CHAPTER NINE

Using Rosetta for RNA homology modeling

Andrew M. Watkins[a], Ramya Rangan[b], Rhiju Das[a,b,*]
[a]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA, United States
[b]Biophysics Program, Stanford University, Stanford, CA, United States
*Corresponding author: e-mail address: rhiju@stanford.edu

Contents

Abstract

The three-dimensional structures of RNA molecules provide rich and often critical information for understanding their functions, including how they recognize small molecule and protein partners. Computational modeling of RNA 3D structure is becoming increasingly accurate, particularly with the availability of growing numbers of template structures already solved experimentally and the development of sequence alignment and 3D modeling tools to take advantage of this database. For several recent "RNA puzzle" blind modeling challenges, we have successfully identified useful template structures and achieved accurate structure predictions through homology modeling tools developed in the Rosetta software suite. We describe our semi-automated methodology here and walk through two illustrative examples: an adenine riboswitch aptamer, modeled from a template guanine riboswitch structure, and a SAM I/IV riboswitch aptamer, modeled from a template SAM I riboswitch structure.

Methods in Enzymology, Volume 623
ISSN 0076-6879
https://doi.org/10.1016/bs.mie.2019.05.026

1. Introduction

RNA plays a host of critical functional roles in cells, from translation regulation to catalysis (Gesteland, Cech, & Atkins, 2005). To achieve these functions, many noncoding RNAs (ncRNAs) take on complex 3D folds, with secondary structure helical elements positioned by structured junctions and tertiary contacts. As experimental characterization of these structures can be challenging and time-consuming, there is strong interest in developing computational strategies to predict these structures using physical modeling or knowledge-based fragment and motif sampling (Das, Karanicolas, & Baker, 2010; Ditzler, Otyepka, Sponer, & Walter, 2010; Laing & Schlick, 2011; Sim, Minary, & Levitt, 2012; Xu & Chen, 2018). To make better use of available information, these computational strategies can be augmented with homology modeling, where a portion of the modeled coordinates is built based on a previously solved homologous structure (Piatkowski et al., 2016; Rother, Rother, Boniecki, Puton, & Bujnicki, 2011).

As the number of experimentally characterized RNA structures grows, the potential for accurate homology modeling for new molecules increases as well. The Protein Data Bank now has over 1500 non-redundant RNA structures at high or medium resolution (4.0 Å resolution or better) (Leontis & Zirbel, 2012), and these deposited structures represent 87 different Rfam families and 21 different Rfam clans (Kalvari et al., 2018). These structures provide a wealth of diverse potential templates for new modeling challenges. Furthermore, the availability of large databases of RNA sequences across species has generated more opportunities to discover homologous sequences. Alignment software can account for covariation in RNA secondary structure, providing more accurate alignments than those based exclusively on sequence (Kalvari et al., 2018; Nawrocki & Eddy, 2013). Expert inspection and biochemical analysis routinely reveal homology between distantly related ncRNA classes even prior to structural characterization. Recent examples include the lariat capping ribozyme, whose catalytic core is homologous to the group I self-splicing intron (Einvik, Nielsen, Westhof, Michel, & Johansen, 1998); the group II self-splicing introns, homologous to the eukaryotic spliceosome (Toor, Keating, Taylor, & Pyle, 2008); extended "sub-motifs" shared between adenosylcobalamin (AdoCbl) and flavin mononucleotide (FMN) riboswitches (Barrick & Breaker, 2007; Jaeger, Verzemnieks, & Geary, 2009); homology between glutamine and

"downstream peptide" riboswitches (Ames & Breaker, 2014); and shared binding sites for S-adenosyl methionine (SAM) across distinct riboswitch classes (Mirihana Arachchilage, Sherlock, Weinberg, & Breaker, 2018; Weinberg et al., 2008).

We have found that riboswitches—genetic control elements that respond to the presence of small molecules (Serganov & Patel, 2012)—are particularly amenable to template identification and then computational homolog modeling of structure. Often, multiple classes of riboswitches bind the same or similar ligand and possess identifiable sequence homology to each other and to solved riboswitch structures. Riboswitches are also excellent use cases for homology modeling because their functional state requires a stably-folded ligand binding site, entailing intricate 3D folds in ligand-recognition domains—folds that are unlikely to be recovered by *de novo* RNA modeling. If ligand binding sites from previously solved riboswitches can be identified for a new riboswitch, borrowing that structural information can greatly improve modeling accuracy.

