An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling

Parin Sripakdeevong[a], Wipapat Kladwang[b], and Rhiju Das[a,b,c,1]

[a]Biophysics Program, Stanford University, Stanford, CA 94305; [b]Department of Biochemistry, Stanford University, Stanford, CA 94305; and [c]Department of Physics, Stanford University, Stanford, CA 94305

Edited by David Baker, University of Washington, Seattle, WA, and approved September 27, 2011 (received for review May 17, 2011)

Atomic-accuracy structure prediction of macromolecules should be achievable by optimizing a physically realistic energy function but is presently precluded by incomplete sampling of a biopolymer's many degrees of freedom. We present herein a working hypothesis, called the "stepwise ansatz," for recursively constructing well-packed atomic-detail models in small steps, enumerating several million conformations for each monomer, and covering all build-up paths. By making use of high-performance computing and the Rosetta framework, we provide first tests of this hypothesis on a benchmark of 15 RNA loop-modeling problems drawn from riboswitches, ribozymes, and the ribosome, including 10 cases that are not solvable by current knowledge-based modeling approaches. For each loop problem, this deterministic stepwise assembly method either reaches atomic accuracy or exposes flaws in Rosetta's all-atom energy function, indicating the resolution of the conformational sampling bottleneck. As a further rigorous test, we have carried out a blind all-atom prediction for a noncanonical RNA motif, the C7.2 tetraloop/receptor, and validated this model through nucleotide-resolution chemical mapping experiments. Stepwise assembly is an enumerative, ab initio build-up method that systematically outperforms existing Monte Carlo and knowledge-based methods for 3D structure prediction.

de novo modeling | tertiary structure | dynamic programming | structure mapping | nucleic acid

Predicting the 3D structures attained by functional macromolecules is a fundamental challenge in computational biophysics and, more generally, in understanding and engineering living systems. There have been numerous recent successes in the high-resolution modeling of small proteins (1–3), protein/RNA complexes (4), and protein/DNA interfaces (5) by optimizing physically realistic energy functions. Nevertheless, rigorous blind trials demonstrate that the predictive power of computational algorithms remains limited, especially if atomic resolution is sought. For essentially all high-resolution modeling problems tackled to date, the shared critical bottleneck of these methods is inefficient sampling of a biopolymer's vast conformational space (1–7). In addition to hindering accurate modeling, poor sampling precludes rigorous tests of the assumed high-resolution energy functions.

To gain insight into the conformational sampling bottleneck, we have been focusing on some of the smallest well-defined biomolecular folding problems: RNA motifs, as short as four nucleotides (nts) in length (8). In addition to offering "toy puzzles" for computational methods (9), these modular loops, junctions, and tertiary interactions are fundamental building blocks of structured noncoding RNAs; they attain well-defined noncanonical conformations that in turn define the positions of the canonical double helices in three dimensions. A previous study presented a fragment assembly of RNA with full-atom refinement (FARFAR) method (10), tested on a benchmark of 32 RNA motifs. Although FARFAR recovered near-atomic-accuracy models in half the cases, the method was unable to consistently sample models within 1.5 Å rmsd of the crystallographic conformation.

Herein we seek to dissect and resolve this conformational sampling bottleneck by focusing on an apparently simpler problem: the structure prediction of single-stranded irregular RNA loops excised from crystallographic models. Modeling these loops is a lock-and-key problem, where the native loop (the key) is the conformation that best fits the surrounding structure (the lock). As with the analogous protein cases, the RNA loop-modeling problem has important practical significance as a critical component of homology-based structure prediction (11, 12) and in the refinement of models generated by coarse-grained algorithms (13–16). As is illustrated below, even the smallest RNA loops are challenging for computational methods, because they are rich in noncanonical interactions, extrahelical bulges, and unusual torsion combinations.

Our major finding is that a recursive stepwise ansatz enables the systematic sampling of RNA loop conformations at atomic resolution and in polynomial computational time. The ansatz is similar in spirit to ab initio "build-up" strategies previously explored in protein modeling (6, 17, 18, 19) but not yet shown to outcompete Monte Carlo or knowledge-based methods (20). Our focus on small RNA loops allows us to revisit and rigorously test these enumerative strategies. After illustrating the limitations of knowledge-based approaches in loop modeling, we describe the motivations for the stepwise ansatz, its potential advantages and disadvantages, and its implementation as the stepwise assembly (SWA) method in the Rosetta framework. We then demonstrate substantial improvements in sampling power and modeling accuracy of the SWA method over prior approaches. As a further rigorous and practical test, we present a blind prediction of an RNA motif of previously unknown structure, the in vitro evolved C7.2 tetraloop/receptor (21, 22), and its experimental validation by subsequent chemical accessibility measurements. We end the paper with discussions of historical precedents for this ansatz as well as extensions of this strategy to multistranded RNA motifs and protein problems.

Results

A Benchmark for the High-Resolution RNA Loop-Modeling Problem. The RNA loop-modeling problem offers small but highly challenging cases for atomic-resolution structure prediction. We compiled a benchmark of 15 single-stranded loops that begin and end at different Watson–Crick double helices, drawn from riboswitches, ribozymes, and other structured noncoding RNAs with crystallographic data (resolution better than 2.85 Å; *SI Appendix*, Table S1). Loop lengths ranged from 4 to 10 nucleotides (longer loops are rare; see *SI Appendix*, Fig. S1). "Hairpin" loops beginning and ending at the same helix as well as multiple-stranded loops can also be treated but are considered separately (see below).

Author contributions: P.S. and R.D. designed research; P.S. and W.K. performed research; P.S. and R.D. analyzed data; and P.S. and R.D. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: rhiju@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1106516108/-/DCSupplemental.

On one hand, these loops assemble into well-defined conformations, forming a significant number of hydrogen bonds—2.6 per nucleotide on average, in the same range as values for an A-form RNA helix (2 to 3). For several cases, independent crystallographic models of the same loop are available and give indistinguishable conformations (*SI Appendix*, Table S2). On the other hand, the loops are highly noncanonical. More than half of the hydrogen bonds are in base-phosphate or base-sugar interactions rather than in base pairs (23). Further, the loop torsions are irregular. Twenty-seven percent of the nucleotide suites are not part of the 46 most commonly observed RNA rotamers (24); and 8 of the 15 loops contain extrahelical bulges. Several loops display sharp turns, exemplified by the J2/4 loop motif that forms a 140° bend in the three-way junction of a thiamine pyrophosphate (TPP) sensing riboswitch (Figs. 1 *A* and *B*). Modeling these intricate loop structures de novo is therefore a well-posed but challenging problem.

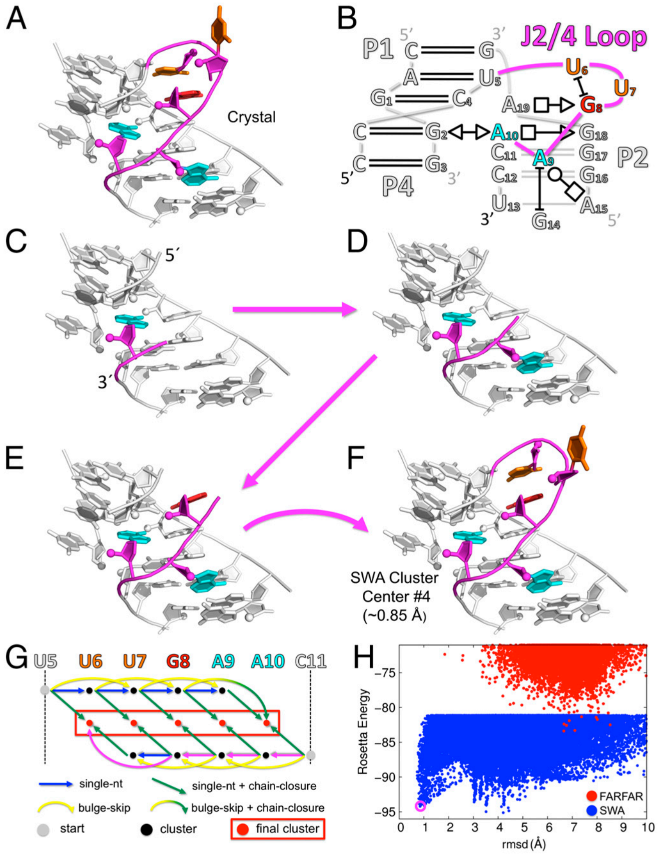


Fig. 1. The stepwise assembly (SWA) structure modeling method. Illustration on the J2/4 loop from the three-way junction of a TPP sensing riboswitch (PDB: 3DV2). (*A*) Crystallographic conformation of the 5-nt loop (shown in color) with surrounding nucleotides from the crystallographic model shown in white. (*B*) Schematic of the three-way junction in the annotation of Leontis and Westhof (23); only nucleotides shown in the 3D structure are numbered. (*C–F*) A build-up path that leads to the experimental conformation; the five nucleotides in the loop are built in a stepwise manner, one at time, starting from the 3′ end. (*G*) A directed acyclic graph delineates the building steps in the SWA method, recursively covering all possible build-up paths. The building steps taken in *C–F* are colored in magenta; other building steps are colored according to type. Gray vertices correspond to the starting point with none of the loop nucleotides built. Black vertices correspond to the partially built subregions; models in each subregion were clustered with the 1,000 lowest energy cluster centers carried forward. Red vertices corresponding the ending points with the loop completely built; all models of the full-length loop were clustered together in a final clustering step. (*H*) Rosetta all-atom energy vs. all-heavy-atom rmsd to the crystallographic conformation for de novo models generated by SWA (blue points) and by the prior method (FARFAR, red points). SWA fourth lowest energy cluster center (purple circle) is within atomic accuracy of the crystallographic model (0.85 Å rmsd).

Limitations of Knowledge-Based Methods. The difficulty of RNA loop modeling is underscored by the poor accuracy of previous methods for RNA structure prediction. For example, a recently developed homology modeling method, RLooM (11), failed to recover near-native models (under 1.5 Å all-heavy-atom rmsd to the crystallographic loop) in 13 of the 15 benchmark cases, unless directly related loop structures from the same species were permitted (*SI Appendix*, *Supporting Results*, and Table S2). As a further test, we updated the high-resolution FARFAR method to carry out loop modeling with chain closure and sampling of extrahelical bulges. FARFAR failed to recover near-native models as one of the five lowest energy cluster centers in more than half of the benchmark cases (11 of 15; ref. 25, Table 1, and *SI Appendix*, Table S3). Some of the problem cases are quite small; for example, the J2/4 loop of the TPP riboswitch was not solvable by FARFAR but is only five nucleotides in length (Figs. 1 *A* and *B*). As in prior work, conformational sampling was the dominant bottleneck. First, for 6 of 11 problem cases, none of the 250,000 models generated gave rmsd accuracy better than 1.5 Å (Table 1). Second, in all cases, this inability to generate near-native structures was traced to the absence of native torsions in the fragment library; the sampling could be rescued by doping native torsions into the fragment library as a "cheat" to aid conformational search (see *SI Appendix*, Table S4). Third, in 10 of 11 cases, the generated models did not achieve near-native energies; the lowest energy of 250,000 models remained higher than the energy of the optimized experimental loops (Table 1). The inability of FARFAR to solve these small loop-modeling problems suggests that one or more basic assumptions of the fragment assembly approach limit its conformational sampling power.

A Stepwise Ansatz. We reasoned that the conformations of RNA loops might be effectively sampled through direct enumeration at high resolution, rather than by restricting the search space to previously known fragments. We discovered that a recursive step-by-step enumeration (Figs. 1 *C–F*) permits efficient de novo sampling of these loops, which we illustrate on one of the FARFAR failures above, the J2/4 loop of the TPP riboswitch.

First, we note that exhaustive enumeration of this 5-nt loop at atomic resolution is not feasible with current computational power. Even building one nucleotide of a loop involves sampling several degrees of freedom, including six backbone torsions, four (coupled) sugar-pucker torsions, the glycosidic torsion, and the 2′-OH torsion. While low-resolution (>3 Å) clustering of exhaustively sampled single-nucleotide conformations results in under 100 "rotamers" (24), clustering with a subangstrom threshold—as is necessary for high-resolution modeling—leads to millions of unique conformations of the nucleotide (*SI Appendix*, Fig. S2). While computing the Rosetta energy for this number of conformations is achievable in less than 1 hour on a single modern central processing unit (CPU), the available conformations multiply exponentially with the RNA length. Thus, combinatorial enumeration of all available conformations of a 5-nt loop would require approximately 10^{23} CPU years, well beyond the computational power achievable in the foreseeable future.

Nevertheless, the feasibility of enumerating the conformations of just one nucleotide suggests an alternative approach to realistic RNA modeling. Enumerative single-nucleotide building permits fine-grained exploration of torsional conformations that form well-packed structures with multiple hydrogen bonds, as is observed in native loops, including rare torsional combinations not covered in the list of consensus rotamers (24). As an illustration, Fig. 1*C* shows the lowest energy conformation for the first 3′ nucleotide of the J2/4 loop, built by exhaustive sampling followed by local energy minimization. The resulting nucleotide

Table 1. Accuracy and conformational sampling efficiency of de novo RNA loop modeling

	Motif properties		Best rmsd* (Å) of five lowest energy clusters†		Lowest rmsd* (Å) achieved		Energy gap to optimized exp. model‡ (RU)	
Motif name	Length	PDB	FARFAR	SWA	FARFAR	SWA	FARFAR	SWA
5′ J1/2, leadzyme	4	1NUJ	1.96	**0.83**	1.66	**0.51**	2.7	**−0.8**
5′ P1, M-box riboswitch	4	2QBZ	**0.72**	**0.96**	**0.53**	**0.61**	2.3	**−0.5**
3′ J5/5a, group I intron	4	2R8S	**0.40**	**0.47**	**0.30**	**0.40**	**0.0**	**0.0**
5′ J5/5a, group I intron	5	2R8S	4.08	**1.04**	**1.05**	**0.66**	0.3	**−0.9**
Hepatitis C virus IRES IIa	5	2PN4	2.11	5.31	**1.04**	**0.71**	**−2.6**	**−5.9**
J2/4, TPP riboswitch	5	3D2V	6.66	**0.85**	1.74	**0.73**	10.8	**−1.0**
23S rRNA (44–49)	6	1S72	**0.69**	**0.73**	**0.47**	**0.71**	2.6	**0.0**
23S rRNA (531–536)	6	1S72	3.18	2.45	2.44	**0.76**	6.9	**−0.6**
J3/1, glycine riboswitch	7	3OWI	**1.13**	**1.35**	**0.71**	**0.64**	2.5	1.3
J2/3, group II intron	7	3G78	1.59	**0.82**	**1.34**	**0.77**	8.5	**−0.2**
L1, SAM-II riboswitch	7	2QWY	2.43	**1.26**	**1.43**	**0.86**	3.8	**−1.3**
L2, viral RNA pseudoknot	7	1L2X	5.44	3.36	**1.35**	**0.91**	3.7	**−4.1**
23S rRNA (2534–2540)	7	1S72	6.39	5.71	3.24	**1.39**	7.3	**−7.3**
23S rRNA (1976–1985)	10	1S72	11.19	7.75	5.06	4.58	9.6	**−10.8**
23S rRNA (2003–2012)	10	1S72	11.36	**0.74**	5.43	**0.64**	41.2	3.2
RMSD < 1.50 Å	—	—	4/15	10/15	9/15	14/15	—	—
Energy Gap < 0.0	—	—	—	—	—	—	2/15	13/15

IRES, internal ribosome entry site; SAM, *S*-adenosylmethionine.

*All-heavy-atom rmsd to the crystallographic loop. Nucleotides found to be extrahelical bulges (both unpaired and unstacked) in the crystallographic model were excluded from the rmsd calculation. Bold text indicates rmsd within 1.5 Å of the crystallographic model.

†Generated models were clustered, such that models with pairwise all-heavy-atom rmsd less than 1.5 Å over the entire loop and less than 2.5 Å over each individual loop nucleotide are grouped (see *SI Appendix, Supporting Methods*). The lowest energy member of each cluster was designated as the cluster center and the five lowest energy cluster centers were considered as the predicted models.

‡Definition of the optimized experimental model is provided in *SI Appendix, Supporting Methods*. Bold text indicates that the lowest energy sampled by the de novo run is lower than the energy of the optimized experimental model (i.e., the energy gap is negative). One Rosetta unit (RU) is approximately equal to 1 k_BT (10, 25).

is positioned with atomic accuracy, giving an rmsd of 0.69 Å from the experimental conformation. We discovered that the entire loop could then be recovered through stepwise enumerative building of each additional nucleotide (Figs. 1 *C*–*F*), carrying forward an ensemble of the lowest energy well-packed, well-hydrogen-bonded conformations from each previous subregion. In addition to standard single-nucleotide building steps, recovering this loop also required a "bulge-skip" building step (to permit the modeling of extrahelical unpaired/unstacked nucleotides) and a chain-closure building step to complete the RNA loop (e.g., Figs. 1 *E*–*F*; see *SI Appendix, Supporting Methods* for complete descriptions of the three types of building steps).

In a de novo structure prediction scenario, we do not know a priori the appropriate order of building steps that will achieve the experimental conformation, and we cannot guarantee that the lowest energy model for a subregion will carry forward into the lowest energy model for the entire loop. Further, the number of such build-up paths grows exponentially with the number of nucleotides. We solved these path-enumeration issues using a recursive strategy, familiar from dynamic programming approaches utilized in sequence alignment (26) and RNA secondary structure prediction (27). We determined a low-energy ensemble of models for each subregion of the loop as modeled from the 5′ end or from the 3′ end and then joined all combinations of these subregions by chain closure. In particular, we modeled each subregion in one of two ways—either by a standard single-nucleotide building step from a subregion one nucleotide shorter, or by a bulge-skip building step from a subregion two nucleotides shorter. We clustered all models for a subregion and carried forward the 1,000 lowest energy cluster centers (which typically included all models within 6 k_BT of the lowest energy state, mimicking conformations accessed by thermal fluctuations). A directed acyclic graph (28) delineates this deterministic, recursive calculation, as shown in Fig. 1*G*. In the case of the J2/4 loop example, searching through all possible paths led to a diverse set of well-packed conformations, including low-energy near-native and nonnative models that were missed by FARFAR (Fig. 1*H*).

This method deterministically enumerates a low-energy subspace of the RNA loop's available conformations through the stepwise, locally optimal building of individual nucleotides, with the hypothesis that the experimentally observed conformation resides within this subspace. We call this method stepwise assembly (SWA) and its underlying working hypothesis, the stepwise ansatz. This ansatz can only be confirmed through empirical tests on naturally occurring biomolecular structures. We have therefore carried out extensive trials of the stepwise ansatz using RNA loop modeling as a biophysically important but unsolved test problem, described next.

Comprehensive Test of the Stepwise Ansatz. To evaluate the validity of the stepwise ansatz, we applied the SWA method on the entire 15-loop benchmark (*SI Appendix*, Table S1). In terms of modeling accuracy, SWA substantially outperformed FARFAR, recovering near-native models (<1.5 Å rmsd) for 10 of 15 test cases, compared to four cases recovered by FARFAR (see Table 1). These included atomic-accuracy models from diverse sources, including a 5-nt loop from the J5/5a hinge in the P4–P6 domain of the group I *Tetrahymena* ribozyme (rmsd of 1.04 Å; Fig. 2*A*); a 7-nt loop connecting helices P2 and P3 of the group II intron (rmsd of 0.82 Å; Fig. 2*B*); and one of the two 10-nt loops in the benchmark, nucleotides 2003–2012 of the large ribosomal subunit from *Haloarcula marismortui* (rmsd of 0.74 Å; Fig. 2*C*). In each of these three cases, the high accuracy of the SWA model is reflected not only in low rmsd to the experimental loops but also complete recovery of the base pair and base stack geometries as classified in the Leontis–Westhof scheme (23) (see *SI Appendix*, Table S5).

