Published in final edited form as:

Nat Methods. 2010 April ; 7(4): 291–294. doi:10.1038/nmeth.1433.

Atomic accuracy in predicting and designing non-canonical RNA structure

Rhiju Das[1], **John Karanicolas**[2], and **David Baker**[3]

[1] Departments of Biochemistry and Physics, Stanford University, B400 Beckman Center, 279 Campus Drive, Stanford, CA 94305

[2] Center for Bioinformatics and Department of Molecular Biosciences, The University of Kansas, 2030 Becker Drive, Lawrence, KS 66045

[3] Howard Hughes Medical Institute and University of Washington, Department of Biochemistry, Box 357350, Seattle, 98195, USA

Abstract

We present a Rosetta full-atom framework for predicting and designing the non-canonical motifs that define RNA tertiary structure, called FARFAR (Fragment Assembly of RNA with Full Atom Refinement). For a test set of thirty-two 6-to-20-nucleotide motifs, the method recapitulated 50% of the experimental structures at near-atomic accuracy. Additionally, design calculations recovered the native sequence at the majority of RNA residues engaged in non-canonical interactions, and mutations predicted to stabilize a signal recognition particle domain were experimentally validated.

RNA is an ancient component of all living systems, whose catalytic prowess, biological importance, and ability to form complex folds have come to prominence in recent years[1]. Methods for inferring an RNA's pattern of canonical base pairs (secondary structure) have been well-calibrated and widely used for decades, often in concert with phylogenetic covariation analysis and structure mapping experiments.[2] A central, unsolved challenge at present is to model how the resulting canonical double helices are positioned into specific tertiary structures. The junctions, loops, and contacts that underlie these tertiary structures are frequently less than ten nucleotides in length and, in some cases, are able to self-assemble into the same microstructures when grafted into other helical contexts.[3,4] A critical requirement for a high-resolution RNA modeling method is its ability to find native-like solutions for the 'jigsaw puzzles' presented by these non-canonical motifs.

Send correspondence to: Phone (650) 723-5976. Fax: (650) 723-6783. rhiju@stanford.edu (RD) Phone: (206) 543-1295. Fax: (206) 685-1792. dabaker@u.washington.edu (DB)..

Author Contributions

R.D., research design, method implementation, data analysis, manuscript preparation; J.K., research design, method implementation; D.B., research design.

The authors declare that they have no competing financial interests with this publication.

Despite their small size, these motifs are often quite complex, with intricate meshes of non-Watson-Crick hydrogen bonds and irregular backbone conformations. Existing *de novo* methods for modeling tertiary structure have largely been limited to low resolution (e.g., Fragment Assembly of RNA (FARNA)[5], DMD[6]) or have required manual atom-level manipulation by expert users (e.g, MANIP[7]). Recent, automated full-atom methods (iFold3D[8], MC-SYM[9]) have described models of impressive quality, but non-canonical regions appear to be either incorrect[8] or take advantage of sequence similarity with homologs of known structure within the method's training database[9]. With respect to RNA design, rational engineering has yielded versatile sensors and nano-structures[10]-[12], but has so far been limited to rearrangements of existing sequence modules rather than designing new non-canonical structures.

In this work, we demonstrate that the Rosetta framework for scoring full-atom models and sampling molecule conformations[13] enables *de novo* structure prediction and design of complex RNAs with unprecedented resolution. Our approach assumes that native RNA structures populate global energy minima; the prediction problem is then to find the lowest energy conformation for a given RNA sequence, and the design problem, to find the lowest energy RNA sequences for a given structure.

Inspired by our experience in protein structure prediction, we hypothesized that the major shortcoming of prior approaches to RNA modeling – poor discrimination of native states by low-resolution energy functions – could be overcome by introducing a high resolution refinement phase driven by an accurate force field for atom-atom interactions (Supplementary Fig. 1). We therefore developed a method for Fragment Assembly of RNA with Full Atom Refinement (FARFAR). This method combines our previous FARNA protocol for low resolution conformational sampling with optimization in the physically realistic full-atom Rosetta energy function.