Recent blind modeling challenges in the RNA Puzzles trials (Miao et al., 2017) support the view that RNA homology modeling can be accurate and biologically useful. We have used identifiable but at times distant homologies to previously solved structures to achieve accurate blind models of numerous riboswitch and other ncRNA targets, including the GIR1 lariat-capping ribozyme, the adenosylcobalamin riboswitch, the glutamine riboswitch, the Zika xrRNA, and the SAM I/IV riboswitch (RNA Puzzles 5, 6, 8, 14, 18; Fig. 1). In several cases, we were further able to correctly predict ligand-binding sites, based on clustering of conserved residues in these 3D folds (Miao et al., 2017). While addressing these blind challenges, we have developed a framework for homology modeling of RNA structures with Rosetta computational tools. Depending on the target and modeling sub-problem, either a fragment assembly algorithm (Fragment Assembly of RNA with Full-Atom Refinement, or FARFAR) (Cheng, Chou, & Das, 2015; Das et al., 2010) or a high resolution fragment-free algorithm called stepwise Monte Carlo (SWM) (Watkins et al., 2018) are best suited to the modeling challenge. FARFAR modeling is the more well-developed approach and the method has been previously reviewed (Cheng, Chou, & Das, 2015), albeit not yet for homology modeling problems. SWM is a newer method that seeks higher resolution. In the case of Zika xrRNA, recent homology modeling with SWM correctly predicted all noncanonical base pairs of the RNA, a previously unmet challenge in RNA computational structural modeling (Fig. 1D) (Watkins et al., 2018).

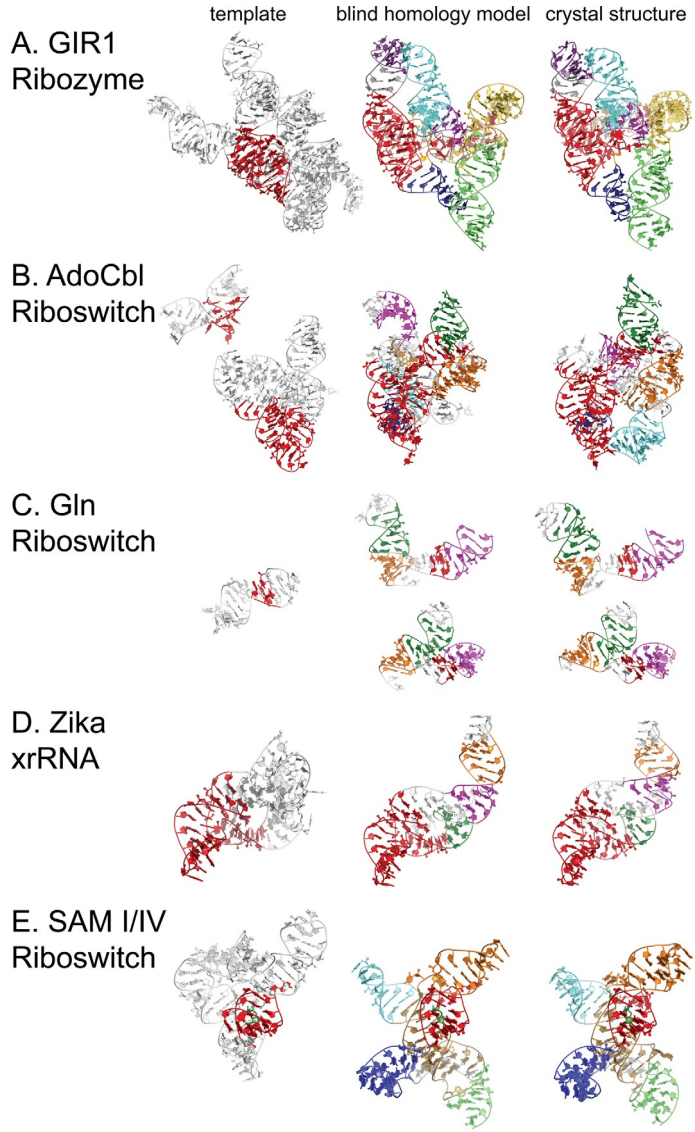

Fig. 1 Prior use of homology modeling in blind challenges using Rosetta algorithms. In RNA Puzzles 5, 6, 14, 18, and 8, previously solved structures provided substantial insight that guided modeling. Each subpanel displays the template structure (left; portion employed in red), the best submitted model (middle), and the eventual crystal structure (right). (A) The GIR1 ribozyme contained a helix arrangement found in the related Azoarcus group I intron. PDB codes: 3BO3 (Lipchock & Strobel, 2008) and 4P8Z (Meyer et al., 2014). (B) The AdoCbl riboswitch contained a T-loop architecture also found in earlier structures of FMN riboswitches, and a kink-turn motif that we obtained