For the remaining five "problem cases," conformational sampling was no longer the major bottleneck. In all five cases, SWA models achieved lower energies than the optimized experimental models (see Table 1 and *SI Appendix*, Fig. S3). Further, in four of the five cases, SWA sampled de novo models within 1.5 Å of the experimental conformation, although these models were not selected as one of the five lowest energy cluster centers. In the last case (a second 10-nt ribosomal loop), the optimized experimental model gave significantly worse energy (by 10.8 k_BT)

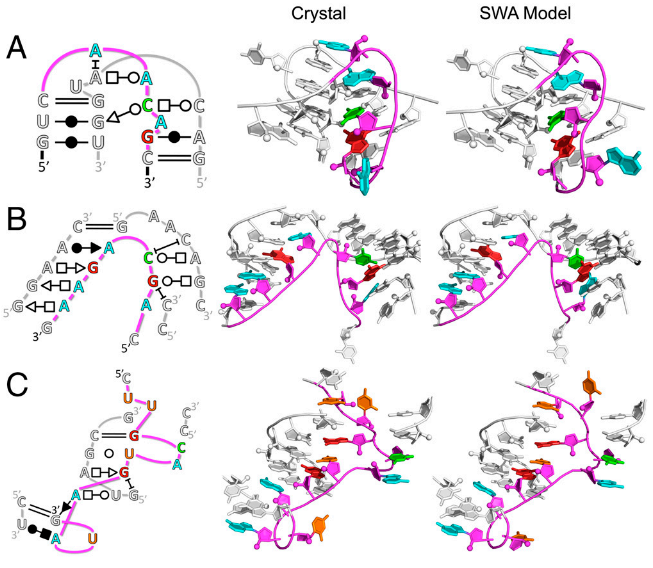


Fig. 2. Comparison of crystallographic and SWA de novo models for three diverse loop motifs. (*A*) Five-nucleotide loop from the J5/5a hinge in the P4–P6 domain of the group I *Tetrahymena* ribozyme (PDB: 2R8S). (*B*) Seven-nucleotide loop connecting helices P2 and P3 of the group II intron (PDB: 3G78). (*C*) Ten-nucleotide loop from the large ribosomal subunit from *H. marismortui* (PDB: 1S72, nucleotides 2003–2012). The modeled loop is shown in color whereas surrounding nucleotides are shown in white. Some surrounding nucleotides are not shown to permit unobstructed view of the modeled loop region. The rmsds to the crystallographic conformations (energy cluster rank) of the displayed SWA models are (*A*) 1.04 Å (fourth), (*B*) 0.82 Å (first), and (*C*) 0.74 Å (second). Two-dimensional schematics apply to both the crystallographic and SWA models.

than the SWA models, explaining the absence of near-native models in the low-energy SWA ensemble. These results demonstrated that the stepwise ansatz is valid in all tested cases, and the absence of atomic-accuracy models among the five lowest energy cluster centers for the five problem cases was due to inaccuracies in the Rosetta all-atom energy function. The results were in strong contrast to the FARFAR results above.

Blind Prediction and Experimental Validation. The most stringent tests for structure prediction algorithms are blind trials. The few prior attempts at blind high-resolution RNA structure modeling have not achieved atomic accuracy [see, e.g., refs. (29–31)]. Encouraged by the strong performance of SWA on the benchmark, we predicted the structure of a tetraloop/receptor motif (the C7.2 mutant; Fig. 3*A*) with no known experimental structure, previously isolated by in vitro selection (21, 22).

This sequence served as an appropriate first blind test because it effectively reduces to a small but challenging loop-modeling problem. Much of the sequence aligns with a widely studied tetraloop/receptor motif whose structure has been determined by crystallography in several different RNAs, including the P4–P6 domain of the *Tetrahymena* ribozyme (32, 33). The main difference is a 3-nt loop (G4-U5-A6) replacing a 2-nt A4-A5 "platform" (*SI Appendix*, Fig. S4). We modeled this loop by SWA, FARFAR, RLooM, and ModeRNA (12). SWA gave the well-packed C7.2 tetraloop-docked receptor model shown in Fig. 3*B* as the lowest energy structure. More extensive SWA calculations modeling eight nucleotides (nucleotides 3–7 and 10–12 in Fig. 3*A*) gave similar structures. In contrast, FARFAR gave models with significantly worse energy (by $>3\ k_BT$) whereas RLooM and ModeRNA gave models with numerous steric clashes (see *SI Appendix*, *Supporting Results*, and Fig. S5).

The SWA model for the C7.2 tetraloop-docked receptor displayed noncanonical features absent in the classic 11-nt receptor (32, 33). The central U5 nucleotide bulged out of the structure.

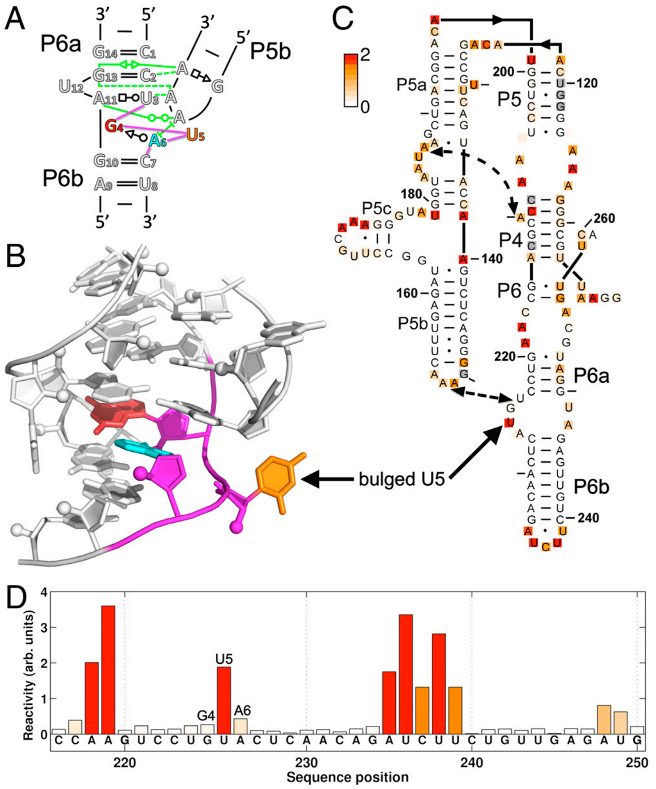


Fig. 3. Blind prediction of the C7.2 tetraloop-docked receptor and validation through single-nucleotide-resolution chemical mapping. (*A*) Two-dimensional schematic of the C7.2 tetraloop/receptor motif; the 3-nt G4-U5-A6 loop at the core of the receptor (shown in color) is different from receptors with previously solved structures. Tertiary interactions between the GAAA tetraloop and the receptor are colored green. (*B*) Three-dimensional model of the C7.2 receptor by SWA. Models from other methods are given in *SI Appendix*, Fig. S5. (*C*) Chemical reactivities of A and C (based on dimethyl sulfate alkylation) and G and U (based on CMCT carbodiimide modification) shown as white-to-red coloring on a mutant of the P4–P6 domain of the *Tetrahymena* ribozyme containing the C7.2 receptor; measurements were acquired in 10 mM $MgCl_2$, 50 mM Hepes, pH 8.0, at 24 °C. (*D*) Bar graph of reactivities for nucleotides near the C7.2 receptor. Sequence positions are given in conventional P4–P6 numbering with one additional nucleotide inserted between positions 225 and 227 to account for the longer length of C7.2 compared to wild type. See *SI Appendix*, Figs. S6 and S7 for full datasets, including both wild type and C7.2 mutant data and error analysis.

Furthermore, the first and third nucleotides of the loop formed a same-stranded *trans* Sugar-edge/Watson–Crick G4-A6 base pair (Fig. 3*B*) that is not isosteric to the *cis* Sugar-edge/Hoogsteen base pair presented in the A-A platform (34). The Find RNA 3D (FR3D) motif search software (35) found only two other instances of this conformation in the entire database of RNA structures, within a malachite green aptamer and in a UUGUAU RNA sequence bound to the human cleavage factor protein Im (see *SI Appendix*, *Supporting Results*). Nevertheless, the neighboring 5′ and 3′ nucleotides in these two precedent structures are positioned differently than in the C7.2 receptor (FR3D geometric discrepancies of 0.72 and 0.89 Å; both higher than the 0.50 Å default cutoff value), explaining the inability of RLooM and ModeRNA to discover these solutions.

The SWA model for the C7.2 tetraloop-docked receptor (Fig. 3*B*) made predictions that were testable by single-nucleotide-resolution chemical modification experiments. We therefore grafted the C7.2 receptor into the J6a/6b and J6b/6a segments of the P4–P6 RNA (Fig. 3*C*) and carried out quantitative chemical mappings with dimethyl sulfate (DMS) and 1-cyclohexyl-3-(2-morpholinoethyl)carbodiimide metho-*p*-toluene sulfonate (CMCT) (36,37). As with the wild-type P4–P6 RNA, the

C7.2-grafted mutant showed clear protections of the L5b tetraloop, J6a/6b tetraloop receptor, and the P5a A-rich bulge upon addition of Mg^{2+}, verifying the attainment of the RNA's global tertiary fold (electrophorograms shown in *SI Appendix*, Fig. S6). Further, as expected, the chemical reactivities of the wild-type RNA and the C7.2 mutant outside the tetraloop/receptor motif were indistinguishable within experimental error (*SI Appendix*, Fig. S7). Within the C7.2 receptor, nucleotides G4 and A6 were both protected from chemical modification, as predicted in the SWA model (nucleotides 225 and 227 in conventional P4–P6 numbering; Figs. 3 *C* and *D*). Most importantly, U5 (nucleotide 226 in conventional numbering) was highly modified by CMCT, with a reactivity value 22 ± 5 times greater than the mean reactivity of Watson–Crick base-paired uridines in the entire P4–P6 RNA. This result provides strong confirmation that U5 is an extrahelical bulge, as predicted. The chemical accessibility data thus validate the de novo SWA model at nucleotide resolution and disfavor first-ranked models from knowledge-based methods (*SI Appendix*, Fig. S5). Subsequent to obtaining these experimental results, we discovered further evidence in support of the SWA model from sequence variations in the original in vitro selection experiment that isolated the C7.2 receptor (21) (summarized in *SI Appendix, Supporting Results*).

Discussion

A Stepwise Ansatz Resolves a Conformational Sampling Bottleneck in Structure Prediction. An inability to guarantee exhaustive conformational sampling has precluded the consistent prediction of biomolecular structure at high resolution (1–6). In Rosetta as well as other frameworks (10–16), potential issues that limit de novo sampling efficiency include these algorithms' dependence on the database of existing experimental structures; the stochasticity of Monte Carlo fragment assembly; and the loss of information due to the use of coarse-grained phases to smooth and reduce the dimensionality of the search space (7, 9, 38). To address these issues, we developed a working hypothesis, called the stepwise ansatz, and its implementation, the SWA method, that enumeratively searches a physically realistic subspace of a molecule's all-atom conformations in polynomial computational time [$O(N)$ where N is the number of nucleotides; see *SI Appendix, Supporting Methods*].

The concept of ab initio step-by-step build-up has been discussed previously, e.g., in enumerative coarse-grained or stochastic all-atom search methods from Dill and coworkers (18, 19), pioneering peptide-modeling work from the 1980s by the Scheraga lab (17), and earlier computational explorations by Levinthal in 1968 (6). However, these prior build-up strategies have not been adopted into the mainstream of structure modeling or shown to outcompete Monte Carlo or knowledge-based methods (19, 20). The prior lack of development appears to stem from the difficulty of searching all possible build-up paths and from the expense of deterministic, enumerative calculations relative to stochastic, knowledge-based methods. For example, modeling a single 5-nt RNA loop herein required 12,000 CPU hours; fortunately, this calculation is now feasible due to the massive parallelization of high-performance computer clusters.

On a challenging benchmark of irregular RNA loop motifs, we have shown that SWA resolves the conformational sampling bottleneck that has hindered knowledge-based methods. In all cases, SWA sampled the experimental loop conformation de novo and/or recovered conformations with energies that surpassed the energy of the optimized experimental loop conformation. Further, in the majority of the cases (10 of 15), the Rosetta all-atom energy function was accurate enough to permit a near-native conformation to be selected as one of the five lowest energy cluster centers. The strongest test of the SWA method is the blind prediction on the C7.2 tetraloop/receptor motif of previously unknown structure. The predicted model includes noncanonical features (including a same-stranded G-A base pair and an extrahelical bulge) and agrees with subsequently measured chemical accessibility data. Further atomic-resolution tests might be achieved if crystals can be obtained for the C7.2 mutant of the P4–P6 RNA.

Stringent Tests of the Rosetta All-atom Energy Function. Prior studies have reported anecdotal cases of failures of the Rosetta all-atom energy function for macromolecule modeling (9, 39), but the work herein is a unique example of a complete high-resolution de novo modeling benchmark in which every failure case can be traced to inaccuracies in the underlying energy function. While prior work has shown that the Rosetta all-atom energy function provides better energetic discrimination than traditional molecular mechanics force fields (10), this work indicates that approximations in the Rosetta all-atom energy function still remain too inaccurate to permit atomic-resolution RNA modeling on a consistent basis. The energy function does not explicitly model metal ions (e.g., see *SI Appendix*, Fig. S3), and water is modeled through a crude solvation term (40). Long-range electrostatic effects, higher-order dispersion effects (41), and hydrogen bond cooperativity are presently neglected. Because of its generality and sampling power, the SWA method should permit stringent tests of more recently developed all-atom energy functions, including those that model polarizable moieties (42). For the same reasons, the SWA approach should be powerful for high-resolution structure determination methods that use limited experimental information as pseudoenergy terms to break degeneracies in physics-based energy functions [see, e.g., refs. (43–45)].

A General Enumerative Strategy for Molecular Modeling. In this work, we have focused mainly on the application of SWA toward single-stranded RNA loop-segments, both to demonstrate the method's conformational sampling power and to solve a basic practical problem that arises in RNA structure prediction. Nevertheless, the strategy should be generally applicable to a diverse class of molecular modeling problems. For example, noncanonical RNA motifs often involve multiple RNA strands interacting with one another or loops returning to the same helix. Extensions of the SWA method to model these motifs appear accurate and computationally tractable (see *SI Appendix*, Fig. S8). With further expected improvements in computational power, de novo atomic-accuracy modeling of RNA motifs with lengths up to 15 nucleotides, a size range that includes many RNA aptamers and catalytic sites, should be feasible. Further, the basic concepts underlying SWA are not specific to RNA structure prediction and should be applicable to other frontier problems in high-resolution macromolecular modeling, including efficient prediction of protein loops and small proteins, rigorous tests of protein and protein/RNA energy functions, and enumerative sequence design of functional protein and RNA loops.

Methods

Both the SWA and FARFAR methods were implemented in C++ in the Rosetta codebase. The software is being made available in the next Rosetta release (3.4). Application of RLooM (database version 12-19-08) and ModeRNA (version 1.6.0) follow the instructions given in the released software. DMS and CMCT modification data of the wild-type P4–P6 RNA and the C7.2 P4–P6 mutant were acquired at single-nucleotide resolution, as described previously (46). Complete description of the SWA method; details on updates to the FARFAR method; explicit command-line examples for RNA loop modeling with SWA, FARFAR, RLooM, and ModeRNA; and details of the experimental method are provided in *SI Appendix, Supporting Methods*.

ACKNOWLEDGEMENTS. We thank the Rosetta community for sharing code. Calculations were carried out on the BioX^2 cluster (National Science Foundation (NSF) Award CNS-0619926). We acknowledge financial support from a Burroughs Wellcome Career Award at the Scientific Interface (R.D.) and a C. V. Starr Asia/Pacific Stanford Graduate Fellowship (P.S.).

1. Bradley P, Misura KM, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309:1868–1871.
2. Jones DT, et al. (2005) Prediction of novel and analogous folds using fragment assembly and fold recognition. *Proteins* 61(Suppl 7):143–151.
3. Qian B, et al. (2007) High-resolution structure prediction and the crystallographic phase problem. *Nature* 450:259–264.
4. Fleishman SJ, et al. (2010) Rosetta in CAPRI rounds 13–19. *Proteins* 78:3212–3218.
5. Ashworth J, et al. (2006) Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* 441:656–659.
6. Levinthal C (1968) Are there pathways for protein folding? *J Chim Phys* 65:44–45.
7. Kim DE, Blum B, Bradley P, Baker D (2009) Sampling bottlenecks in de novo protein structure prediction. *J Mol Biol* 393:249–260.
8. Jucker FM, Heus HA, Yip PF, Moors EH, Pardi A (1996) A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J Mol Biol* 264:968–980.
9. Das R (2011) Four small puzzles that Rosetta doesn't solve. *PLoS One* 6:e20044.
10. Das R, Karanicolas J, Baker D (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat Methods* 7:291–294.
11. Schudoma C, May P, Nikiforova V, Walther D (2010) Sequence-structure relationships in RNA loops: Establishing the basis for loop homology modeling. *Nucleic Acids Res* 38:970–980.
12. Rother M, Rother K, Puton T, Bujnicki JM (2011) ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res* 39:4007–4022.
13. Das R, Baker D (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci USA* 104:14664–14669.
14. Parisien M, Major F (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 452:51–55.
15. Flores SC, Altman RB (2010) Turning limited experimental information into 3D models of RNA. *RNA* 16:1769–1778.
16. Cao S, Chen SJ (2011) Physics-based de novo prediction of RNA 3D structures. *J Phys Chem B* 115:4216–4226.
17. Gibson KD, Scheraga HA (1987) Revised algorithms for the build-up procedure for predicting protein conformations by energy minimization. *J Comput Chem* 8:826–834.
18. Hockenmaier J, Joshi AK, Dill KA (2007) Routes are trees: The parsing perspective on protein folding. *Proteins* 66(1):1–15.
19. Shell MS, Ozkan SB, Voelz V, Wu GA, Dill KA (2009) Blind test of physics-based prediction of protein structures. *Biophys J* 96:917–924.
20. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A (2009) Critical assessment of methods of protein structure prediction—round VIII. *Proteins* 77(Suppl 9):1–4.
21. Costa M, Michel F (1997) Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: Comparison with in vivo evolution. *EMBO J* 16:3289–3302.
22. Geary C, Baudrey S, Jaeger L (2008) Comprehensive features of natural and in vitro selected GNRA tetraloop-binding receptors. *Nucleic Acids Res* 36:1138–1152.
23. Leontis NB, Westhof E (2001) Geometric nomenclature and classification of RNA base pairs. *RNA* 7:499–512.
24. Richardson JS, et al. (2008) RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA* 14:465–481.
25. Kortemme T, Kim DE, Baker D (2004) Computational alanine scanning of protein-protein interfaces. *Sci STKE* 2004:pl2.
26. Smith TF, Waterman MS, Fitch WM (1981) Comparative biosequence metrics. *J Mol Evol* 18:38–46.
27. Zuker M, Sankoff D (1984) RNA secondary structures and their prediction. *B Math Biol* 46:591–621.
28. Maurer SB (2003) Directed acyclic graphs. *Handbook of Graph Theory*, eds JL Gross and J Yellen (CRC, Boca Raton, FL), pp 142–155.
29. Leontis NB, Westhof E (1998) The 5S rRNA loop E: Chemical probing and phylogenetic data versus crystal structure. *RNA* 4:1134–1153.
30. Lemieux S, Chartrand P, Cedergren R, Major F (1998) Modeling active RNA structures using the intersection of conformational space: Application to the lead-activated ribozyme. *RNA* 4:739–749.
31. Harris S, Schroeder SJ (2010) Nuclear magnetic resonance structure of the prohead RNA E-loop hairpin. *Biochemistry* 49:5989–5997.
32. Cate JH, et al. (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* 273:1678–1685.
33. Ye JD, et al. (2008) Synthetic antibodies for specific recognition and crystallization of structured RNA. *Proc Natl Acad Sci USA* 105:82–87.
34. Stombaugh J, Zirbel CL, Westhof E, Leontis NB (2009) Frequency and isostericity of RNA base pairs. *Nucleic Acids Res* 37:2294–2312.
35. Sarver M, Zirbel C, Stombaugh J, Mokdad A, Leontis N (2008) FR3D: Finding local and composite recurrent structural motifs in RNA 3D structures. *J Math Biol* 56:215–252.
36. Tijerina P, Mohr S, Russell R (2007) DMS footprinting of structured RNAs and RNA-protein complexes. *Nat Protoc* 2:2608–2623.
37. Stern S, Moazed D, Noller HF (1988) Structural analysis of RNA using chemical and enzymatic probing monitored by primer extension. *Methods Enzymol* 164:481–489.
38. Beauchamp K, Sripakdeevong P, Das R (2011) Why can't we predict RNA structure at atomic resolution? *RNA 3D Structure Analysis and Prediction*, eds N Leontis and E Westhof (Springer), in-press.
39. Mandell DJ, Coutsias EA, Kortemme T (2009) Subangstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Methods* 6:551–552.
40. Rohl CA, Strauss CEM, Misura K, Baker D (2004) Protein structure prediction using Rosetta. *Methods Enzymol* 66–93.
41. Sato T, Tsuneda T, Hirao K (2005) A density-functional study on pi-aromatic interaction: Benzene dimer and naphthalene dimer. *J Chem Phys* 123:104307-1–104307-10.
42. Ponder JW, et al. (2010) Current status of the AMOEBA polarizable force field. *J Phys Chem B* 114:2549–2564.
43. Shen Y, et al. (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* 105:4685–4690.
44. Vallurupalli P, Hansen DF, Kay LE (2008) Structures of invisible, excited protein states by relaxation dispersion NMR spectroscopy. *Proc Natl Acad Sci USA* 105:11766–11771.
45. Zuo X, et al. (2010) Solution structure of the cap-independent translational enhancer and ribosome-binding element in the 3′ UTR of turnip crinkle virus. *Proc Natl Acad Sci USA* 107:1385–1390.
46. Kladwang W, Cordero P, Das R (2011) A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. *RNA* 17:522–534.

