



# Automated *de novo* prediction of native-like RNA tertiary structures

Rhiju Das and David Baker\*

Department of Biochemistry and Howard Hughes Medical Institute, University of Washington, Box 357350, Seattle, WA 98195

Edited by Ignacio Tinoco, Jr., University of California, Berkeley, CA, and approved July 10, 2007 (received for review April 25, 2007)

RNA tertiary structure prediction has been based almost entirely on base-pairing constraints derived from phylogenetic covariation analysis. We describe here a complementary approach, inspired by the Rosetta low-resolution protein structure prediction method, that seeks the lowest energy tertiary structure for a given RNA sequence without using evolutionary information. In a benchmark test of 20 RNA sequences with known structure and lengths of  $\approx 30$  nt, the new method reproduces better than 90% of Watson–Crick base pairs, comparable with the accuracy of secondary structure prediction methods. In more than half the cases, at least one of the top five models agrees with the native structure to better than 4 Å rmsd over the backbone. Most importantly, the method recapitulates more than one-third of non-Watson–Crick base pairs seen in the native structures. Tandem stacks of “sheared” base pairs, base triplets, and pseudoknots are among the noncanonical features reproduced in the models. In the cases in which none of the top five models were native-like, higher energy conformations similar to the native structures are still sampled frequently but not assigned low energies. These results suggest that modest improvements in the energy function, together with the incorporation of information from phylogenetic covariance, may allow confident and accurate structure prediction for larger and more complex RNA chains.

*ab initio* | energy-based | fragment assembly | nucleic acid | Rosetta

The biological roles of RNA molecules range from carrying simple messages to sensing, modifying, and creating a wide array of biomolecules (1). The latter tasks typically require the attainment of complex, three-dimensional structures, and it has long been noted that the problem of predicting the folds of stable, structured RNA molecules should be significantly easier than the analogous puzzle for proteins (2). A limited alphabet of chemically similar side-chains ensures that a fairly complete picture of the common conformations and preferred interactions for each nucleotide can be obtained. Further, the accuracy of secondary structure prediction algorithms (3) effectively reduces the RNA folding problem to one of determining the non-Watson–Crick base pairs and the backbone trajectories that interconnect canonical Watson–Crick double helices. There has been much recent progress in the careful classification of base-pairing interactions (4–6) and of backbone conformations (7, 8). Along with these insights and advances, several powerful software packages have been developed to model RNA tertiary folds (see refs. 9–15; reviewed in ref. 3).

The success of each of these RNA fold prediction algorithms relies on harnessing experimental data, evolutionary information, and interactive input from expert users to select and position noncanonical tertiary features in the final models. In principle, however it should be possible to predict RNA tertiary structures by minimizing the free energy estimated for each chain configuration. Indeed, widely used methods for RNA secondary structure prediction are founded on such a “thermodynamic hypothesis” for RNA structure (16). Such a purely energy-based prediction of tertiary structure has perhaps not been attempted because of concerns about sufficient conformational sampling and a sufficiently accurate energy function for noncanonical RNA interactions.

In this study, we explore a fully automated and energy-based approach to RNA tertiary structure prediction inspired by the Rosetta low-resolution protein structure prediction method. Fragment assembly of RNA (FARNA) guided by a knowledge-based energy function takes into account both the backbone conformational preferences and side-chain interaction preferences seen in experimentally determined RNA structures. The FARNA methodology is a *de novo* approach in the sense that phylogenetic information, secondary structure predictions, experimental data, and structures of direct homologs are not used as inputs to the method. We present an initial benchmark of the method on 20 small RNA sequences. The results indicate that the method effectively samples and frequently selects the canonical and noncanonical features found in the native structures. We find excellent recapitulation and discrimination of native structures for approximately half of the test set, and sampling of near-native structures for nearly the whole set. Most importantly, better than one-third of noncanonical base pairs, the crucial interactions that define RNA tertiary motifs, are recovered in the benchmark.

## Results

Tertiary structure prediction for biopolymers requires an effective method for efficiently sampling plausible conformations and an energy function which approximates the physics underlying folding. We first describe a fragment assembly strategy that greatly simplifies conformational sampling and then the components of a simple energy function to guide this sampling. To illustrate the approach, we describe how each of these ingredients contributes to an accurate all-atom structure for a small model system, a hairpin with a GCAA tetraloop. We then describe results of applying this automated method to a larger benchmark of RNA sequences. We conclude by discussing prospects for extending the methodology to higher resolution and longer sequences.

**Assembling Subfragments from RNA Structures to Limit Conformational Space.** At first glance, RNA folding seems to involve an “astronomical” number of conformations, analogous to Levinthal’s paradox for protein folding. With seven torsion angles per residue and a deformable ribose ring, the conformational space available to a small 12-residue RNA chain comprises nearly one hundred dimensions. Even if only two potential states are assumed per torsion angle, there are  $\approx 10^{28}$  potential conformations to search, in the absence of correlations between the torsion angles. In reality, however, there are strong limitations on the sets of sampled nucleotide torsion angles because of covalent closure of the sugar

Author contributions: R.D. and D.B. designed research; R.D. performed research; R.D. analyzed data; and R.D. and D.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Abbreviation: FARNA, fragment assembly of ribonucleic acids.

\*To whom correspondence should be addressed. E-mail: dabaker@u.washington.edu.

This article contains supporting information online at [www.pnas.org/cgi/content/full/0703836104/DC1](http://www.pnas.org/cgi/content/full/0703836104/DC1).