We tested FARFAR on a benchmark set of 32 motifs observed in high-resolution crystallographic models of ribozymes, riboswitches, and other non-coding RNAs (Supplementary Fig. 2). The conformational search made use of fragments of similar sequence drawn from a single crystallographic model, the large ribosomal subunit from *Haloarcula marismortui*[14]. We mimicked a true prediction scenario by ensuring that regions with evolutionary kinship to our test motifs were either absent or excised from the database. Unlike previous work that included canonical double helical regions that were straightforward to model[5],[6],[9] (see Supplementary Fig. 3), we focused on the conformations of non-canonical regions. The tests specified single canonical base pairs immediately adjacent to the motifs, as they provided necessary boundary conditions. The total computational time for fragment assembly and refinement of a single model of a twelve-nucleotide motif was 21 seconds on an Intel Xeon 2.33 GHz processor.

Out of the 32 targets, 14 cases gave at least one of five final models with better than 2.0 Å all-heavy-atom RMSD to the experimentally observed structure (Table 1 and Supplementary Fig. 4). Successes included widely studied RNAs such as the bulged-G motif of the sarcin-ricin loop, the most conserved domain of the signal recognition particle RNA, the bacterial loop E motif, and the kink-turn motif (Figs. 1a-d). Most strikingly, in nearly all of these

cases (11 of 14), the cluster center or lowest energy member recovered all the native non-canonical base pairs, recapitulating not only which residues were interacting but also the exact base edges making contact (Table 1). Several cases of incomplete base pair recovery appeared due to well-known ambiguities in automated pair assignments.[15] Finally, in an additional two cases with slightly higher RMSDs (see, e.g., Fig. 1e), *de novo* models recovered all the non-canonical base pairs. Thus the FARFAR method achieved high accuracy in 16 of 32 test cases. (Excluding targets used in optimizing weights of the energy function gave slightly better results, with high accuracy achieved in 9 of 16 cases; see Methods.) The Rosetta energy function was critical to the success of the approach. Refinements with the previous knowledge-based energy function (FARNA) and with molecular mechanics force fields (AMBER, CHARMM) and standard implicit solvent models led to worse discrimination (Supplementary Table 1). An upcoming generation of polarizable force fields with explicit treatments of water and ions, combined with novel free energy estimation methods, may eventually provide increased accuracy, albeit at much higher computational expense.

For the cases in which the current FARFAR method failed to achieve high resolution, symptoms of poor conformational sampling were observed: non-convergence of the lowest energy models, the inability to sample conformations near the native conformation, and the inability to reach energies as low as the native state (see cluster center size and closest-approach RMSD in Table 1; and energy gaps in Supplementary Table 1, respectively). In particular, each of these metrics became worse for larger motifs, with major difficulty encountered in the sampling of motifs with more than 12 residues (Fig. 1f).

Beyond structure prediction, we subjected the Rosetta full-atom energy function to an orthogonal test that is also a critical precedent for rational biomolecule engineering: the optimization of sequence to match a desired molecular backbone. This “inverse folding problem” was readily solved for even large RNAs by sequence design algorithms available in the Rosetta framework. For fifteen whole high-resolution RNA crystal structures (Supplementary Table 2), we stripped away the base atoms and remodeled them *de novo* by combinatorial optimization of base identities (A, C, G, or U) and rotameric conformations. The overall sequence recovery was 45%, well above the 25% expected by chance. Further, non-canonical sequences (not Watson-Crick or G·U) were recovered at a much higher rate of 65% (Fig. 2a). We observed poorer recovery with the previously developed low resolution FARNA score function (Fig. 2a & Supplementary Table 2).

Some sequence preferences that differed between natural RNA sequences and the Rosetta redesigns suggested that functional constraints besides folding stability exist for natural sequences, such as binding of protein partners or conformational switching. The availability of a “gold standard” sequence alignment of signal recognition particle RNAs from all three kingdoms of life permitted the robust identification of such discrepancies between natural and computed sequence profiles. Sequence changes *I* and *II* (see Fig. 2b) in this RNA's most conserved domain were calculated to stabilize this motif; their scarcity in the natural consensus may be due to binding of the protein Ffh. We tested the Rosetta prediction by chemical structure mapping experiments. In a folding buffer of 10 mM $MgCl_2$, 50 mM Na-HEPES, pH 8.0, both double mutant and wild type constructs gave indistinguishable patterns

of dimethyl sulfate modification that were consistent with the predicted tertiary structure (Figs. 2c,d). Further, the mutated construct exhibited increased folding stability compared to the wild type sequence, with less Mg^{2+} required to undergo the folding transition (Fig. 2e); the difference in free energy of folding, −1.2±0.5 kcal/mol, agreed with the predicted value of −1.6 kcal/mol (see Supplementary Fig. 5 for energy calibration). Tests of the single mutations also were in agreement with the Rosetta predictions (Supplementary Fig. 6). These same two sequence changes were previously suggested to be compatible with the SRP structure in an insightful visual comparison of the SRP motif and the loop E motif[15], although no predictions were made regarding stability.