Supporting Information for "An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling"

Parin Sripakdeevong[1], Wipapat Kladwang[2], Rhiju Das[1,2,3,*]

[1]Biophysics Program, Stanford University, Stanford, CA 94305, USA

[2]Department of Biochemistry, Stanford University, Stanford, CA 94305, USA

[3]Department of Physics, Stanford University, Stanford, CA 94305, USA

[*] To whom correspondence should be addressed. Phone: (650) 723-5976. Fax: (650) 723-6783. E-mail: rhiju@stanford.edu.

List of Supporting Text sections:

List of Supporting Figures:

List of Supporting Tables:

Supporting Methods

Stepwise assembly in Rosetta

The stepwise assembly method proceeds through the recursive building of single nucleotides. This section describes (a) the general recursive scheme, (b) the single-nucleotide building step, (c) the bulge-skip building step, (d) the chain-closure step, and (e) the models clustering step used in the method.

(a) General recursive scheme

The stepwise assembly method determines low-energy models for each sub-region of the loop, culminating in models of the entire loop, through recursive application of four kinds of steps (see main text Fig. 1G). Given the loop nucleotides i through j [denoted here (i, j)], the sub-regions consisted of loop pieces built from the 5′-end (i, k), with $i \leq k \leq j$, and loop pieces built from the 3′-end (l, j) with $i \leq l \leq j$. The recursion was defined as follows:

i. For a 5′-loop piece (i, k), models were generated by applying the *single-nucleotide building step* on models available for the smaller loop piece $(i, k–1)$. Similarly, for a 3′-loop piece (l, j), models were built from the models of the smaller loop piece $(l+1, j)$. The recursion proceeds until the loop is completed [i.e, the full-length loop (i, j)].

ii. An additional set of models for each loop piece was generated by *bulge-skip building steps* to permit the incorporation of bulged nucleotides. For a 5′-loop piece (i, k), models were created by building nucleotide k with an extra-helical bulge at $k–1$ from the models of the smaller loop piece $(i, k–2)$. Similarly, for a 3′-loop piece (l, j), models were built from the models of the smaller loop piece $(l+2, j)$. The recursion proceed until the loop is completed [i.e, the full-length loop (i, j)].

iii. Additional models of the full-length loop were generated by combining models of loop piece $(i, k–1)$ with models for loop piece $(k+1, j)$ and applying the *single-nucleotide building step* to build nucleotide k from either the 5′ edge $(k-1)$ or the 3′ edge $(k+1)$.

iv. For each partially built sub-region, all available models were grouped in a *clustering step*, with the lowest 1000 energy models carried forward.

v. All full-length loop models were clustered in a final *clustering step.*

The initial conditions for the recursion were defined by the crystallographic model with the entire loop (i, j) excised, as well as the P, O1P, O2P, O5′ atoms of neighboring nucleotide $j+1$ removed (to erase information about the "take-off" direction for the loop). This initial model also contained no hydrogens, as these atoms are not typically included in crystallographic models; 2′-OH hydrogens for the entire structure were sampled in the building steps (see below). The individual building steps are described next.

(b) Single-nucleotide building step

Here, the nucleotide to be built was directly adjacent to the 5′ or 3′ edge of the previous sub-region. Sub-Angstrom enumeration of the nucleotide's conformational space was achieved by sampling all the backbone torsions connecting the nucleotide to its 5′ or 3′ neighbor [the backbone "suite" $\varepsilon, \zeta, \alpha, \beta, \gamma$ (1)], the nucleobase's glycosidic torsion (χ) in 20° intervals, and the nucleotide ribose sugar (torsions ν_0, ν_1, ν_2, ν_3/δ, and ν_4) in its north (3′-endo) and south (2′-endo) puckered conformations. 2′-OH torsion sampling was carried out separately (see below). The α, γ, and ζ torsions were allowed to sample the full 360° range. Other torsions were restricted to sterically allowed regions: β in the *trans* region (80°–280°) and ε to the *trans* and *gauche*$^+$ regions (170°–290° if the sugar pucker was north; 182°–302° if south). For both pyrimidines and purines, the χ torsion was sampled in the *anti* region (179°-219° if north; 217°-257° and south). Additionally, for purines, the χ torsion was sampled in the *syn* region (49°-89° if north; 50°-90° if south). This led to the generation of 5,388,768 unique conformations for purines and 2,694,384 unique conformations for pyrimidines that were spaced from each other, on average, by 0.6 Å all-heavy-atom RMSD (Fig. S2).

After the enumeration above, every atom in the nucleotide was positioned except for the single 2′-OH hydrogen atom. The Rosetta *packer* algorithm (2, 3) originally written for protein side-chain packing and design, was used to determine the optimal orientation of these 2′-OH hydrogen atoms in all loop nucleotides as well as in all surrounding nucleotides through efficient combinatorial sampling, similar to REDUCE (4). The resulting conformations were then finely clustered with an all-atom RMSD cutoff of 0.5 Å over all the sampled atoms and by similarity of the newly built nucleotide's sugar pucker. The lowest energy member of each cluster was retained. The 108 lowest energy conformations (typically including all conformations within 8 k_BT of the very lowest energy conformation in the calculation) were subject to continuous minimization over all loop torsions with the Davidson–Fletcher–Powell algorithm (the Rosetta *minimizer*) and then re-clustered.

The calculations above, for even a single nucleotide, require comparing energies of millions of conformations in order to select out the lowest energy models. To accelerate the computation, we took advantage of the working hypothesis that the newly built nucleotides should form at least one favorable and no unfavorable interactions with previous nucleotides. We imposed the following filters. First, conformations were discarded in which the new nucleotide could not potentially form either base-stacking or base-pairing interactions with at least one other nucleotide in the structure, as assessed by geometric criteria that we defined based on interactions observed in the RNA crystallographic database [see also (5)]. Define d as the displacement vector between two base centroids; z_1 and z_2 are the vector projections of d onto the first base normal and second base normal; ρ_1 and ρ_2 are the vector projections of d onto the first base plane and second base plane; and θ is the angle between the base normals. The base-stacking criteria were defined to loosely include geometries for co-axially stacked bases: $|d| < 6.4$ Å; 2.5 Å $< |z_1| < 4.5$ Å and 2.5 Å $< |z_2| < 4.5$ Å; $|z_1|/|d| > 0.707$ and $|z_2|/|d| > 0.707$; and $\cos\theta > 0.707$. The base-pairing criteria were $|d| < 12.0$ Å; $|\rho_1| < 5.0$ Å and $|\rho_2| < 5.0$ Å; $|z_1| < 5.0$ Å and $|z_2| < 5.0$ Å; $|z_1|/|d| < 0.5$ and $|z_2|/|d| < 0.5$; and $\cos\theta > 0.866$. As a second filter, conformations were discarded if the attractive component (*fa_atr*) of the computed van der Waals interactions between the new nucleotide and other nucleotides in the structure was not better than –1.0 Rosetta units (~1 k_BT) or if the sum of the attractive and repulsive component (*fa_rep*) was worse than 0.0 Rosetta units. Approximately 1% of all the conformations passed the filters; computing the full Rosetta all-atom energy score for the resulting tens of thousands of conformations required only tens of minutes on an Intel Core i7 2.66 GHz processor. Test calculations (on single-nucleotide building steps) run without the filters gave indistinguishable results, but at much greater computational expense.

(c) Bulge-skip building step

To permit modeling of single extra-helical bulges, a bulge-skip building step was implemented, involving two nucleotides. This algorithm involved finding all possible positions of a new nucleotide that is covalently linked to previously built nucleotides by a single extra-helical bulge (unpaired and unstacked) nucleotide. The new nucleotide was sampled by carrying out a grid search over the nucleobase's six rigid-body degrees of freedom. The translation component (*x, y,* and *z* coordinates of the base centroid) was sampled in 1.0 Å intervals over a grid encapsulating all previously built nucleotides; the rotation component (the three Euler angles: azimuthal ϕ, bend angle θ, and second azimuthal ψ) was sampled over 20° intervals over the entire 0°–360° range for ϕ and ψ, and in intervals of 0.05 between –1.0 and 1.0 over $\cos\theta$ in 0.05 intervals, to prevent oversampling of polar regions (6). For each of the resulting base positions/orientations, the nucleotide's ribose sugar was built by sampling the glycosidic torsion (χ) in 20° intervals and by sampling the ribose sugar in both the north and south conformations. Exactly as above, the 2′-OH hydrogen atoms were sampled by the Rosetta *packer*; conformations were clustered; and the 108 lowest energy conformations were then selected. All loop torsions (and the position and orientation of the newly built centroid) were then subjected to continuous minimization, and the conformations were re-clustered. Again, a series of filters accelerated the calculation. Firstly, conformations were discarded if the distance between the newly built nucleotide and the previously built nucleotide would disallow building of the intervening extra-helical bulge nucleotide and chain closure (distance between the C5′ atom of the 3′ loop nucleotide and the O3′ atom of the 5′ nucleotide beyond 11.4 Å). Secondly, a coarse filter discarded nucleotide conformations with more than 2 atoms that clashed with previously built atoms (distance less than sum of the atom van der Waals radii minus 0.8 Å). Thirdly, base-stacking and base-pairing filters similar to the ones above in the single-nucleotide building step were applied. In this case, conformations that did not satisfy either the above base-stacking criterion (but with the cutoffs on $|z|/|d|$ and cos θ increased to 0.90) or satisfy both the above base-stacking and base-pairing criteria (but with the cutoff on cos θ of base-pairing loosened to 0.707) were discarded.

Finally, the extra-helical bulge nucleotide suite (the nucleotide plus the phosphate group of its 3′ neighbor) was built to connect the newly built nucleotide and the previously built nucleotides. This was accomplished by using the sampling procedure described in the single-nucleotide building step [see part *(b)*] to sample the extra-helical bulge nucleotide until a conformation that was clash-free and satisfied chain-closure [see part *(d)*] was found. The extra-helical bulge nucleotide was built of the previously built nucleotide and CCD chain-closure [see part *(d)*] was applied to connect to the newly built nucleotide. Building the extra-helical bulge ensured that a viable bulge conformation existed (if none existed, then the model was discarded). The extra-helical bulge nucleotide suite is then assigned the Rosetta "*virtual_rna_residue*" variant type, which causes all the atoms in the bulge nucleotide suite to become Rosetta "virtual" atom. The effect of this action was to ignore the normal energetic contribution of the atoms in the bulge nucleotide suite and to instead give a fixed energy bonus to account for conformational entropy (see below). Furthermore, this ensured that the atoms in the bulge nucleotide suite would not sterically impede other nucleotides including ones built to subsequent steps.

(d) Chain closure, for building steps that result in full-length loops.

Chain closure was required at several points in the calculations. When the single-nucleotide or the bulge-skip building step corresponded to building the last remaining nucleotide in the loop (i.e. completing the loop), the chain will need to be closed. Both building steps were modified as follows to allow for chain closure. First, for the *chain-closure+single-nucleotide building step*, conformations that failed the filter for base-stacking or base-pairing interactions were *not* discarded, to allow for the possibility that the new nucleotide is an extra-helical bulge. Second, for both the *chain-closure+single-nucleotide building step* and the *chain-closure+bulge-skip building step*, the van der Waals interaction filter was modified so that a nucleotide conformation was discarded only if the repulsive component (*fa_rep*) of the van der Waals interaction was worse than 10.0 Rosetta units. Chain closure was carried out via cyclic coordinate descent (CCD) (7) on the torsion angles of closing nucleotide's backbone suite ($\varepsilon, \zeta, \alpha, \beta, \gamma$). Two additional filters were applied to discard conformations that did not have the chain properly closed. First, before CCD chain closure was applied, a fast distance filter was used to discard conformations in which the distance between the C5′ atom of the nucleotide 3′ of the chain-break and O3′ atom of the nucleotide 5′ of the chain-break was greater than the theoretical maximum distance (4.63 Å) or less than the theoretical minimum distance (2.00 Å). After the attempted chain-closure with CCD, nucleotides with chain conformations that were not properly closed were discarded, i.e., if the O3′-P bond distance deviated greater than 0.15 Å from the ideal value (1.593 Å) or $(\Delta\theta_1/8.5°)^2 + (\Delta\theta_2/5.7°)^2 \geq 5.0$, where $\Delta\theta_1$ and $\Delta\theta_2$ are the deviations of the C3′-O3′-P and O3′-P-O5′ bond angles from their ideal values (119.8° and 109.0°, respectively). Subsequent minimization of loop torsion angles (see above) with the Rosetta *linear_chainbreak* energy term (see below) further improved the geometry at the closed nucleotide suite.

(e) Clustering

For each partially built sub-region, the generated models were clustered together with stringent criteria. Two models were grouped together if their all-heavy-atom RMSD over the whole loop motif was less than 0.7 Å, their all-heavy-atom RMSD computed over each individual nucleotide was less than 1.0 Å, and if both their north/south sugar pucker classification and Rosetta "*virtual_rna_residue*" variant type (see above) classification matched over each nucleotide. For the final *clustering step*, all the models from every full-length loop regions were clustered together with more relaxed criteria. Two models were grouped together if their all-heavy-atom RMSD over the whole loop motif was less than 2.0 Å, their all-heavy-atom RMSD computed over each individual nucleotide was less than 2.5 Å, and if their Rosetta "*virtual_rna_residue*" variant type classification matched over each nucleotide. Loop nucleotides with Rosetta "*virtual_rna_residue*" variant type were also excluded from the RMSD calculations. The lowest energy member of each cluster was then designated as the cluster center. For each partially built sub-region, the 1000 lowest energy cluster centers were kept and used as starting structures for the next building step along the pathway. For the final *clustering step*, the five lowest energy cluster centers were considered as the final predicted models.

Time complexity of stepwise assembly

The time complexity of the stepwise assembly method is *O(N)*, meaning that the computational time required by the method grows linearly with *N*, the number of nucleotides in the loop. From the above description of the building steps of the Stepwise assembly method**,** we find that for a loop of length *N*, there are (2*N*−2), (4*N*−2) and (2*N*−1) instances of the *single-nucleotide building step* (including + *chain closure*), the *bulge-skip building step* (including + *chain closure*) and *clustering step* respectively. Hence the number of steps grows linearly with *N*. The computational time required to run each *single-nucleotide building step* and the *bulge-skip building step* are comparable and both are much greater than the computational time required to run the *clustering step*. Since only the lowest 1000 energy models are carried forward at each sub-region, there is a fixed upper bound on the number of input structures for each building step. This means that the computational time required to run each building step remains approximately constant and since the total number of building steps (both *single-nucleotide* and *bulge-skip)* increases linearly with *N*, the total computational time increases linearly with *N* as well.

FARFAR

FARFAR models were generated by fragment assembly followed by full-atom refinement in the Rosetta framework, as described previously (2); the fragment source was the large ribosomal subunit of *H. marismortuii* (PDB: 1JJ2), filtered to remove loops with evolutionary kinship to targets in our benchmark. 250,000 models were generated per motif, and clustered together as in the final *clustering step* of SWA. The lowest energy member of each cluster was designated as the cluster center. The five lowest energy cluster centers were then considered as the predicted models. To ensure a rigorous comparison to the SWA results above, the same torsion angles were sampled (ε and ζ of the nucleotide 5′ to the loop; all torsions inside the loop, which was built with Rosetta ideal bond lengths and bond angles; α*,* β*,* and γ of the nucleotide 3′ to the loop; and all 2′-OH torsions). Fragment assembly of loops requires a transient chainbreak that is iteratively closed after each move (8, 9); as in SWA, we chose cutpoint locations at any of the possible loop suites with equal probability, and carried out CCD loop closure. Explicit command line examples for FARFAR loop modeling are below [see ***Generating FARFAR models (command lines)*** subsection].

Generation of the optimized experimental model

To fairly compare the computed energies of crystallographic (experimental) models and *de novo* models, both models need to be optimized with the Rosetta all-atom energy function over the same degrees of freedom with similar amounts of computational power. We optimized the experimental loop structure using three different methods. First, 50,000 FARFAR models were generated using only fragments derived from the crystallographic loop. Second, 50,000 FARFAR models were generated using the standard fragment library doped with crystallographic fragments. Third, SWA was carried out on the loops, but focused on conformations near the crystallographic models; a filter was imposed at each building step requiring models to be within 2 Å RMSD of the crystallographic conformation. Finally, all models generated by these three methods that were within 1.5 Å all-heavy-atom RMSD of the crystallographic model were selected out and the lowest energy model among them was taken as the optimized experimental model (also referred to in the main text as the optimized experimental loop conformation).

Updates to the Rosetta all-atom energy function

Updates were made to the Rosetta all-atom energy function (2). The energy unit reported herein is in Rosetta units (RU), which is used internally by the Rosetta program to store evaluated energy values. Comparisons to RNA Watson-Crick helix thermodynamic parameters (10) indicate that 1 Rosetta unit is approximately equal to 1 k_BT (2) [the structure modeling results given in the main text do not depend on the absolute scale of this energy unit]. Compared to prior work (2), the internally used Rosetta weighting factor of two energy terms were modified:

a. *Torsional Potential term*: The Rosetta weighting factor of the torsional potential term *rna_torsion* was poorly constrained in the original optimization, with weighting factors ranging from 0.1 to 5 giving similar accuracies in FARFAR *de novo* modeling (2). The original weighting factor was chosen to be 0.1 and with this weighting factor, the torsional potential gave negligible energetic contribution; for example the energetic difference between the energy minima and the maxima for each torsional angle was on the order of 0.1 RU for most torsions. In this work, we increased the Rosetta weighting factor of this energy term from 0.1 to 2.9.

b. *Side-chain-side-chain and long-range backbone-side-chain hydrogen bond terms*: The Rosetta weighting factor for both of the terms *hbond_sc* and *hbond_bb_sc* were set to 2.4, instead of 3.4 in (2); this change in the weighting factor was found to slightly improve the results obtained in (2).