In this review, we describe our approach for homology modeling, from obtaining an appropriate template to the computational strategies for completing missing coordinates, as outlined in Fig. 2. We then illustrate this approach by walking through worked examples for two challenges: first, an artificial challenge modeling an adenine riboswitch based on a well-known guanine riboswitch structure; second, an actual blind challenge (RNA Puzzle 8) where we modeled the SAM I/IV riboswitch. Before describing this Rosetta-focused methodology, we note that other groups have also developed excellent tools outside Rosetta for RNA homology modeling, and we advise the reader to try multiple tools and check for consensus (Piatkowski et al., 2016; Popenda et al., 2012; Xu & Chen, 2018). We also note that current homology modeling methods can be computationally expensive, especially if only a fraction of a target RNA has an identifiable template in a previously solved structure and the rest must be modeled *de novo*. Readers should be aware that computer clusters with ~100 CPUs are therefore required for the methods below. Last, we note that, at this time, computational strategies for RNA structural modeling must be augmented with human insight to identify and best use homology information; several specific examples of the value of expert inspection are described below.

2. Method

Identifying a homologous solved structure can substantially reduce the unknown portion of a structure that must be modeled through computationally expensive algorithms, allowing for the generation of models with higher accuracy with the same amount of computer time. For illustration, in the allied field of protein structural modeling, modeling based on

from the structure of the MBP-L30e-mRNA complex. PDB codes: 2YIE (Vicens, Mondragon, & Batey, 2011); 4GXY (Peselis & Serganov, 2012); and 1T0K (Chao & Williamson, 2004). (C) Both bound (top) and unbound (bottom) structures of the glutamine riboswitch contain a sarcin-ricin loop. PDB codes: 1Q9A (Correll, Beneken, Plantinga, Lubbers, & Chen, 2003); 5DDO (Ren et al., 2015); and 5DDP (Ren et al., 2015). (D) The Zika xrRNA features substantial homology to Murray Valley Encephalitis xrRNA, but the presence of a pseudoknot not found in the original crystal changes the overall fold architecture. PDB codes: 4PQV (Chapman et al., 2014) and 5TPY (Akiyama et al., 2016). (E) The SAM I/IV riboswitch, later discussed as a "worked example" here, was RNA Puzzle 8; in the original blind modeling, we used a template region around the SAM binding site. PDB codes: 2GIS (Montange & Batey, 2006) and 4OQU (Trausch et al., 2014).