The power of full-atom refinement demonstrated herein, combined with the ease of ascertaining RNA secondary structure, the small size of tertiary motifs, and the limited RNA alphabet, now permit atomic resolution *de novo* modeling and thermostabilization of non-canonical RNA motifs. Unsolved problems remain, including the blind prediction of previously unseen RNA motifs, the incorporation of small molecule ligands and explicit metal ions, and the prediction and design of larger RNA folds with new functionalities. Improvements in conformational sampling as well as incorporation of even modest experimental data should enable computational methods to meet these critical next challenges. The Rosetta code base is freely available for download at http://www.rosettacommons.org/.

Methods

All computational methods were implemented in the Rosetta 3.1. Full documentation, explicit command lines, and example files necessary to model the structure of the most conserved domain of the signal recognition particle (PDB *1LNT*) and to redesign all of its residues are included in the “manual” and “rosetta_demos” directories that are part of the release, freely available for download at http://www.rosettacommons.org.

Identification of RNA motifs

An automated algorithm to parse non-canonical segments (i.e., residues forming base pairs besides Watson-Crick or G-U pairs), along with “bounding” canonical base pairs, was applied to RNA crystal structures with diffraction resolutions of 3 Å or better, with a focus on ribozymes and riboswitches. Candidate motifs that did not interact with other regions of the structure and had lengths of 20 nucleotides or less were selected. This subset was then further filtered to remove sequence-redundant motifs. A final set of thirty-two sequence motifs and the assumed canonical base pairs (which form “boundary conditions” for each motif) are illustrated in Supplementary Fig. 2.

De novo modeling

Generation of *de novo* models was carried out by Fragment Assembly of RNA (FARNA), as described previously [5], starting from extended chains with ideal bond lengths and bond angles. Minor improvements to the FARNA score function were made to model base-backbone and backbone-backbone interactions at a coarse-grained level, as described in Supplementary Fig. 7. Further, small improvements in the conformational search were

implemented. Rather than using three-residue fragments, the fragment length was made finer, from 3 to 2 to 1, in successive stages of Monte Carlo fragment assembly. In addition, variations in sugar bond-length and bond-angle geometries were recorded in the fragment library and copied during fragment insertion moves to ensure sugar ring closure.

Most of the motifs herein involved multiple chains connected by at least one Watson-Crick base pair. These canonical base pairs were assumed to form, because they are typically known *a priori* in RNA modeling and because without these double-helical boundary constraints, RNA sequences often form alternative structures (see, e.g., ref.[18]). The energy function was supplemented with harmonic constraints placed between Watson-Crick edge atoms in the two residues that were assumed to form each bounding canonical base pair (see Supplementary Fig. 2). Further, each *de novo* run was seeded with a random subset of $N - 1$ Watson-Crick base pairs to define the connections between N chains by a tree-like topology for coordinate kinematics[19],[20]; every ten fragment insertions, alternative base-pairing geometries, drawn from an RNA database, were tested as an additional type of Monte Carlo move. The source of both the torsion fragments and the base pairing geometries was the refined structure of the archaeal large ribosomal subunit (1JJ2[14]), with the sarcin-ricin loop and the kink-turn motifs excluded. Using an alternative ribosome crystal structure for the fragment source (1VQ8) gave indistinguishable results for, e.g., Z-scores (see next section).

50,000 FARNA models were optimized in the context of the Rosetta full-atom energy function. This energy function is a simple and transferrable function that represents an approximate free energy (minus the conformational entropy) for each molecular state. Interactions between non-bonded atoms are modeled by pair-wise, distance-dependent potentials for van der Waals forces, hydrogen bonds, the packing of hydrophobic groups, and the desolvation penalties for burying polar groups[13]. Based on recent work in the Rosetta community on proteins and DNA, three additional non-bonded terms (Supplementary Fig. 8) were incorporated here and reweighted through an iterative calibration: (1) a potential for weak carbon hydrogen bonds, previously investigated for membrane proteins, (2) an alternative orientation-dependent model for desolvation based on occlusion of protein moeities, and (3) a term to approximately describe the screened electrostatic interactions between phosphates. Because subtle, bond-specific quantum effects complicate the general derivation of torsional potentials, we derived preferred values for RNA torsion angles and their corresponding spring constants from the ribosome crystal structure (Supplementary Fig. 9). More sophisticated treatments of electrostatics and the site-specific binding of water and multivalent metal ions, which are expected to be important for some RNA molecules[21], will be explored in future work.