Additionally, three new energy terms were introduced:

a. *Linear chainbreak term*: This term was introduced to penalize conformations in which the chain does not properly close at the chain-closure location. Analogous to loop closure in Rosetta protein modeling (8, 9), chain-closure was assessed and optimized by computing *linear_chainbreak,* the summed distances of virtual atoms OVL1 and OVL2 appended to the 5′ nucleotide with the P and O5′ of the 3′ nucleotide, and of virtual atom OVU1 prepended to the 3′ nucleotide with O3′of the 5′ nucleotide.

b. *Base-stacking term*: Previously, the only term that energetically favored base-stacking conformations was the attractive component of the van der Waals interaction. The term *fa_stack* was introduced to approximately model the energetic contribution of π-π dispersion interactions beyond attraction already captured in the Rosetta van der Waals term; its functional form was set to approximately reproduce quantum mechanical calculations on parallel benzene-benzene dimers (11). For two heavy atoms in two different nucleobase, define d to be the inter-atom displacement vector, z to be the vector projection of d onto the first base normal, and $\cos \varphi = |z|/|d|$. An attractive potential $g(\cos \varphi) \times f(|d|)$ was applied, where $g(\cos \varphi) = \cos^2\varphi$ ensures parallel stacking of bases, and $f(|d|) = -0.025$ for $|d| < 4$ Å, interpolating by a standard cubic spline to $f(|d|) = 0.0$ for $|d| > 6$ Å. A reciprocal term was applied for the second base.

c. *Extra-helical bulge bonus term*: Conformational entropy is not explicitly calculated in the Rosetta energy function; in prior protein and nucleotide work, conformational entropies were assumed to be similar for all well-packed conformations. Nevertheless, RNA loops and non-canonical motifs often contain extra-helical bulges, which retain conformational fluctuations compared to nucleotides that are involved in base-pairing and/or base-stacking interactions. To account for this, extra-helical bulge nucleotides were given a fixed bonus *rna_bulge* (here, –4.5 RU; similar results were achieved with –3.0 to –6.0 RU). In SWA, a nucleotide suite (the nucleotide plus the phosphate group of its 3′ neighbor) was assigned to be a bulge if (1) the nucleotide was the bulge in the *bulge-skip building step* or (2) if the nucleotide was built in the *single-nucleotide + chain-closure building step* and did not pass the filter for base-stacking or base-pairing interactions. In FARFAR, a nucleotide suite was assigned to be a bulge if the fixed-bulge bonus outweighed the total energetic contribution of the nucleotide suite (assessed by assigning the Rosetta "*virtual_rna_residue*" variant type to the nucleotide suite and recalculating the Rosetta energy of the structure). We did not model consecutive extra-helical bulges in either SWA or FARFAR, due to their rarity in experimental structures (see ***Modeling consecutive extra-helical bulges*** subsection of Supporting Results below**)**.

Modeling peripheral regions

To optimize the computational speed of both the SWA and FARFAR algorithms, especially for cases taken from the large ribosomal subunit, we modeled peripheral regions well outside the loop motif only as steric (nucleotides in which every atoms were beyond distance d from the loop; d =10 Å for the rRNA cases and d =5 Å for the other cases). We retained a 3D grid of these "steric-only" regions and screened out conformations in which any loop nucleotide contained more than 2 atom-atom pair clashes (distance less than sum of the atom van der Waal radii minus 1.2 Å) with any of these peripheral nucleotides. This optimization was applied to both SWA and FARFAR to ensure comparability between the two methods. On three test cases (J2/4 loop of the TPP riboswitch, 3′ J5/5a loop of group I intron and J2/3 group II intron), we ran SWA with and without this optimization and found excellent agreement. For example, the pair-wise loop RMSD between the best cluster center of the SWA run with and without the optimization was less than 0.25 Å for all three cases.

Extending the SWA method to treat hairpins and multiple-stranded loops

The stepwise assembly method was extended to treat hairpins (loop beginning and ending at the same helix) as well as multiple-stranded loops. Briefly, the overall motif building calculation was ordered into N^2 building stages corresponding to each continuous subregion of the target motif of length N nucleotides. The same single-nucleotide building, bulge-skip building, chain-closure, and the models clustering steps described above were implemented. The canonical base pairs at the edges of the motif were assumed to be known *a priori* (2) and were built using idealized geometry (12). An extensive benchmark testing the SWA method on hairpins and multiple-stranded loops is underway; initial results are presented in Fig. S8.

Generating stepwise assembly models (command line scripts):

Documentation of the stepwise assembly code is being made available within the Rosetta software. For completeness, we provide here explicit examples of command line scripts for modeling the three-nucleotide loop in the C7.2 tetraloop receptor. The same scripts were used in modeling loop motifs in the benchmark.

a. The entire SWA loop building process can be formulated as a directed acyclic graph (DAG) (13) with the command files for each individual building step automatically set up with the Python script `setup_SWA_RNA_dag_job_files.py`. For the C7.2 tetraloop receptor case, the setup script command line is:

```
setup_SWA_RNA_dag_job_files.py -s template.pdb –fasta C7_2_target.fasta -sample_res 11 12 13 -
nstruct 1000 -num_slave_nodes 250 -single_stranded_loop_mode True
```

The "`-s`" flag specifies the template structure. In this case, the known crystallographic model of the P4-P6 *Tetrahymena* group I intron (PDB: 2R8S) was used as the template. Specifically, coordinates of nucleotides in the L5b tetraloop and J6a/J6b receptor regions of 2R8S were copied into `template.pdb` [nts 149-154, 221-224, 227-228 and 246-252 in conventional P4-P6 numbering (see Fig. S7C)]. The A-A platform (nts 225 and 226) was omitted since it will be replaced by the three-nucleotide GUA loop in the C7.2 model. We then mutate the G227-U247 base pair into a C-G base pair (to match the C7.2 sequence) by performing base mutations inside Rosetta (preserving the glycosidic torsion and the sugar-phosphate backbone). The "`–fasta`" flag specifies the fasta file containing the full sequence of the target structure:

```
> C7_2_target.fasta
ggaaacuccuguacuagaugga
1   5    10   15   20
```

The "`-sample_res`" flag specifies the missing nucleotides in the template structure that will be modeled *de novo*; the nucleotides "`11 12 13`" corresponds to the three-nucleotide GUA loop. The "`-nstruct`" flag specifies the number of cluster centers to be carried forward, the "`-num_slave_nodes"` flag specifies the number of CPUs to be allocated to the job, and the "`-single_stranded_loop_mode`" flag specifies that the motif is a single-stranded loop.

Once the DAG files are set up, the Python script `dagman_continuous.py` is then called. This script controls the workflow of the DAG and automatically queues the individual building steps on a high-performance Linux cluster using either the Load Sharing Facility or Condor queuing systems. Examples of the automatically generated command lines for each individual step are provided below.

b. Example of an automatically generated command line for the single-nucleotide building step (building G11 from the 5′-end):

```
rna_swa_test.<exe> -algorithm rna_sample -database <path to database> -s template.pdb -
out:file:silent REGION_0_1/START_FROM_REGION_0_0/region_0_1_sample.out -cluster:radius 0.5 -
score:weights rna_loop_hires_04092010.wts -fixed_res 1 2 3 4 5 6 7 8 9 10 14 15 16 17 18 19 20
21 22 -rmsd_res 11 12 13 -jump_point_pairs 1-22 -alignment_res 1-22 –fasta C7_2_target.fasta -
global_sample_res_list 11 12 13 -sample_res 11 -input_res 1 2 3 4 5 6 7 8 9 10 14 15 16 17 18
19 20 21 22
```

c. Example of an automatically generated command line for the bulge-skip building step (building U12 with G11 as a extra-helical bulge from the 5′-end):

```
rna_swa_test.<exe> -algorithm rna_sample -database <path to database> -s template.pdb -
out:file:silent REGION_0_2/START_FROM_REGION_0_0/region_0_2_sample.out -cluster:radius 0.5 -
score:weights rna_loop_hires_04092010.wts -fixed_res 1 2 3 4 5 6 7 8 9 10 14 15 16 17 18 19 20
21 22 -rmsd_res 11 12 13 -jump_point_pairs 1-22 -alignment_res 1-22 -fasta C7_2_target.fasta -
global_sample_res_list 11 12 13 -sample_res 12 11 -floating_base true -input_res 1 2 3 4 5 6 7
8 9 10 14 15 16 17 18 19 20 21 22
```

d. Example of an automatically generated command line for the clustering step (after building G11 from the 5′-end):

```
rna_swa_test.<exe> -algorithm rna_cluster -database <path to database> -nstruct 1000 -
clusterer_min_struct 1000 -suite_cluster_radius 1.0 -loop_cluster_radius 0.7 -
clusterer_quick_alignment true  -score:weights rna_loop_hires_04092010.wts -fixed_res 1 2 3 4
5 6 7 8 9 10 14 15 16 17 18 19 20 21 22 -rmsd_res 11 12 13 -jump_point_pairs 1-22 -
alignment_res 1-22 -fasta C7_2_target.fasta -sample_res 11 -input_res 1 2 3 4 5 6 7 8 9 10 14
15 16 17 18 19 20 21 22 -in:file:silent REGION_0_1/start_from_region_0_0_sample_filtered.out -
in:file:silent_struct_type binary_rna -out:file:silent region_0_1_sample.cluster.out -
silent_read_through_errors
```

Generating FARFAR models (command lines):

a. The entire FARFAR loop building process can also be formulated as a directed acyclic graph (DAG), albeit a simple one since the FARFAR method creates models in an ‘embarrassingly parallel’ fashion with only one building step needed to complete the whole loop. The python script `setup_FARFAR_RNA_dag_job_files.py` is used to set up the dag files. For the C7.2 tetraloop receptor case, the setup script command line is:

```
setup_FARFAR_RNA_dag_job_files.py -s template.pdb —fasta C7_2_target.fasta -sample_res 11 12
13 -nstruct 250000 -num_slave_nodes 250 -single_stranded_loop_mode True
```

The “`-nstruct`” flag differs from the SWA case above in that it instead specifies here, the total number of models to be generated by FARFAR in an embarrassingly parallel fashion. Once the DAG files are set up, the Python script `dagman_continuous.py` is then called. This script controls the workflow of the DAG and automatically queues the individual building steps on a high-performance Linux cluster using either the Load Sharing Facility or Condor queuing systems.

b. Example of an automatically generated fragment assembly command line to build the entire loop:

```
rna_denovo.<exe> -database <path to database> -fasta C7_2_target.fasta -params_file
cutpoint_closed_13/params -nstruct 200 -cycles 10000 -output_virtual -heat true -close_loops
true -minimize_rna true -out:file:silent cutpoint_closed_13/DAG_ID_0/0/silent_file.out -
score:weights rna_loop_hires_04092010.wts -in:file:silent_struct_type binary_rna -fixed_res 1
2 3 4 5 6 7 8 9 10 14 15 16 17 18 19 20 21 22 -virtual_phosphate_list 1 -allow_bulge_mode true
-allow_bulge_res_list 11 12 13 -rmsd_res 11 12 13 -native_alignment_res 1 22 -s
cutpoint_closed_13/FARFAR_start_cutpoint_13.out -start_silent_tag FARFAR_start_cutpoint_13
```

The file `cutpoint_closed_13/params` contains information regarding which nucleotides are sampled and the loop cutpoint location:

```
OBLIGATE   PAIR 1 22 H H A
ALLOW_INSERT  11 13
CUTPOINT_CLOSED 13
```

In this case, the loop cutpoint location is at the phosphate backbone (O3′-P bond) between nucleotides A13 and C14. The possible `CUTPOINT_CLOSED` values for this 3-nucleotide loop are 10, 11, 12 and 13.

Additional modeling the C7.2 tetraloop receptor with SWA and FARFAR

We first modeled the three-nucleotide GUA loop in the C7.2 tetraloop receptor *de novo* using the standard SWA and FARFAR loop-modeling procedure as described above. While the standard SWA run led to satisfactory loop conformations, the generated models contain an incorrect geometry in the C227-G247 closing base pair (in conventional P4-P6 numbering). This incorrect geometry arises because in the standard loop-modeling procedure, we mutated the G227-U247 base pair into a C227-G247 base pair (to match C7.2 sequence) even though G-U wobble and Watson-Crick base pairs are not perfectly isosteric (14). Furthermore, in the standard loop-modeling procedure, the generated models inherit the relative position and orientation between the P6a and the P6b helix (see Fig. 3C) from the P4-P6 crystallographic model (PDB: 2R8S); however, the correct relative position and orientation between the two helices might be different for the C7.2 mutant. To resolves these issues, we carried out additional runs as follows. First, we replaced the two crystallographic base pairs G227-U247 and U228-A246 in the P6b helix with an idealized two-base-pair C-G/U-A helix. Second, we sampled the relative position and orientation between the P6a and the P6b helix. For SWA, this is accomplished by including building steps that sampled the A6-C7 (see Fig. 3A) loop backbone suite and closed the chain with CCD (7) at the G10-A11 inter-helix backbone suite, and vice versa. This modified run gave similar loop conformations to the standard SWA loop-modeling run. For FARFAR, the G4-U5-A6 loop (including the U3-G4 and A6-C7 backbone suite) was sampled with fragment assembly and the chain was closed with CCD at the G10-A11 inter-helix backbone suite. The SWA and FARFAR models presented in main text Fig. 3B and Fig. S5 were generated using this modified procedure.

Generating RLooM models:

The website (http://rloom.mpimp-golm.mpg.de/) provides a web interface for homology RNA loop modeling with RLooM (database version 12-19-08) (15, 16). The modeling procedure involves two steps.

Step 1: Upload the template PDB to be used for modeling onto the web-server. For modeling the three-nucleotide loop in the C7.2 tetraloop receptor, the known crystallographic model of the P4-P6 *Tetrahymena* group I intron (PDB: 2R8S) was used as the template PDB. The A-A platform (nucleotides 225 and 226 in the conventional P4-P6 numbering) was removed from the template PDB and replaced with the three-nucleotide G-U-A loop in the extended conformation. For modeling the loop motifs in the benchmark, the template PDB is a parsed out segment of the crystallographic model containing the native loop and its surroundings.

Step 2: Input the command line. For example, the command line for modeling the three-nucleotide loop in the C7.2 tetraloop receptor is:

```
<segment>
<anchor>122 :R</anchor>
<anchor>126 :R</anchor>
<query k=ΔH>GUA</query>
</segment>
```

Nucleotides 122 and 126 (224 and 227 in the conventional P4-P6 numbering) are the 5′ and 3′ anchor respectively. GUA is the sequence of the target loop. Δ_H is the maximum allowed sequence dissimilarity as given by the number of differing bases between the homolog loops found by RLooM and the target loop sequence (we started the search at Δ_H=0 and increase this parameter until at least a total of five models were generated).

The default value for the maximum anchor-atom-RMSD (4.00 Å) and clash thresholds (5.00 Å) were used. By default, RLooM does not mutate the sequence of the generated loops to match the sequence of the target loop, so we performed this base mutation step inside Rosetta (preserving the glycosidic torsion and the sugar-phosphate backbone). The RLooM models were ranked first by sequence similarity to the target loop (from high to low) and second by anchor-atom-RMSD (from low to high). The top five ranked models were then taken as the predicted models. For modeling the loop motifs in the benchmark, we included every structure in the PDB as possible homology candidates except for directly related loop structures from the same species as the target loop, i.e., the homolog loop came from the same motif of the same biomolecule of the same species as the native loop. The "0.5 Å with sequence identity" cluster set was used, following ref (15).

Generating ModeRNA models

The ModeRNA (17) modeling python module (version 1.6.0) was downloaded from http://genesilico.pl/moderna/. ModeRNA models were then generated following the instructions located at http://genesilico.pl/moderna/tutorial. Inside the python interpreter:

```
>from moderna import *
>t = load_template('template.pdb', 'chain_ID')
>a = load_alignment('alignment.fasta') m = create_model(t,a)
>m.write_pdb_file('ModeRNA_model.pdb')
```

ModeRNA was used to model the three-nucleotide G4-U5-A6 loop (see Fig. 3A) in the C7.2 tetraloop receptor. The template structure was the known crystallographic model of the P4-P6 *Tetrahymena* group I intron (PDB: 2R8S), with the A-A platform (nucleotides 225 and 226 in the conventional P4-P6 numbering) removed. The inputted `chain_ID` was ‘`R`’ and the following alignment sequence was used:

```
>target (C7.2 mutant):
GGAAUUGCGGGAAAGGGGUCAACAGCCGUUCAGUACCAAGUCUCAGGGGAAACUUUGAGAUGGCCUUGCAAAGGGUAUGGUAAUAAGCUGACGG
ACAUGGUCCUAACACGCAGCCAAGUCCUGUACUCAACAGAUCUUCUGUUGAGAUGGAUGCAGUUCA

>template (PDB:2R8S with A-A platform removed):
GGAAUUGCGGGAAAGGGGUCAACAGCCGUUCAGUACCAAGUCUCAGGGGAAACUUUGAGAUGGCCUUGCAAAGGGUAUGGUAAUAAGCUGACGG
ACAUGGUCCUAACACGCAGCCAAGUCCU---GUCAACAGAUCUUCUGUUGAUAUGGAUGCAGUUCA
```

Fig. S5F shows the ModeRNA model for the C7.2 mutant at the remodeled tetraloop receptor region.

Generating atom-atom pair clash list with Molprobity

The *clashlist* shell script [downloaded from http://kinemage.biochem.duke.edu/software/scripts.php, also implemented as part of the MolProbity program (18, 19)], was used to identify clashed atom-atom pairs in the C7.2 tetraloop receptor models generated by SWA, FARFAR, RLooM and ModeRNA. First, the original hydrogen atoms in the PDB file (if any existed) were removed and the program REDUCE (4) was then used to add hydrogen atoms back into the models. Second, MolProbity was used to revise the PDB to version 2.3 standards. Third, the *clashlist* shell script was run with the optional “–stdbonds” flag [*clashlist* does not identify severe steric clashes, where the two atom-centers are very close together, if this flag is omitted (19)].

Generating single nucleotide conformations for clusters counting

The conformations used for the clusters counting exercise (see Fig. S2) were generated in the following way. First, an anchor adenosine nucleotide was built in the A-form conformation. A ‘moving’ adenosine nucleotide was then attached to the 3′-end of the anchor nucleotide. All backbone torsions between the anchor and moving nucleotide (α, β, γ, ε ,ζ) and the glycosidic torsion (χ) of the moving nucleotide were sampled in 20° intervals using the sampling procedure of the single-nucleotide building step (see ***Stepwise assembly in Rosetta*** subsection), leading to the generation of 5,388,768 conformations. The 2′-OH torsion was neglected since it only determines the position of the 2′-OH hydrogen, a non-heavy atom. These conformations were then filtered for steric clashes both intra-nucleotide and between the moving nucleotide and the A-form anchor (a conformation was discarded if four or more atom-pairs have their van der Waal radii overlap by more than 0.5 Å) leaving a total 2,183,676 clash-free conformations, which were then clustered.

RNA structure mapping through chemical modification

DMS and CMCT reactivity data of the wild type P4-P6 RNA and the C7.2 P4-P6 mutant were acquired at single-nucleotide resolution, as described previously (20). Briefly, preparation of DNA templates, *in vitro* transcription of RNAs, DMS and CMCT chemical mapping, and capillary electrophoresis were carried out in 96-well format, accelerated through the use of magnetic bead purification steps, as has been described previously. Data were analyzed with the HiTRACE software package (21); background subtraction and correction for attenuation of reverse transcription products were carried out as in (20); and figures prepared in MATLAB. The P4-P6 sequence used in the chemical accessibility experiments were:

1. Wild type P4-P6 domain of *Tetrahymena* Group I Intron:

5′-ggccaaaacaacGGAAUUGCGGGAAAGGGGUCAACAGCCGUUCAGUACCAAGUCUCAGGGGAAACUUUGAGAUGGCCUUGC
AAAGGGUAUGGUAAUAAGCUGACGGACAUGGUCCUAAC**C**ACGCAGCCAAGUCCUAAGUCAACAGAUCUUCUGUUGAUAUGGAUG
CAGUUCAaaaccaaaccaaagaaacaacaacaacaac-3′

2. C7.2 mutant:

5′-ggccaaaacaacGGAAUUGCGGGAAAGGGGUCAACAGCCGUUCAGUACCAAGUCUCAGGGGAAACUUUGAGAUGGCCUUGC
AAAGGGUAUGGUAAUAAGCUGACGGACAUGGUCCUAAC**C**ACGCAGCCAAGUCCUGUACUCAACAGAUCUUCUGUUGAGAUGGAU
GCAGUUCAaaaccaaaccaaagaaacaacaacaacaac-3′

The upper-case regions correspond to the P4-P6 domain, and flanking sequences designed to avoid base pairing interactions are shown in lowercase. The primer binding site is underlined. Two of the six replicate measurements were carried out on the more thermostable P4-P6 variants (22, 23) in which C209 (bold and red) was deleted; no changes were observed in chemical reactivity beyond nucleotides 208-210, and these data were averaged with data for C209-containing P4-P6 variants.

