Atomic accuracy in predicting and designing noncanonical RNA structure

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

We present fragment assembly of RNA with full-atom refinement (FARFAR), a Rosetta framework for predicting and designing noncanonical motifs that define RNA tertiary structure. In a test set of thirty-two 6–20-nucleotide motifs, FARFAR recapitulated 50% of the experimental structures at near-atomic accuracy. Sequence redesign calculations recovered native bases at 65% of residues engaged in noncanonical interactions, and we experimentally validated mutations predicted to stabilize a signal recognition particle domain.

© 2010 Nature America, Inc. All rights reserved.

RNA is an ancient component of all living systems, and its catalytic prowess, biological importance and ability to form complex folds have recently come to prominence[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 10 nucleotides long and, in some cases, can 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 noncanonical motifs.

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 (for example, fragment assembly of RNA (FARNA)[5] and discrete molecular dynamics (DMD)[6]) or have required manual atom-level manipulation by expert users (for example, Manip[7]). Recent automated full-atom methods (iFold3D[8] and MC-Sym[9]) have described models of impressive quality, but noncanonical regions appear to be either incorrect[8] or take advantage of sequence similarity with homologs of known structure in the method's training database[9]. With respect to RNA design, rational engineering has yielded versatile sensors and nanostructures[10–12] but has so far been limited to rearrangements of existing sequence modules rather than designing new noncanonical 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 is 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 noncoding 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] (**Supplementary Fig. 3**), we focused on the conformations of noncanonical 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 12-nucleotide motif was 21 s 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 r.m.s. deviation to the experimentally observed structure (**Table 1** and **Supplementary Fig. 4**). These included widely studied RNAs such as the bulged-G motif of the sarcin-ricin loop, the most conserved domain of the signal recognition particle (SRP) RNA,

[1]Departments of Biochemistry and Physics, Stanford University, Stanford, California, USA. [2]Center for Bioinformatics and Department of Molecular Biosciences, The University of Kansas, Lawrence, Kansas, USA. [3]Howard Hughes Medical Institute and University of Washington, Department of Biochemistry, Seattle, Washington, USA. Correspondence should be addressed to R.D. (rhiju@stanford.edu) or D.B. (dabaker@u.washington.edu).

RECEIVED 8 SEPTEMBER 2009; ACCEPTED 12 JANUARY 2010; PUBLISHED ONLINE 28 FEBRUARY 2010; DOI:10.1038/NMETH.1433

the bacterial loop E motif and the kink-turn motif (**Fig. 1a–d**). In nearly all of these cases (11 of 14), the cluster center or lowest energy member recovered all the native noncanonical 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 occurred because of well-known ambiguities in automated pair assignments[15]. Finally, in two more cases with slightly higher r.m.s deviations (**Fig. 1e**), *de novo* models recovered all the noncanonical 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; Online 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 (from assisted model building with energy refinement (AMBER) and chemistry at Harvard Molecular Mechanics (CHARMM) packages) 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 new free energy estimation methods, may eventually provide increased accuracy, albeit at much higher computational expense[16].

For the cases in which the current FARFAR method did not achieve high resolution, we observed symptoms of poor

Table 1 | Attainment of native-like structure by *de novo* FARFAR

	Motif properties		Clustering statistics		Cluster center		Lowest energy cluster member		Lowest r.m.s. deviation sampled (Å)
	Residues	Chains	Cluster rank	Cluster size	R.m.s. deviation (Å)[a]	NWC[b]	R.m.s. deviation (Å)[a]	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 RNase 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 thermophila* 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
SRP 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
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 and its receptor, P4-P6 domain, *Tetrahymena* 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, *Tetrahymena* 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 internal ribosome entry site	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

The lowest-energy 500 of 50,000 refined conformations were clustered with a model-model heavy-atom r.m.s. deviation cutoff of 2.0 Å. The five lowest energy clusters were taken as the *de novo* models; features of the best cluster (lowest r.m.s. deviation to the experimental structure) are listed. See **Supplementary Figure 2** for motif definitions.

[a]Heavy-atom r.m.s. deviation (in Å) to crystal structure. [b]Recovered/total number of non–Watson-Crick (NWC) base pairs. Assignment of base pairing followed an automated method based on the RNAview algorithm; counts of correct base pairings are lowered owing to ambiguities in assigning bifurcated base pairs, pairs connected by single hydrogen bonds or pairs that are not completely co-planar.

© 2010 Nature America, Inc. All rights reserved.

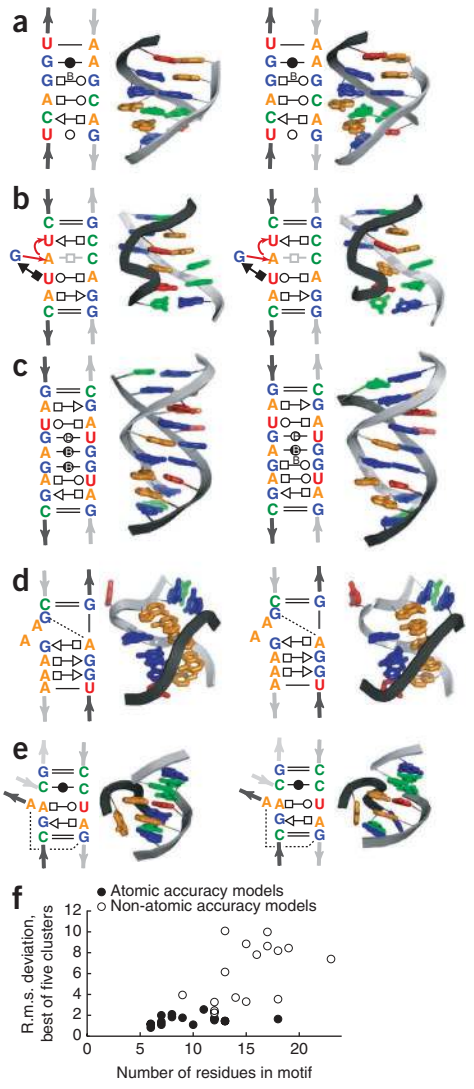


Figure 1 | *De novo* modeling of noncanonical RNA structure with FARFAR. (**a**–**e**) Two-dimensional annotations[15] and three-dimensional representations for the *Escherichia coli* SRP domain IV RNA (**a**), the bulged-G motif from the *E. coli* sarcin-ricin loop (**b**), the *E. coli* loop E motif (**c**), the kink-turn motif from the SAM-I riboswitch (*Thermoanaerobacter tengcongensis*) (**d**) and the hook-turn motif (**e**). (Protein Data Bank (PDB) codes are 1LNT, 1Q9A, 354D, 2GIS and 1MHK, respectively.) Depicted are the experimentally observed structures (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. Symbols and letters are as in ref. 15. (**f**) All-heavy-atom r.m.s. deviation for the best of five final predictions (low-energy cluster centers) plotted against the number of residues in the modeled motif.

conformational sampling: nonconvergence 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 r.m.s. deviation in **Table 1** and energy gaps in **Supplementary Table 1**). 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 15 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, noncanonical 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** and **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 SRP 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 (**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 $MgCl_2$ and 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 (**Fig. 2c**,**d**). Additionally, 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^{-1}, agreed with the predicted value of -1.6 kcal mol^{-1} (see **Supplementary Fig. 5** for energy calibration). Data from tests of the single mutations also agreed with the Rosetta predictions (**Supplementary Fig. 6**). These same two sequence changes previously had been 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 had been made regarding stability.

The power of full-atom refinement demonstrated here, 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 noncanonical RNA motifs. Unsolved problems remain, including prediction of previously unseen RNA motifs in a blinded fashion, incorporation of small-molecule ligands and explicit metal ions, and 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/.

© 2010 Nature America, Inc. All rights reserved.