Combinatorial sampling of 2′-OH torsions was followed by continuous, gradient-based optimization of all internal degrees of freedom by the Davidson-Fletcher-Powell method. Constraints were included to maintain bond lengths and angles within 0.02 Å and 2°, respectively, of ideal values and to tether atoms near their starting positions (with harmonic constraints penalizing a 2 Å deviation by 1 unit). After removing the latter set of tethers, a second stage of 2′-OH torsion optimization and minimization was carried out. After this process, steric clashes and bond geometry deviations were reduced to the level seen in

experimental RNA structures, as assessed by the independent MolProbity toolkit (see Supplementary Table 3 for a complete overview).

To test the AMBER99 force field, the TINKER module *minimize* with the GBSA keyword (implementing the Born radii of Still et al.[22]) was applied to the models that had been refined with the full-atom Rosetta energy function. To test the CHARMM27 force field, the CHARMM molecular mechanics program[23] was applied, using the nucleic acid force field (PARAM27) [24]. The generalized born molecular volume (GBMV) method[25],[26] was used as an implicit representation of the solvent. Default parameters for minimization and GBMV were taken from the MMTSB tool set[27]. Current molecular mechanics packages do not offer the prospect of continuous minimization of model coordinates in the context of the computationally expensive non-linear Poisson-Boltzmann treatment of counterions; as a first estimate of the effects of ion screening, we minimized models with the ion-free GBMV model, and then recomputed solvation energies with the Poisson-Boltzmann solver available in MMTSB. In principle, the explicit treatment of counterions and water in molecular mechanics calculations can provide increased accuracy, although the precise and efficient estimation of free energy differences between different molecular conformations remains an unsolved challenge in biomolecular simulation.

Base pairs of models and experimental structures were carried out with an automated annotation method based on RNAVIEW, but implemented in the Rosetta framework. The automated pair assignments were not entirely unambiguous. As an example, an ambiguity occurred for the SRP motif; base pair assignments from RNAVIEW[28] disagreed with the authoritative manual annotation[15] by giving different interacting edges to a central bifurcated G-G base pair and assigning an extra hydrogen bond between two (non-planar) C residues (see Supplementary Fig. 2). Fig. 1 shows the manual annotation.

Iterative optimization of weights of the energy function

Half of the thirty-two RNA motifs were randomly selected to optimize the weights on the tested score functions. Two thousand RNA models were generated by *de novo* fragment assembly, and two thousand additional native-like models were obtained by using a library of fragments drawn from the native structure rather than from the ribosome. Weights on the different components of the force field (12 parameters for the Rosetta energy function) were optimized with the fminsearch method in MATLAB to maximize the sum of the Z-score over the training set motifs, with the weights on the van der Waals term fixed. The Z-score for the force field was computed as the mean score of non-native decoys minus the mean score of the ten lowest-energy near-native models, divided by the standard deviation of non-native decoy scores. In this computation, non-native decoys with anomalously poor scores (higher than three standard deviations from the mean) were filtered out.

Results for large-scale *de novo* modeling for both training and test sets are given in Table 1. Because weight fitting can lead to unfair bias, we also carried out our analyses on the training and test sets separately. Results on the withheld test set were in fact better than for the training set (mean Z-scores of 3.61 vs 3.28;number of cases with positive energy gaps of 10 vs. 8; median rmsd for best of five clusters of 2.28 Å vs. 2.34 Å; and recovery of non-Watson-Crick base pairs of 43% vs. 38%), indicating that weight over-parametrization did

not occur. Furthermore, final results were largely independent of chosen weights. We recomputed the mean Z-scores for native state discrimination after changing the weights of each energy function term by ± 50% and optimizing weights of the other scores. Final Z-scores changed by less than 5% despite these large perturbations, indicating a robustness to the choice of weights; we have observed similar results in protein structure prediction (R.D., D.B., unpublished data).