Supporting Results

Analysis of RLooM models

RLooM successfully modeled 2 of the 15 loop motifs in the benchmark (Table S2). RLooM modeled these 2 success cases by taking advantage of the similarity between the native loops (from 23s rRNA of *H. marismortui*) and homologous loops (at corresponding positions in 23s rRNA of *E. coli* and *D. radiodurans*). However, the failure to recover the native loop structure in the majority of the cases (13 of 15) suggested that homology RNA loop modeling is not a generally applicable strategy (given the limited number of RNA structures currently in the PDB). Some RNA structures in the benchmark including the SAM-II riboswitch and the M-box riboswitch currently do not contain homologous structures in the PDB (excluding directly related structures from the same species; last checked on April 27, 2011). Furthermore, even in cases where homologous structures exist in the PDB, the native and corresponding homolog loops might differ significantly sequence-wise and/or structure-wise. The corresponding homolog loop might also be completely missing in the homologous structures. For example, the J5/5a hinge motif in the benchmark was taken from the structure of the *Tetrahymena* Group I Intron. Homologous *Twort* and *Azoarcus* Group I Intron structures exist in the PDB, however, the J5/Ja hinge motif does not exist in these homologs. Another example is the J2/4 loop in the *A. thaliana* TPP riboswitch structure. The homologous *E. coli* TPP riboswitch structure contains a loop at the corresponding position, however significant sequence differences (UUGAA vs. UAUCA) prevent this homolog loop from being correctly identified. Lastly, for three of the five 23S rRNA loops in the benchmark (loop nucleotides 531-536, 1976-1985 and 2534-2540), RLooM fails to detect the correct homolog loops in the PDB. While correct homolog loops do in fact exist in the RLooM database, they are defined as sub-regions of larger loops in the database.

For modeling the C7.2 tetraloop receptor, the five top-ranked RLooM models were generated. However, all the models were either invalidated by the chemical accessibility data and/or contained severe steric clashes and anomalous empty cavities in the core of the structure. Fig. S5E shows the first-ranked RLooM model for the C7.2 mutant at the remodeled tetraloop receptor region. MolProbity identified 63 atom-atom pair clashes in this model (excluding those inherited from the crystallographic model).

In modeling the three-nucleotide loop, the 5′ anchor (U3) does not form a Watson-Crick or G-U wobble base pair and this might affect RLooM's performance (for rationale, see (15)). We therefore also modeled the 4-nucleotide U3-G4-U5-A6 loop (see nucleotide numberings in Fig. 3A). However, each of the five top-ranked RLooM models generated by this alternative method also contained severe steric clashes and/or was not consistent with the chemical accessibility data.

Modeling consecutive extra-helical bulges

The stepwise assembly currently does not model consecutive extra-helical bulges. While it is feasible to include additional building steps to model consecutive extra-helical bulges, we choose not to do so because of the rarity of consecutive bulges in experimental structures. For example, the entire 23s rRNA structure (PDB: 1S72; chain 0) contains only 12 instances of consecutive bulges (nts 1029-1030, nts 1604-1605, nts 1652-1653, nts 2344-2345, nts 2588-2589 and nts 2849-2850), constituting less than 0.5% of the total 2754 nucleotides. For this analysis, extra-helical bulges were annotated by the program MC-annotate (24) (i.e. search for nucleotides that are both unpaired and unstacked).

Confidence selection of the C7.2 tetraloop receptor model

In assessing the performance of the SWA and FARFAR methods on the 15-loop benchmark, the five lowest energy cluster centers were considered as the predicted models. In contrast, we selected out the lowest energy model (i.e. the first cluster center) as the single predicted model for the C7.2 tetraloop receptor blind prediction. The following evidences provide justifications for why we can confidently select out the SWA (but not necessarily the FARFAR) lowest energy model as our single predicted C7.2 model:

i.) Accurate energy discrimination among short loops: We observe that there is accurate energy discrimination among the SWA models when the loops are sufficiently short (≤4 nts). For example, there is good agreement between the SWA lowest energy model and the crystallographic conformation for all 3 short (≤4 nts) loops in the benchmark (≤ 1.5 Å RMSD, see Table S5). The analogous condition does not hold for FARFAR models (≤ 1.5 Å RMSD for only 1 of 3 loops, see Table S3).

ii.) Control run on the A-A platform: As a consistency check, we have also carried out a control run on the A-A platform from the classic 11-nt tetraloop receptor (PDB: 2R8S). We found that for this control loop, there is again good agreement between the SWA lowest energy model and the crystallographic model (RMSD=0.28 Å; see Fig. S4). We note that there is good agreement between the FARFAR lowest energy model and the crystallographic model as well (RMSD=0.58 Å).

iii.) An extensive SWA C7.2 run: We have also carried out a more extensive SWA calculation modeling eight nucleotides of the C7.2 tetraloop receptor (nts 3-7 and 10-12 in Fig. 3A). This serves as another consistency check since the more extensive 8-nt SWA model should agree with the 3-nt SWA model presented in the main text if the model is indeed correct. We found that there is good agreement between the lowest energy model of the extended 8-nt SWA run and the 3-nt SWA run (pairwise RMSD= 1.19 Å). The extend 8-nt SWA model recovered the crucial *trans* Sugar-edge/Watson-Crick G4-A6 base pair and the extra-helical bulged U5 found in the 3-nt SWA model. In contrast, the lowest energy model of the extended 8-nt FARFAR run and the 3-nt FARFAR run do not agree (pairwise RMSD= 2.90 Å).

Lastly, we note that the SWA modeling occurred *before* comparisons to chemical accessibility data. The chemical accessibility data were not used as input in the selection process and hence the agreement between the SWA lowest energy model and the subsequently measured chemical accessibility data serve as an experimental validation to our blind prediction.

FR3D search on the C7.2 tetraloop receptor

The FR3D program (25) can search motifs using both geometric (RMSD-like) and/or symbolic (relation-based) criteria. The search for the same stranded G4-A6 motif in the C7.2 tetraloop receptor was carried out in three stages. In the first stage, we performed a purely symbolic search for all *trans* Sugar-edge/Watson G-A base pairs (including near-matches). The search was carried out on a non-redundant list of RNA crystallographic models in the PDB with resolution better than 4 Å (from the FR3D website) and found 253 candidates. In the second stage, an additional requirement that the G and the A nucleotide be sequentially separated by exactly 1 nucleotide (i.e. *5´-GNA-3´*) was imposed and this reduced the number of candidates to 13. In the third stage, we then performed a geometric search to calculate the geometric discrepancy between the coordinates of the G4 and A6 nucleotides in the SWA C7.2 tetraloop receptor model and the 13 candidates. Only 2 candidates have geometric discrepancy below the 0.50 Å default cutoff value. These two candidates, the malachite green aptamer (PDB: 1F1T) and the UUGUAU RNA bound to the human cleavage factor protein Im (PDB: 3MDG) were reported in the main text.

The constraint that the G and the A nucleotide be sequentially separated by exactly 1 nucleotide in the FR3D search was necessary to ensure that template-based methods (RLooM and ModeRNA) would be able to discover the template loop in the PDB database. These template-based methods work by inserting the conformation of entire loops excised from template structures into the target structure. Hence the length (and to a lesser extend, the sequence identity) of the template loop needs to match that of the target loop. Finally, RLooM and ModeRNA also use the coordinates of the nucleotides located immediately 5′ and 3′ of the loop as anchor points and hence in the final FR3D search, we have included the coordinates of the 5′ and 3′ anchor nucleotides (U3 and C7) in the FR3D geometric discrepancy calculation.

Analysis of the DMS and CMCT chemical accessibility data

DMS and CMCT chemical accessibility measurements (see Fig. S7D) were acquired in 10 mM $MgCl_2$, 50 mM HEPES, pH 8.0, at 24 °C. Two groups of nucleotides were excluded from all analysis. The first group consisted of nucleotides at the 5′-end and 3′-end of the P4-P6 molecule [G102, G103, A104, A105, U106, U259, C260 and A261 in conventional P4-P6 numbering (see Fig. 3C)], since their native conformation in solution is unlikely to be accurately represented by their crystallographic conformation [they are likely unstructured single-stranded region in solution but are observed to form significant artificial crystal contacts in the P4-P6 crystallographic model (PDB: 2R8S)]. The second group consisted of nucleotides that have highly variable reactivity values between replicates and/or high reactivity values to both the CMCT and DMS probe (due to high background stops). This includes G118, G119, U120, G150, C208 and C213 (colored gray in see Fig. S7D). The chemical accessibility data were averaged over six replicate measurements and then were normalized so that the mean reactivity of all A and C nucleotides in each DMS dataset and all G and U nucleotides in each CMCT dataset equals 1.0 unit.

i.) Uridines with high CMCT reactivity at the N3 position: We found that a CMCT reactivity above than 1.25 units indicated that the uridine nucleotide is unpaired and unstacked (i.e. an extra-helical bulge). In the folded wild type P4-P6 RNA CMCT dataset, we found that there were seven uridine nucleotides with CMCT reactivity above than 1.25 units [U130, U179, U185, U199, U236, U238 and U239]. All seven were observed to be extra-helical bulges in the P4-P6 crystallographic model (PDB: 2R8S). The folded C7.2 P4-P6 mutant CMCT dataset also detected the same seven uridine nucleotides with the 1.25 units cutoff. The 1.25 units cutoff is 15 times higher than the mean reactivity of all internal Watson-Crick base-paired uridines in the two datasets (mean reactivity=0.08 ± 0.03 units; U142, U144, U157, U182, U190, U221, U228, U241, U243 and U244). Focusing on the new C7.2 G4-U5-A6 loop (see Fig. 3A), the high CMCT reactivity of U5 (1.88 ± 0.40) indicates that U5 is an extra-helical bulge. This constraint rejects the first-ranked ModeRNA model (Fig. S5F) in favors of the first-ranked SWA, FARFAR and RLooM model (Figs. S5A, B and E).

As further support for the high CMCT reactivity corresponding to extra-helical bulge, we note that the high reactivities of the seven extra-helical bulges were consistent with the high-calculated accessible surface area (ASA) at the N3 atom (using a 3.0-Å radius for the CMCT probe; see next section). The crystallographic conformation of each of these seven nucleotides gives a calculated ASA value greater than 15.0 $Å^2$ with a mean ASA value of 21.5 $Å^2$ (see next section for the calculation details).

ii.) Adenosine with low DMS reactivity at the N1 position: We found that a DMS reactivity below 0.60 units indicated that the adenosine's N1 atom at the Watson-Crick edge is protected [i.e. has zero accessible surface area by the DMS probe, using a 1.8-Å radius probe; see next section]. In the folded C7.2 P4-P6 mutant DMS dataset excluding the teraloop/receptor region, there are eleven adenosine nucleotides with DMS reactivity below 0.60 units (A133, A136, A159, A161, A192, A196, A230, A231, A233 and A246). All eleven were all true positives and have 0.0 $Å^2$ ASA at the N1 position as their Watson-Crick edges are protected by base-pairing interactions in the P4-P6 crystallographic model (PDB: 2R8S).

Outside the tetraloop/receptor region, the folded wild type P4-P6 CMCT dataset agrees with folded C7.2 P4-P6 mutant dataset to within the experimental errors and gave similar results. All eleven true positives from the C7.2 dataset were detected in the wild type dataset. The wild type DMS dataset, however, does detect two additional adenosines, A187 and A252, with DMS reactivity below 0.50 units. A252 (WT: 0.32 ± 0.16, C72: 1.04 ± 0.48) forms a Watson-Crick base pair and have 0.0 $Å^2$ ASA at N1. A187 (WT: 0.50 ± 0.26, C72: 0.67 ± 0.45), however, does have its Watson-edge exposed (3.69 $Å^2$ ASA at N1) and thus is either a false positive (or a difference between the crystallographic and solution structure). Hence, there are a total of 12 unique true positives and 1 false positive in the two datasets.

We noticed that for each of the 12 true positive cases, the adenosine not only base pairs with its Watson-Crick edge but also forms base-stacking interactions *both* below and above its base plane (i.e. internally stacked). The false positive A187 also internally stack as well. However, internal stacking alone appears to be insufficient to protect the adenosine nucleotide from DMS methylation. Aside the false positive A187, there are thirteen other adenosines that are internally stacked but have their Watson-edge unpaired in the crystallographic model (A113, A114, A115, A139, A140, A171, A173, A183, A198, A207, A218, A219, A256). All thirteen have DMS reactivity above the 0.60 units cutoff and hence serve as true negative controls. The mean reactivity among these thirteen negative controls is also high (2.60 ± 0.30 in the C7.2 mutant dataset and 2.40 ± 0.25 in the wild type dataset). The high reactivity is consistent with the observation that stacking alone does not completely occlude the N1 position; all thirteen adenosine have non-zero ASA at N1 and a mean ASA value of 6.76 Å^2, which is about one-third the maximum possible ASA at N1 (18 Å^2).

With thresholds set for DMS reactivity, we focused on the C7.2 G4-U5-A6 loop (see Fig. 3A). The low DMS reactivity of A6 (0.43 ± 0.11 units) indicated that its N1 atom should be protected and has zero accessible surface area by the DMS probe. This constraints further rejects the first-ranked FARFAR and ModeRNA models (1.42 Å^2 and 1.63 Å^2 ASA at the N1 position of A6 respectively; see Figs. S5B and E) and favors the first-ranked SWA model (see Fig. 3B/Fig. S5A). In the first-ranked SWA model (i.e. the global lowest energy model), the N1 atom of A6 forms a hydrogen bond with the 1H2 amino proton of G4 (as part of a *trans* Sugar-edge/Watson-Crick base pair) and has 0.0 Å^2 ASA.

iii.) Guanosine with low CMCT reactivity at the N1 position. We found that a CMCT reactivity below 0.35 units indicated that the guanosine's N1 atom at the Watson-Crick edge was protected [i.e. has zero accessible surface area by the CMCT probe, using 3.0 Å probe radius; see next section]. In the folded C7.2 P4-P6 mutant DMS dataset excluding the tetraloop/receptor region, there were twenty-nine nucleotides with DMS reactivity below 0.35 units [G108, G116, G117, G129, G134, G141, G147, G148, G158, G160, G163, G164, G175, G180, G181, G188, G191, G194, G195, G200, G201, G212, G215, G220, G227, G234, G242, G245 and G254]; all twenty-nine are true positives and have 0.0 Å^2 ASA at the N1 position.

Outside the tetraloop/receptor region, the folded wild type P4-P6 CMCT dataset again agreed with folded C7.2 P4-P6 mutant dataset to within experimental errors and gave similar results. Twenty-seven of the twenty-nine true positives from the C7.2 dataset have reactivity below the 0.35 units cutoff in the wild type dataset as well. The exceptions were G195 (WT: 0.53 ± 0.49, C72: −0.02 ± 0.83) and G200 (WT: 0.38 ± 0.30, C72: 0.26 ± 0.49). The wild type dataset also detected five other guanosines with DMS reactivity below 0.35 units [G110 (WT: 0.25 ± 0.06, C72: 0.54 ± 0.18), G111 (WT: 0.24 ± 0.10, C7.2: 0.39 ± 0.19), G126 (WT: 0.19 ± 0.12, C7.2: 0.38 ± 0.24), G174 (WT: 0.23 ± 0.11, C7.2: 0.37 ± 0.19) and G251 (WT: 0.23 ± 0.13, C72: 0.61 ± 0.20)]; again all five are true positives and have 0.0 Å^2 ASA at the N1 position. Hence, there are a total of thirty-four unique true positives and zero false positives from the two datasets.

Among the thirty-four true positives, the majorities have their N1 atom at the Watson–Crick edge protected through a Watson-Crick or a G-U wobble base-pairing interaction. There are five exceptions, G126, G163, G164, G234 and G254, which have their N1 atom occluded in other ways. G126 forms a non-canonical *trans* W.C/W.C base pair with A196. The N1 atom of G163 and G254 are occluded by the phosphate group of A139 and A217 respectively (forming hydrogen bonds between the guanosines' NH1 imino proton and the phosphates' O2P oxygen atom). The hydroxyl oxygen of C138 and the amino group at the N3 position of A178 occlude the N1 atom of G164. Lastly, the N1 atom of G234 is occluded by the base of C240.

Finally, we would like to note that the evidences in support of the guanosine relationship are as statistically significant as those supporting the relationships for uridine and adenosine in cases *i.)* and *ii.)*. The reason is that, unlike in the two prior relationships, there are only two negative controls available for this relationship. In the entire P4-P6 crystallographic model, there are only two guanosines with non-zero ASA at N1, G150 (ASA=1.16 Å^2) and G169 (ASA= 11.87 Å^2). Furthermore, G150 have highly variable reactivity values between replicates and high reactivity values to both the CMCT and DMS probe (due to high background stops; see Fig. S7D). While both G150 and G169 do have CMCT reactivity above the cutoff (WT: 1.59 ± 0.53, C7.2: 4.134 ± 1.13 and WT: 0.48 ± 0.13, C72: 0.53 ± 0.14 respectively), we would have more confidence in the guanosine relationship if there were more negative controls.

With the thresholds set for guanosine CMCT reactivities, we focused on the C7.2 G4-U5-A6 loop (see Fig. 3A). The low DMS reactivity of G4 (0.27 ± 0.13 units) indicates that its N1 atom should be protected and has zero accessible surface area by the CMCT probe. The only remaining model, the first-ranked SWA model (Fig. S5A), is also consistent with this final constraint (0.0 Å^2 ASA at N1 of G4); in the first-ranked SWA model, the N1 atom of G4 is occluded by the ribose of A10 and A11 and the intervening phosphate group (forming a hydrogen bond between the G4 NH1 imino proton and the A11 ribose's O4 oxygen atom)

Accessible surface area (ASA) calculations

All the accessible surface area (ASA) calculations in this paper were carried out on the GETAREA webserver (26). Based on reference (27), an effective radius of 3.0 Å was assumed for the CMCT probe in the ASA calculations at the N1 and N3 atom of guanosine and uridine respectively. Based on the modeling of methyl cation which is the attacking species of DMS [see (28)], an effective radius of 1.8 Å was assumed for the DMS probe in the ASA calculations at the N1 atom of adenosine. Lastly, during the calculation process, we noticed that extra-helical bulges sometime artificially occlude and reduce the calculated ASA of atoms in neighboring nucleotides. Since extra-helical bulges are unpaired and unstacked, the conformation that appear in a static structure such as a crystallographic model would represent only one possible 'snapshot' of a wide range of conformations in which the bulge can occupies. For this reason, we decided to remove all extra-helical bulges from the structures before the ASA calculations. The extra-helical bulges removed are A125, U130, U179, U185, U199, U236, C237, U238, U249 and C255 in conventional P4-P6 numbering (see Fig. 3C) and for the C7.2 receptor models, any additional extra-helical bulge that exists in the G4-U5-A6 loop (see Fig. 3A). The only exception is when the ASA calculation is on a nucleotide that is itself an extra-helical bulge, in which case that particular nucleotide was included back into the structure.

Variations in the tetraloop receptor sequence in the vitro selection experiment

Subsequent to carrying out the prediction of the C7.2 tetraloop receptor structure with SWA and experimentally validating this prediction through chemical mapping measurements, we discovered further evidence in support of the SWA model from sequence variations in the original *in vitro* selection experiment that isolated the C7.2 receptor (29). The predicted *trans* Sugar-edge/Watson-Crick G4-A6 base pair belongs to the isoteric group I6.2 (14) and is only isosteric to two other base pairs (tSW A-C and tSW C-A), both of which have only a single base-base hydrogen bond. The G-A base pair with two base-base hydrogen bonds is expected to be more stable and hence we would expect to see few, if any, substitutions at both the G4 and A6 positions. In contrast, the predicted U5 extra-helical bulge is unpaired and thus should be replaceable by other nucleotides. Examining all four sequenced clones in class IC [which contains the C7.2 mutant; see (29)], we find that indeed, G4 and A6 remained invariant, while sequence substitution of U5 with A and C were observed, giving further validation of the SWA model.

References

1. Richardson JS*, et al.* (2008) RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). RNA 14(3):465.
2. Das R, Karanicolas J, & Baker D (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. Nat Methods 7(4):291-294.
3. Leaver-Fay A*, et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 487:545-574.
4. Word JM, Lovell SC, Richardson JS, & Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 285(4):1735-1747.
5. Das R & Baker D (2007) Automated de novo prediction of native-like RNA tertiary structures. Proceedings of the National Academy of Sciences 104(37):14664-14669.
6. Kuffner JJ (2004) Effective sampling and distance metrics for 3D rigid body path planning. *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, pp 3993-3998 Vol.3994.
7. Canutescu AA & Dunbrack RL, Jr. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci 12(5):963-972.
8. Mandell DJ, Coutsias EA, & Kortemme T (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods 6(8):551-552.