© 2007 by The National Academy of Sciences of the USA

ring, strong steric penalties against atom overlap within each nucleotide and between adjacent nucleotides (see, e.g., ref. 17 and references therein), and other physical/chemical factors like inter-nucleotide hydrogen bonding. In this study, we model these sequence-dependent contributions by assuming that the distribution of conformations observed in known RNA structures for a given trinucleotide sequence provides a reasonable approximation to the conformations sampled by the sequence during folding, following the approach used in Rosetta protein structure prediction (18). This approach captures local conformational correlations, including those between side chain and backbone degrees of freedom, in a straightforward empirical manner.

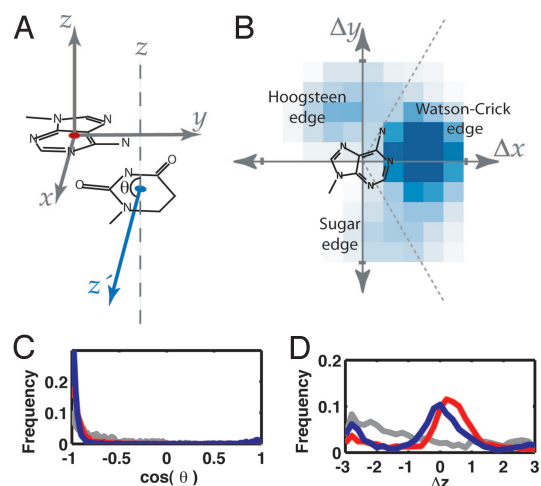
In benchmarks of *de novo* protein folding algorithms, great care must be taken to avoid contamination of fragment libraries by proteins that are related by evolution to the targets of interest. In the RNA case, the problem is conveniently avoided by drawing fragments from a single crystal structure containing just over 2,700 ribonucleotides from the large ribosomal subunit from *Haloarcula marismortui* [1FFK (19)]. Because of the limited four-letter alphabet of RNA (here further reduced to a two-letter pyrimidine/purine alphabet to diversify the fragment library), this single source still provides >300 potential three-residue fragments for each position of any new RNA sequence. Relative to what is sampled by isolated RNA chains, the library is presumably enriched for diverse, noncanonical conformations because of the RNA/protein interactions in the ribosome [see, e.g., studies of DNA double helix conformation in the presence and absence of proteins (20)]. We note that peptide fragments for protein structure prediction are similarly drawn from molecules that interact with different partners, including ligands, nucleic acids, and other proteins.

**Deriving a Knowledge-Based Energy Function for RNA Tertiary Structure.** To guide conformational sampling, an energy function is required that encodes the physical interactions most important for stabilizing RNA tertiary structure with reasonable accuracy and at a resolution appropriate to the molecular representation. As in protein structure prediction (18), the RNA energy function used herein includes a term weakly favoring compactness (proportional to radius-of-gyration) and a term to penalize steric clashes between atoms [see *Methods* and [supporting information \(SI\) Fig. 5](#)]. The remaining terms of the potential are specially designed for RNA interactions.

Even before the earliest crystal structures of long folded RNA chains were obtained, base pairing and base stacking were recognized as critical interactions that stabilize native nucleic acid structures. Taking into account their effects led to astounding predictions of three-dimensional structures for molecules including double-stranded DNA (21) and transfer RNA (22). To capture these important interactions in an internally consistent manner, we have developed a knowledge-based base-pairing potential and a base-stacking potential, similar to the potential proposed by Sykes and Levitt (6). In particular, rather than modeling hydrogen bonds in atomic detail, the potential has a coarse-grained form whose resolution has been chosen to match the coarse resolution offered by assembly of discrete fragments.

A coordinate system is set up on each base (Fig. 1A), with the origin at the centroid of the base heavy atoms, the *x* axis passing through the Watson-Crick edge at the N1 atom (purines) or N3 atom (pyrimidines), and the *z* axis perpendicular to the base plane. Base pairs in the ribosome crystal structure 1FFK generally have coplanar bases ( $|\Delta z| < 3 \text{ \AA}$ ) whose centroids are close to each other ( $\sqrt{\Delta x^2 + \Delta y^2} < 8 \text{ \AA}$ ). To encode the observed pairing geometries, we constructed a low-resolution knowledge-based interaction potential proportional to the logarithm of the frequency of finding the  $\Delta x$  and  $\Delta y$  of the base configurations in the ribosome crystal structure.

As an example, the free energy for finding a uracil base at a given



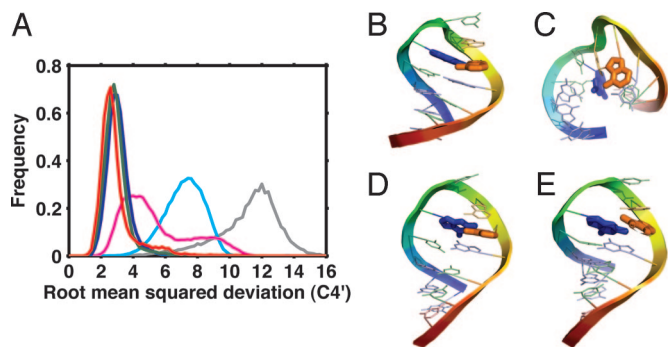
**Fig. 1.** A simple energy function for RNA fragment assembly. (A) Coordinate system set up on one base to define the potential. (B) Distribution of  $\Delta x$ ,  $\Delta y$  coordinates for a uridine residue near adenosine residues in the ribosome crystal structure, smoothed with a 2- $\text{\AA}$  Gaussian filter. The logarithm of this distribution provides a knowledge-based potential for *de novo* RNA structure prediction. Interactions with the three different edges of adenosine (Watson-Crick, Hoogsteen, and sugar) correspond to positions in the three sectors of the map demarcated by the dotted lines and the negative *x* axis. (C and D) Distributions of angle between base planes (C) and relative stagger of base planes (D) for base pairs observed in the large ribosome crystal structure (blue) and models produced without and with coplanarity terms (gray and red, respectively).