Fixed backbone design

Tests of side-chain and sequence recovery were carried out on RNA crystal structures with resolutions better than 2.5 Å without close interactions to protein partners and with bases stripped from the structures (Supplementary Table 2). Using the same core routines as in protein side chain packing and design, the optimization of side-chain conformation and identity was carried out simultaneously at all residues; rapid simulated annealing was aided by pre-computation of all rotamer-rotamer pairwise energies. The nucleobase rotamers were constructed with the glycosidic torsion angle X set at its most probable *anti* value and at −1, −1/2, +1/2, and +1 standard deviations from this central value. The central value and standard deviations were computed based on RNA residues in the ribosome crystal structure for 2′-endo and 3′-endo sugar puckers separately. For purines, *syn* rotamers for X were analogously sampled. The placement of the 2′-OH hydrogen was also simultaneously optimized with the base rotamer; the torsion angle defined by the C3′-C2′-O2′-HO2′ atoms was sampled at six torsion angles (−140°, −80°, −20°, 40°, 100°, and 160°).

Structure mapping

A newly developed high-throughput RNA preparation, chemical modification, and capillary electrophoresis readout protocol was used for thermodynamic and structure mapping experiments and is briefly summarized here. SRP-motif RNA constructs were prepared with sequence GGCUACGCAAGUAAAACAAAUUACUCAGGUCCGGAAGGAAGCAGGUAAAAACCAAACC**AAAGAAACAACAACAACAAC** (primer binding site in bold), or with the mutations shown in the text. DNA templates including the 20 nt T7 primer sequence (TTCTAATACGACTCACTATA) were prepared by extension (Phusion, Finnzymes, MA) of 60 nucleotide sequences (Integrated DNA Technologies, IA), purified in Qiaquick columns (Qiagen, CA), and used as templates for *in vitro* transcription with T7 polymerase (New England Biolabs, MA). RNA was purified by phenol and chloroform extraction and buffer-exchanged into deionized water with P30 RNAse-free spin columns (BioRad, CA). The RNA (0.5 pmol) was incubated at 44 °C in a Hybex incubator with 50 mM Na-HEPES, pH 8.0, with varying concentrations of $MgCl_2$; after 1 minute, dimethyl sulfate (freshly diluted into water) was added to a final concentration of 0.25% and final volume of 20 μL. Repeat reactions with a final volume of 100 μL gave indistinguishable results for free energy differences between variants. After 15 minutes of modification, reactions were quenched with 0.25 volumes of 2-mercaptoethanol, oligo-dT beads (poly(A) purist, Ambion, CA), and 5′-rhodamine-green labeled primer (AAAAAAAAAAAAAAAAAAAAGTTGTTGTTGTTTCTTT, 0.125 pmol), and purified by magnetic separation. Reverse transcriptase reactions were carried out using Superscript III (Invitrogen, CA) and 10 mM dNTPs (with 2-deoxyinosine triphosphate

replacing dGTP), and purified by alkaline hydrolysis of the RNA and magnetic separation. Fluorescent DNA products, with a co-loaded Texas-Red-labeled reference ladder, were separated by capillary electrophoresis on an ABI3100 DNA sequencer and analyzed with specialized versions of the SAFA analysis scripts [29]. Plots and fits of fraction folded were carried out in MATLAB (MathWorks, MA), with errors estimated by bootstrapping. Free energy differences between variants with fitted $MgCl_2$ midpoints K_1 and K_2 and apparent Hill coefficients n_1 and n_2 were calculated as $\Delta\Delta G = (1/2)\,(n_1+n_2)\,k_B T \log(K_1/K_2)$. This expression corresponds to a model in which the additional number of Mg^{2+} associated to the RNA upon folding can vary linearly with log [$MgCl_2$].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank contributors to the current Rosetta codebase, local computer administrators D. Alonso and K. Laidig, the BioX^2 cluster (National Science Foundation award CNS-0619926) and TeraGrid computing resources for enabling rapid development of macromolecular modeling methods; and K. Sjölander for suggesting the acronym FARFAR. This work was supported by the Jane Coffin Childs and Burroughs-Wellcome Foundations (to R.D.), the Damon Runyon Cancer Research Foundation (J.K.), and the Howard Hughes Medical Institute (D.B.).