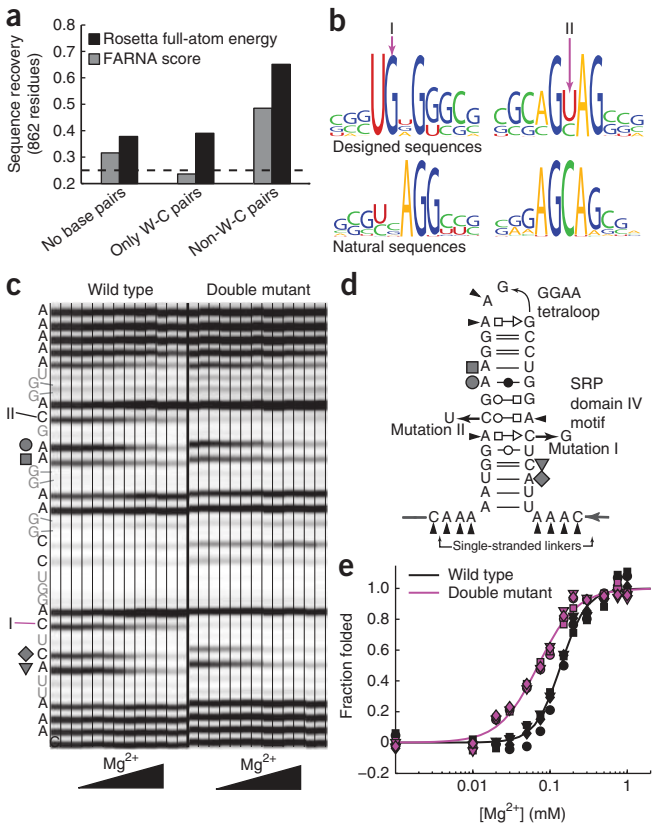


Figure 2 | Computational and experimental tests validating sequence design and thermostabilization. (**a**) Sequence recovery over 15 high-resolution side-chain–stripped RNA structures optimizing the Rosetta full-atom energy compared to random recovery (25%, dashed line) and compared to tests with the FARNA score function. W-C, Watson-Crick base pairs. (**b**) Sequence preference predicted from 1,000 redesigns (top) compared to an alignment of SRP domain IV RNA sequences drawn from all three kingdoms of life[17], in Sequence Logo format[18]. Two mutations (I and II) predicted by the Rosetta redesigns to stabilize folding are indicated. (**c**) Dimethyl sulfate (DMS) modification data probing the structure and thermodynamics of the wild-type SRP motif and a double-mutant variant. Sites of chemical modification were read out by reverse transcription of modified RNA with fluorescently labeled DNA primers, separated by multiplexed capillary electrophoresis. (**d**) Wild-type and mutant sequences as a two-dimensional annotation (see ref. 15) of tertiary contacts. Wedges mark residues that remained accessible to dimethyl sulfate in high Mg^{2+} folding conditions for the wild-type RNA. (**e**) Folding isotherms by Mg^{2+} titration for four separate residues involved in the SRP motif's noncanonical structure (square, circle, diamond and triangle in **c** and **d**). The left-most symbols represent conditions without Mg^{2+}.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank contributors to the current Rosetta codebase, local computer administrators D. Alonso and K. Laidig, the BioX^2 cluster (US 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 (R.D.), the Damon Runyon Cancer Research Foundation (J.K.) and the Howard Hughes Medical Institute (D.B.).

AUTHOR CONTRIBUTIONS

R.D. designed research, implemented the method, analyzed data and prepared the manuscript; J.K. designed research and implemented the method; and D.B. designed research.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Gesteland, R.F., Cech, T.R. & Atkins, J.F. *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2006).
2. Shapiro, B.A., Yingling, Y.G., Kasprzak, W. & Bindewald, E. *Curr. Opin. Struct. Biol.* **17**, 157–165 (2007).
3. Moore, P.B. *Annu. Rev. Biochem.* **68**, 287–300 (1999).
4. Brion, P. & Westhof, E. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 113–137 (1997).
5. Das, R. & Baker, D. *Proc. Natl. Acad. Sci. USA* **104**, 14664–14669 (2007).
6. Ding, F. *et al. RNA* **14**, 1164–1173 (2008).
7. Massire, C. & Westhof, E. *J. Mol. Graph Model.* **16**, 197–205 (1998).
8. Sharma, S., Ding, F. & Dokholyan, N.V. *Bioinformatics* **24**, 1951–1952 (2008).
9. Parisien, M. & Major, F. *Nature* **452**, 51–55 (2008).
10. Breaker, R.R. *Nature* **432**, 838–845 (2004).
11. Win, M.N. & Smolke, C.D. *Proc. Natl. Acad. Sci. USA* **104**, 14283–14288 (2007).
12. Jaeger, L., Westhof, E. & Leontis, N.B. *Nucleic Acids Res.* **29**, 455–463 (2001).
13. Rohl, C.A., Strauss, C.E., Misura, K.M. & Baker, D. *Methods Enzymol.* **383**, 66–93 (2004).
14. Klein, D.J., Schmeing, T.M., Moore, P.B. & Steitz, T.A. *EMBO J.* **20**, 4214–4221 (2001).
15. Leontis, N.B. & Westhof, E. *RNA* **7**, 499–512 (2001).
16. Boas, F.E. & Harbury, P.B. *Curr. Opin. Struct. Biol.* **17**, 199–204 (2007).
17. Larsen, N. & Zwieb, C. *Nucleic Acids Res.* **19**, 209–215 (1991).
18. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. *Genome Res.* **14**, 1188–1190 (2004).

© 2010 Nature America, Inc. All rights reserved.

ONLINE METHODS

Software implementation. All computational methods were implemented in Rosetta 3.1. Full documentation, explicit command lines and example files necessary to model the structure of the most conserved domain of SRP (PDB code 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/manuals/archive/rosetta3.1_user_guide/.

Identification of RNA motifs. An automated algorithm to parse noncanonical segments (that is, 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 filtered to remove sequence-redundant motifs. A final set of 32 sequence motifs and the assumed canonical base pairs (which form 'boundary conditions' for each motif) are illustrated in **Supplementary Figure 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 (**Supplementary Fig. 7**). 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 had 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, for example, ref. 19). 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 (**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[20,21]; every 10 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; ref. 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, for example, Z scores (see below).

We optimized 50,000 FARNA models 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 nonbonded 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 nonbonded terms (**Supplementary Fig. 8**) were incorporated here and reweighted through an iterative calibration: (i) a potential for weak carbon hydrogen bonds, previously investigated for membrane proteins, (ii) an alternative orientation-dependent model for desolvation based on occlusion of protein moieties, and (iii) 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[22], 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 in ref. 23) was applied to the models that had been refined with the full-atom Rosetta energy function. To test the CHARMM27 force field, the chemistry at Harvard Molecular Mechanics (CHARMM) program[24] was applied, using the nucleic acid force field (PARAM27)[25]. The CHARMM generalized Born molecular volume (GBMV) method[26,27] was used as an implicit representation of the solvent. Default parameters for minimization and GBMV were taken from the multiscale modeling tools for structural biology (MMTSB) tool set[28]. Current molecular mechanics packages do not offer the prospect of continuous minimization of model coordinates in the context of the computationally expensive nonlinear 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[29], 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 disagreed with the authoritative manual annotation[15]

© 2010 Nature America, Inc. All rights reserved.

by giving different interacting edges to a central bifurcated G-G base pair and assigning an extra hydrogen bond between two (nonplanar) C residues (**Supplementary Fig. 2**). **Figure 1** shows the manual annotation.

Iterative optimization of weights of the energy function. Half of the 32 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 nonnative decoys minus the mean score of the 10 lowest-energy near-native models, divided by the s.d. of nonnative decoy scores. In this computation, nonnative decoys with anomalously poor scores (higher than three s.d. 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 versus 3.28; number of cases with positive energy gaps of 10 versus 8; median r.m.s. deviation for best of five clusters of 2.28 Å versus 2.34 Å; and recovery of non–Watson-Crick base pairs of 43% versus 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. and 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 precomputation of all rotamer-rotamer pairwise energies. The nucleobase rotamers were constructed with the glycosidic torsion angle χ set at its most probable anti value and at −1, −1/2, +1/2 and +1 s.d. from this central value. The central value and s.d. 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 χ 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 5′-GGCUACGCAAGUAA AACAAAUUACUCAGGUCCGGAAGGAAGCAGGUAAAAACCA AACCAAAGAAACAACAACAACAAC-3′ (the last 20 nucleotides form the primer binding site), or with the mutations discussed in the main text. DNA templates including the 20 nucleotide T7 primer sequence (5′-TTCTAATACGACTCACTATA-3′) were prepared by extension (Phusion; Finnzymes) of 60-nucleotide sequences (Integrated DNA Technologies), purified on Qiaquick columns (Qiagen) and used as templates for *in vitro* transcription with T7 polymerase (New England Biolabs). RNA was purified by phenol and chloroform extraction and buffer-exchanged into deionized water with P30 RNAse-free spin columns (BioRad). 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 min, dimethyl sulfate (freshly diluted in water) was added to a final concentration of 0.25% (vol/vol) 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 min of modification, reactions were quenched with 0.25 volumes of 2-mercaptoethanol, oligo-dT beads (poly(A) purist; Ambion) and 5′-rhodamine-green labeled primer (5′-AAAAAAAAAAAAAAAA AAAAGTTGTTGTTGTTTCTTT-3′, 0.125 pmol), and purified by magnetic separation. Reverse transcriptase reactions were carried out using Superscript III (Invitrogen) 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[30]. Plots and fits of fraction folded were carried out in Matlab (MathWorks), 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 [$MgCl_2$].

19. Baeyens, K.J., De Bondt, H.L., Pardi, A. & Holbrook, S.R. *Proc. Natl. Acad. Sci. USA* **93**, 12851–12855 (1996).
20. Bradley, P. & Baker, D. *Proteins* **65**, 922–929 (2006).
21. Das, R. & Baker, D. *Annu. Rev. Biochem.* **77**, 363–382 (2008).
22. Draper, D.E., Grilley, D. & Soto, A.M. *Annu. Rev. Biophys. Biomol. Struct.* **34**, 221–243 (2005).
23. Qiu, D., Shenkin, P.S., Hollinger, F.P. & Still, W.C. *J. Phys. Chem. A* **101**, 3005–3014 (1997).
24. Brooks, B.R. *et al. J. Comput. Chem.* **4**, 187–217 (1983).
25. MacKerell, A.D.J. *et al. J. Phys. Chem. B* **102**, 3586 (1998).
26. Lee, M.S., Salsbury, F.R.J. & Brooks, C.L.I. *J. Chem. Phys.* **116**, 10606 (2002).
27. Lee, M., Feig, M., Salsbury, F.J. & Brooks, C.R. *J. Comput. Chem.* **24**, 1348–1356 (2003).
28. Feig, M., Karanicolas, J. & Brooks, C. *J. Mol. Graph. Model.* **222**, 377–395 (2004).
29. Yang, H. *et al. Nucleic Acids Res.* **31**, 3450–3460 (2003).
30. Das, R., Laederach, A., Pearlman, S.M., Herschlag, D. & Altman, R.B. *RNA* **11**, 344–354 (2005).