position relative to an adenosine base is shown in Fig. 1B, with the full base-pairing potential described in SI Fig. 6. This potential encodes the geometries and relative strengths of possible interactions with the Hoogsteen, sugar, and Watson-Crick edges of adenosine (4, 6); each of these types of interactions is visible in Fig. 1B. The resulting potential is necessarily approximate as it ignores the correlations between neighboring base pairs that shape the distributions observed in the ribosome. In principle, the extensive data on base-pairing energies from duplex melting experiments (16, 23) could also be used to calibrate this potential, as in secondary structure prediction algorithms. However, the scarcity of thermodynamic data on non-Watson-Crick interactions, as well as ambiguities in dissecting experimental energies into base stacking, base pairing, and entropic components, led us to take a simple knowledge-based approach, similar to the strategies long used in protein structure prediction (18, 24, 25) and supported by the apparently limited number of ways that bases pair with each other (4, 6).

Use of the above base-pairing potential alone leads to pairs that are not coplanar. We therefore include terms that are dependent on the stagger between the bases ( $\Delta z$  in Fig. 1A) and on the dot product of the two base normals ( $\cos \theta$ ). These terms were derived from the log-odds ratio of the distributions of these values in 1FFK versus a set of fragment-assembled decoys without any such coplanarity terms (see gray and blue lines in Fig. 1C and D and SI Fig. 7). Addition of these terms brings the model distributions in good agreement with the distributions seen in the ribosome crystal structure (compare red and blue lines in Fig. 1C and D).

Finally, in addition to this base-pairing potential, base doublets with  $\sqrt{x^2 + y^2} < 4 \text{ \AA}$ , and  $3 \text{ \AA} < |\Delta z| < 6.2 \text{ \AA}$  are given a bonus of  $-1 \text{ kT}$ , as a reward for stacking. In practice, halving or doubling the weight on this potential had little effect on the results (data not shown).

**Fragment Assembly Monte Carlo Tested on a Model System.** We give the overall automated procedure of Monte Carlo fragment assem-



**Fig. 2.** Finding the native structure of a small RNA hairpin by FARNA. (A) Histograms of rmsd (calculated over C4' atoms) between models generated by 5,000 cycles of Monte Carlo fragment assembly without the influence of any energy terms (gray) and with successive addition of the following terms to the energy function: radius-of-gyration (cyan); steric penalties (magenta); Watson-Crick-edge component of the base-pairing term (blue); Hoogsteen and sugar-edge components of base pairing and base stacking (green); and coplanarity terms (red). (B) Native structure of the hairpin [first model from NMR ensemble 1ZIH (26)]. (C–E) Lowest energy structures from simulations with radius-of-gyration term and steric penalties (C), plus Watson-Crick-edge component of the base-pairing term (D), and the full energy function (E). The residues G5 and A8, which form a sheared base pair in the native structure, are highlighted. In this figure and following figures, the coloring scheme shows rainbow coloring for the backbone (cartoons); and adenosine, cytidine, guanosine, and uridine bases are orange, green, blue, and red, respectively. Residues discussed in *Results* are rendered with thicker lines. Figures of molecules prepared in Pymol (Delano Scientific).

bly guided by a simple energy function the acronym of “FARNA,” for fragment assembly of RNA. This section illustrates how the addition of each of the energy components described above contributes to improved predictions by FARNA for a small model system. The system, a 12-residue sequence GGGCGCAAGCCU, forms a short, stable hairpin capped by a GCAA tetraloop and is well characterized by NMR spectroscopy [Fig. 2; (26)].

Fig. 2A displays a histogram of backbone rmsd (computed over C4' atoms) from the native state for conformations produced by Monte Carlo fragment assembly (gray line) beginning with a fully extended chain. Remarkably, favoring compaction by use of a term proportional to the chain's radius of gyration produces a measurable fraction of models with global shape similarity to the native state, with rmsd < 4 Å to the native state (cyan line in Fig. 2A). Further, disallowing clashes between RNA atoms produces conformations that are even more native-like, with the most probable rmsd improving to 4 Å (magenta line in Fig. 2A; see Fig. 2C). In this population, approximately one out of a thousand conformations has a nearly atomic-resolution backbone trace (rmsd < 2 Å). This is a far higher frequency than the rate of  $\approx 10^{-28}$  expected from the naive estimate given above and attests to the power of backbone conformational preferences, generic compaction, and sterics in favoring native conformations.

Inclusion of just the Watson-Crick-edge component of the base-pairing potential gave a dramatic shift of nearly the entire population to native-like structures (rmsd < 4 Å; see blue line in Fig. 2A). The lowest energy conformations were nearly indistinguishable from the native state (Fig. 2C). These models reproduce not only the four canonical base pairs in the stem but also the “sheared” G-A base pair (blue and orange bases in Fig. 2B–E) and the stacking pattern of the loop. Despite the absence of terms that might directly favor the sugar-edge/Hoogsteen-edge G-A interaction and base stacking, the native conformation is still selected because of conformational preferences for the tetraloop present in the ribosome-derived library. Inclusion of the database-derived sugar-edge and Hoogsteen-edge base-pairing components and a base stacking term slightly improves the population (green line in

Fig. 2A). A further improvement is obtained by including terms that favor coplanarity of the two interacting base pairs (red line in Fig. 2A; see Fig. 2E).