References

1. Gesteland, RF.; Cech, TR.; Atkins, JF. The RNA world : the nature of modern RNA suggests a prebiotic RNA world. Cold Spring Harbor Laboratory Press; Cold Spring Harbor, NY: 2006.
2. Shapiro BA, Yingling YG, Kasprzak W, Bindewald E. Curr Opin Struct Biol. 2007
3. Moore PB. Annu Rev Biochem. 1999; 68:287. [PubMed: 10872451]
4. Brion P, Westhof E. Annu Rev Biophys Biomol Struct. 1997; 26:113. [PubMed: 9241415]
5. Das R, Baker D. Proc Natl Acad Sci U S A. 2007; 104:14664. [PubMed: 17726102]
6. Ding F, et al. RNA. 2008; 14:1164. [PubMed: 18456842]
7. Massire C, Westhof E. J Mol Graph Model. 1998; 16:197. [PubMed: 10522239]
8. Sharma S, Ding F, Dokholyan NV. Bioinformatics. 2008; 24:1951. [PubMed: 18579566]
9. Parisien M, Major F. Nature. 2008; 452:51. [PubMed: 18322526]
10. Breaker RR. Nature. 2004; 432:838. [PubMed: 15602549]
11. Win MN, Smolke CD. Proc Natl Acad Sci U S A. 2007; 104:14283. [PubMed: 17709748]
12. Jaeger L, Westhof E, Leontis NB. Nucleic Acids Res. 2001; 29:455. [PubMed: 11139616]
13. Rohl CA, Strauss CE, Misura KM, Baker D. Methods Enzymol. 2004; 383:66. [PubMed: 15063647]
14. Klein DJ, Schmeing TM, Moore PB, Steitz TA. EMBO J. 2001; 20:4214. [PubMed: 11483524]
15. Leontis NB, Westhof E. RNA. 2001; 7:499. [PubMed: 11345429]
16. Larsen N, Zwieb C. Nucleic Acids Res. 1991; 19:209. [PubMed: 1707519]
17. Crooks GE, Hon G, Chandonia JM, Brenner SE. Genome Res. 2004; 14:1188. [PubMed: 15173120]
18. Baeyens KJ, De Bondt HL, Pardi A, Holbrook SR. Proc Natl Acad Sci U S A. 1996; 93:12851. [PubMed: 8917508]
19. Bradley P, Baker D. Proteins. 2006; 65:922. [PubMed: 17034045]
20. Das R, Baker D. Annu Rev Biochem. 2008; 77:363. [PubMed: 18410248]
21. Draper DE, Grilley D, Soto AM. Annu Rev Biophys Biomol Struct. 2005; 34:221. [PubMed: 15869389]

22. Qiu D, Shenkin PS, Hollinger FP, Still WC. J. Phys Chem. A. 1997; 101:3005.
23. Brooks BR, et al. J. Comput. Chem. 1983; 4:187.
24. MacKerell ADJ, et al. J. Phys. Chem B. 1998; 102:3586. [PubMed: 24889800]
25. Lee MS, Salsbury FRJ, Brooks CLI. J. Chem. Phys. 2002; 116:10606.
26. Lee M, Feig M, Salsbury FJ, Brooks C.r. J Comput Chem. 2003; 24:1348. [PubMed: 12827676]
27. Feig M, Karanicolas J, Brooks C.r. J Mol Graph Model. 2004; 222:377. [PubMed: 15099834]
28. Yang H, et al. Nucleic Acids Res. 2003; 31:3450. [PubMed: 12824344]
29. Das R, Laederach A, Pearlman SM, Herschlag D, Altman RB. RNA. 2005; 11:344. [PubMed: 15701734]

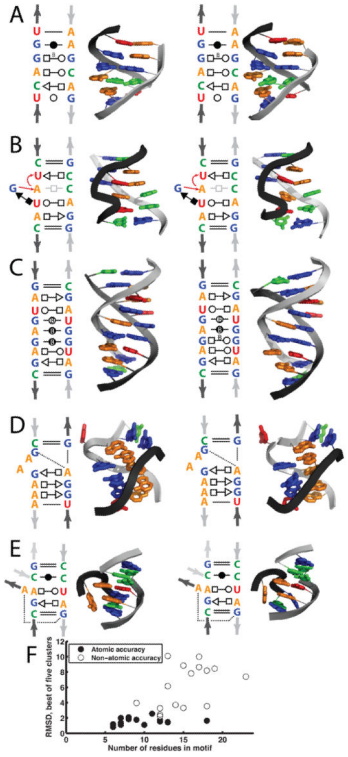

Figure 1.

Successes of *de novo* modeling of non-canonical RNA structure with Fragment Assembly of RNA with Full Atom Refinement (FARFAR). Two-dimensional annotations[15] and three-dimensional representations are shown for (**a**) the E. coli signal recognition particle Domain IV RNA, (**b**) the bulged-G motif from the E. coli sarcin-ricin loop, (**c**) the E. coli loop E motif, (**d**) the kink-turn motif from the SAM-I riboswitch (T. tengcongensis), and (**e**) the hook-turn motif. (PDB codes are 1LNT, 1Q9A, 354D, 2GIS, and 1MHK respectively.) Each panel depicts the experimentally observed structure (left) and the best of five low-energy cluster centers (right). In (**a**), a conserved A-C interaction that was missed by automated annotation is shown in gray. (**f**) All-heavy-atom RMSD for the best of five final predictions (low-energy cluster centers) plotted against the number of residues in the modeled motif. Filled symbols denote atomic accuracy models (see text).