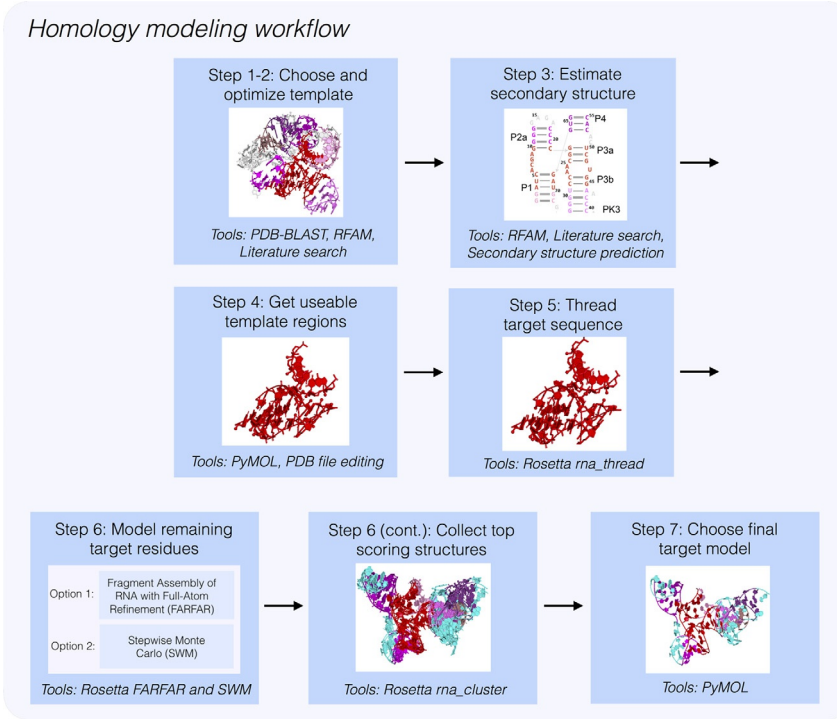

Fig. 2 Workflow for homology modeling with the Rosetta framework. Preparing homologous regions requires selecting an appropriate template structure and refining that structure within the Rosetta score function using ERRASER-Phenix. Here, the template is colored at helical regions (shades of purple and pink) and in an elaborately folded ligand-binding core (red) that serves to seed modeling runs. The secondary structure of the target sequence should be estimated. Useable template regions should be selected that lack experimental artifacts and are sufficiently similar to the modeled structure. The desired sequence is then threaded onto the template structure. Finally, remaining regions of the structure are modeled *de novo* in the context of this threaded structure using modeling with either FARFAR or SWM. New non-template coordinates are shown in teal.

templates has yielded substantially more accurate models than *de novo* modeling, even for proteins with limited similarity to existing structures (Moult, Fidelis, Kryshtafovych, Schwede, & Tramontano, 2018). However, after identifying a solved structure with substantial homology, modeling a new RNA molecule remains challenging. Judgments must be made as to which portions of the template should be used for the new structure, and significant portions of the structure may remain incomplete, requiring *de novo* modeling. Indeed, sequence discrepancies between the modeled structure and

template in non-helical regions can critically impact the global fold. In this section, we provide guidelines to help address these challenges, describing a series of steps for homology modeling from selecting a template structure to retrieving final coordinates.

2.1 Selecting candidate template structures

Homology modeling begins with identifying a candidate template structure. Often the target molecule is already known to fall into a known ncRNA family, and a literature search can provide an expert sequence alignment for that family based on manual curation. This scenario is typically the case for new riboswitches, which are indeed often identified on the basis of such sequence alignments and careful, manual detective work leveraging expert knowledge of biochemistry and molecular biology (Weinberg et al., 2017). If the function of the target molecule is known, a simple PDB search (https://www.rcsb.org) for the name of the riboswitch class, e.g., "guanine riboswitch," will often uncover templates with previously solved structure.

If homologies of the target molecule to previously solved structures are not immediately apparent, an automated back-up method is to use the PDB's advanced search, which includes an option to BLAST a target sequence. If any structures result, the PDB will automatically identify an approximate sequence alignment. If a PDB-BLAST fails, the Rfam database can carry out a sophisticated search of an input sequence against its large archive of ncRNA families and return a sequence alignment. The Rfam alignment can also help identify additional members of the same family from the PDB.

Sometimes these homology searches result in multiple possible template structures for the target. Although multiple templates may be used to seed parallel modeling runs, it is still a good idea to choose one template, or at least a subset of those available to optimize use of computational time. The two most important parameters in making this choice are the resolution of the structure in question and the sequence similarity to the target. As a rule of thumb, as long as the resolution is sufficient to place atomic coordinates, it should be a secondary concern; sequence similarity is more important in choosing between possible templates.

Finally, we note that template structures do not have to be large or complex to substantially improve or accelerate modeling. Many RNA subsequences form structural *modules* which form extremely similar folds,

varying little with their structural context (Miao & Westhof, 2017; Westhof, Masquida, & Jossinet, 2011). As a result, small sub-structures of a handful of nucleotides, which would never appear in any homology search, may still be of use. By far the most valuable such sub-structures to identify are tertiary contacts, because they can anchor the relative orientation of local structures that are distal in the RNA sequence and secondary structure. Two common tertiary contact modules are the intercalated T-loop (Chan, Chetnani, & Mondragon, 2013) and the tetraloop/receptor interaction (Jaeger, Michel, & Westhof, 1994; Wu, Chai, Fraser, & Zimmerly, 2012). If sequences compatible with either of these interactions may be found, it is likely worthwhile to attempt some modeling runs with these interactions seeded in, using a T-loop containing PDB (such as a tRNA structure) or a tetraloop/receptor (such as the *Tetrahymena* ribozyme's P4–P6 domain) as a miniature template. Packages like RMdetect and FR3D can discover some simple RNA motifs automatically, including kink turns (Cruz & Westhof, 2011; Sarver, Zirbel, Stombaugh, Mokdad, & Leontis, 2008), although we typically still recommend inspecting sequence-conserved regions of the predicted secondary structure of target molecules for evidence of tertiary motifs like T-loops and tetraloop receptors, which are not well-captured by any currently available automated program.