**A Benchmark for Fragment Assembly Monte Carlo.** Do the results on the simple hairpin model system generalize to other RNA sequences? We tested FARNA on a benchmark of 20 diverse sequences with stable structures (27) that have been experimentally characterized at high resolution, shown in Table 1. These RNA structures contain non-Watson-Crick base pairs, triplets, and unusual backbone trajectories, and most have lengths < 30 residues, to ensure reasonable sampling. Because the database of RNA single-chain structures solved by high-resolution crystallography is small, the benchmark includes cases solved by NMR as well as several crystallographic cases involving more than one chain. For several of the multiple-chain cases, the separated chains are known to form alternative single-chain structures in isolation but rearrange into oligomer complexes at the high effective concentrations sampled by crystallography (28–31). To avoid sampling the monomer configurations, the relative rigid body orientation of a single inter-chain base pair (see Table 1) was held fixed in the simulation, similar to a procedure recently developed for enforcing beta strand pairings in proteins (32). We then assessed the subsequent recapitulation of other canonical and noncanonical features.

We first discuss the overall accuracy of the models, and then describe individual examples. The assessment of RNA structures requires analysis of both base-pairing patterns (4, 5) and backbone conformation (7, 8). The number of native Watson-Crick and non-Watson-Crick base pairs and the backbone rmsd to the native state are given in Table 1 for the best of five largest clusters of models generated by fragment assembly (similar to the procedure used in evaluating protein structure predictions; see *Methods*). We first note that the majority of Watson-Crick base pairs (92%) are recapitulated for the best of five models across the benchmark; assessing the top cluster center, rather than the best of five, reduces this value to 86%. These rates for FARNA are comparable with the rate for state-of-the-art secondary structure prediction algorithms for this set, e.g., 94% using Unafold (16). Such accurate secondary structures might be expected to lead to excellent global backbone shapes, as modeling regions predicted to be A-form double helices is straightforward. Indeed, FARNA models for 11 of the 20 benchmark sequences agreed with the native state within a backbone rmsd of 4.00 Å (Table 1), with even better agreement if just subsets of residues making Watson-Crick base pairs are considered (SI Table 2).

The most interesting features of native RNA structures are noncanonical backbone conformations and non-Watson-Crick base pairs, but prediction of these features is difficult and typically requires signatures from phylogenetic covariance culled by human inspection (3, 5, 22, 33). It is therefore encouraging that the automated FARNA methodology finds accurate conformations for noncanonical regions for 13 of the 20 benchmark sequences (SI Table 2). Low rmsd values, however, can be achieved by RNA conformations with incorrect base interactions (see, e.g., 1KD5 in Table 1) or can partly follow from the assumed pairings in multi-chain cases. Thus, the most important result of this FARNA benchmark is the accurate prediction of native noncanonical base pairs, including information on which two base edges are interacting in each base pair, at a significant rate of 36% (Table 1).

**Accurate Prediction of Noncanonical Features.** Fig. 3A shows an example of a stack of four noncanonical G-A and A-A base pairs observed in the native structure 283D (31) that is accurately recapitulated in the FARNA model. The convergence of FARNA to this model is particularly noteworthy because the symmetry of this duplex was not imposed during fragment assembly. Further, each chain is known to form a stable GAAA-tetraloop-capped hairpin in isolation (31) (see, e.g., Fig. 2B), and indeed FARNA

**Table 1. Benchmark of 20 RNA molecules**

PDB	Method	Len	Pairing	Cut	Native*			Cluster center			Lowest RMS model			Models		
					WC	NWC	BUL	RMS	WC	NWC	BUL	RMS	WC		NWC	BUL
157D	X-ray	24	7–18	12	10	2	0	2.96	10	2	0	1.15	10	2	0	53679
1A4D	NMR	41	—	—	12	7	1	6.48	11	1	0	3.43	4	3	0	28949
1CSL	X-ray	28	12–14	13	9	3	2	4.03	9	2	2	2.26	8	2	1	45441
1DQF	X-ray	19	9–10	9	9	0	1	2.75	9	0	1	1.31	9	0	1	66481
1ESY	NMR	19	—	—	6	3	4	3.98	6	1	0	1.44	6	1	0	69103
119X	X-ray	26	13–14	13	12	0	2	4.46	12	0	0	1.93	12	0	2	51267
1J65	X-ray	24	†	†	0	24	0	13.99	0	5	0	2.17	0	13	0	46815
1KD5	X-ray	22	10–12	11	6	6	0	3.58	4	1	0	1.61	3	0	0	59896
1KKA	NMR	17	—	—	5	3	0	4.14	5	1	0	2.08	5	0	0	81492
1L2X	X-ray	27	—	—	8	5	3	3.88	7	1	0	3.11	7	1	0	47958
1MHK	X-ray	32	1–26, 14–31	12, 26	10	4	0	10.53	10	3	0	3.83	10	3	0	38179
1Q9A	X-ray	27	—	—	6	6	0	6.11	6	2	0	2.65	5	3	0	48817
1QWA	NMR	21	—	—	8	1	2	3.71	6	0	0	2.01	6	0	0	65977
1XJR	X-ray	46	—	—	13	9	3	9.82	10	4	0	6.25	11	1	2	24646
1ZIH	NMR	12	—	—	4	1	0	1.71	4	1	0	1.03	4	1	0	117104
255D	X-ray	24	12–13	12	10	2	0	1.68	10	2	0	1.31	10	2	0	54701
283D	X-ray	24	12–13	12	8	4	0	2.61	8	4	0	1.65	8	2	0	53062
28SP	NMR	28	—	—	7	6	1	3.20	7	3	0	2.31	6	3	0	46034
2A43	X-ray	26	—	—	7	4	4	4.93	4	1	1	2.79	6	0	1	49972
2F88	NMR	34	—	—	13	2	2	3.63	13	1	0	2.41	10	0	0	36664
Total†					154	89	25		142	32	4		131	34	7	
Freq.‡					1	1	1		0.92	0.36	0.16		0.85	0.38	0.28	