9. Wang C, Bradley P, & Baker D (2007) Protein-protein docking with backbone flexibility. J Mol Biol 373(2):503-519.

10. Xia T, *et al.* (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry 37(42):14719-14735.

11. Sato T, Tsuneda T, & Hirao K (2005) A density-functional study on pi-aromatic interaction: benzene dimer and naphthalene dimer. J Chem Phys 123(10):104307.

12. Lu XJ & Olson WK (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res 31(17):5108-5121.

13. Maurer SB (2003) Directed Acyclic Graphs. *Handbook of Graph Theory*, eds Gross JL & Yellen J (CRC Press, Boca Raton), pp 142-155.

14. Stombaugh J, Zirbel CL, Westhof E, & Leontis NB (2009) Frequency and isostericity of RNA base pairs. Nucleic Acids Res 37(7):2294-2312.

15. Schudoma C, May P, Nikiforova V, & Walther D (2010) Sequence–structure relationships in RNA loops: establishing the basis for loop homology modeling. Nucleic Acids Research 38(3):970-980.

16. Schudoma C, May P, & Walther D (2010) Modeling RNA loops using sequence homology and geometric constraints. Bioinformatics 26(13):1671-1672.

17. Rother M, Rother K, Puton T, & Bujnicki JM (2011) ModeRNA: a tool for comparative modeling of RNA 3D structure. Nucleic Acids Research.

18. Chen VB, *et al.* (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 66(Pt 1):12-21.

19. Word JM, *et al.* (1999) Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol 285(4):1711-1733.

20. Kladwang W, Cordero P, & Das R (2011) A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. RNA 17(3):522-534.

21. Yoon S, *et al.* (2011) HiTRACE: high-throughput robust analysis for capillary electrophoresis. Bioinformatics 27(13):1798-1805.

22. Ye JD, *et al.* (2008) Synthetic antibodies for specific recognition and crystallization of structured RNA. Proc Natl Acad Sci U S A 105(1):82-87.

23. Juneau K & Cech TR (1999) In vitro selection of RNAs with increased tertiary structure stability. RNA 5(8):1119-1129.

24. Gendron P, Lemieux S, & Major F (2001) Quantitative analysis of nucleic acid three-dimensional structures. J Mol Biol 308(5):919-936.

25. Sarver M, Zirbel C, Stombaugh J, Mokdad A, & Leontis N (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. Journal of Mathematical Biology 56(1):215-252.

26. Fraczkiewicz R & Braun W (1998) Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. Journal of Computational Chemistry 19(3):319-333.

27. Holbrook SR & Kim SH (1983) Correlation between chemical modification and surface accessibility in yeast phenylalanine transfer RNA. Biopolymers 22(4):1145-1166.

28. Lavery R & Pullman A (1984) A new theoretical index of biochemical reactivity combining steric and electrostatic factors. An application to yeast tRNAPhe. Biophys Chem 19(2):171-181.

29. Costa M & Michel F (1997) Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: comparison with in vivo evolution. EMBO J 16(11):3289-3302.

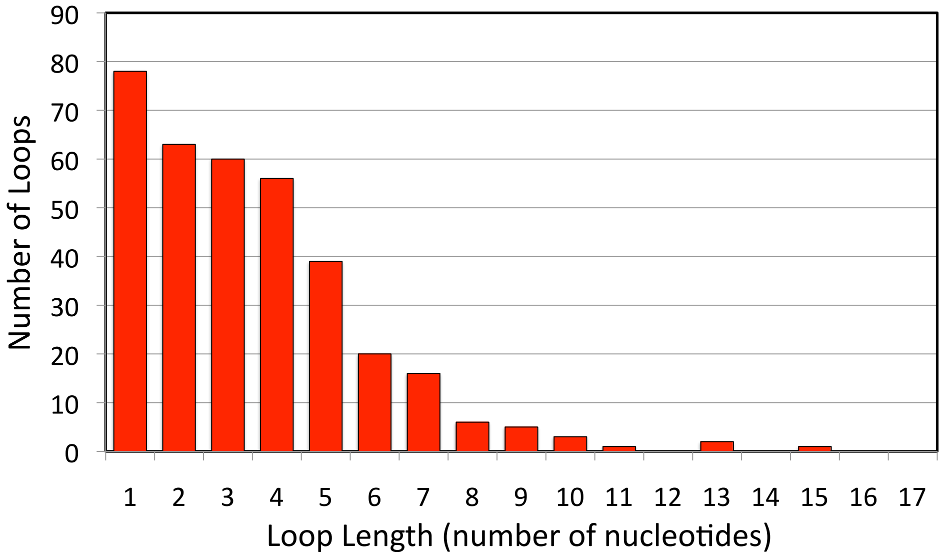


Figure S1. RNA loop length distribution in the 23S rRNA. A total of 350 single-stranded RNA loop segments separated by Watson-Crick and/or G-U wobble base pair(s) are found in the structure of the *Haloarcula marismortui* large ribosomal subunit (PDB: 1JJ2). The mean loop length is 3.5 nucleotides and loops longer than 10 nucleotides in length are rare (occur in <1%; 4 out of 350).

A

All-atom RMSD cluster radius (Å)	# clash-free cluster centers
4.0	69
3.5	129
3.0	294
2.5	803
2.0	2802
1.5	13784
1.0	124483
0.9	219,177
0.8	416,801
0.7	863,743
0.6	2,003,118
0.5	5,417,396
0.4	18,307,070
0.3	87,982,647
0.2	804,098,805
0.1	35,318,316,979

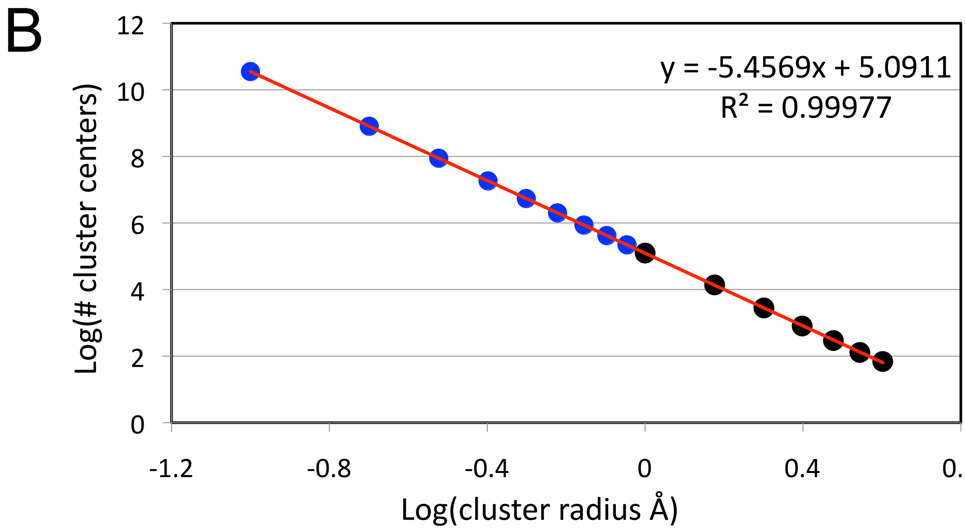


Figure S2. Number of conformation cluster centers in a nucleotide versus the RMSD cluster radius. (A) All sterically available conformations of a single RNA nucleotide were clustered together and the table reports the number of cluster centers as a function of the all-heavy-atom RMSD clustering radius. Clustering with sub-Angstrom threshold – as is necessary for high resolution modeling – leads to millions of unique clusters of single-nucleotide conformations. (B) Log-log plot of the number of cluster center versus the all-atom RMSD cluster radius. Data with cluster radius greater than or equal to 1.0 Å were directly generated (black font in table and black points in plot). However, for smaller cluster radius values, the computation became infeasible and hence was determined through a least-squares interpolation (blue font in table and blue points in plot). Complete description of how the conformations used for this clusters counting exercise were generated is provided in the Supporting Methods: ***Generating single nucleotide conformations for clusters counting*** subsection.

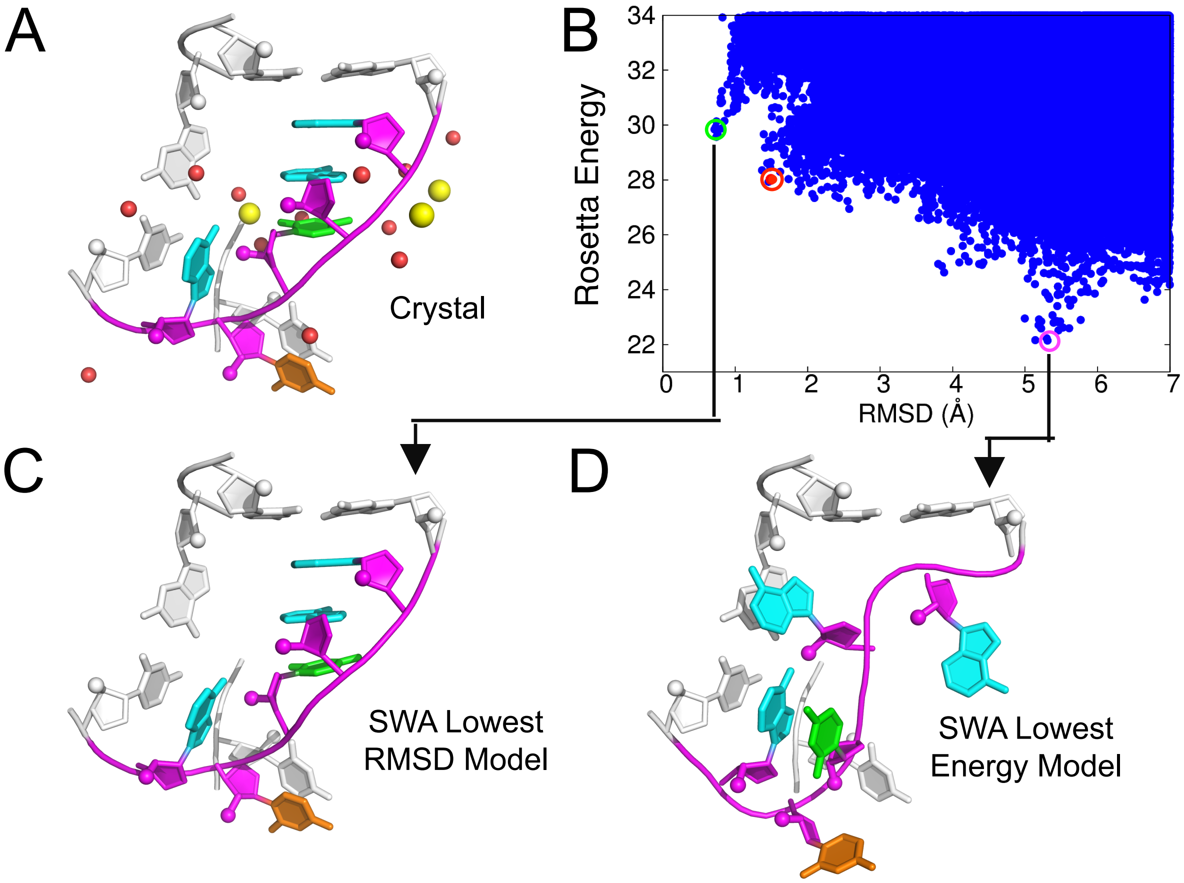


Figure S3. Limitations in modeling accuracy are no longer due to conformational sampling. Modeling of a five-nucleotide loop from the hepatitis C virus IRES subdomain IIa. (A) Crystallographic model (PDB: 2PN4). Bound divalent Sr^{2+} cations (colored yellow) are proposed to stabilize the loop through both direct hydrogen bonds as well as through hydrogen bonds mediated by tightly bound water molecules (O atoms colored red) [see ref, Acta Crystallogr D Biol Crystallogr. 2008 Apr;64(Pt 4):436-43]. (B) Stepwise assembly successfully samples the native conformation of the loop to atomic accuracy (0.71 Å RMSD), but the model (C) is not ranked within one of the five lowest energy clusters. (D) The Rosetta all-atom energy function incorrectly assigns numerous non-native models (>5.0 Å RMSD) with significantly lower energies (~6 k_BT) than the optimized experimental model (red point in B).

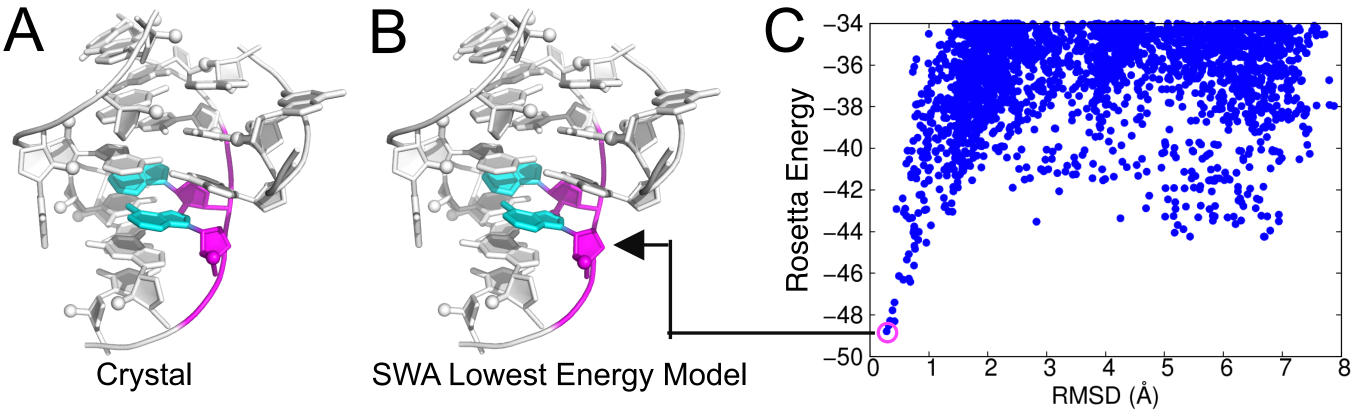


Figure S4. SWA modeling of the A-A platform in the wild type P4-P6 RNA. (A) Crystallographic model of the A-A platform in the classic 11-nt tetraloop receptor of the P4-P6 domain of the *Tetrahymena* ribozyme (PDB: 2R8S). (B-C) The stepwise assembly lowest energy model is within atomic accuracy (0.28 Å RMSD) of the crystallographic model. The A-A platform modeling serves as a control before modeling the GUA loop in the C7.2 P4-P6 mutant.

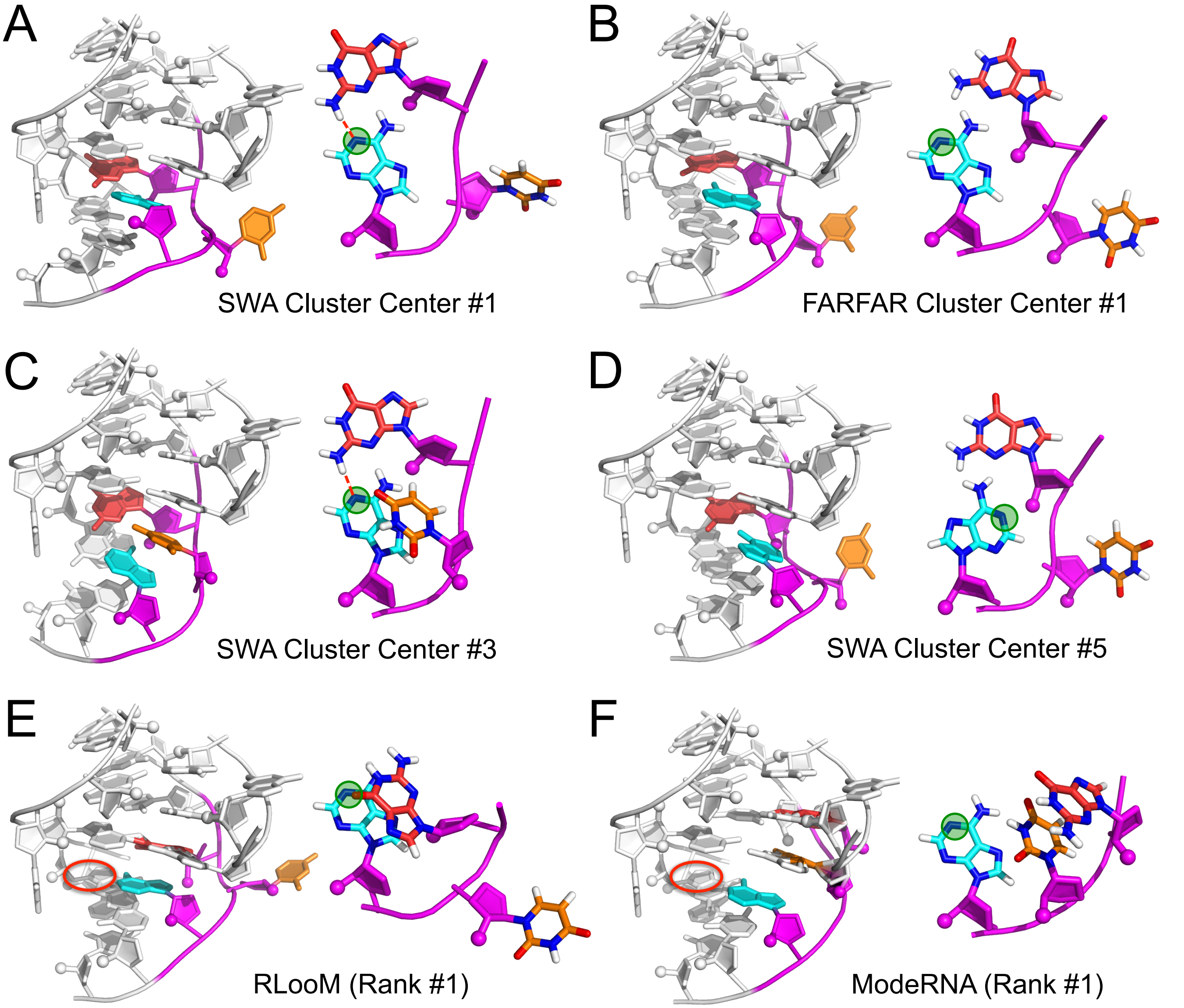
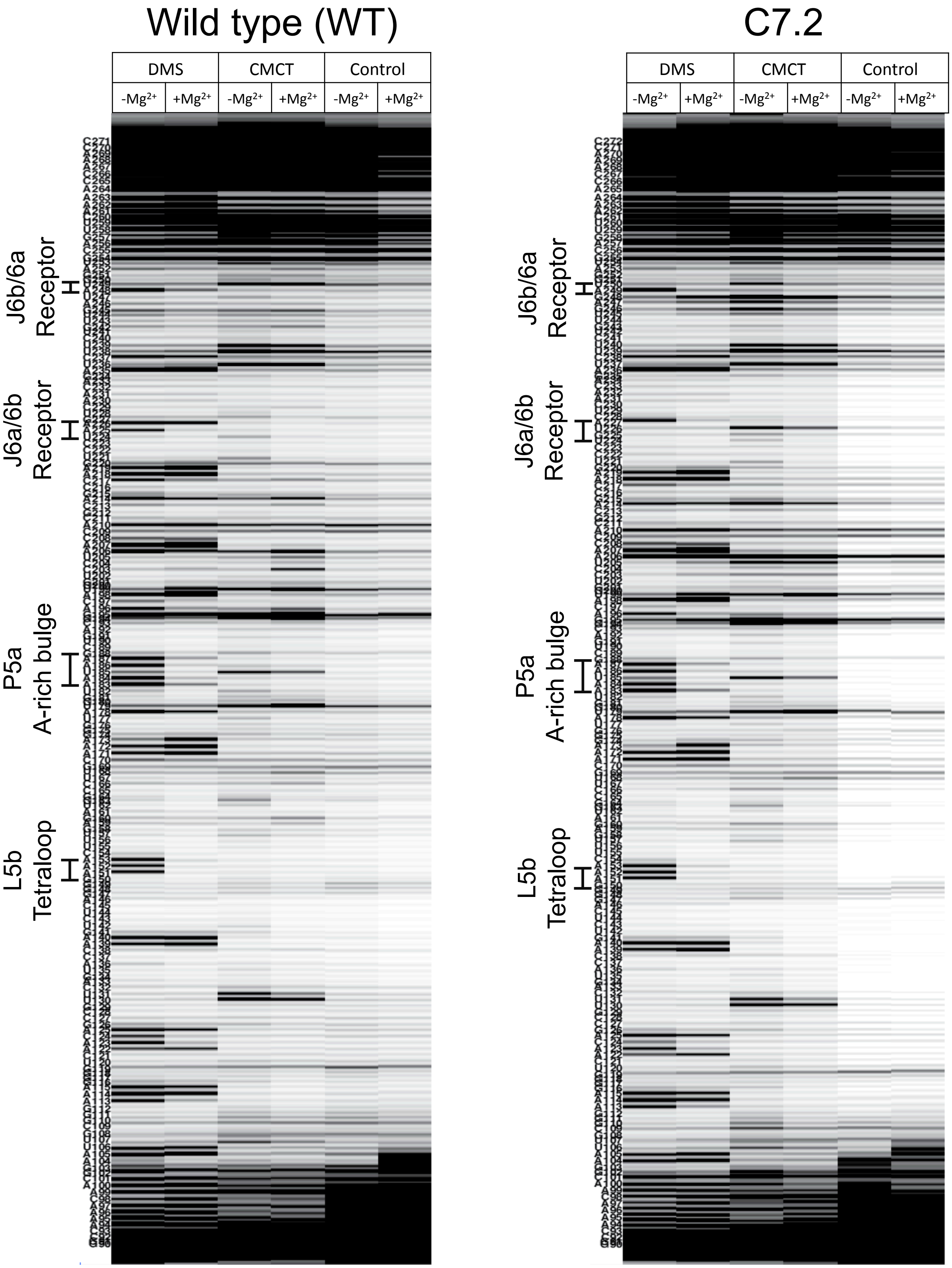


