Modeling Small Non-canonical RNA Motifs with the Rosetta FARFAR Server

Joseph D. Yesselman[1], Rhiju Das[1,2,*]

Affiliations:
[1]Biochemistry Department, Stanford University, Stanford, CA 94305, USA
[2]Physics Department, Stanford University, Stanford, CA 94305, USA
[*]Correspondence to: Rhiju Das (rhiju@stanford.edu)

Abstract

Non-canonical RNA motifs help define the vast complexity of RNA structure and function, and in many cases, these loops and junctions are on the order of only 10 nucleotides in size. Unfortunately, despite their small size, there is no reliable method to determine the ensemble of lowest energy structures of junctions and loops at atomic accuracy. This chapter outlines straightforward protocols using a webserver for Rosetta Fragment Assembly of RNA with Full Atom Refinement (FARFAR) (http://rosie.rosettacommons.org/rna_denovo/submit) to model the 3D structure of small non-canonical RNA motifs for use in visualizing motifs and for further refinement or filtering with experimental data such as NMR chemical shifts.

Key words: RNA 3D structure prediction, RNA Motifs.

1. Introduction. RNA plays critical roles in all living systems through its ability to adopt complex 3D structures and perform chemical catalysis (1). RNA structure appears highly modular in nature, defined through base pairing interactions. Nucleotides can either form highly structured helices composed of canonical Watson-Crick base pairs or small unpaired or non-canonical base paired regions in the form of junctions and loops (motifs) (2-4). Helices are, for the most part, structurally similar to each other, leaving non-canonical motifs to define the vast complexity of RNA structure and function. These non-canonical elements define the topology of the 3D structure of RNA by orienting the helices to which they connect and by forming long-range tertiary contacts that can lock specific global RNA conformations in place. In addition to defining the overall 3D structure of RNA (5,6), non-canonical motifs are the sites of small molecule binding and chemical catalysis (7-10). Many non-canonical motifs are on the order of only 10 nucleotides in size. Unfortunately, despite their small size, there is no reliable method to determine the ensemble of lowest energy structures of junctions and loops at near

1

atomic accuracy. Nevertheless, to model RNA at high resolution, it is critical to achieve accurate solutions for these small motifs.

When their structures are solved experimentally, most motifs turn out to form complex arrangements of non Watson-Crick hydrogen bonds and a wide range of backbone conformations. Due to the large number of interactions possible and each nucleotide's many degrees of internal freedom, it remains difficult to determine the lowest energy conformation (11). Recently, Fragment Assembly of RNA with Full Atom Refinement (FARFAR) was developed to help address this problem, FARFAR adapted the well-developed Rosetta framework for proteins to predict and design RNA non-canonical motifs (12). Out of a 32-target test set, 14 cases gave at least one out of five models that were better than 2.0 Å all-heavy-atom RMSD to the experimentally observed structure. While not perfect, this level of accuracy can be combined with even sparse experimental data, such as ^{1}H chemical shifts, to obtain high confidence structural models, as was demonstrated recently in blind predictions with the CS-ROSETTA-RNA method (13). The motif models can also form building blocks for modeling more complex RNAs and has been tested in the RNA Puzzles trials (14). Our FARFAR method for large RNAs with complete folds has been reviewed recently (15). The current bottleneck for some of these motifs and for larger RNAs is the difficulty of complete conformational sampling (11). On-going work with stepwise assembly (SWA) attempts to resolve this issue (16), but this more advanced procedure requires greater computational expense and a complex workflow that is not yet straightforward to implement on a public server, except in the special case of refinement one-nucleotide-at-a-time crystallographic refinement (17). Stepwise assembly is available in the main Rosetta codebase, but will not be further discussed here.

This chapter outlines straightforward protocols that should enable anyone to use FARFAR through a simple web server. FARFAR (RNA De Novo) is part of the Rosetta Online Server that Includes Everyone (ROSIE) software, a push to give wide access to the algorithms found in the Rosetta 3.x framework (18). The web server requires no initial setup for the user; all that is needed is to supply a sequence and an optional secondary structure to obtain all-atom models of an RNA motif of interest.