For multi-chain targets, the Pairing column refers to residues that were connected by a Watson–Crick base pair (drawn from the ribosome crystal structure 1FFK) throughout the simulation (see, e.g., ref. 33), and the Cut column refers to the residue after which a new chain begins. RMS refers to rmsd in angstroms from the native structure, calculated over C4' atoms. The shown cluster center is the best of five (in terms of the number of recapitulated non-Watson–Crick base pairs) obtained when clustering the lowest energy 1% of all models. WC, the number of native Watson–Crick base pairs (here including G–U wobble pairs); NWC, the number of native non-Watson–Crick base pairs; BUL, the number of native bulged residues recapitulated in the model; Len, number of residues.

\*For NMR models, the first model of the ensemble was taken as the reference state for rmsd calculations.

†For 1J65, assumed pairings were not Watson–Crick pairings but instead sugar-edge/Watson–Crick-edge G–U pairings (between residues 1,14; 8,13; and 7,20) taken from the native crystal structure. Cut points were at 6, 12, and 18 for this four-chain complex.

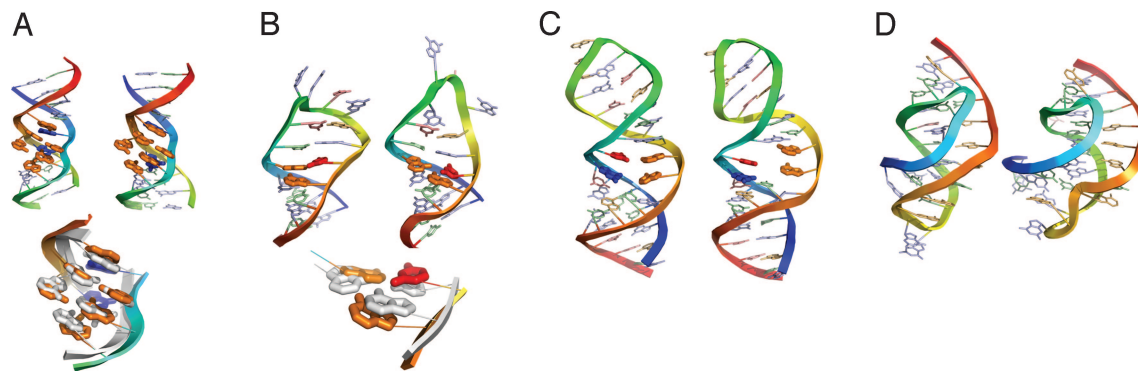
‡Total and frequency rows do not include base pairs assumed during the simulations to bring multi-chain complexes together (see Pairing column).

reproduces this alternative structure when a single chain is modeled (data not shown; see also Fig. 3C).

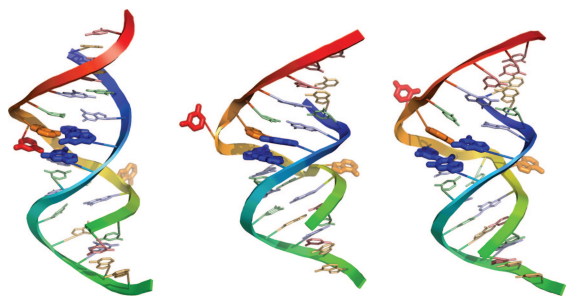
Platform motifs, in which two adjacent nucleotides are base-paired, are commonly seen in RNA structures as mediators of RNA interactions. Fig. 3B shows an example of an A–U platform forming a part of an A–U–A base triplet from the NMR structure 1ESY of the stem loop SL2 of the HIV-1 PSI packaging signal (34). The FARNa model reproduces this triplet motif and the resulting local distortion of the helix. Atomic resolution details of this base triplet such as hydrogen bonds and a potential coordinated water (34) will likely require a description of RNA conformation and energetics more finely grained than the current Monte Carlo fragment moves and coarse-grained potential. The nucleotides at the top of the

structure seem to be incorrectly modeled by FARNa as stacked rather than directed into solution as in the NMR structure; however, the backbone conformation is reasonably well modeled, and the NMR ensemble displays large conformational variance for the region.

The match-making tendency of FARNa to find base pairing or base stacking partners for all nucleotides is not an absolute rule. Fig. 3C shows the NMR structure for Domain 5 from a group II intron [2F88 (35)]. There is a “hole” in the middle of the structure where the strands do not base pair across the helix. Further, the observation of different conformations in this same region for structures solved for homologous molecules (35–38) suggests that the region is dynamic in solution. The FARNa model leaves the bases in this



**Fig. 3.** Best of five FARNa cluster centers (Left in each panel) and native structures (Right in each panel) for a curved RNA helix incorporating an internal loop with G–A and A–A non-Watson–Crick base pairing [283D (31)] (A); stem loop SL2 of the HIV-1 PSI RNA packaging signal [1ESY (34)] (B); domain 5 from the *Pyiella littoralis* group II intron [2F88 (35)] (C); and the frameshifting RNA pseudoknot from beet western yellow virus [1L2X (39)] (D). (A and B) Magnified superpositions of noncanonical base pairs (native in white; model in color) are displayed (Lower).