© 2010 Nature America, Inc. All rights reserved.

doi:10.1038/nmeth.1433

nature | methods

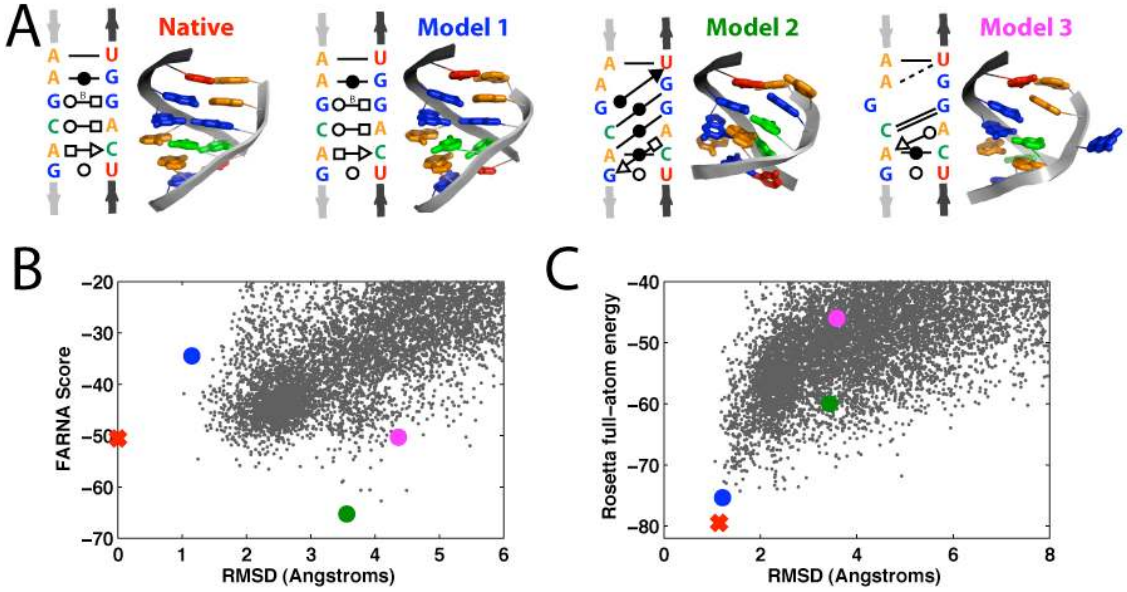
Atomic accuracy in predicting and designing noncanonical RNA structure

Rhiju Das, John Karanicolas & David Baker

Supplementary figures and text:

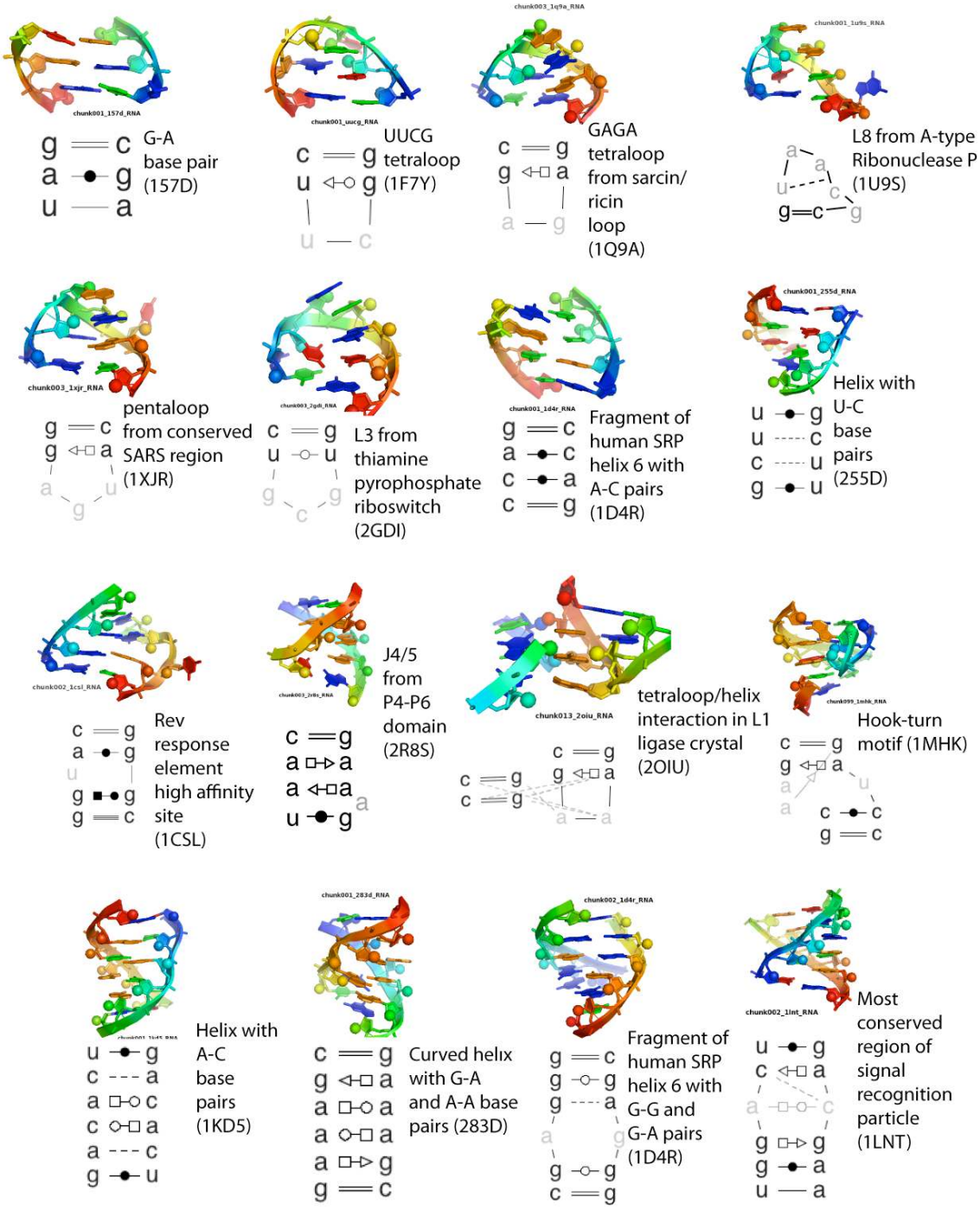
Supplementary Figure 1	Full-atom refinement for near-atomic accuracy
Supplementary Figure 2.	Tertiary structure and secondary structure representations of the 32 motifs used to test *de novo* structure prediction.
Supplementary Figure 3.	Example of structure modeling for a molecule including both canonical and non-canonical base pairs.
Supplementary Figure 4.	Rosetta full-atom energy versus RMSD to the experimental motif structure
Supplementary Figure 5.	Structure prediction and thermodynamic comparisons for folding 32 RNA duplexes.
Supplementary Figure 6.	Structural and thermodynamic characterization of the SRP Domain IV motif and variants predicted to thermostabilize its folding.
Supplementary Figure 7.	Additional terms added to the FARNA low-resolution score function.
Supplementary Figure 8.	Terms added to the previously published full-atom Rosetta energy function
Supplementary Figure 9.	Histograms of RNA torsion values derived from the refined crystal structure of the large ribosomal subunit (PDB ID: 1JJ2).
Supplementary Table 1.	Discrimination of native-like (< 2.0 Å heavy-atom rmsd) from non-native (> 3 Å) models.
Supplementary Table 2.	Recovery of native sequences through redesign of base conformation and identity with two force-fields.
Supplementary Table 3.	The quality of refined *de novo* models compares well to experimental structures.

Supplementary Figure 1. Full-atom refinement for near-atomic accuracy, illustrated with the signal recognition particle (SRP) Domain IV RNA. (A) Two-dimensional Leontis/Westhof annotation and crystal structure of the isolated SRP motif (PDB: 1LNT), here shown with two bounding canonical base pairs, compared to three models generated by fragment assembly of RNA (FARNA). Adenosine, cytidine, guanosine, and uracil bases are shown in orange, green, blue, and red, respectively. Three-dimensional representations were prepared in PyMol (Delano Scientific). (B) The knowledge-based FARNA score incorrectly ranks the native structure (red) and near-native Model 1 (blue) as worse in score compared to non-native Models 2 (green) and 3 (magenta). (C) Full-atom refinement with the Rosetta full-atom energy function corrects the ranking, enabling near-atomic resolution *de novo* modeling for this motif.