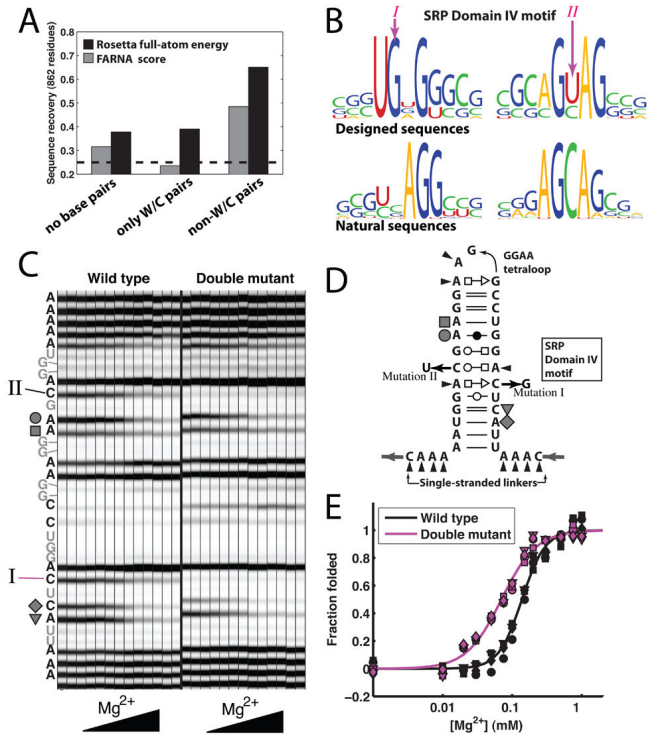

Figure 2.

Computational and experimental tests validate sequence design and thermostabilization. (**a**) Sequence recovery over 15 high resolution side-chain-stripped RNA structures optimizing the Rosetta full-atom energy (black bars) was better than chance (25%, dashed line) and better than tests with the FARNA score function (gray bars). (**b**) Sequence preference predicted from 1000 redesigns (top) compared to an alignment of SRP Domain IV RNA sequences drawn from all three kingdoms of life [16], in sequence logo format [17]. Two mutations (I and II) predicted by the Rosetta redesigns to stabilize folding are indicated. (**c**) Dimethyl sulfate (DMS) modification data probes the structure and thermodynamics of the SRP motif and variants. Sites of chemical modification were read out by reverse transcription of modified RNA with fluorescently labeled DNA primers, separated by multiplexed capillary electrophoresis. (**d**) Schematic of the construct's tertiary structure. Wedges mark residues that remained accessible to dimethyl sulfate in high Mg^{2+} folding conditions for the wild type RNA; the pattern for the mutant construct is indistinguishable except at the sites of mutation. (**e**) Folding isotherms by Mg^{2+} titration for four separate residues involved in the SRP motif's noncanonical structure (cf. symbols in **c** & **d**) overlay well and indicate that the Rosetta-predicted double mutant folds more stably than the wild type sequence. The left-most symbols represent conditions without Mg^{2+}. Full electrophoretic profiles and single mutant fits are presented in Supplementary Fig. 6.

Table 1

Attainment of native-like structure by *de novo* Fragment Assembly of RNA with Full Atom Refinement (FARFAR), using the full-atom Rosetta energy function. The lowest energy 500 of 50,000 refined conformations were clustered with a model-model heavy-atom RMSD cutoff of 2.0 Å. The five lowest energy clusters were taken as the *de novo* models; features of the best cluster (lowest RMSD to the experimental structure) are listed. See Supplementary Fig. 2 for motif definitions.