Figure S5. Comparisons of C7.2 tetraloop receptor models. (A) In the SWA lowest energy model (cluster center #1), the N1 atom (highlighted with green circle) of the A6 nucleotide (blue) forms a hydrogen bond with the 1H2 amino proton of the G4 nucleotide (red) and hence is protected from DMS methylation. (B) FARFAR lowest energy model (cluster center #1) adopts a different loop conformation from the SWA first cluster center and is worse in energy by $3.0k_BT$. In this model, the N1 atom of A6 is exposed and hence accessible to DMS methylation. (C) SWA cluster center #3 (worst in energy than the SWA first cluster center by $3.6k_BT$). In this model, the N1 atom of A6 forms a hydrogen bond similar to the SWA first cluster center but U5 (orange) internally stacks, which is inconsistent with nucleotide's high CMCT reactivity. (D) SWA cluster center #5 (worse in energy than the SWA first cluster center by $6.7k_BT$). Aside from the SWA first cluster center, this is the only other SWA cluster center among the top five that is consistent with the chemical accessibility data. In this model, G4 and A6 forms a *cis* Sugar-edge/Watson-Crick base pair and the A6 base adopts the syn conformation. The N1 atom of A6 is occluded by the hydroxyl group of G4 (forming a hydrogen bond with the 2′-OH hydrogen). Details on the CMCT and DMS data analysis and comparisons to accessible surface area calculations are provided in the Supporting Results, ***Analysis of the DMS and CMCT chemical accessibility data*** subsection. (E) RLooM model. (F) ModeRNA model. Both the RLooM and ModeRNA models are unlikely to be correct due to significant steric clashes. MolProbity identified 63 and 133 atom-atom clashes in the RLooM and the ModeRNA model respectively (excluding clashes inherited from the crystallographic model). In contrast, MolProbity identified only 1 atom-atom clash in the SWA model (A) and only 3 atom-atom clashes in the FARFAR model (B). The RLooM and ModeRNA models also have anomalous empty cavities in their cores (red circle).


Wild type (WT)

C7.2

Figure S6. Capillary electropherograms of chemical mapping experiments on the wild type and C7.2-substituted P4-P6 RNAs. Modifications were read out by high throughput reverse transcription with 5´-fluorescently labeled radiolabeled primers and capillary electrophoresis. Raw fluorescence traces (arbitrary units) are shown after automated alignment and normalization to mean intensity. Shorter products (higher electrophoretic mobility) appear at the top. Both DMS alkylation and CMCT carbodiimide modification measurements were acquired at 24 °C in 50 mM HEPES, pH 8.0, and for the +Mg^{2+} lanes, 10 mM $MgCl_2$. The 'Control' lanes contained no chemical modifier and provided background estimates for the measurements.

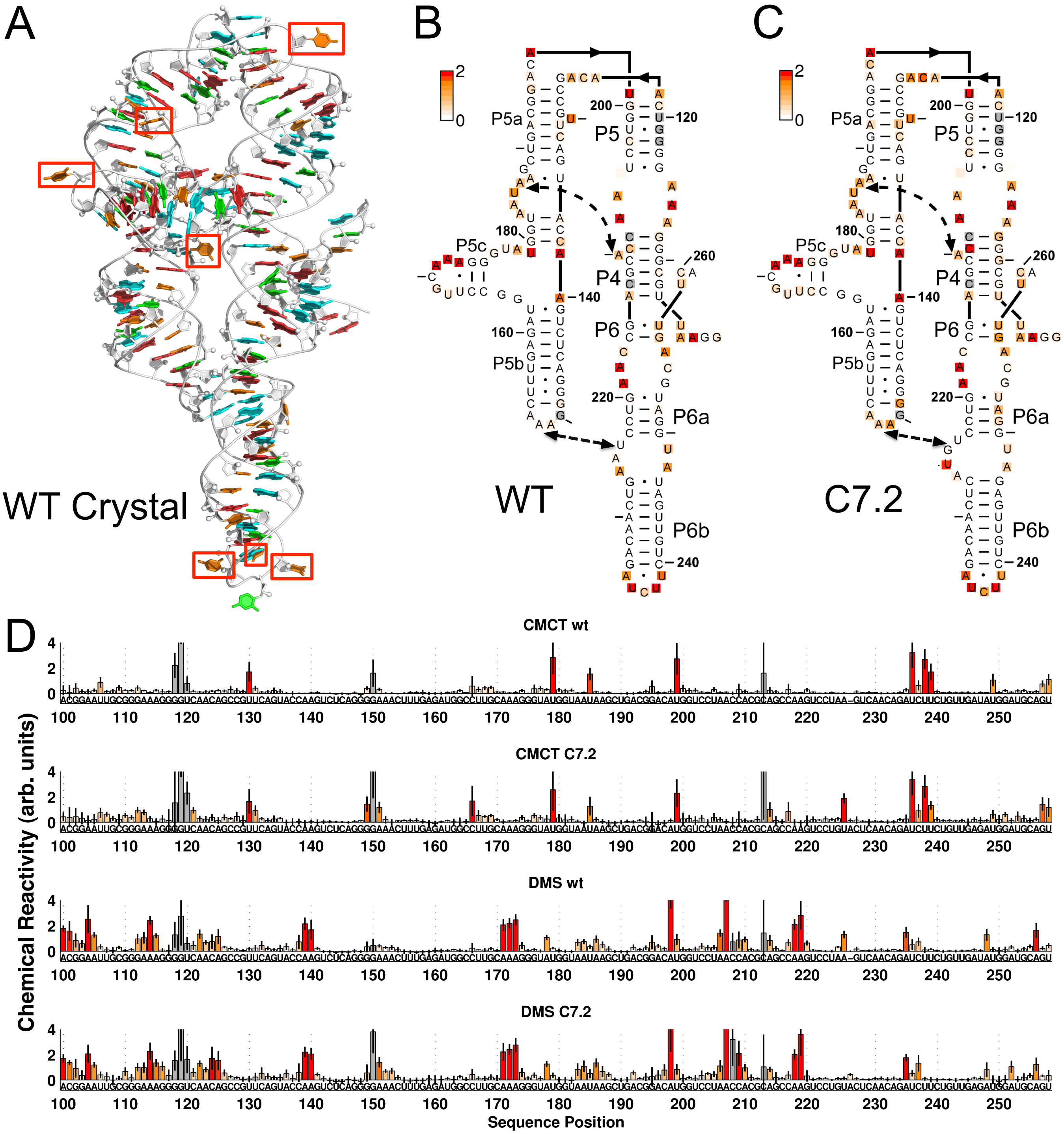


Figure S7. Chemical reactivity of the C7.2 tetraloop receptor, grafted into the P4-P6 RNA. (A) Crystallographic model of the P4-P6 domain of the *Tetrahymena* ribozyme (PDB: 2R8S), with the seven uridines outside the receptor region that achieve CMCT modification rates of at least 15-fold above background highlighted in red boxes. (B) Chemical reactivities of A and C (based on DMS alkylation) and G and U (based on CMCT carbodiimide modification) for the wild type P4-P6 RNA; measurements were acquired in 10 mM $MgCl_2$, 50 mM HEPES, pH 8.0, at 24 °C. Data were normalized so that the mean reactivity of all A and C nucleotides in each DMS dataset and all G and U nucleotides in each CMCT dataset equals 1.0 unit. Nucleotides with high variability between replicates and/or high reactivities to both the CMCT and DMS probe (due to high background stops) are colored in gray. (C) Chemical reactivities of the C7.2 mutant of the P4-P6 RNA. (D) DMS and CMCT reactivities plotted as separate bar graphs, with same coloring as (B) & (C). Sequence positions are given in conventional P4-P6 numbering. Error bars give standard deviations over six replicate measurements.

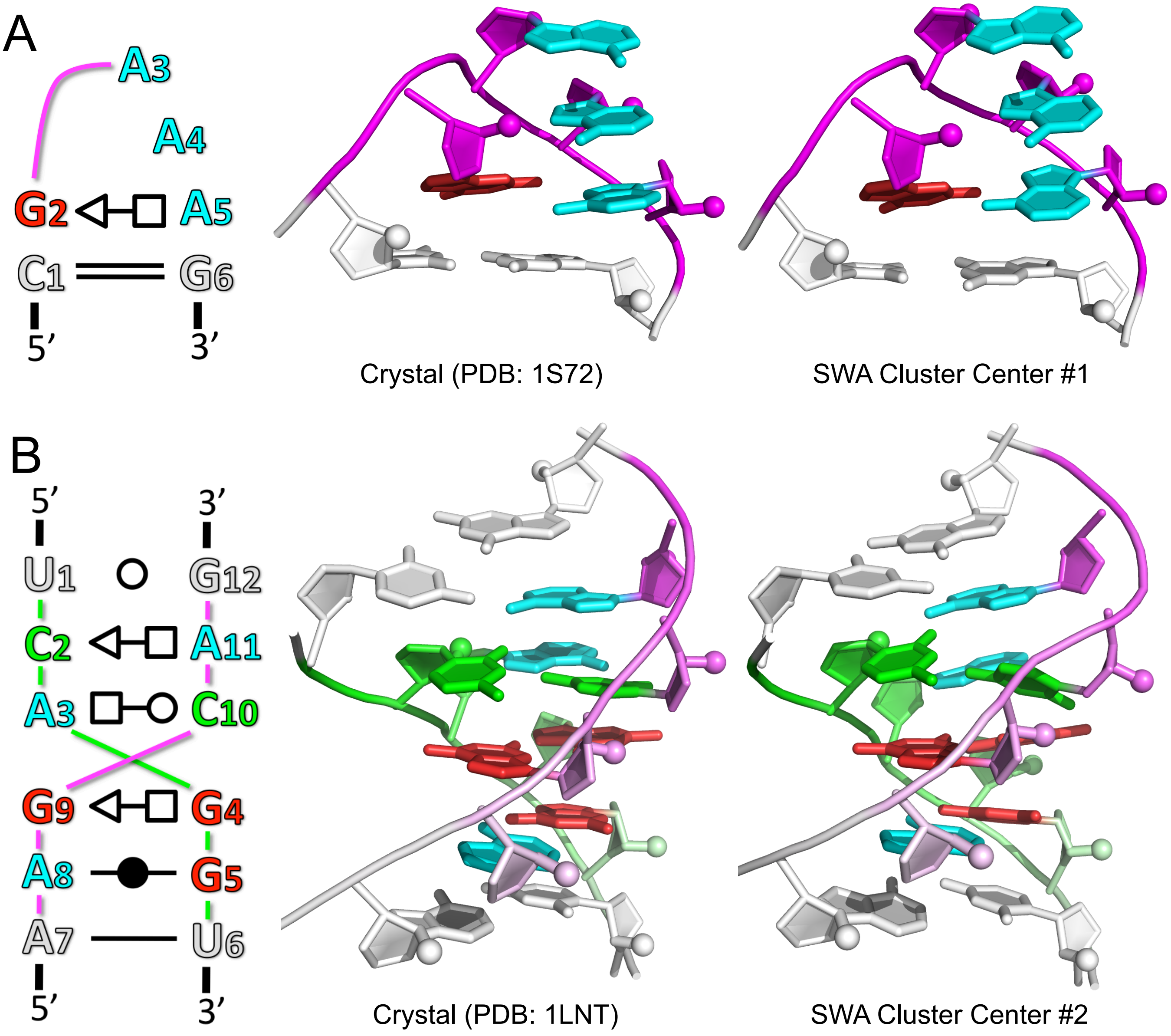


Figure S8. Extension of the SWA method to treat hairpins and multiple-stranded loops. SWA was able recapitulate the native conformation of (A) a 4-nt GAAA tetraloop hairpin (0.75 Å RMSD) and (B) a 8-nt double-stranded internal loop from the signal recognition particle RNA (0.98 Å RMSD). For the 4-nt GAAA tetraloop hairpin (PDB: 1S72), one correct building pathway is (C1-G6)→G2→A3→A4→A5. For the 8-nt double-stranded internal loop from the signal recognition particle RNA (PDB: 1LNT), one correct building pathway is: (U1-G12) →C2→A11→C10→A3→G4→G9→G5→A8→(U6-A7). Description of the extension of the SWA method to build hairpins and multiple-stranded loops *de novo* is provided in the Supporting Methods.

Table S1. RNA loop motifs benchmark information.

Motif Name	Source		# Nucleotides			# Base Pairs[a]		# Potential Hydrogen Bonds[b]				
	PDB:Chain (Nucleotide segment)	Resolution[c]	Total	Bulge[d]	Outlier[e]	NWC	WC or GU	Total	Base-Base	Base-2′OH	Base-Phos[f]	Others[g]
5′ J1/2, Leadzyme	1NUJ:E (23-26)	1.80 Å	4	0	0	4	0	14	8	2	4	0
5′ P1, M-Box Riboswitch	2QBZ:X (164-167)	2.60 Å	4	1	3	3	0	11	5	3	1	2
3′ J5/5a, Group I Intron	2R8S:R (196-199)	1.95 Å	4	1	1	3	0	10	6	2	0	2
5′ J5/5a, Group I Intron	2R8S:R (122-126)	1.95 Å	5	1	3	4	0	11	7	2	0	2
Hepatitis C Virus IRES IIa	2PN4:A (53-57)	2.32 Å	5	0	1	0	0	4	0	4	0	0
J2/4, TPP Riboswitch	3D2V:B (40-44)	2.00 Å	5	1	2	4	0	17	7	3	5	2
23S rRNA (44-49)	1S72:0 (44-49)	2.40 Å	6	1	0	6	0	21	10	6	5	0
23S rRNA (531-536)	1S72:0 (531-536)	2.40 Å	6	0	1	3	0	16	5	3	5	3
J3/1, Glycine Riboswitch	3OWI:B (75-81)	2.85 Å	7	0	0	3	0	7	3	2	2	0
J2/3, Group II Intron	3G78:A (287-293)	2.80 Å	7	1	3	5	0	20	10	5	1	4
L1, SAM-II Riboswitch	2QWY:B (8-14)	2.80 Å	7	0	2	7	0	18	11	4	0	3
L2, Viral RNA Pseudoknot	1L2X:A (19-25)	1.25 Å	7	1	2	4	0	15	6	6	2	1
23S rRNA (2534-2540)	1S72:0 (2534-2540)	2.40 Å	7	0	2	4	0	19	9	2	5	3
23S rRNA (1976-1985)	1S72:0 (1976-1985)	2.40 Å	10	0	3	6	0	25	8	6	7	4
23S rRNA (2003-2012)	1S72:0 (2003-2012)	2.40 Å	10	2	5	6	2	37	14	8	11	4
Total			94	9	28	62	2	245	109	58	48	30
Per nucleotide			1.00	0.10	0.30	0.66	0.02	2.61	1.16	0.62	0.51	0.32

[a] Base pairs are automatically annotated using the program *MC-annotate* [Nucleic Acids Res. 2002 Oct 1;30(19):4250-63] and follows the scheme of Leontis and Westhof [RNA. 2001 Apr;7(4):499-512]. The non-Watson-Crick base pair column ("NWC") excludes canonical A-U and G-C base pairs as well as G-U wobble base pairs.

[b] Potential polar hydrogen bonds are identified using the following simple geometric criteria: (a) the distance between the hydrogen atom and acceptor atom is less than 2.7 Å and (b) the donor-hydrogen-acceptor angle is greater than 130°.

[c] Resolution of the x-ray diffraction crystallographic data.

[d] Extra-helical bulge nucleotide found to be both unpaired and unstacked in the crystallographic model as annotated by the program *MC-annotate*.

[e] An outlier is a nucleotide suite that is not part of the 46 most commonly observed RNA rotamers defined by the RNA Ontology Consortium [RNA. 2008 Mar;14(3):465-81].

[f] Base-Phos: Polar hydrogen bonds between the base and the phosphate group.

[g] Others: Polar hydrogen bonds between (a) base and O4′ atom, (b) 2′-OH group and 2′-OH group, (c) 2′-OH group and phosphate group or (d) 2′-OH group and O4′ atom.

Table S2. Application of RLooM homology modeling to the loop motifs benchmark.

Motif name	Motif properties		Top five RLooM models[a]		Excluded native source PDB[c]
	Length	PDB	Lowest RMSD[b] (Å)	Models information (template PDB, sequence dissimilarity, anchor-atoms-RMSD (Å))	
5′ J1/2, Leadzyme	4	1NUJ	4.35	(1XJR, 0, 0.94), (1LDZ, 0, 0.94), (1VOW, 0, 1.26), (2B9N, 0, 1.60), (1SML, 1, 0.85)	none
5′ P1, M-Box Riboswitch	4	2QBZ	3.83	(2GYC, 0, 1.27), (1VOW, 0, 1.48), (2GIO, 0, 1.91), (2HGH, 1, 1.29), (1NJN, 1, 1.42)	2QBZ
3′ J5/5a, Group I Intron	4	2R8S	6.97	(3DLL, 0, 1.68), (2ZJP, 0, 1.69), (2VHN, 0, 2.72), (2GYC, 0, 3.20), (3CC2, 1, 1.80)	2R8S, 1X8W, 1GID, 1GRZ, 1L8V, 1HR2
5′ J5/5a, Group I Intron	5	2R8S	5.31	(1PNU, 1, 1.63), (2AW4, 1, 1.64), (1NKW, 1, 1.78), (1NJP, 1, 1.79), (1NJO, 0, 1.813)	2R8S, 1HR2, 1GID, 1L8V, 1X8W, 1GRZ
Hepatitis C Virus IRES IIa	5	2PN4	3.30	(1P5P, 0, 0.65), (1NKW, 1, 2.09), (2O43, 1, 2.19), (2RKJ, 1, 2.38), (2HGJ, 1, 2.518)	none
J2/4, TPP Riboswitch	5	3D2V	2.64	(2GYC, 0, 7.54), (2GYC, 1, 1.29), (1C2W, 1, 1.45), (1C2W, 1, 1.90), (2HGJ, 1, 3.49)	3D2X, 2CKY
23S rRNA (44-49)	6	1S72	**1.42**	(1JZX, 0, 0.18), (1P9X, 0, 0.23), (1NJP, 0, 0.26), (1Z58, 0, 0.28), (2OGN, 0, 0.29),	1YHQ, 1Q86
23S rRNA (531-536)	6	1S72	8.45	(2FEY, 1, 5.85), (1VQ5, 1, 6.25), (1YL3, 2, 1.98), (2JL8, 2, 2.02), (1VSP, 2, 2.03)	none
J3/1, Glycine Riboswitch	7	3OWI	6.76	(2VHP, 1, 4.91), (1C2W, 1, 12.95), (2V47, 2, 2.34), (1YI2, 2, 2.35), (2J01, 2, 2.40)	none
J2/3, Group II Intron	7	3G78	13.33	(2QP0, 1, 5.24), (2VHO, 1, 5.25), (3BBN, 2, 2.77), (2GY9, 2, 2.93), (1N36, 2, 3.03)	none
L1, SAM-II Riboswitch	7	2QWY	8.19	(2VHM, 2, 2.29), (2GYC, 2, 2.95), (1C2W, 2, 7.68), (2VHM, 2, 9.40), (2QBE, 2, 9.48)	none
L2, Viral RNA Pseudoknot	7	1L2X	6.69	(3EGZ, 2, 2.11), (3BBO, 2, 2.65), (1VSP, 2, 2.98), (1FKA, 2, 3.09), (3D5B, 2, 3.85)	none
23S rRNA (2534-2540)	7	1S72	10.19	(2B64, 3, 1.99), (1JZX, 3, 2.06), (1YL4, 3, 2.30), (1KQS, 3, 2.43), (1P9X, 3, 2.65)	none
23S rRNA (1976-1985)	10	1S72	10.09	(1VOR, 5, 2.78), (1VP0, 5, 2.97), (2V46, 5, 4.18), (1P9X, 5, 5.81), (3BBO, 5, 8.27)	none
23S rRNA (2003-2012)	10	1S72	**0.87**	(2AW4, 2, 0.48), (2GYA, 2, 0.73), (2GYC, 2, 1.01), (1C2W, 2, 7.20), (1P9X, 3, 0.32)	1M90
RMSD < 1.50 Å			2/15		

[a] The top five models as ranked first by sequence similarity (from high to low) and second by anchor-atom-RMSD (from low to high). See Supporting Methods for explicit command-line examples used to generate the models. A complete analysis of the results is also given in Supporting Results.