2.2 Optimizing the templates

Due to the challenges associated with manually fitting atomic coordinates, especially with the medium resolution (2.5–3.5 Å diffraction resolution) often obtained for RNA structures, experimental structures almost always contain flaws and geometric errors that can impede accurate modeling. These errors are particularly pervasive for coordinates deposited before 2014, after which the PDB began its new deposition and validation system (Chou, Echols, Terwilliger, & Das, 2016; Chou, Sripakdeevong, Dibrov, Hermann, & Das, 2013; Keating & Pyle, 2012; Read et al., 2011; Wang et al., 2008). Sometimes these flaws are innocuous, but other times they can change functional conclusions, and it is difficult to identify what will happen *a priori*. While the PDB keeps records of MolProbity analyses to catalog clashes and geometrical outliers (Chen et al., 2010), rather than merely selecting the least flawed option, there are facilities to eliminate geometrical outliers while maintaining minimal deviations from the parent structure. These include ERRASER-Phenix (Chou et al., 2013, 2016), RNABC (Wang et al., 2008), and RCrane (Keating & Pyle, 2012). For the Rosetta

homology modeling methods described herein, we recommend application of ERRASER-Phenix as at least a final step, because ERRASER can refine the structure into the same force field as used in subsequent Rosetta modeling.

2.3 Estimating target secondary structure

To choose relevant regions of the template structure and to complete 3D modeling of new regions of the target structure, the target's secondary structure provides important, often critical, constraints. For regions of the target molecule homologous to a template of known structure, the secondary structure can be inferred through homology. For other parts, various tools are available for predicting the secondary structure of an RNA from its sequence. The most powerful source of secondary structure information are expert analyses in the literature, which typically integrates all available biochemical, evolutionary, and prior structure information about the target. If experimental facilities are available, multidimensional chemical mapping experiments offer reliable RNA secondary structures (Tian & Das, 2016; Cheng, Kladwang, Yesselman, & Das, 2017); it is worth noting here that SHAPE and DMS-guided approaches are simpler but have been less reliable (Kladwang, VanLang, Cordero, & Das, 2011; Tian & Das, 2016; Miao et al., 2017, 2015). If an Rfam family or an expert-curated sequence alignment exists for the target sequence, the secondary structure constructed from a sequence alignment by tools like Infernal (Nawrocki & Eddy, 2013) will be the most reliable estimate for the target secondary structure. If an Rfam family or extensive sequence alignment is not available, tools for secondary structure prediction including RNAstructure (Reuter & Mathews, 2010), NUPACK (Dirks & Pierce, 2003), ViennaRNA (Lorenz et al., 2011), and CONTRAfold (Do, Woods, & Batzoglou, 2006) are appropriate although none reliably return information on pseudoknots. All these tools can be used to generate secondary structures for the sequence at hand, and the resulting predictions should be compared to find consensus helices.

2.4 Identifying useable portions of the template

Once a template structure is optimized and the target's secondary structure is predicted, sections of the template that will be used for further modeling should be identified. One possible template inaccuracy is the difference between crystallization and biologically relevant conditions. Regions of the template with crystal artifacts should be excluded in further modeling.

For instance, in our homology modeling of the Zika xrRNA (Fig. 1D) (Watkins et al., 2018), we identified a previously solved structure of the Murray Valley Encephalitis xrRNA, but it crystallized as a dimer, replacing an internal pseudoknot with an intermolecular contact, meaning that despite high sequence similarity, almost half of the template structure had to be discarded.

The sequence match to the homologous structure will be the primary determinant for which remaining regions of the template structure to keep. In addition, any sections in which the new sequence's assumed secondary structure disagrees with the template structure should not be included, as illustrated by the Zika xrRNA case described above. Any non-helical regions should be left out if replacing with the new sequence would break hydrogen bonding patterns. For example, if a GCAA tetraloop in the template is replaced with a UUCG tetraloop, it is better to model this tetraloop *de novo* or to find a template from another structure, for instance by searching the RNA 3D motif atlas (Petrov, Zirbel, & Leontis, 2013). With these exclusions, some regions of the structure may no longer be connected by conserved tertiary contacts or junctions; an example occurred in the adenosylcobalamin riboswitch where a kink turn and a complex T-loop/pseudoknot motif were drawn from two unconnected templates (separate red regions in Fig. 1B). Separated regions can be included as distinct rigid bodies in later modeling steps and allowed to reorient with respect to each other as intervening loops and helices are remodeled.

