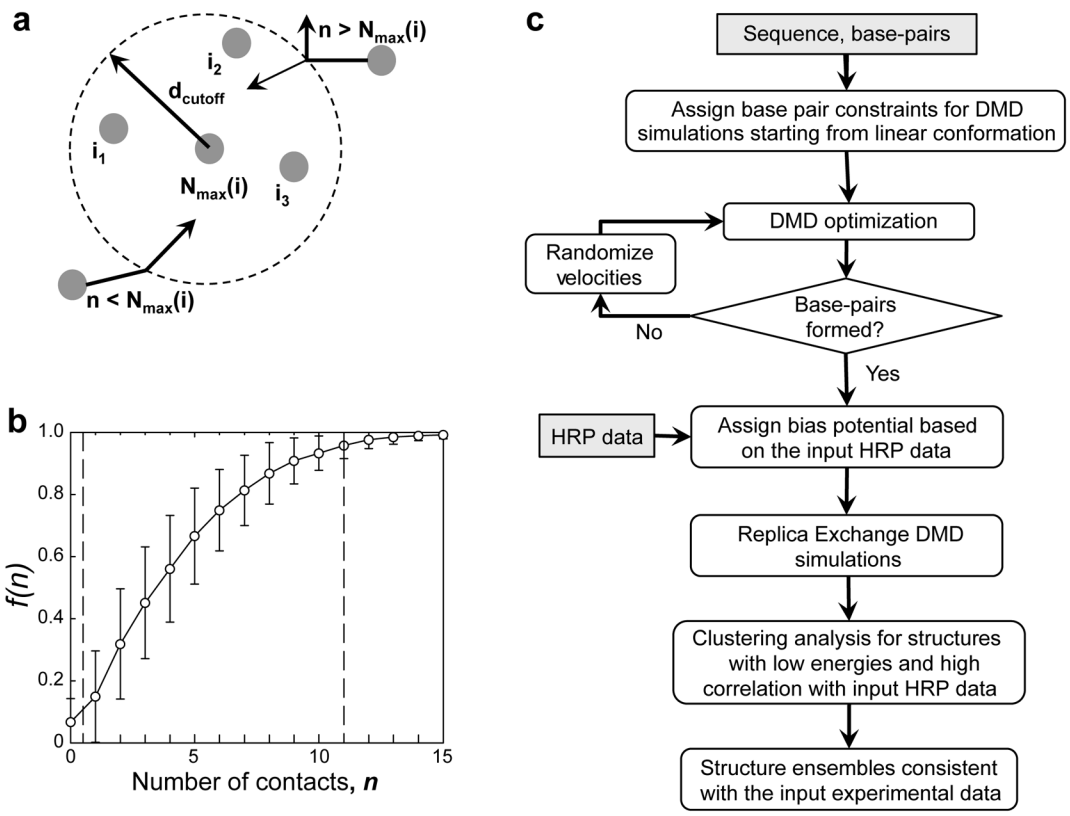

Figure 2. Assignment of potentials for incorporating HRP reactivities into DMD simulations

(a) Scheme for modeling the number of allowed contacts. 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 over-packing (**Online Methods**). Otherwise, the nucleotide reflects back without forming a new contact (denoted by the outward arrow). (b) Fraction of nucleotides, *f*(*n*), forming at most a given number of contacts, *n*. Mean (open circles) and standard deviations (error bars) were computed over all single-chain RNA structures in the RCSB database. Adjacent and same-helix nucleotide neighbors were excluded from the number of contacts calculation. Vertical dashed lines correspond to the minimal and maximal number of contacts, 0.5 and 11, respectively. (c) HRP-directed DMD simulation algorithm.