1.1 FARFAR Calculation: The FARFAR structure-modeling algorithm is based on two discrete steps. First, the RNA is assembled using 1 to 3 nucleotide fragments from existing RNA crystal structures whose sequences match subsequences of the target RNA. Fragment Assembly of RNA (FARNA) uses a Monte Carlo process guided by a low-resolution knowledge-based energy function (19). Afterwards, these models can be further refined in an all-atom potential to yield more realistic structures with cleaner hydrogen bonds and fewer clashes; the resulting energies are also better at discriminating native-like conformations from non-native conformations. The two-stage protocol is called Fragment Assembly of RNA with Full Atom Refinement (FARFAR).

2. Materials: FARFAR (RNA De Novo) is a webserver implementation of the Rosetta RNA fragment assembly algorithm server using the ROSIE framework. ROSIE is a web front-end for Rosetta 3 software suite, which provides experimentally tested and rapidly evolving tools for the high resolution 3D modeling of nucleic acids, proteins, and other biopolymers. FARFAR (RNA De Novo) can be reached using any of the standard web browsers such as Apple Safari, Microsoft Internet Explorer, Mozilla Firefox and Google Chrome here: http://rosie.rosettacommons.org/rna_denovo/submit. [**Note to the editor -- address will change**]

3. Methods: This protocol outlines the steps to use the FARFAR (RNA De Novo) webserver located on the ROSIE website. Although it is possible to submit jobs without creating an account, having an account yields numerous benefits, such as email alerts when jobs are finished, as well as the ability to create private jobs that are not visible to other users. It is highly recommended to create an account when first visiting ROSIE. In addition to the FARFAR webserver, ROSIE also hosts many other Rosetta based applications with a continuous stream of novel applications in development.

3.1.1 Main Page Form: This demonstration of FARFAR (RNA De Novo) uses the GCAA tetraloop; the whole structure was determined through NMR spectroscopy by *Jucker et al.* (PDB 1ZIH) (20). This tetraloop has a sequence of `gggcgcaagccu` and secondary structure of `((((....))))` in dot parentheses notation (**Figure 1**). **Figure 2** shows the main submission form for the RNA De Novo server. The only required input is the sequence, from 5′ to 3′. This is typically in lower-case letters, but upper-case are acceptable and will be converted. Use a space, *, or + between strands. Note that this

3

sequence is treated as RNA so that any Ts that appear in the sequence are automatically converted to Us for the calculation. Next, enter the secondary structure, in dot-parentheses notation. This is optional for single-stranded motifs, but required for multi-strand motifs. Note that even if a location is 'unpaired' in the input secondary structure (given by a dot, ' . '), it is not forced to remain unpaired. Although this is optional for single stranded motifs, the results improve with the addition of the correct secondary structure. If uncertain about the secondary structure, consider utilizing the Vienna RNAfold webserver (21) (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) or other utilities described in this book. Alternatively, use chemical mapping techniques to estimate the structure; these methods have been recently tested in blind trials for their accuracy (22,23). In addition, note that there is currently a size submission limit of 32 nucleotides for FARFAR (RNA De Novo), as the amount of computation greatly increases as a function of number of residues.

There are two more optional arguments. First is a file containing the ^{1}H chemical shifts determined by NMR spectroscopy. The format of this file follows the STAR v2.1 format used by the Biological Magnetic Resonance Data Bank (BMRB) (24). An example of the format is displayed in **Figure 3** with an explanation of each column. In addition, it is also possible to supply a native structure for the server to compare to for rmsd calculations. This file must be in PDB format, and for this case it is possible to download the structure from http://pdb.org/pdb/explore/explore.do?structureId=1zih. To supply a native structure, click the "Choose file" button next to native PDB-formatted file and select the appropriate file from your local hard drive.