**Fig. 4.** Best of five FARNA cluster centers (*Left*), native structure (*Center*), and model with best recovery of non-Watson–Crick base pairs for overall FARNA population (*Right*) for the HIV-1 Rev response element high-affinity site [1CSL (52)]. Residues discussed in *Results* are rendered with thicker lines.

region appropriately unpaired and predicts the other parts of the structure accurately, including a GAAA tetraloop.

Finally, we investigated the ability of FARNA to model nontrivial backbone trajectories. The test RNA molecules discussed so far are hairpins or double helices with distortions due to noncanonical interactions. More complex structures typically involve >40 nt, beyond our current sampling ability (see *Discussion*). Among smaller RNA molecules, however, pseudoknots are recurring motifs with complex topologies, typically involving a single chain folding into two coaxial helices connected by loops, as shown in Fig. 3D [1L2X (39)]. The FARNA methodology recapitulates this backbone conformation, guided by noncanonical base pairs that form between loop residues and the Watson–Crick base pairs of the helices.

**Limits of the Current Methodology: Sampling or Energy Function?** In the FARNA benchmark, models for nine of the twenty RNA sequences are not native-like, with rmsd to the native state >4.00 Å. Failure of a Monte Carlo prediction methodology can generally be traced to deficiencies in one or both of its intrinsic components, conformational sampling and the energy function.

A hallmark of poor sampling is the inability of Monte Carlo moves to find conformations that are lower in energy than a known reference state. Inspection of energy versus rmsd plots (shown as **SI Fig. 8**) immediately highlights three cases where the current sampling strategy is not able to find any models with lower energy than the native structure. A potential reason for insufficient sampling might simply be an insufficient number of computational cycles for searching the conformation space. For one case, 1J6S, an intricate network of quadruplex interactions may hinder efficient conformational sampling of this 24-nt motif. The other two examples, 1A4D and 1XJR, are the largest sequences in our test set with lengths of 41 and 46 residues, respectively. Thus, more comprehensive or more efficient sampling may lead to successful recapitulation of the native structures for these cases.

The remaining six problem cases involve nonnative FARNA models with energies lower than the native structure, pointing to inaccuracies in the simple energy function that guides the Monte Carlo procedure. In each of these cases, however, the population of  $\approx 50,000$  models produced by FARNA contains at least one structure (typically several structures) with backbone rmsd within 4.00 Å of the native state (Table 1 and **SI Figs. 8 and 9**). Thus, the assumed energy landscape indeed contains these near-native structures as local minima. The conformations can be reached by the fragment assembly procedure but are not visited frequently enough or given low enough energies to be selected as one of the final five candidate cluster centers.

Fig. 4 shows an example of such a case from the crystal structure of the HIV-1 Rev response element (RRE) high-affinity site (1CSL; Fig. 4 *Center*). The FARNA cluster center (Fig. 4 *Left*) fails

to predict one of two bulged nucleotides seen in the native structure. On one hand, the crystal structure may reflect a nonphysiological conformation; unpaired nucleotides that are bulged in crystal structures of other molecules have been shown through NMR experiments to be stacked into helices in solution (compare refs. 40 and 41 and compare refs. 42 and 43; see also **SI Fig. 10** for an example of crystal contacts influencing a “hook-turn motif” and its prediction by FARNA). On the other hand, NMR studies of the RRE high-affinity site in Fig. 4 support both bulges observed in the crystal structure (44, 45). There are members of the FARNA population that do contain these bulges (Fig. 4 *Right*), but they do not have energies as low as the cluster center. The underprediction of bulged nucleotides seems to be a general issue for FARNA (Table 1) and may be ameliorated by a more realistic energy function including side chain entropy and bound water molecules that stabilize such backbone kinks.

## Discussion

**Prospects for a Highly Accurate Prediction Method for RNA.** Based on these initial results, the fundamental bottleneck for predicting structures of RNA molecules <40 nt seems to be not conformational sampling but the development of a more sophisticated energy function. The energy function presented herein already allows sampling of native structures and base pairs, so the problem becomes a tractable one of choosing from a few thousand conformations rather than from the astronomical number of structures that is theoretically possible.

We propose that the addition of a few fine-grained energy terms may be sufficient to solve this problem. Currently, electrostatic terms are not modeled. Use of Poisson–Boltzmann calculations, or knowledge-based approximations, would provide an approximate energy for these general effects of the counterion atmosphere (46); direct interactions with metal ions might be treated by using a rotamer sampling approach used for water in protein designs (47). Explicit hydrogen bonds are also not modeled. An orientation-dependent hydrogen bond potential [as is used for proteins (47)] and explicit sampling of local water positions could supplant the current coarse-grained base-pairing potential and would allow modeling of 2'-OH and phosphate hydrogen bonds that are presently ignored. Each of these additions would likely require continuous minimization of the RNA chain's torsional degrees of freedom rather than the coarse fragment moves used in FARNA. The computational expense of such high-resolution minimization would be justified, however, because an initial stage of FARNA could be used to deliver a population of starting conformations that already contain native-like (but inaccurately scored) conformations. A similar philosophy underlies the current Rosetta approach to protein structure prediction (48, 49).