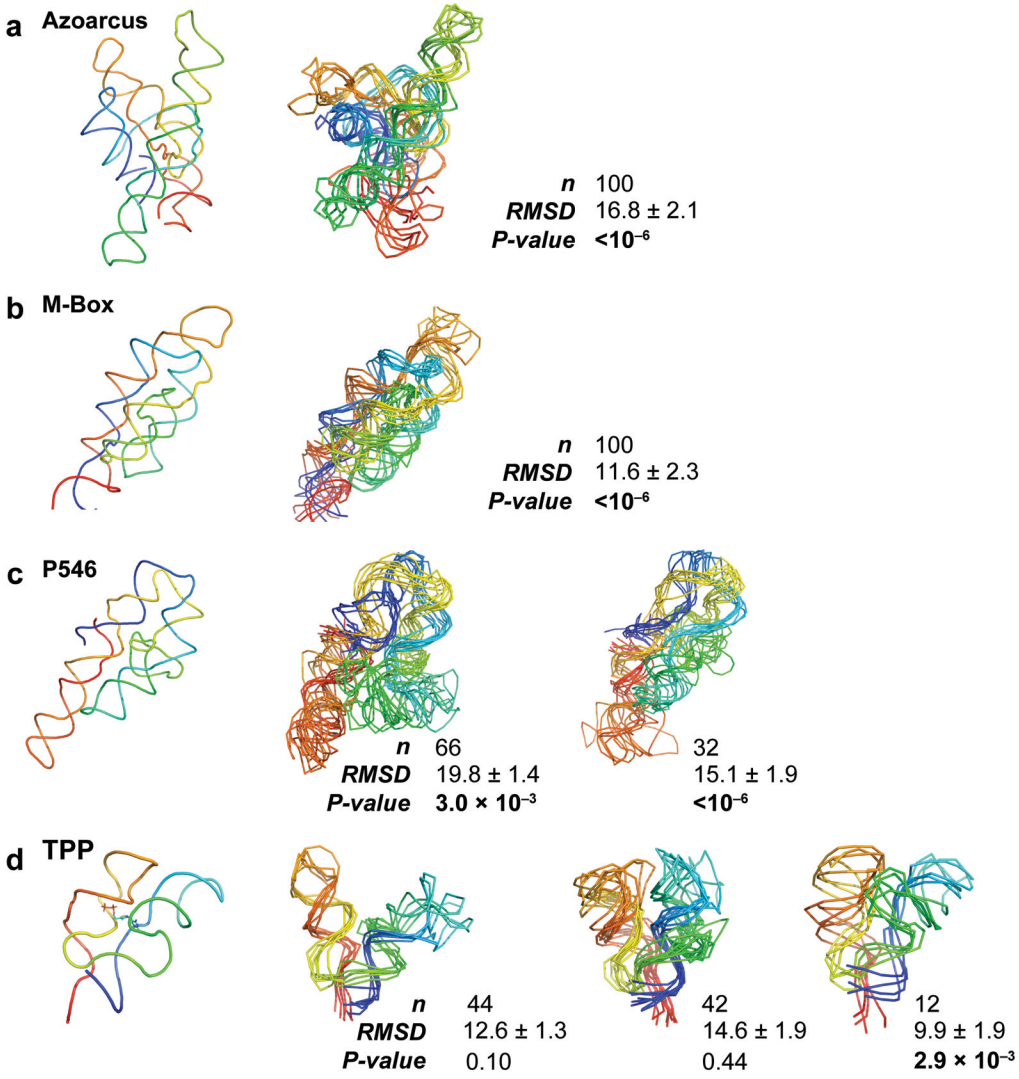

Figure 3. HRP-directed RNA fold refinement for the training set

RNAs are shown with backbone traces. The left-most panel shows the accepted structure for each RNA. Right-hand panels show representative structures for each highly populated cluster. Small clusters (with 1 or 2 structures) are not shown. 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[16] are emphasized in bold.

Table 1

Summary of HRP-directed RNA fold refinement for the studied RNAs

The first six RNAs comprise the training set used for algorithm optimization and applicability determination: yeast $tRNA^{Asp}$ (ref 27), thiamine pyrophosphate (TPP) riboswitch[28], specificity domain of ribonuclease P[29]), P546 domain of the *Tetrahymena thermophila* group I intron[19], M-Box riboswitch[24], and *Azoarcus* group I intron[30]. The last four RNAs were used for testing the performance: glmS ribozyme[31], lysine riboswitch[32], catalytic domain of ribonuclease P[33], and *Oceanobacillus iheyensis* group II intron[34]. The fraction of highly protected nucleotides, $f_{0.25}$, was computed using only the experimental HRP data; $f_{0.25}$ values above 0.25 are in bold. The structure-reactivity correlation, $C_{S\text{-}R}$, was calculated with reference to the accepted experimental structure. One hundred selected structures were clustered by pairwise RMSD (Online **Methods**). Small clusters (1 or 2 structures) were excluded. For each cluster, *P*-values were calculated based on the average RMSD with respect to the accepted experimental structure[16]; highly significant predictions ($P < 0.01$) are in bold.

RNA	Length (nts)	$f_{0.25}$	$C_{S\text{-}R}$	Number of clusters	Large Clusters		
					n **(out of 100)**	***RMSD*** **(Å)**	***P*****-value**
Azoarcus 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

RNA	Length (nts)	$f_{0.25}$	$C_{S\text{-}R}$	Number of clusters	Large Clusters		
					***n* (out of 100)**	***RMSD* (Å)**	***P*-value**
Group II intron	412	0.21	−0.30	-	-	-	-
RNase P catalytic domain	231	**0.28**	−0.50	4	46	19.2±1.6	$<\mathbf{10^{-6}}$
					41	21.6±2.1	$<\mathbf{10^{-6}}$
					8	25.0±0.7	$\mathbf{1.3 \times 10^{-4}}$
					5	24.4±2.0	$\mathbf{3.5 \times 10^{-5}}$
Lysine riboswitch	174	**0.36**	−0.57	3	57	12.0±1.6	$<\mathbf{1.0^{-6}}$
					42	18.1±1.0	$\mathbf{1.6 \times 10^{-6}}$
glmS ribozyme	152	**0.35**	−0.55	2	74	16.6±1.7	$\mathbf{1.5 \times 10^{-5}}$
					26	8.5±1.3	$<\mathbf{10^{-6}}$