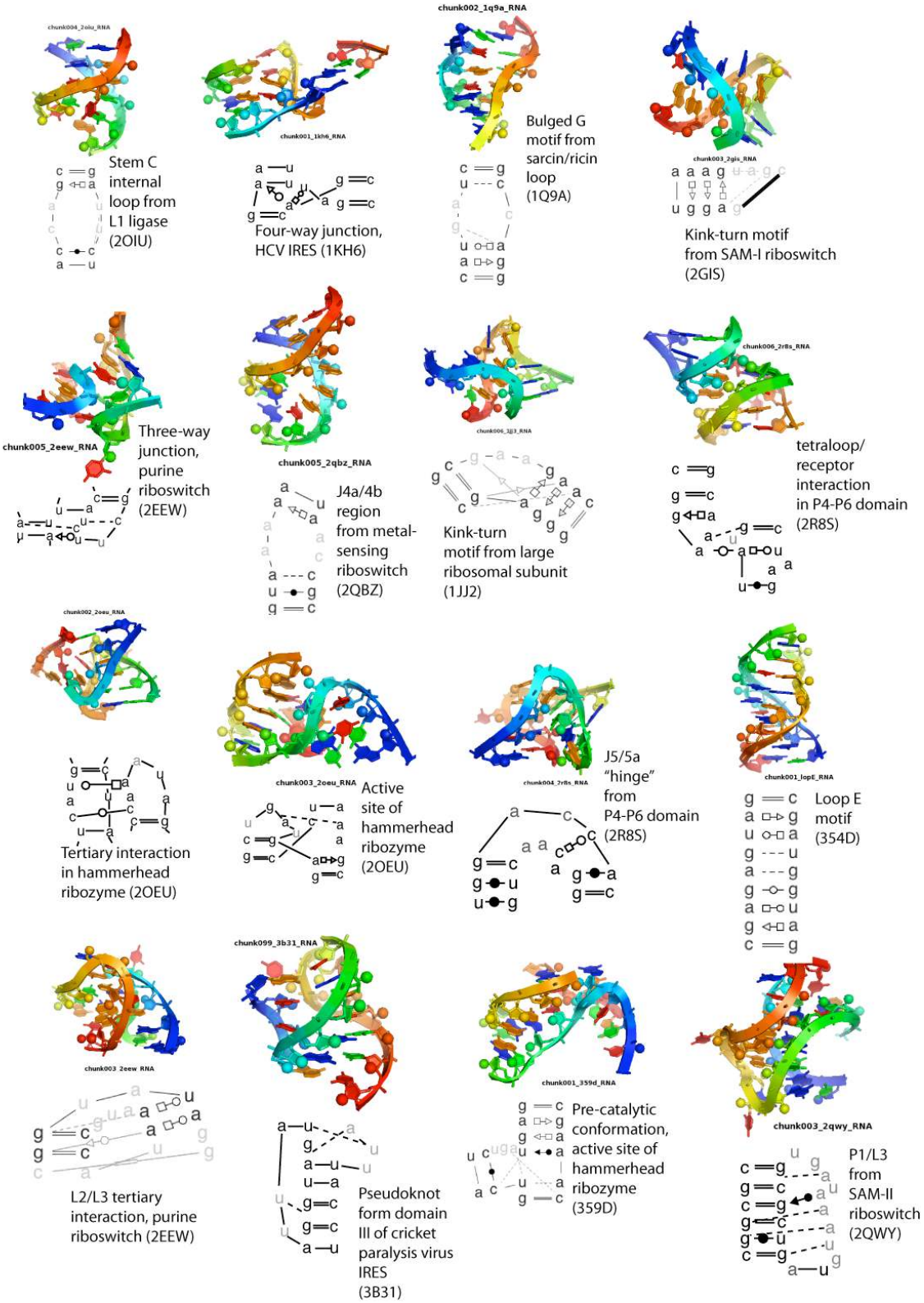


Nature Methods: doi:10.1038/nmeth.1433

Supplementary Figure 2. Tertiary structure and secondary structure representations of the 32 motifs used to test *de novo* structure prediction. Secondary structure representations follow the convention of Leontis and Westhof [RNA (2001):7, 499-512] and were prepared either automatically in RNAmlView [Yang et al., *NAR* (2003) 31: 3450-60] or manually in cases that could not be parsed or clearly represented in RNAmlView. Note that the secondary structure diagrams are not always in the same orientation as the tertiary structure graphics. Watson-Crick base pairs (assumed to be known a priori) connect separate chains of each motif, except for tetraloop/receptor interactions, in which all possible inter-chain base-pairs were tested.

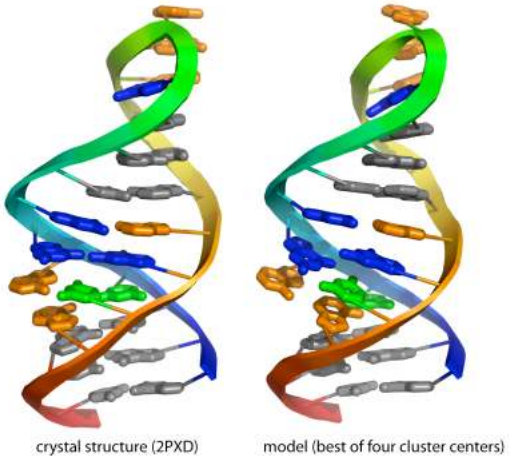


Nature Methods: doi:10.1038/nmeth.1433



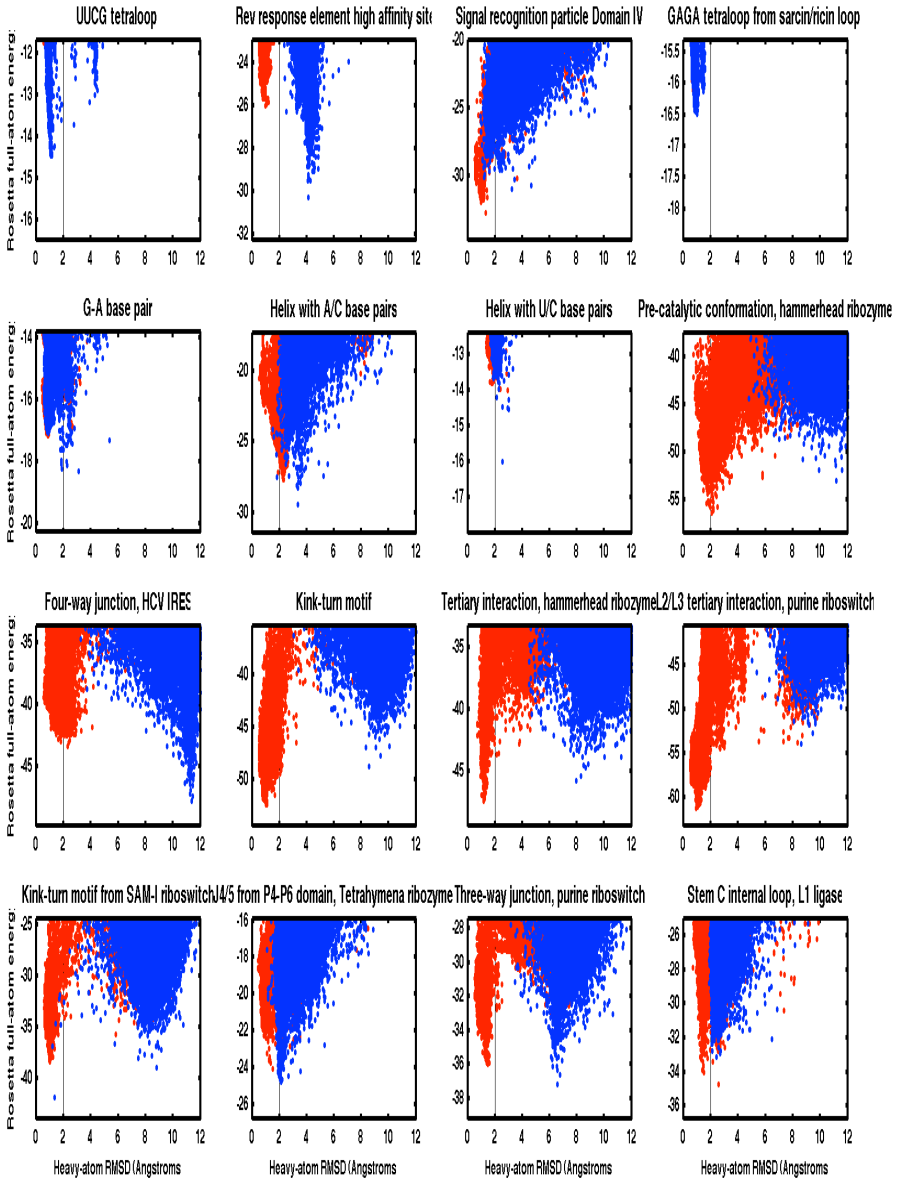
Nature Methods: doi:10.1038/nmeth.1433

Supplementary Figure 3. Example of structure modeling for a molecule including both canonical and non-canonical base pairs. A 24-nucleotide-long domain of the *E. coli* signal recognition particle RNA (*2PXD,* left) is recapitulated at atomic resolution (1.6 Å all-heavy-atom RMSD) by one of four cluster centers (right). After separate FARFAR runs to generated 2,000-model ensembles of two double helical stems (gray) and two non-canonical motifs (colored bases), the four separate regions were assembled in the Rosetta structure modeling framework.