There are two ways of running a FARFAR (RNA De Novo) job. The first is a trial run, which generates only one structure with a limited number of fragment assembly steps. This is for testing purposes only, and allows you to confirm that the job is set up properly. The second is a full run that takes more computational time to complete and produces thousands of models. It is advised when setting up a job for a new sequence and secondary structure to always run the job as a trial. Then, using PyMOL or your favorite viewer, open the PDB file. This is particularly important if you have a multi-stranded motif -- check that the strands are separated, and that any specified Watson-Crick pairs are reasonably paired. Once this is set up, go to the bottom of the page and click "Submit FARFAR (RNA De Novo) job". Upon submission, a temporary status page will load (**Figure 4**).

4

3.1.2 Advanced Options: In addition to the options discussed above, there are a few additional options that may be used occasionally. First is "Vary bond lengths and angles"; typically each residue has a set of bond lengths and angles between atoms that are based on idealized parameters. Checking this option will allow these parameters to vary slightly based on the Rosetta force field energy. This can increase conformational search space if you are interested in a specific interaction between residues and was used in previous benchmark studies, but requires more computational time (12).

When checked, "High resolution, optimize RNA after fragment assembly" will perform the all-atom refinement after fragment assembly; it is not recommended to uncheck this unless you are interested in quickly seeing the initial results or would like to perform your own high resolution optimization. "Allow bulge (include entropic score term to favor extra-helical bulge conformations)", will include conformations with residues bulged out and not interacting with other residues. If a residue is known to be extruded from the helix, this might be a good option to try to reduce the conformational space searched. When "Allow bulge (include entropic score term to favor extra-helical bulge conformations)" is checked, please note that residues that are bulged out will not be present in the final pdb model. "Number of structures to generate", will change the number of final models, which will also greatly increase the time each run takes. "Number of Monte Carlo cycles", controls the quality of each model; if models generated for a specific run have wildly different structures, then FARFAR has poor confidence in the accuracy (see next section). Increasing the number of Monte Carlo cycles can increase convergence, at the expense of greater computation.

3.3 Server Results:

The server returns pictures of the best-scoring models from the five best-scoring clusters from the run in rank order by energy (**Figure 5**). The clustering radius is 2.0 Å by default. Click on the [Model-N] link to download the PDB file. The server returns cluster centers (without pictures) for the next 95 clusters as, as well as the top 20 lowest-energy structures. These may be valuable if you are filtering models based on experimental data. The server also returns a 'scatter plot' of the energies of all the models created. The x-axis is a distance measure from the native/reference model in RMSD (root mean-squared deviation) over all heavy atoms; if a reference model is not provided, then the RMSD is computed relative to the lowest energy model discovered by FARFAR. The y-

5

axis is the score (energy) of the structure. In runs where a native structure is not supplied, the x-axis is a distance measure from the best scoring model found. As with nearly every Rosetta application, a hallmark of a successful run is convergence, visible as an energetic "funnel" of low-energy structures clustered around a single position. That is, near the lowest energy model there are additional models within ~ 2 Angstrom RMSD. In such runs, the lowest energy cluster centers have a reasonable chance of covering native-like structures for the motif, based on our benchmarks. A hallmark of an unsuccessful run is a lack of convergence -- few structures within 2 Å RMSD of the lowest energy model. Below the scatter plot, there is a detailed table of all the score terms used to calculate the final score as well as the RMSD to the native structure (if supplied). A description of the meaning of each term can be found in **Table 1**.