[b] All-heavy-atom RMSD to the crystallographic loop. Nucleotides found to be an extra-helical bulge (both unpaired and unstacked) in the native crystallographic model were excluded from the RMSD calculation. Bold text indicates RMSD within 1.5 Å of the crystallographic model.

[c] We included every structure in the PDB as possible homology candidates except for structures representing the same motif of the same biomolecule of the same species as the target native loop structure. The “0.5 Å with sequence identity” cluster set was used, following ref. [Nucleic Acids Res. 2010 Jan;38(3):970-80].

Table S3. Supplemental benchmark data for *de novo* loop modeling with FARFAR.

Motif Name	Motif Properties		Best of Five Lowest Energy Cluster Centers					Lowest RMSD Model			Lowest Energy Sampled	
	Length	PDB	Cluster Rank	RMSD (Å)	Base-pair Recovery[a]	Base-stack Recovery[a]	Rosetta Energy (RU)	RMSD (Å)	Base-pair Recovery[a]	Base-stack Recovery[a]	Rosetta Energy (RU)	E-Gap to Opt. Exp.[b] (RU)
5′ J1/2, Leadzyme	4	1NUJ	1	1.96	3/4	4/6	-53.6	1.66	3/4	3/6	-53.6	2.7
5′ P1, M-Box Riboswitch	4	2QBZ	3	**0.72**	3/3	4/7	129.0	**0.53**	3/3	5/7	128.9	2.3
3′ J5/5a, Group I Intron	4	2R8S	1	**0.40**	3/3	6/6	-21.4	**0.30**	3/3	6/6	-21.4	**0.0**
5′ J5/5a, Group I Intron	5	2R8S	2	4.08	2/4	3/5	-51.3	**1.05**	3/4	5/5	-52.5	0.3
Hepatitis C Virus IRES IIa	5	2PN4	3	2.11	0/0	6/7	27.3	**1.04**	0/0	6/7	25.4	**-2.6**
J2/4, TPP Riboswitch	5	3D2V	1	6.66	2/4	2/6	-83.4	1.74	2/4	5/6	-83.4	10.8
23S rRNA (44-49)	6	1S72	1	**0.69**	6/6	8/10	-163.8	**0.47**	6/6	8/10	-163.8	2.6
23S rRNA (531-536)	6	1S72	1	3.18	1/3	3/5	-253.8	2.44	0/3	3/5	-253.8	6.9
J3/1, Glycine Riboswitch	7	3OWI	1	**1.13**	3/3	13/14	94.8	**0.71**	3/3	12/14	94.8	2.5
J2/3, Group II Intron	7	3G78	1	1.59	4/5	9/10	-59.0	**1.34**	4/5	9/10	-59.0	8.5
L1, SAM-II Riboswitch	7	2QWY	5	2.43	3/7	5/11	13.4	**1.43**	5/7	9/11	5.3	3.8
L2, Viral RNA Pseudoknot	7	1L2X	4	5.44	1/4	3/9	-183.4	**1.35**	1/4	9/9	-185.6	3.7
23S rRNA (2534-2540)	7	1S72	5	6.39	1/4	3/10	-215.1	3.24	0/4	1/10	-218.0	7.3
23S rRNA (1976-1985)	10	1S72	1	11.19	0/6	3/15	-283.5	5.06	0/6	5/15	-283.5	9.6
23S rRNA (2003-2012)	10	1S72	5	11.36	0/8	2/14	-250.7	5.43	1/8	0/14	-252.9	41.2
AVERAGE	6.3	–	2.3	3.95	2.1/4.3	4.9/9.0	-90.3	1.85	2.3/4.3	5.7/9.0	-91.5	6.6
RMSD < 1.50 Å	–	–	–	4/15	–	–	–	9/15	–	–	–	–
Energy Gap < 0.0	–	–	–	–	–	–	–	–	–	–	–	2/15

[a] Number of native base-pairs and native base-stacks correctly recovered by the *de novo* model. Base pairs and the base stacks are automatically annotated using the program *MC-annotate* [J Mol Biol. 2001 May 18;308(5):919-36]. Base-pairing annotation follows the Leontis and Westhof nomenclature [RNA. 2001 Apr;7(4):499-512] and recovery entails having the correct edge-to-edge interaction (Watson-Crick, Hoogsteen, or Sugar-edge) and local strand orientation (*cis* or *trans*). Counts of correctly recovered base pairs are lowered owing to ambiguities in assigning bifurcated base pairs, pairs connected by single hydrogen bonds and pairs that are not completely co-planar. Base-stacking are classified as either upward, downward, outward or inward [RNA. 2009 Oct;15(10):1875-85] and recovery entails having the correct base-stacking type.

[b] The Rosetta energy of the experimental loop structure was optimized through three different methods (see Supporting Methods for details), and the lowest energy model derived from all three methods was taken as the optimized experimental model. Bold text indicates that the lowest energy sampled by the *de novo* run is lower than the energy of the optimized experimental model (i.e. the energy gap is negative).

Table S3. Supplemental benchmark data for *de novo* loop modeling with FARFAR (continue).

Motif Name	Motif Properties		Cluster Center #1		Cluster Center #2		Cluster Center #3		Cluster Center #4		Cluster Center #5	
	Length	PDB	RMSD (Å)	Rosetta Energy (RU)	RMSD (Å)	Rosetta Energy (RU)	RMSD (Å)	Rosetta Energy (RU)	RMSD (Å)	Rosetta Energy (RU)	RMSD (Å)	Rosetta Energy (RU)
5′ J1/2, Leadzyme	4	1NUJ	1.96	-53.6	2.05	-53.0	2.05	-52.8	2.30	-51.9	2.00	-51.6
5′ P1, M-Box Riboswitch	4	2QBZ	2.85	128.9	3.73	129.0	**0.72**	129.0	5.43	130.7	1.80	130.7
3′ J5/5a, Group I Intron	4	2R8S	**0.40**	-21.4	3.81	-15.7	4.19	-15.4	2.90	-13.4	**1.47**	-12.3
5′ J5/5a, Group I Intron	5	2R8S	4.10	-52.5	4.08	-51.3	4.27	-51.0	5.12	-50.8	6.23	-50.0
Hepatitis C Virus IRES IIa	5	2PN4	6.23	25.4	4.73	26.9	2.11	27.3	4.84	27.3	3.73	27.7
J2/4, TPP Riboswitch	5	3D2V	6.66	-83.4	7.10	-83.3	6.77	-82.6	8.53	-82.5	6.67	-82.4
23S rRNA (44-49)	6	1S72	**0.69**	-163.8	3.77	-161.6	2.44	-160.8	**0.83**	-159.4	**1.31**	-159.0
23S rRNA (531-536)	6	1S72	3.18	-253.8	5.13	-251.9	5.27	-251.5	3.37	-251.2	3.31	-251.1
J3/1, Glycine Riboswitch	7	3OWI	**1.13**	94.8	2.23	95.7	3.38	96.3	**1.21**	98.6	2.19	99.6
J2/3, Group II Intron	7	3G78	1.59	-59.0	2.83	-58.1	4.40	-57.6	3.61	-57.1	4.03	-56.7
L1, SAM-II Riboswitch	7	2QWY	2.96	5.3	2.96	11.0	3.46	11.4	3.23	12.0	2.43	13.4
L2, Viral RNA Pseudoknot	7	1L2X	6.01	-185.6	5.77	-185.5	5.90	-184.7	5.44	-183.4	5.70	-183.4
23S rRNA (2534-2540)	7	1S72	6.74	-218.0	9.34	-216.2	9.19	-215.6	9.37	-215.1	6.39	-215.1
23S rRNA (1976-1985)	10	1S72	11.19	-283.5	12.90	-283.1	12.99	-283.0	15.74	-282.4	13.76	-282.2
23S rRNA (2003-2012)	10	1S72	12.92	-252.9	13.69	-251.9	12.42	-251.5	14.68	-251.1	11.36	-250.7
AVERAGE	6.3	–	4.57	-91.5	5.61	-89.9	5.30	-89.5	5.77	-88.6	4.82	-88.2
RMSD < 1.50 Å	–	–	3/15	–	0/15	–	1/15	–	2/15	–	2/15	–
Energy Gap < 0.0	–	–	–	–	–	–	–	–	–	–	–	

Table S4. Performance of FARFAR with and without doping native fragments into the fragment library.

Motif name	Motif properties		Lowest RMSD model	
			Exclude native fragments (standard)	Dope in native fragments (cheat[a])
	Length	PDB	RMSD (Å)	RMSD (Å)
5′ J1/2, Leadzyme	4	1NUJ	1.66	**0.30**
5′ P1, M-Box Riboswitch	4	2QBZ	**0.53**	**0.37**
3′ J5/5a, Group I Intron	4	2R8S	**0.30**	**0.30**
5′ J5/5a, Group I Intron	5	2R8S	**1.05**	**0.42**
Hepatitis C Virus IRES IIa	5	2PN4	**1.04**	**0.98**
J2/4, TPP Riboswitch	5	3D2V	1.74	**0.45**
23S rRNA (44-49)[b]	6	1S72	**0.47**	**0.38**
23S rRNA (531-536)[b]	6	1S72	2.44	**0.46**
J3/1, Glycine Riboswitch	7	3OWI	**0.71**	**0.49**
J2/3, Group II Intron	7	3G78	**1.34**	**0.54**
L1, SAM-II Riboswitch	7	2QWY	**1.43**	**0.76**
L2, Viral RNA Pseudoknot	7	1L2X	**1.35**	**0.43**
23S rRNA (2534-2540)[b]	7	1S72	3.24	**0.49**
23S rRNA (1976-1985)[b]	10	1S72	5.06	**0.69**
23S rRNA (2003-2012)[b]	10	1S72	5.43	**0.48**
RMSD < 1.50 Å			9/15	15/15

[a] Fragments of the native loop were doped/added into FARFAR's standard fragment library.

[b] FARFAR's standard fragment library is composed of RNA fragments extracted from a single structure of the archaeal large ribosomal subunit (PDB: 1JJ2). To mimick a true *denovo* modeling scenario, we ensure that regions with evolutionary kinship to our benchmark motifs were either absent or removed from the fragment library. For example, 5 motifs in the benchmark came from the PDB: 1S72 which is another archaeal large ribosomal subunit (in fact 1S72 is a revised structure of 1JJ2). Hence when modeling these 5 ribosomal loops, the correspond loop regions in the PDB: 1JJ2 were excised from the fragment library.

Table S5. Supplemental benchmark data for *de novo* loop modeling with SWA.

Motif Name	Motif Properties		Best of Five Lowest Energy Cluster Centers					Lowest RMSD Model			Lowest Energy Sampled	
	Length	PDB	Cluster Rank	RMSD (Å)	Base-pair Recovery[a]	Base-stack Recovery[a]	Rosetta Energy (RU)	RMSD (Å)	Base-pair Recovery[a]	Base-stack Recovery[a]	Rosetta Energy (RU)	E-Gap to Opt. Exp.[b] (RU)
5′ J1/2, Leadzyme	4	1NUJ	3	**0.83**	4/4	6/6	-55.6	**0.51**	4/4	6/6	-57.1	**-0.8**
5′ P1, M-Box Riboswitch	4	2QBZ	1	**0.96**	3/3	6/7	126.1	**0.61**	3/3	7/7	126.1	**-0.5**
3′ J5/5a, Group I Intron	4	2R8S	1	**0.47**	3/3	6/6	-21.4	**0.40**	3/3	6/6	-21.4	**0.0**
5′ J5/5a, Group I Intron	5	2R8S	4	**1.04**	4/4	5/5	-52.8	**0.66**	3/4	5/5	-53.7	**-0.9**
Hepatitis C Virus IRES IIa	5	2PN4	1	5.31	0/0	1/7	22.1	**0.71**	0/0	7/7	22.1	**-5.9**
J2/4, TPP Riboswitch	5	3D2V	4	**0.85**	4/4	6/6	-94.2	**0.73**	4/4	6/6	-95.1	**-1.0**
23S rRNA (44-49)	6	1S72	1	**0.73**	6/6	8/10	-166.4	**0.71**	6/6	8/10	-166.4	**0.0**
23S rRNA (531-536)	6	1S72	5	2.45	2/3	4/5	-260.3	**0.76**	1/3	5/5	-261.3	**-0.6**
J3/1, Glycine Riboswitch	7	3OWI	1	**1.35**	2/3	13/14	93.5	**0.64**	3/3	14/14	93.5	1.3
J2/3, Group II Intron	7	3G78	1	**0.82**	4/5	9/10	-67.7	**0.77**	4/5	9/10	-67.7	**-0.2**
L1, SAM-II Riboswitch	7	2QWY	5	**1.26**	5/7	9/11	3.5	**0.86**	6/7	9/11	0.2	**-1.3**
L2, Viral RNA Pseudoknot	7	1L2X	5	3.36	3/4	9/9	-192.0	**0.91**	4/4	9/9	-193.3	**-4.1**
23S rRNA (2534-2540)	7	1S72	2	5.71	2/4	3/10	-232.6	**1.39**	3/4	8/10	-232.6	**-7.3**
23S rRNA (1976-1985)	10	1S72	5	7.75	0/6	3/15	-301.2	4.58	0/6	5/15	-303.9	**-10.8**
23S rRNA (2003-2012)	10	1S72	2	**0.74**	7/8	14/14	-290.1	**0.64**	7/8	14/14	-290.9	3.2
AVERAGE	6.3	–	2.7	2.24	3.3/4.3	6.8/9.0	-99.3	0.99	3.4/4.3	7.9/9.0	-100.1	-1.9
RMSD < 1.50 Å	–	–	–	10/15	–	–	–	14/15	–	–	–	–
Energy Gap < 0.0	–	–	–	–	–	–	–	–	–	–	–	13/15

[a] Number of native base-pairs and native base-stacks correctly recovered by the *de novo* model. Base pairs and the base stacks are automatically annotated using the program *MC-annotate* [J Mol Biol. 2001 May 18;308(5):919-36]. Base-pairing annotation follows the Leontis and Westhof nomenclature [RNA. 2001 Apr;7(4):499-512] and recovery entails having the correct edge-to-edge interaction (Watson-Crick, Hoogsteen, or Sugar-edge) and local strand orientation (*cis* or *trans*). Counts of correctly recovered base pairs are lowered owing to ambiguities in assigning bifurcated base pairs, pairs connected by single hydrogen bonds and pairs that are not completely co-planar. Base-stacking are classified as either upward, downward, outward or inward [RNA. 2009 Oct;15(10):1875-85] and recovery entails having the correct base-stacking type.

[b] The Rosetta energy of the experimental loop structure was optimized through three different methods (see Supporting Methods for details), and the lowest energy model derived from all three methods was taken as the optimized experimental model. Bold text indicates that the lowest energy sampled by the *de novo* run is lower than the energy of the optimized experimental model (i.e. the energy gap is negative).

Table S5. Supplemental benchmark data for *de novo* loop modeling with SWA (continue).

Motif Name	Motif Properties		Cluster Center #1		Cluster Center #2		Cluster Center #3		Cluster Center #4		Cluster Center #5	
	Length	PDB	RMSD (Å)	Rosetta Energy (RU)	RMSD (Å)	Rosetta Energy (RU)	RMSD (Å)	Rosetta Energy (RU)	RMSD (Å)	Rosetta Energy (RU)	RMSD (Å)	Rosetta Energy (RU)
5′ J1/2, Leadzyme	4	1NUJ	**1.43**	-57.1	3.97	-55.8	**0.83**	-55.6	2.23	-55.3	1.90	-54.8
5′ P1, M-Box Riboswitch	4	2QBZ	**0.96**	126.1	3.63	129.7	5.36	130.0	5.58	130.2	**1.31**	130.2
3′ J5/5a, Group I Intron	4	2R8S	**0.47**	-21.4	4.64	-16.1	3.51	-15.0	4.09	-14.6	**1.38**	-14.5
5′ J5/5a, Group I Intron	5	2R8S	5.30	-53.7	4.52	-53.4	5.44	-53.4	**1.04**	-52.8	3.96	-52.6
Hepatitis C Virus IRES IIa	5	2PN4	5.31	22.1	8.27	23.1	7.66	23.5	7.00	23.6	6.82	23.9
J2/4, TPP Riboswitch	5	3D2V	4.21	-95.1	3.52	-95.0	4.44	-94.8	**0.85**	-94.2	3.76	-92.6
23S rRNA (44-49)	6	1S72	**0.73**	-166.4	4.32	-162.2	4.56	-161.5	4.54	-161.0	**1.44**	-160.7
23S rRNA (531-536)	6	1S72	5.96	-261.3	5.79	-261.0	4.28	-260.5	4.07	-260.4	2.45	-260.3
J3/1, Glycine Riboswitch	7	3OWI	**1.35**	93.5	2.53	93.7	2.62	94.1	3.39	94.3	3.74	94.4
J2/3, Group II Intron	7	3G78	**0.82**	-67.7	1.57	-66.5	1.91	-65.3	**1.25**	-65.2	4.09	-64.9
L1, SAM-II Riboswitch	7	2QWY	1.84	0.2	4.04	1.3	3.53	3.0	2.39	3.0	**1.26**	3.5
L2, Viral RNA Pseudoknot	7	1L2X	4.66	-193.3	3.69	-192.8	5.89	-192.6	4.99	-192.6	3.36	-192.0
23S rRNA (2534-2540)	7	1S72	6.38	-232.6	5.71	-232.6	7.52	-232.5	7.46	-232.2	7.36	-232.0
23S rRNA (1976-1985)	10	1S72	8.33	-303.9	8.45	-302.1	8.18	-301.6	8.72	-301.2	7.75	-301.2
23S rRNA (2003-2012)	10	1S72	4.93	-290.9	**0.74**	-290.1	2.86	-289.1	5.03	-289.0	3.45	-289.0
AVERAGE	6.3	–	3.51	-100.1	4.36	-98.7	4.57	-98.1	4.18	-97.8	3.60	-97.5
RMSD < 1.50 Å	–	–	6/15	–	1/15	–	1/15	–	3/15	–	4/15	–
Energy Gap < 0.0	–	–	–	–	–	–	–	–	–	–	–	–