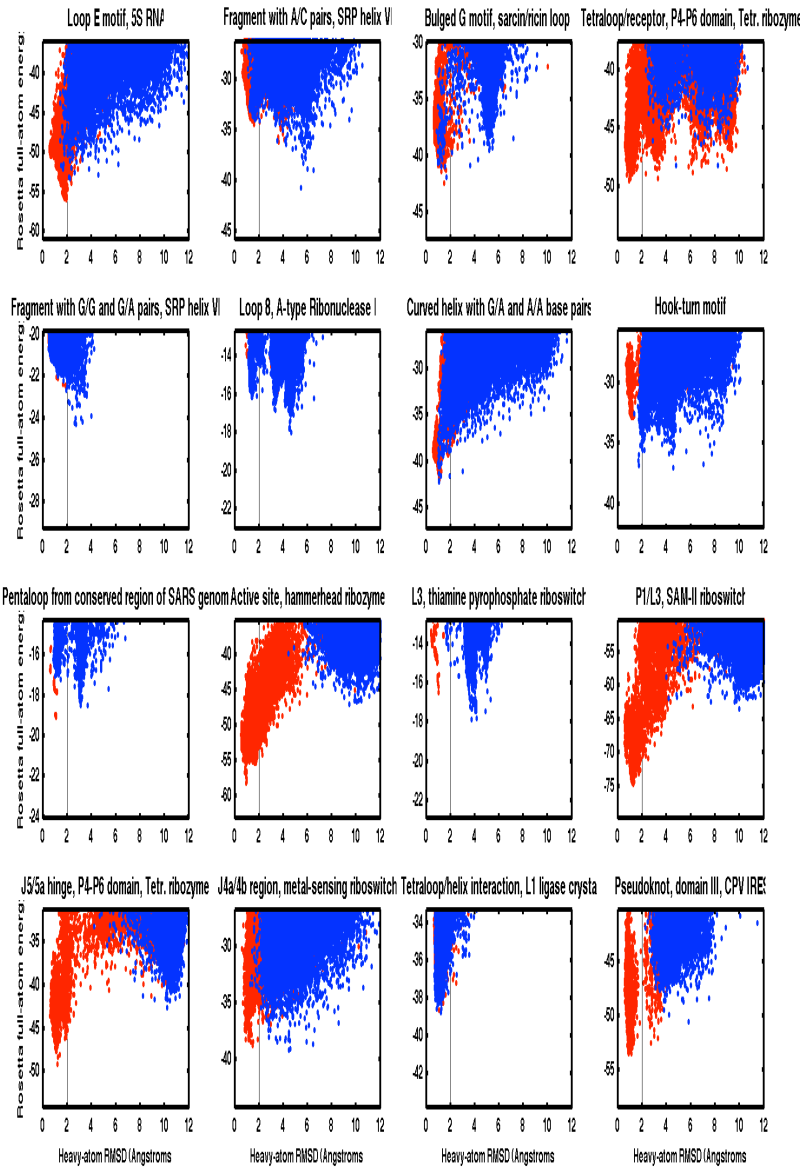


Nature Methods: doi:10.1038/nmeth.1433

Supplementary Figure 4. Rosetta full-atom energy vs. RMSD to the experimental motif structure for 50,000 *de novo* models (blue) compared to models constructed only with native fragments (red). Motif definitions are given in Supplementary Fig. 2.