Visual representation of convergence of the models generated by FARFAR (RNA De Novo) can be found in **Figure 6**. As the figure demonstrates, there is high convergence in the top models found throughout the run. In addition, if one has ^{1}H chemical shift data as mentioned it can also be supplied; this can greatly increase the convergence and accuracy of an FARFAR prediction run. This can be demonstrated through a simple GA tandem motif, first generating models without ^{1}H chemical shift data (**Figure 6D**) yields the correct overall fold of the structure while incorrectly predicting the GA base pairs to be sheared instead of forming hydrogen bonds through their Watson and Crick edge (25). The ^{1}H chemical shift data adds sufficient restraints to resolve the base pairing discrepancy, with all top 20 models having the correct base pairing as the NMR solved structure. Both the native pdb and the chemical shift file can be downloaded from http://rosie.rosettacommons.org/documentation/rna_denovo.

6
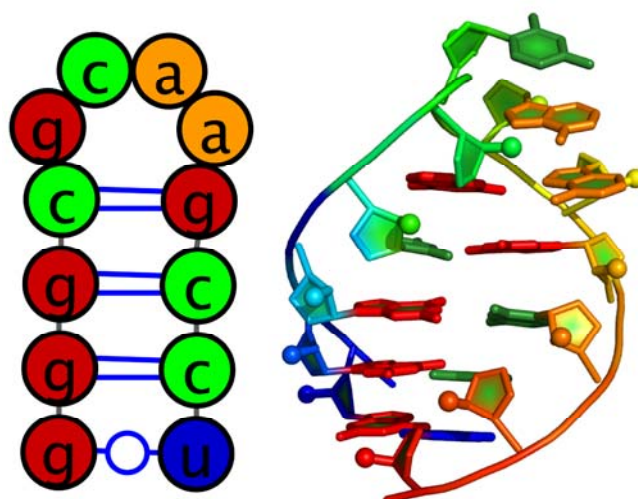
Figure 1: (Left) secondary structure of GCAA tetraloop. (Right) 3D structure of GCAA tetraloop (PDB: 1ZIH).

Figure 2: Main page of the FARFAR (RNA De Novo) webserver. Here the user can enter a sequence and secondary to submit a job to generation an all atom model of their construct.

```
 1   2   2 G H1'  H    5.62 . .
 2   2   2 G H2'  H    3.74 . .
 3   2   2 G H3'  H    4.75 . .
 4   2   2 G H4'  H    4.37 . .
 5   2   2 G H5'  H    4.57 . .
 6   2   2 G H5'' H    4.14 . .
 7   2   2 G H8   H    7.55 . .
 8   3   3 G H1'  H    5.89 . .
 9   3   3 G H2'  H    4.93 . .
10   3   3 G H3'  H    4.85 . .
11   3   3 G H4'  H    4.52 . .
12   3   3 G H5'  H    4.50 . .
13   3   3 G H5'' H    4.18 . .
14   3   3 G H8   H    8.02 . .
```

Figure 3: Example chemical shift data. Column description is as follows. 1) Atom entry number. 2) Residue author sequence code. 3) Residue sequence code. 4) Residue label. 5) Atom name. 6) Atom type. 7) Chemical shift value. 8) Chemical shift value error. 9) Chemical shift ambiguity code.

Figure 4: The status page for a submitted FARFAR (RNA De Novo) job.

Results

Cluster centers models, five lowest–energy structures created in rna_denovo run:

[Model–1] [Model–2] [Model–3] [Model–4] [Model–5]
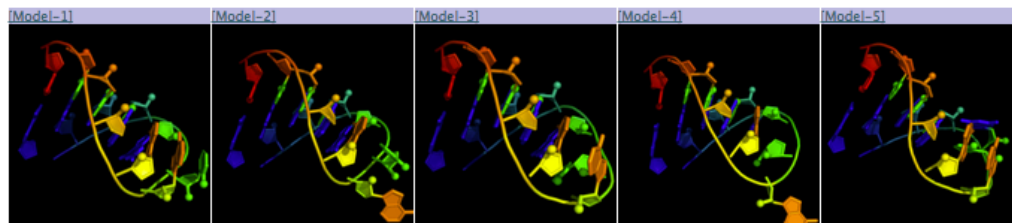
Rest of Cluster–centers top models:
[C–06] [C–07] [C–08] [C–09] [C–10] [C–11] [C–12] [C–13] [C–14] [C–15] [C–16] [C–17] [C–18] [C–19] [C–20] [C–21] [C–22] [C–23] [C–24] [C–25]
[C–26] [C–27] [C–28] [C–29] [C–30] [C–31] [C–32] [C–33] [C–34] [C–35] [C–36] [C–37] [C–38] [C–39] [C–40] [C–41] [C–42] [C–43] [C–44] [C–45]
[C–46] [C–47] [C–48] [C–49] [C–50] [C–51] [C–52] [C–53] [C–54] [C–55] [C–56] [C–57] [C–58] [C–59] [C–60] [C–61] [C–62] [C–63] [C–64] [C–65]
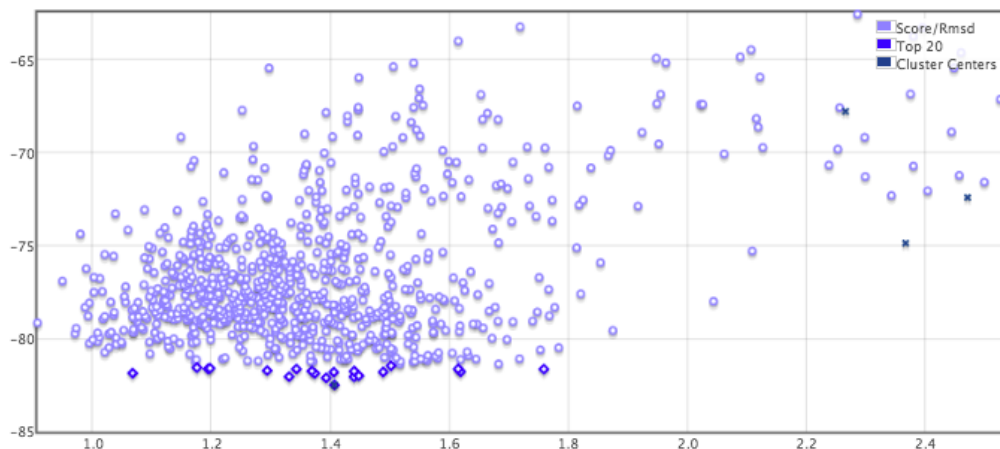[C–66] [C–67] [C–68] [C–69] [C–70] [C–71] [C–72] [C–73] [C–74] [C–75] [C–76] [C–77] [C–78] [C–79] [C–80] [C–81] [C–82] [C–83] [C–84] [C–85]
[C–86] [C–87] [C–88] [C–89] [C–90] [C–91] [C–92] [C–93] [C–94] [C–95] [C–96] [C–97] [C–98] [C–99]

Top–20 lowest–energy structures created in rna_denovo run:
[M–1] [M–2] [M–3] [M–4] [M–5] [M–6] [M–7] [M–8] [M–9] [M–10] [M–11] [M–12] [M–13] [M–14] [M–15] [M–16] [M–17] [M–18] [M–19] [M–20]

Score/RMSD

The only 5000 lowest energy models are shown. Hover over the graph points to see exact score and download particular result file.

Score data [Download original score file]

decoy	score	fa_atr	fa_rep	fa_intra_rep	lk_nonpolar	fa_elec_rna_	ch_bond	rna_torsion	rna_sugar_cl	hbond_sr_
R_000615	-82.49	-31.883	2.034	0.259	1.293	0.14	-17.639	0.447	0.899	0
R_000734	-82.112	-30.343	1.862	0.261	1.124	0.244	-16.851	0.442	0.988	0
R_000495	-82.056	-32.341	3.023	0.257	1.083	0.142	-17.409	0.335	0.663	0
R_000808	-82.045	-31.847	2.223	0.262	1.17	0.32	-17.427	0.498	0.915	0
R_000948	-81.993	-31.779	2.109	0.263	1.229	0.289	-16.982	0.295	0.853	0
R_000297	-81.872	-31.48	2.055	0.258	1.296	0.188	-17.423	0.365	0.965	0
R_000493	-81.856	-32.577	2.3	0.282	1.556	0.616	-17.698	1.072	1.232	0
R_000809	-81.805	-32.292	2.29	0.255	1.092	0.124	-17.483	0.488	0.77	0
R_000583	-81.794	-31.912	2.963	0.26	0.93	0.169	-16.274	0.413	0.571	0
R_000921	-81.776	-30.669	1.77	0.269	0.929	0.255	-16.266	0.579	0.771	0
R_000594	-81.746	-31.847	2.115	0.261	1.247	0.392	-17.131	0.49	1.063	0