	Motif properties		Clustering statistics		Cluster center		Lowest energy cluster member		Lowest RMSD sampled
	No. res.	No. chains	Clust Rank	Cluster size	RMSD[a]	f_{NWC}[b]	RMSD[a]	f_{NWC}[b]	
G-A base pair	6	2	1	471	1.19	1/ 1	1.89	0/ 1	0.54
UUCG tetraloop	6	1	1	498	1.12	1/ 1	1.14	1/ 1	0.64
GAGA tetraloop from sarcin/ricin loop	6	1	1	500	0.82	1/ 1	1.00	1/ 1	0.52
Loop 8, A-type Ribonuclease P	7	1	5	27	1.38	0/ 0	1.41	0/ 0	1.13
Pentaloop from conserved region of SARS genome	7	1	3	237	1.10	1/ 1	1.48	1/ 1	0.88
L3, thiamine pyrophosphate riboswitch	7	1	4	6	2.00	0/ 1	2.68	0/ 1	1.44
Fragment with A-C pairs, SRP helix VI	8	2	1	284	1.83	2/ 2	2.74	1/ 2	0.48
Helix with U-C base pairs	8	2	2	491	2.10	2/ 2	2.56	1/ 2	1.11
Rev response element high affinity site	9	2	2	4	3.95	1/ 2	4.42	0/ 2	1.96
J4/5 from P4-P6 domain, Tetrahymena ribozyme	9	2	1	335	1.76	1/ 2	2.12	1/ 2	1.09
Tetraloop/helix interaction, L1 ligase crystal	10	3	1	500	1.10	1/ 3	1.21	2/ 3	0.69
Hook-turn motif	11	3	5	121	2.56	3/ 3	2.06	3/ 3	1.37
Helix with A-C base pairs	12	2	2	242	2.45	1/ 4	1.81	2/ 4	1.53
Curved helix with G-A and A-A base pairs	12	2	1	205	1.74	2/ 4	1.06	4/ 4	0.96
Fragment with G-G and G-A base pairs, SRP helix VI	12	2	3	98	3.27	0/ 5	4.25	0/ 5	0.86
Signal recognition particle Domain IV	12	2	4	321	1.54	2/ 5	1.22	4/ 5	0.93
Stem C internal loop, L1 ligase	12	2	1	489	2.24	2/ 3	2.42	2/ 3	1.88
Four-way junction, HCV IRES	13	4	3	30	10.09	1/ 4	10.63	1/ 4	2.99
Bulged G motif, sarcin/ricin loop	13	2	1	81	1.46	4/ 4	1.66	3/ 4	0.86

	Motif properties		Clustering statistics		Cluster center		Lowest energy cluster member		Lowest RMSD sampled
	No. res.	No. chains	Clust Rank	Cluster size	RMSD[a]	f_{NWC}[b]	RMSD[a]	f_{NWC}[b]	
Kink-turn motif from SAM-I riboswitch	13	2	1	7	1.43	3/ 3	1.36	3/ 3	1.22
Three-way junction, purine riboswitch	13	3	3	24	6.15	0/ 3	6.10	0/ 3	3.16
J4a-4b region, metal-sensing riboswitch	14	2	3	4	3.71	0/ 2	3.52	0/ 2	1.27
Kink-turn motif	15	2	2	25	8.85	1/ 3	9.43	2/ 3	3.05
Tetraloop/receptor, P4-P6 domain, Tetr. Ribozyme	15	3	4	13	3.31	2/ 5	2.89	2/ 5	2.21
Tertiary interaction, hammerhead ribozyme	16	3	2	4	7.82	0/ 3	8.50	1/ 3	4.37
Active site, hammerhead ribozyme	17	3	4	5	8.64	1/ 3	9.28	1/ 3	4.41
J5-5a hinge, P4-P6 domain, Tetr. Ribozyme	17	2	3	12	9.99	0/ 4	10.12	0/ 4	4.23
Loop E motif, 5S RNA	18	2	2	40	1.64	3/ 6	2.16	6/ 6	1.43
L2-L3 tertiary interaction, purine riboswitch	18	2	2	10	8.19	0/ 7	8.08	0/ 7	5.04
Pseudoknot, domain III, CPV IRES	18	2	4	11	3.55	0/ 0	3.90	0/ 0	2.29
Pre-catalytic conformation, hammerhead ribozyme	19	3	5	2	8.44	1/ 4	7.66	0/ 4	4.80
P1-L3, SAM-II riboswitch	23	2	5	5	7.40	0/ 1	7.47	0/ 1	3.99

[a] Heavy-atom RMSD to crystal structure.

[b] Number of non-Watson-Crick base pairs in crystal structure recovered in the model. Assignment of base pairing followed an automated method based on the RNAVIEW algorithm; counts of correct base pairings are lowered due to ambiguities in assigning bifurcated base pairs, pairs connected by single hydrogen bonds, or pairs that are not completely coplanar.