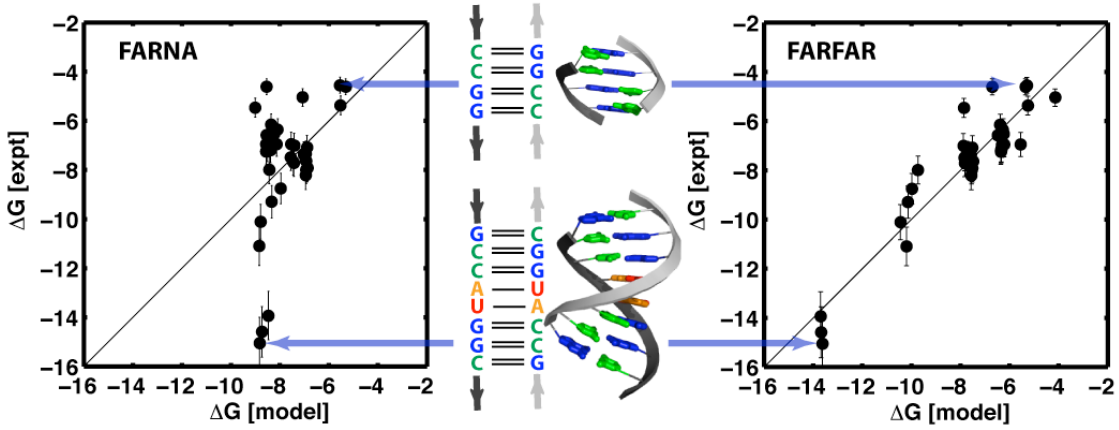


Nature Methods: doi:10.1038/nmeth.1433



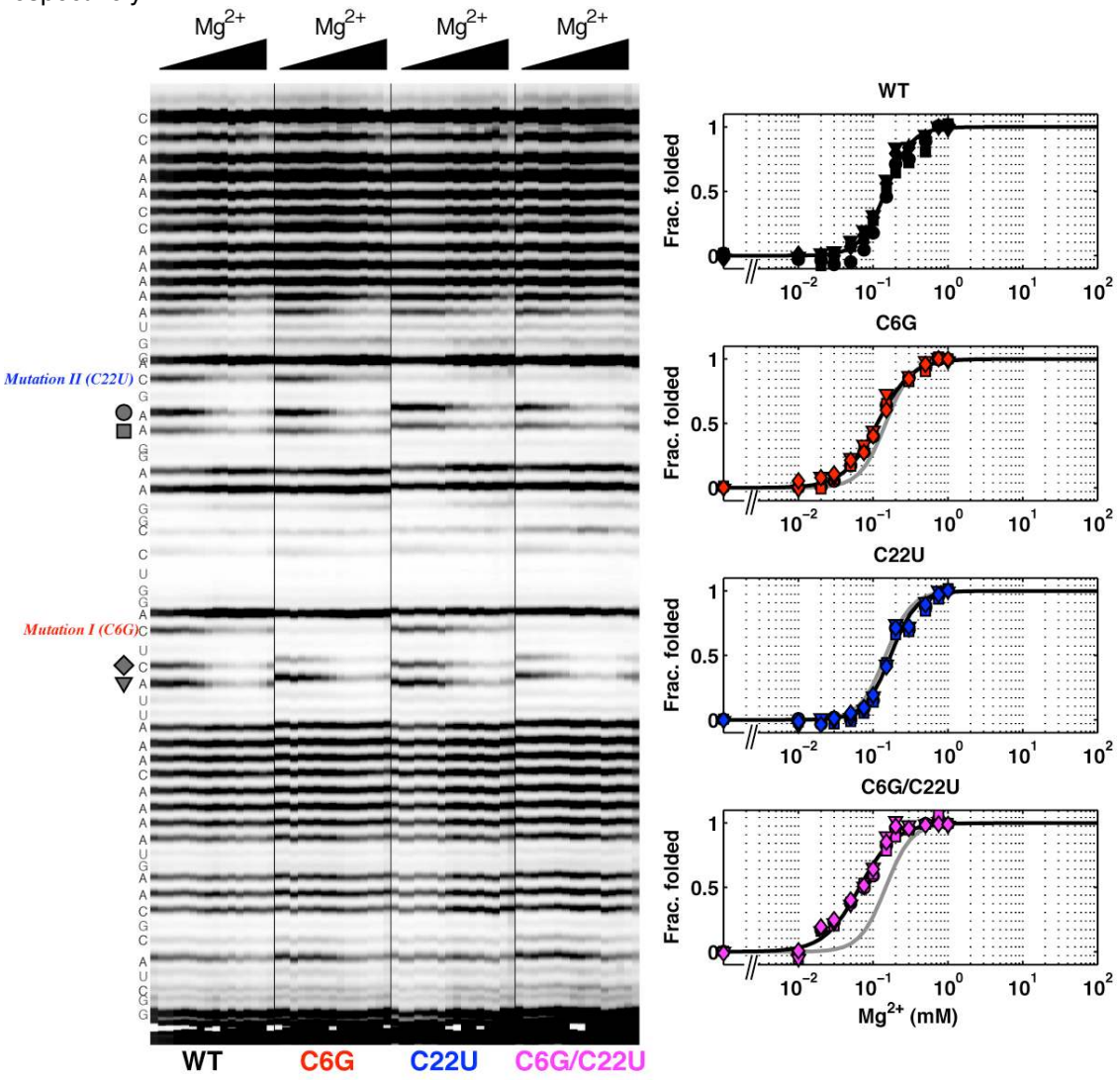
Nature Methods: doi:10.1038/nmeth.1433

Supplementary Figure 5. Structure prediction and thermodynamic comparisons for folding 32 RNA duplexes. An extensive thermodynamic data set is available from the classic work of Turner and colleagues on RNA duplexes with lengths between 4 and 9 base pairs [Xia et al. (1998), *Biochemistry* 37: 14719.], and we have generated FARFAR structures for these sequences. The calculated full-atom energies for these molecules cannot be directly compared to experimental free energies; the Rosetta energy scale is *a priori* unknown and two additional parameters, the free energies of entropy loss upon strand association and upon forming each base pair, are difficult to model. Nevertheless, allowing these three unknown parameters to be varied yields reasonable agreement between Rosetta full-atom energies and measured values (right panel). Correlation of predicted duplex energies and energies measured by optical melting (at reference temperature 37 °C and strand concentrations of 0.1 mM) are shown. The mean unsigned error of the fit is 0.7 kcal/mol, providing an estimate of the error incurred by approximations inherent to the Rosetta full-atom energy function. In contrast, the knowledge-based score function FARNA gives a dramatically worse fit (left panel), with an error of 1.6 kcal/mol. Example FARFAR structures for the least and most stable sequences are shown in the center. Fit parameters for the energy scale factor, entropy loss on strand association, and entropy loss per base pair were 0.06, –2.59, and 0.09 kcal/mol (FARNA); and 0.43, 4.1, and 4.9 kcal/mol (FARFAR).



Nature Methods: doi:10.1038/nmeth.1433

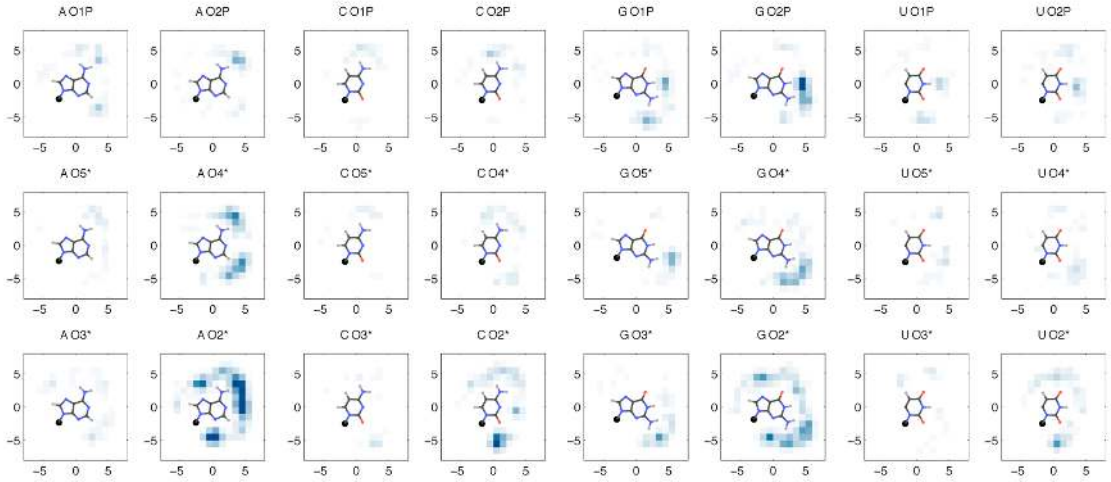
Supplementary Figure 6. Structural and thermodynamic characterization of the SRP Domain IV motif and variants predicted to thermostabilize its folding. (A) Fluorescent time traces display the reverse-transcripase readout of dimethyl sulfate (DMS) modification of each construct, resolved by capillary electrophoresis. Modification was carried out in 0.25% DMS, 50 mM Na-HEPES, pH 8.0, at 44 °C with increasing amounts of $MgCl_2$ to promote tertiary folding. Protection patterns for DMS (as well as SHAPE and CMCT experiments; not shown) indicate that in the Mg^{2+}-free conditions, only a GAAA tetraloop with G-C base pairs are formed; addition of Mg^{2+} gives the fully expected secondary and tertiary structure of each construct (see also Fig. 2D, main text). (B) After quantitation of the time traces by peak-fitting, folding isotherms at four residues involved in the SRP motif's noncanonical structure overlay well (to aid visual comparison, the wild type isotherm fit is shown as a gray curve on each plot). Fitting parameters for the isotherms are: n = 2.5±0.3, K = 0.149±0.011 mM (WT); n = 1.9±0.1, K = 0.115±0.003 mM (C6G); n = 2.7±0.1, K = 0.168±0.012 mM (C22U); n = 1.9±0.2, K = 0.064±0.007 mM (C6G/C22U). The experimentally measured free energy changes ($\Delta\Delta G$) are calculated to be –0.37±0.19 kcal/mol (C6G), 0.20±0.20 kcal/mol (C22U), and –1.22±0.49 kcal/mol (C6G/C22U). [Repeat measurements with the RNA diluted by 5-fold during DMS modification gave agreeing $\Delta\Delta G$ values of –0.32±0.016 kcal/mol, 0.47±0.25 kcal/mol, and –1.39±0.73.] The measurements are in agreement with the Rosetta predictions of –1.6 kcal/mol, –0.1 kcal/mol, and –1.7 kcal/mol, respectively.



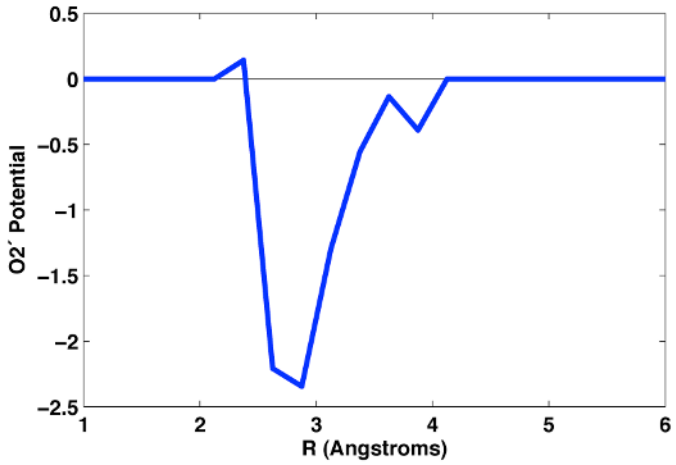
Nature Methods: doi:10.1038/nmeth.1433

Supplementary Figure 7. Additional terms added to the FARNA low-resolution score function. (A) The original FARNA score function [Das and Baker, *PNAS* (2007): 104:14664-14669] did not include specific interactions between bases and backbone oxygen atoms mediated by hydrogen bonds. The panels give the frequency of occurrence of each backbone oxygen atom lying within 2.5 Å of the base plane in the large ribosomal subunit crystal structure. A potential proportional to the logarithm of these frequencies has been added to the FARNA score function. [Note that the counts have been convolved with a two-dimensional Gaussian function with width 0.5 Å to create a smooth potential appropriate for coarse-grained fragment assembly.] (B) A new potential, applied between O2′ groups, was derived from the logarithm of the ratio of the number of backbone oxygen interactions in low resolution FARNA models vs. crystal structures of fragments of the large ribosomal subunit. The same potential is applied at half strength between O2′ groups and the O1P and O2P oxygen acceptors of phosphates, again based on their absence in FARNA models compared to experimental structures.