Figure 5: Results page for a RNA Denovo job.

11
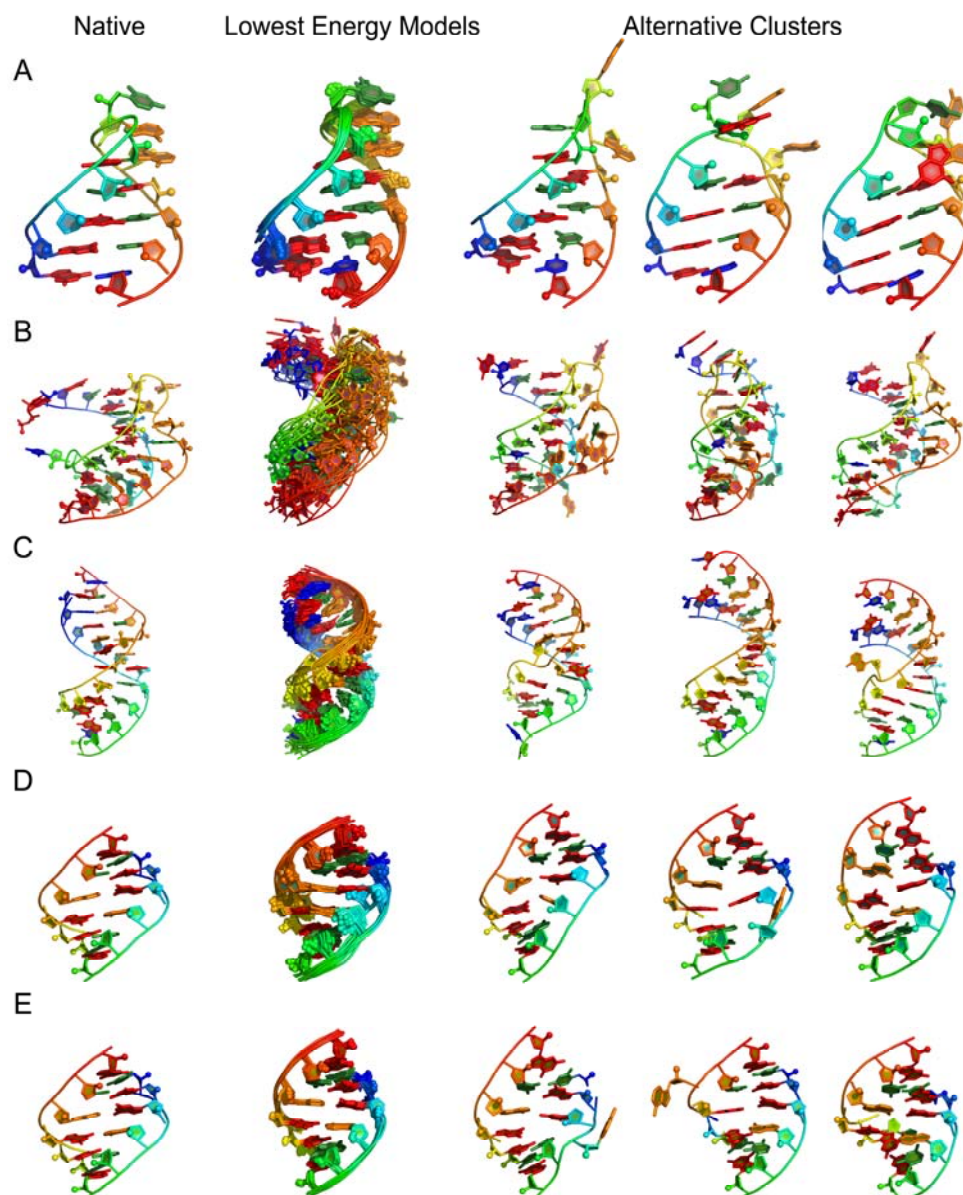
Figure 6: A) GCAA tetraloop (1ZIH), RNA Denovo lowest energy models displays a high level of convergence. B) Pseudoknot (1L2X) (26), less converged then tetraloop but also larger, still within 3Å heavy-atom rmsd for top model. C) 4x4 internal loop solved by NMR at PDB ID 2L8F (27), converges despite presenting 4 non-canonical base pairs. D) Tandem GA (1MIS) (25) without application of ^{1}H chemical shifts. E) Tandem GA with ^{1}H chemical shifts, demonstrates the improved convergence with the addition of ^{1}H chemical shifts.