Finally, the most interesting functional RNA molecules, including ribozymes and riboswitches, are composed of sequences longer than 40 nt, and their structures involve interactions between multiple double helices. Whereas the Monte Carlo sampling used by FARNA is more computationally efficient than enumerative sampling strategies (14), sufficient sampling by FARNA still becomes difficult for sequence lengths beyond 40 nt (**SI Fig. 9**). Further optimization of the code will likely allow an order-of-magnitude more sampling for these large constructs. More generally, however, we propose that seeding these fragment assembly simulations with potential Watson–Crick base pairs given by secondary-structure prediction algorithms (3) will be a powerful strategy for limiting the conformational space that needs to be explored by FARNA. Further, combining constraints from phylogenetic covariance with the *de novo* methodology presented herein offers exciting prospects for inferring functional structures of even the largest RNAs.

We have presented a fully automated algorithm for RNA structural modeling based on FARNA guided by a simple energy function. Even in this first study, canonical and noncanonical

features of 20 RNA molecules have been recapitulated at significant rates of 92% and 36%, respectively. Smaller RNAs in the test set are accurately reproduced with a resolution of better than 4 Å. In the remaining cases, we are encouraged to find that such near-native structures are still sampled with reasonably high frequency by this methodology.

Our efforts in RNA modeling have largely been guided by insights drawn from the large and lively field of *de novo* protein structure prediction. One of the important steps in the protein field has been the publication of “decoy sets” that have challenged investigators to find energy functions and refinement strategies to robustly discriminate native-like structures from nonnative structures. In this spirit, we are making the population of “decoys” from our study freely available, with the hope that other investigators will join the search for a more accurate and sophisticated energy function for RNA structures.

The prospect of blind, accurate structure prediction for small RNAs, relying solely on the minimization of free energy, seems feasible. Once this challenge is met, modeling efforts (potentially combining the *de novo* strategies described herein and phylogenetic information) become tantalizing possibilities for large ribozymes, riboswitches, and protein/RNA complexes.

## Methods

**Fragment Library.** For each position of the target RNA sequence, a library of trinucleotide torsion angles ( $\alpha$ ,  $\beta$ ,  $\delta$ ,  $\gamma$ ,  $\epsilon$ ,  $\zeta$ ,  $\chi$ , and sugar pucker amplitude; see e.g., ref. 50) is drawn from three-nucleotide segments of 1FFK that match the target in terms of the pattern of pyrimidines and purines. For one case, 1Q9A (the sarcin/ricin loop), the homologous sequence (residues 2684–2710) was excised from the 1FFK structure before choosing this torsional library.

**Energy Function.** The energy function is a sum of six terms. The first term, favoring the compact conformations seen in experimental RNA structures, is the radius-of-gyration (in Å), with a weight of 1 kT/Å. The second term penalizes steric clashes between several representative atoms on each nucleotide with steric radii inferred from the third smallest distance observed in the ribosome crystal structure 1FFK between atoms  $i$  and  $j$ , similar to the criteria used to derive the Rosetta low-resolution energy function for proteins; see SI Fig. 5. The third component of the energy function is a

base-pairing potential dependent on coordinates  $\Delta x$  and  $\Delta y$  (see Fig. 1 and SI Fig. 6). The fourth and fifth components of the energy function enforce coplanarity of pairing bases and are dependent on the variables  $\Delta z$  and  $\theta$  shown in Fig. 1A (see SI Fig. 7). The final component rewards base stacking, as described in Results.

**Fragment Assembly.** Simulations were initialized with an extended chain. Ideal bond lengths and bond angles were taken from the Nucleic Acid Databank website ([http://ndbserver.rutgers.edu/standards/ideal\\_geometries.html](http://ndbserver.rutgers.edu/standards/ideal_geometries.html)). At each Monte Carlo step, a random position was chosen in the chain, and torsions for three residues were replaced with those from a randomly chosen fragment; the move was accepted or rejected based on the classic Metropolis criterion (see, e.g., ref. 51). After an initial “heating” cycle of 1,000 random fragment insertions with no energy function, 50,000 fragment insertions were carried out with the RNA energy function, with the weight on coplanarity terms set to zero, half weight, and full weight for the first third, second third, and final third of the simulation, respectively. Generation of a single model takes  $\approx 45$  s for a 30-nt RNA on a Macintosh Intel 2 GHz processor, similar to the computational expense for low-resolution *de novo* structure prediction of proteins of comparable lengths (18). Increasing the number of fragment insertions by 10-fold to 500,000, made possible by the distributed computing network Rosetta@home, produces a slight improvement in model quality for larger RNAs; data from these runs are presented in Table 1. The best 1% by energy for  $\approx 50,000$  models were clustered with a 3 Å pairwise rmsd threshold, and the five largest clusters were assessed.

The procedure and energy function are implemented as part of Rosetta, whose source code and executable are available to academic users free of charge. Models are available at [http://faculty.washington.edu/rhiju/FARNA/farna\\_decoys.gz](http://faculty.washington.edu/rhiju/FARNA/farna_decoys.gz).

We thank Phil Bradley and Jim Havranek for advice on nucleic acid representations within Rosetta, Mike Tyka and John Karanicolas for helpful comments on the manuscript, Keith Laidig and Chance Reschke for excellent administration of computational resources, and the users of Rosetta@home for enabling rapid tests of the presented ideas (top users are listed in SI Table 3). We acknowledge the National Institutes of Health, the Howard Hughes Medical Foundation, and a Jane Coffin Childs Fellowship (to R.D.) for funding.