A

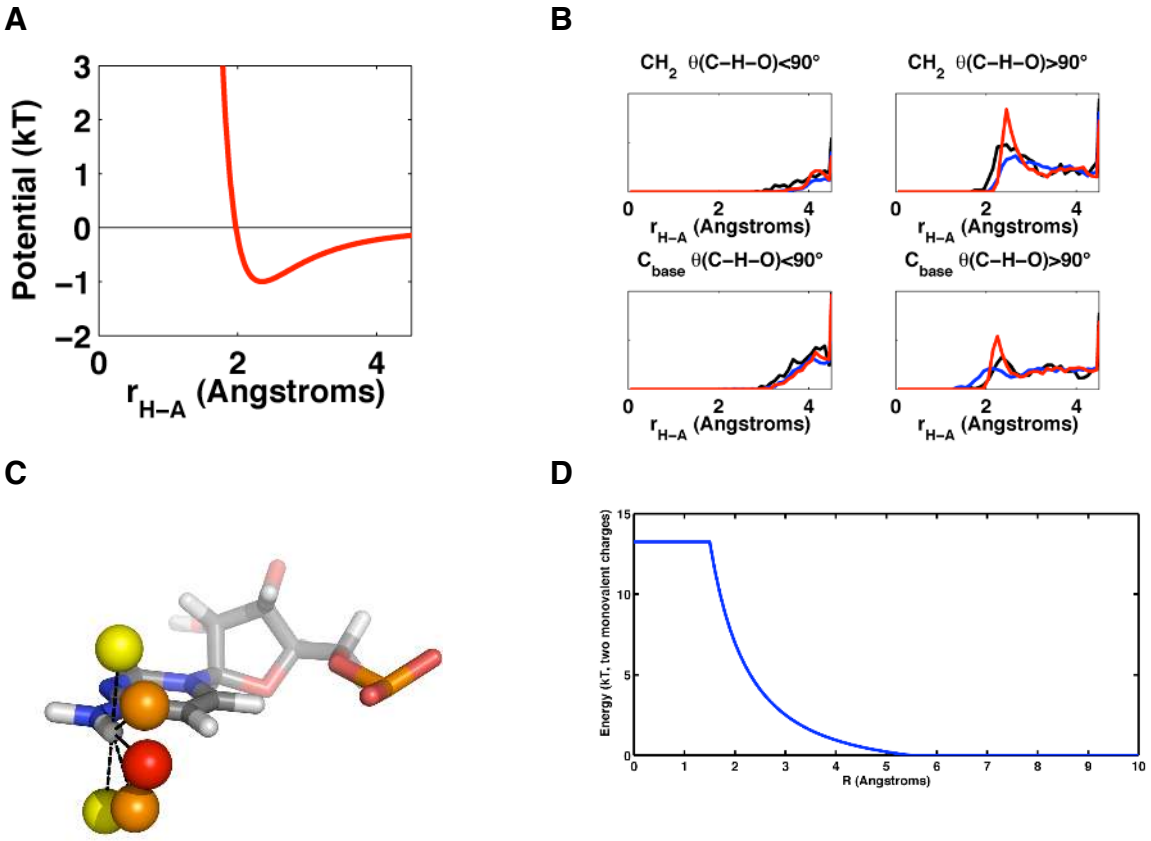


B



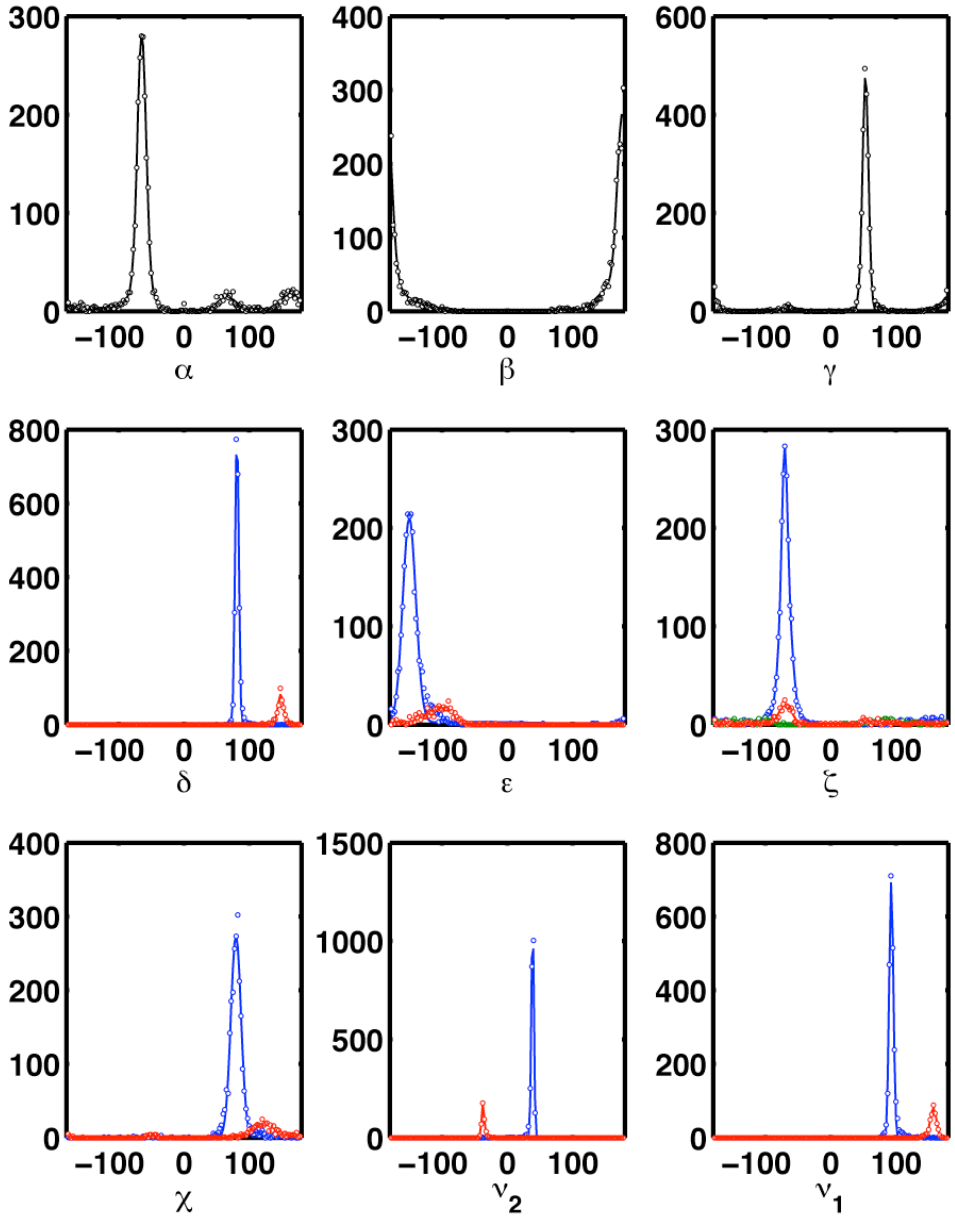
Nature Methods: doi:10.1038/nmeth.1433

Supplementary Figure 8. Terms added to the previously published full-atom Rosetta energy function [Rohl, Strauss, Misura, & Baker, Methods Enzymol. 383: 66-93]. A potential for weak carbon hydrogen bonds (CH…O) is implemented for C-H-O angles less than 60°, with the shown distance dependence (A); its implementation improves distance distributions observed in Rosetta models (cf. blue and red lines in B) to better match distributions in the crystal structure of the archaeal large ribosomal subunit (black line in B). (C) A newly tested solvation potential for polar atoms estimates the desolvation cost as the sum of hydrogen bond energies to virtual water molecules placed at the locations of each occluding atom, calculated as per Kortemme et al. [*JMB* (2003) 326: 1239-1259]. Spheres demonstrate change in solvation penalty (from yellow to red) as the location of an occluding atom changes around an N-H donor. In the previous angle-independent solvation potential of Laziridis and Karplus [*Proteins* (1999) 35: 133-52], each of the spheres would incur the same desolvation penalty from the N-H donor. (D) The functional form of a simple electrostatic model to prevent close approach of phosphates is assumed to follow Coulomb's law with a distance-dependent dielectric [$\varepsilon = 10\ (r/\text{Å})$], vanishing at 5.5 Å (the maximum distance of interaction in Rosetta) and saturating below 1.5 Å (well below the van-der-Waal radii of full-atoms). Partial charges for the P, O1P, O2P, O5´, and O3´ atoms follow parametrization used in CHARMM and for Rosetta DNA applications [e.g., Ashworth et al. *Nature* (2006) 441: 656-9]. The steep drop in the repulsion beyond ~3 Å is consistent with expected behavior at high counterion concentrations (e.g., > 1 M NaCl) in which the Debye length drops below this distance scale.



Nature Methods: doi:10.1038/nmeth.1433

Supplementary Figure 9. Histograms of RNA torsion values derived from the refined crystal structure of the large ribosomal subunit (PDB ID: 1JJ2). In each case, fits are to the sums of two to three Gaussians. For δ, ε, χ, and the sugar torsions ν_1, and ν_2, separate fits have been carried out for 3´-endo (blue) and 2´-endo (red) sugar puckers. For ζ, separate fits have been carried out depending on whether α of the following residue is *gauche*$^-$ (blue), *trans* (red), or *gauche*$^+$ (green). The torsional potential assumed in the full-atom Rosetta energy function is proportional to the logarithm of the fitted histogram. No torsional potential was assumed for additional sugar torsions ν_0, ν_3, and ν_4, as they are constrained by maintaining ideal bond lengths and angles in the sugar ring.



Nature Methods: doi:10.1038/nmeth.1433

Supplementary Table 1. Discrimination of native-like (< 2.0 Å heavy-atom rmsd) from non-native (> 3 Å) models. For 32 RNA motifs, 2000 *de novo* models and 2000 native-like models were generated and refined in four force fields. The Z-score is defined as the difference between the mean non-native model score and the lowest native model score, divided by the standard deviation of the non-native model scores. The energy gap ("Egap") is defined as the difference (in kcal/mol) between the lowest scoring non-native model and the lowest scoring native model. Positive scores indicate successful discrimination. See Supplementary Fig. 2 for motif definitions. For four motifs, models with > 3 Å rmsd from the crystal structure were not generated by FARNA and were not included in the assessment of Z-scores or energy gaps.