Term	Definition
Score	Final total score
fa_atr	Lennard-jones attractive between atoms in different residues
fa_rep	Lennard-jones repulsive between atoms in different residues
fa_intra_rep	Lennard-jones repulsive between atoms in the same residue
lk_nonpolar	Lazaridis-karplus solvation energy, over nonpolar atoms
hack_elec_rna_phos_phos	Simple electrostatic repulsion term between phosphates
ch_bond	Carbon hydrogen bonds
rna_torsion	RNA torsional potential
rna_sugar_close	Term that ensures that ribose rings stay closed during refinement
hbond_sr_bb_sc	Backbone-sidechain hbonds close in primary sequence
hbond_lr_bb_sc	Backbone-sidechain hbonds distant in primary sequence
hbond_sc	Sidechain-sidechain hydrogen bond energy
linear_chainbreak	For 'temporary' chainbreaks, penalty term that keeps chainbreaks closed
N_WC	Number of Watson-Crick base pairs
N_NWC	Number of non-Watson-Crick base pairs
N_BS	Number of base stacks
Following are provided if the user gives a native structure	
rms	All-heavy-atom RMSD to the native structure
rms_stem	All-heavy-atom RMSD to helical segments in the native structure
f_natWC	Fraction of native Watson-Crick base pairs recovered
f_natNWC	Fraction of native non-Watson-Crick base pairs recovered
f_natBP	Fraction of native base pairs recovered

Table 1: Score terms reported on RNA Denovo results page.

References

1.   Cech TR, Steitz JA. The Noncoding RNA Revolution— Trashing Old Rules to Forge New Ones. Cell. Elsevier Inc; 2014 Mar 27;157(1):77–94.

2.   Leontis NB, Westhof E. Analysis of RNA motifs. Current Opinion in Structural Biology. 2003 Jun;13(3):300–8.

3.   Hendrix DK, Brenner SE, Holbrook SR. RNA structural motifs: building blocks of a modular biomolecule. Quart Rev Biophys. 2006 Jul 3;38(03):221.

4.   Leontis NB, Lescoute A, Westhof E. The building blocks and motifs of RNA architecture. Current Opinion in Structural Biology. 2006 Jun;16(3):279–87.

5.   Moore PB. Structural motifs in RNA. Annu Rev Biochem. 1999;68(1):287–300.

6.   Brion P, WESTHOF E. Hierarchy and dynamics of RNA folding. Annu Rev Biophys Biomol Struct. 1997;26(1):113–37.

7.   Lauhon CT, Szostak JW. RNA aptamers that bind flavin and nicotinamide redox cofactors. J Am Chem Soc. 1995 Feb 1;117(4):1246–57.