- Gesteland RF, Cech TR, Atkins JF (2006) *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World* (Cold Spring Harbor Lab Press, Cold Spring Harbor, NY).
- Tinoco I, Jr, Bustamante C (1999) *J Mol Biol* 293:271–281.
- Shapiro BA, Yingling YG, Kasprzak W, Bindewald E (2007) *Curr Opin Struct Biol* 17:157–165.
- Leontis NB, Westhof E (2001) *RNA* 7:499–512.
- Lescoute A, Leontis NB, Massire C, Westhof E (2005) *Nucleic Acids Res* 33:2395–2409.
- Sykes MT, Levitt M (2005) *J Mol Biol* 351:26–38.
- Murray LJ, Arendall WB, 3rd, Richardson DC, Richardson JS (2003) *Proc Natl Acad Sci USA* 100:13904–13909.
- Duarte CM, Wadley LM, Pyle AM (2003) *Nucleic Acids Res* 31:4755–4761.
- Wang R, Alexander RW, VanLoock M, Vladimirov S, Bukhtiyarov Y, Harvery SC, Cooperman BS (1999) *J Mol Biol* 286:521–540.
- Macke T, Case D (1998) in *Molecular Modeling of Nucleic Acids*, eds Leontis NB, SantaLucia JJ (Am Chem Soc, Washington, DC), pp. 379–393.
- Zwieb C, Müller F (1997) *Nucleic Acids Symp Ser* 36:69–71.
- Massire C, Westhof E (1998) *J Mol Graphics Model* 16:197–205, 255–7.
- Jossinet F, Westhof E (2005) *Bioinformatics* 21:3320–3321.
- Major F (2003) *Computing in Science & Engineering* 5:44–53.
- Yingling YG, Shapiro BA (2006) *J Mol Graphics Model* 25:261–274.
- Mathews DH, Sabina J, Zuker M, Turner DH (1999) *J Mol Biol* 288:911–940.
- Murthy VL, Srinivasan R, Draper DE, Rose GD (1999) *J Mol Biol* 291:313–327.
- Simons KT, Kooperberg C, Huang E, Baker D (1997) *J Mol Biol* 268:209–225.
- Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) *Science* 289:905–920.
- Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB (1998) *Proc Natl Acad Sci USA* 95:11163–11168.
- Watson JD, Crick FH (1953) *Nature* 171:737–738.
- Levitt M (1969) *Nature* 224:759–763.
- Xia T, SantaLucia JJ, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH (1998) *Biochemistry* 37:14719–14735.
- Jernigan RL, Bahar I (1996) *Curr Opin Struct Biol* 6:195–209.
- Zhang Y, Skolnick J (2004) *Proc Natl Acad Sci USA* 101:7594–7599.
- Jucker FM, Heus HA, Yip PF, Moors EH, Pardi A (1996) *J Mol Biol* 264:968–980.
- Suhnel J (1997) *Trends Genet* 13:206–207.
- Leonard GA, McAuley-Hecht KE, Ebel S, Lough DM, Brown T, Hunter WN (1994) *Structure (London)* 2:483–494.
- Szep S, Wang J, Moore PB (2003) *RNA* 9:44–51.
- Holbrook SR, Cheong C, Tinoco I, Jr, Kim SH (1991) *Nature* 353:579–581.
- Baeyens KJ, De Bondt HL, Pardi A, Holbrook SR (1996) *Proc Natl Acad Sci USA* 93:12851–12855.
- Bradley P, Baker D (2006) *Proteins* 65:922–929.
- Lehnert V, Jaeger L, Michel F, Westhof E (1996) *Chem Biol* 3:993–1009.
- Amarasinghe GK, De Guzman RN, Turner RB, Summers MF (2000) *J Mol Biol* 299:145–156.
- Seetharaman M, Eldho NV, Padgett RA, Dayie KT (2006) *RNA* 12:235–247.
- Zhang L, Doudna JA (2002) *Science* 295:2084–2088.
- Sashital DG, Cornilescu G, McManus CJ, Brow DA, Butcher SE (2004) *Nat Struct Mol Biol* 11:1237–1242.
- Sigel RK, Sashital DG, Abramovitz DL, Palmer AG, Butcher SE, Pyle AM (2004) *Nat Struct Mol Biol* 11:187–192.
- Eglin M, Minasov G, Su L, Rich A (2002) *Proc Natl Acad Sci USA* 99:4302–4307.
- Joshua-Tor L, Rabinovich D, Hope H, Frolow F, Appella E, Sussman JL (1988) *Nature* 334:82–84.
- Patel DJ, Kozlowski SA, Marky LA, Rice JA, Broka C, Itakura K, Breslauer KJ (1982) *Biochemistry* 21:445–451.
- Miller M, Harrison RW, Wlodawer A, Appella E, Sussman JL (1988) *Nature* 334:85–86.
- Roy S, Sklenar V, Appella E, Cohen JS (1987) *Biopolymers* 26:2041–2052.
- Battiste JL, Mao H, Rao NS, Tan R, Muhandiram DR, Kay LE, Frankel AD, Williamson JR (1996) *Science* 273:1547–1551.
- Peterson RD, Feigon J (1996) *J Mol Biol* 264:863–877.
- Woodson SA (2005) *Curr Opin Chem Biol* 9:104–109.
- Jiang L, Kuhlman B, Kortemme T, Baker D (2005) *Proteins* 58:893–904.
- Bradley P, Misura KM, Baker D (2005) *Science* 309:1868–1871.
- Das R, Qian B, Raman VS, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Sheffler V, et al. (2007) *Proteins*, in press.
- Bloomfield VA, Crothers DM, Tinoco I, Jr (1999) *Nucleic Acids: Structure, Properties and Functions* (University Science Books, Sausalito, CA).
- Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1995) *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge Univ Press, Cambridge, UK).
- Ippolito JA, Steitz TA (2000) *J Mol Biol* 295:711–717.