RNA description	FARNA	FARNA refined	AMBER 99*	CHARM M27*	FAR-FAR	FARNA	FARNA refined	AMBER 99*	CHARM M27*	FAR-FAR
Training set										
UUCG tetraloop	2.89	1.16	5.24	2.42	3.11	0.54	0.29	11.25	-0.17	1.71
Rev response element high affinity site	-0.46	-0.77	1.00	2.74	1.17	-2.91	-4.72	-7.25	3.96	-2.61
J4/5 from P4-P6 domain, Tetrahymena ribozyme	2.39	2.01	2.77	4.65	3.28	0.20	-1.83	1.62	14.56	1.24
Helix with A/C base pairs	0.90	1.56	0.80	1.47	2.26	-0.71	-0.71	-16.43	3.13	0.93
Signal recognition particle Domain IV	1.30	2.37	1.98	3.51	4.48	-1.33	0.67	-1.92	24.05	3.35
Stem C internal loop, L1 ligase	1.21	1.23	1.06	0.74	2.75	-0.82	0.17	1.91	5.20	2.23
Four-way junction, HCV IRES	2.00	1.69	0.16	2.29	2.98	3.49	-2.38	-9.37	-11.58	0.20
Kink-turn motif from SAM-I riboswitch	3.19	4.33	1.18	0.82	3.44	3.28	2.65	0.00	-26.13	1.05
Three-way junction, purine riboswitch	2.03	2.87	1.16	1.94	3.58	1.06	0.51	0.00	-10.65	2.26
Kink-turn motif	2.27	4.90	3.59	2.84	4.74	3.95	5.00	13.25	12.80	7.39
Tertiary interaction, hammerhead ribozyme	1.87	2.38	3.28	3.32	3.35	3.03	2.43	-57.37	15.80	2.21
L2/L3 tertiary interaction, purine riboswitch	2.42	3.42	3.03	2.97	4.53	7.76	3.43	0.52	-8.73	10.01
Pre-catalytic conformation, hammerhead ribozyme	2.57	2.78	0.89	2.25	2.98	3.21	4.78	-10.09	3.66	3.60
MEAN Z-score	1.89	2.30	2.01	2.46	3.28					
Egap > 0 kcal						9	9	5	8	12
Test set										
Loop 8, A-type Ribonuclease P	0.33	1.07	0.46	1.89	1.31	-1.19	-1.21	-2.54	2.35	-0.84
Pentaloop from conserved region of SARS genome	2.62	2.10	1.70	1.49	3.42	-0.68	-0.37	-1.69	-6.09	0.98
L3, thiamine pyrophosphate riboswitch	0.00	0.18	1.18	2.17	2.98	-1.51	-1.07	-2.97	-0.58	-1.10
Tetraloop/helix interaction, L1 ligase crystal	3.94	2.53	1.85	2.77	3.13	2.61	3.61	7.07	19.16	5.04
Hook-turn motif	2.49	1.56	2.95	1.87	2.20	0.33	-1.08	2.65	3.87	-0.28
Curved helix with G/A and A/A base pairs	3.14	1.59	1.06	3.70	4.43	0.52	1.57	-10.31	19.63	6.43
Fragment with G/G and G/A pairs, SRP helix VI	0.46	1.42	0.40	2.41	2.89	-2.26	-2.20	-4.10	4.38	1.81
Bulged G motif, sarcin/ricin loop	1.92	1.48	1.82	2.51	3.03	-1.83	-0.15	-0.88	8.57	1.23
J4a/4b region, metal-sensing riboswitch	1.50	1.56	0.87	2.52	3.14	-0.43	0.07	9.17	2.47	1.31
Tetraloop/receptor, P4-P6 domain, Tetr. ribozyme	2.36	2.61	2.45	2.23	2.81	-0.72	0.94	1.85	-1.71	3.26
Active site, hammerhead ribozyme	3.25	3.07	3.22	3.55	5.74	3.50	5.47	2.00	17.05	14.20
J5/5a hinge, P4-P6 domain, Tetr. ribozyme	2.11	1.46	1.46	0.48	4.66	-1.50	0.40	-7.49	-33.44	8.20
Loop E motif, 5S RNA	2.29	1.42	2.09	2.96	3.63	1.69	1.51	-2.43	20.31	5.90
Pseudoknot, domain III, CPV IRES	5.58	2.45	5.01	3.61	5.55	4.26	0.59	20.30	16.64	5.55
P1/L3, SAM-II riboswitch	3.05	1.98	3.00	1.17	5.21	6.90	7.69	30.86	-38.99	9.22
MEAN Z-score	2.34	1.76	1.97	2.36	3.61					
Egap > 0 kcal						7	9	7	10	12
Whole set										
MEAN Z-score	2.13	2.01	1.99	2.40	3.46					
Egap > 0 kcal						16	18	12	18	24

·Molecular mechanics calculations were carried out without taking into account counterion screening, due to the difficulty of calculating free energies with explicit modeling of discrete ions and to the present unavailability of accurate implicit treatments of ions [e.g., the Poisson-Boltzmann (PB) equation] during minimization. CHARMM27-minimized decoys were re-scored with the PB solvation model; native state discrimination was significantly worse (mean Z scores of –0.4,–0.6, and –0.6) for monovalent salt concentrations of 0 M, 0.01 M and 1 M. Rescaling of the PB weight did not improve discrimination over the GB-based calculation.

Nature Methods: doi:10.1038/nmeth.1433

Supplementary Table 2. Recovery of native sequences through redesign of base conformation and identity with two force-fields; backbones are taken from high-resolution RNA crystal structures (PDB ID is given).

	Residues forming no base pairs		Residues forming only Watson/Crick base pairs		Residues forming non-canonical base pairs		All residues	
	Number residues	Fraction recovered	Number residues	Fraction recovered	Number residues	Fraction recovered	Number residues	Fraction recovered
1CSL	4	0.25	18	0.15	6	0.31	28	0.20
1ET4	11	0.47	15	0.10	9	0.60	35	0.34
1F27	7	0.44	21	0.32	2	0.50	30	0.36
1L2X	6	0.44	11	0.49	10	0.41	27	0.45
1LNT	0	0.00	16	0.23	8	0.34	24	0.27
1Q9A	4	0.36	12	0.22	11	0.61	27	0.40
1U8D	6	0.38	40	0.17	21	0.45	67	0.28
1X9C	4	0.16	35	0.21	21	0.50	60	0.31
1XPE	12	0.33	80	0.23	0	0.00	92	0.24
2GCS	22	0.17	85	0.22	35	0.43	142	0.26
2GDI	15	0.26	37	0.33	26	0.53	78	0.38
2OEU	11	0.46	40	0.20	10	0.48	61	0.30
2R8S	29	0.34	87	0.28	43	0.55	159	0.36
354D	2	0.07	10	0.23	12	0.51	24	0.35
UUCG*	2	0.00	4	0.06	2	0.05	8	0.05
Overall	135	0.32	511	0.24	216	0.49	862	0.31
Low resolution (FARNA)								
1CSL	4	0.50	18	0.35	6	0.66	28	0.44
1ET4	11	0.47	15	0.41	9	0.43	35	0.44
1F27	7	0.45	21	0.30	2	0.82	30	0.37
1L2X	6	0.33	11	0.39	10	0.66	27	0.48
1LNT	0	0.00	16	0.49	8	0.63	24	0.53
1Q9A	4	0.59	12	0.22	11	0.54	27	0.41
1U8D	6	0.28	40	0.33	21	0.44	67	0.36
1X9C	4	0.31	35	0.41	21	0.67	60	0.50
1XPE	12	0.17	80	0.43	0	0.00	92	0.40
2GCS	22	0.34	85	0.39	35	0.64	142	0.44
2GDI	15	0.35	37	0.37	26	0.70	78	0.48
2OEU	11	0.48	40	0.28	10	0.57	61	0.37
2R8S	29	0.40	87	0.47	43	0.74	159	0.53
354D	2	0.00	10	0.33	12	0.83	24	0.55
UUCG*	2	1.00	4	0.23	2	1.00	8	0.62
OVERALL	135	0.38	511	0.39	216	0.65	862	0.45
High resolution (Rosetta)								

* UUCG tetraloop drawn from 1F7Y.

Nature Methods: doi:10.1038/nmeth.1433

Supplementary Table 3. The quality of refined *de novo* models compares well to experimental structures. For the 32 tested RNA motifs, the listed features are compared between the experimental crystallographic models and 10 low-resolution models (best scoring based on FARNA score) and 10 full-atom-refined models (best scoring by Rosetta energy). Base interaction features are calculated within Rosetta; other features use the automated tools developed by the Richardsons and colleagues [Davis et al. NAR (2007): 35, W375-83]. For the bond-angle and bond-length outlier tests, only a web server tool is available, and so only the lowest scoring model for each RNA motif was assayed.

	Crystal structures	**FARNA models (low res.)**	**FARFAR models (refined)**
Base interaction features (normalized to # residues)			
Watson/Crick base pairs	0.21	0.24	0.23
Non-Watson/Crick pairs	0.23	0.22	0.21
Base stacks	0.74	0.75	0.67
MolProbity metrics			
Clash score	17.16	76.69	18.74
Frac. res., bond length outliers	0.04	0.00	0.00
Frac. res., bond angle outliers	0.18	0.09	0.00
Suiteness			
Fraction triaged rotamers	0.088	0.006	0.086
Fraction outlier rotamers	0.061	0.061	0.090
Suiteness of assigned rotamers	0.457	0.554	0.316

Nature Methods: doi:10.1038/nmeth.1433