8.   Paige JS, Wu KY, Jaffrey SR. RNA mimics of green fluorescent protein. Science. American Association for the Advancement of Science; 2011 Jul 29;333(6042):642–6.

9.   Doudna JA, Lorsch JR. Ribozyme catalysis: not different, just worse. Nat Struct Mol Biol. 2005 May;12(5):395–402.

10.   Lilley DM. Structure, folding and mechanisms of ribozymes. Current Opinion in Structural Biology. 2005 Jun;15(3):313–23.

11.   Sripakdeevong P, Beauchamp K, Das R. Why Can't We Predict RNA Structure At Atomic Resolution? Nucleic Acids and Molecular Biology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. pp. 43–65.

12.   Das R, Karanicolas J, Baker D. Atomic accuracy in predicting and designing noncanonical RNA structure. 2010 Feb 28;7(4):291–4.

13.   Sripakdeevong P, Cevec M, Chang AT, Erat MC, Ziegeler M, Zhao Q, et al. Structure determination of noncanonical RNA motifs guided by 1H NMR chemical shifts. Nat Meth. 2014 Mar 2;11(4):413–6.

14.   Cruz JA, Blanchet MF, Boniecki M, Bujnicki JM, Chen SJ, Cao S, et al. RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. RNA. 2012 Mar 19;18(4):610–25.

15.   Cheng CY, Chou FC, Das R. Modeling Complex RNA Tertiary Folds with Rosetta. Meth Enzymol. 2015.

16.   Sripakdeevong P, Kladwang W, Das R. An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling. Proc Natl Acad Sci USA. 2011 Dec

20;108(51):20573–8.

17. Chou F-C, Sripakdeevong P, Dibrov SM, Hermann T, Das R. Correcting pervasive errors in RNA crystallography through enumerative structure prediction. 2013 Jan;10(1):74–6.

18. Lyskov S, Chou F-C, Conchúir SÓ, Der BS, Drew K, Kuroda D, et al. Serverification of Molecular Modeling Applications: The Rosetta Online Server That Includes Everyone (ROSIE). Uversky VN, editor. PLoS ONE. 2013 May 22;8(5):e63906.

19. Das R, Baker D. Automated de novo prediction of native-like RNA tertiary structures. Proc Natl Acad Sci USA. 2007 Sep 11;104(37):14664–9.

20. Jucker FM, Heus HA, Yip PF, Moors EH, Pardi A. A network of heterogeneous hydrogen bonds in GNRA tetraloops. Journal of Molecular Biology. 1996 Dec 20;264(5):968–80.

21. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA websuite. Nucleic Acids Research. 2008 Jul 1;36(Web Server issue):W70–4.

22. Kladwang W, Cordero P, Das R. A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. RNA. 2011 Feb 10;17(3):522–34.

23. Miao Z, Adamiak RW, Blanchet M-F, Boniecki M, Bujnicki JM, Chen S-J, et al. *RNA-Puzzles Round II: Assessment of RNA structure prediction programs applied to three large RNA structures.* RNA.

24. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, et al. BioMagResBank. Nucleic Acids Research. 2008 Jan;36(Database issue):D402–8.

25. Wu M, Turner DH. Solution structure of (rGCGGACGC)2 by two-dimensional NMR and the iterative relaxation matrix approach. Biochemistry. 1996 Jul 30;35(30):9677–89.

26. Egli M, Minasov G, Su L, Rich A. Metal ions and flexibility in a viral RNA pseudoknot at atomic resolution. Proc Natl Acad Sci USA. National Acad Sciences; 2002 Apr 2;99(7):4302–7.

27. Lerman YV, Kennedy SD, Shankar N, Parisien M, Major F, Turner DH. NMR structure of a 4 x 4 nucleotide RNA internal loop from an R2 retrotransposon: identification of a three purine-purine sheared pair motif and comparison to MC-SYM predictions. RNA. 2011 Sep;17(9):1664–77.

15