In deciding on useable portions of the template, it is helpful to additionally consider all available data on the sequence to be modeled. Biochemical data such as mutational analyses may point to divergence from the homologous structure. If a sequence alignment is available, covariance analysis can suggest new contacts or delineate regions in which the template is not applicable. Last, chemical mapping data, particularly new multidimensional experiments (Tian & Das, 2016), can be informative in detecting or confirming homologies of the target molecule to previously solved structures. As noted above for secondary structure, literature analyses of the RNA class, if available, remain the most valuable source for template identification.

2.5 Threading the target sequence onto template structure

Even for regions of the template that can be retained for the modeling task, the target sequence being modeled may deviate from the template structure's sequence. For example, an AU base pair in the template structure

may be replaced with a GC base pair in the target molecule. As a first step, the target sequence must be threaded onto the template, such that the RNA backbone remains constant while the corresponding nucleotides change to the target sequence. If the sequence alignment is ambiguous, this step and further steps should be applied to each reasonable threading—or correspondence—of the modeled sequence to the template. To obtain the threaded structure, one can use the `rna_thread` Rosetta application (see Appendix for example command-line).

Depending on the sequence changes being made, small-scale optimization of the threaded structure may be necessary prior to further modeling. If all the changes are in helix sequences, then the local geometry is likely agnostic to threading. In particular, helix structure will be unaltered when the sequence changes between Watson Crick pairs. For helix sequence changes that convert Watson-Crick pairs to G-U wobble pairs and *vice versa*, the user may wish to leave those base pairs or the entire helix out of the template so that they are rebuilt in the next step.

2.6 Modeling the target structure

Once the threaded template is optimized, the task remaining is to model the regions of the structure that lack homology to the template; doing so will set the relative orientation of any templated sections and provide 3D coordinates for every atom in the target RNA. As noted in the Introduction, two algorithms are available in Rosetta for modeling RNA folds: a medium-resolution method (Fragment Assembly of RNA with Full-Atom Refinement, or FARFAR) (Das et al., 2010) and a method that can achieve higher resolution but at high computational expense (stepwise Monte Carlo, or SWM) (Watkins et al., 2018). Detailed command-lines for these two approaches will be presented when discussing the worked examples.

Choosing between these algorithms involves three considerations:

a. The size of the target complex, and of the region that requires new modeling.

If high resolution is desired, SWM is the method of choice, but involves significantly more computation per model than FARFAR. Large structures and extensive remodeling challenges are possible but computationally expensive with SWM. As a rule of thumb, SWM modeling scales with the number of nucleotides or fixed helices that need to be rebuilt. Each such element requires 50 cycles of SWM simulation (typically a few minutes) to solve, up to problems with around

20 elements (several hours) per model. For example, if a problem involves three helices of known structure, interconnected by three loops with a total of 10 nucleotides, 650 cycles will be required to complete SWM models, corresponding to about an hour of simulation time per model. Ideally, hundreds of models need to be created, so even with a computer cluster with 100 CPUs, such a SWM modeling problem will require overnight runs. Challenges of more than 30 residues are better approached using FARFAR, as the lower resolution achieved will be offset by the greater number of completed trajectories.

b. The number of helices with length changes between template and target.

FARFAR has an efficient method for accurately recapitulating realistic helix flexibility; SWM samples one residue at a time and has an inefficient method for adding base-paired residues. Therefore, if there are major helix length changes between template and target, FARFAR is currently recommended.

c. The presence of multiway junctions and tertiary contacts.

Multiway junctions or tertiary contacts often form the structural core of complex RNA folds and can feature intricate combinations of non-Watson-Crick base pairs whose high-resolution geometry is essential to an accurate global fold. Ideally, as many as possible of these interactions may be found in the template structure. Otherwise, these features should be modeled using SWM. If there are many such complex interactions in a fold, one may decompose the structure into independent SWM jobs, and subsequently assemble successful sub-problem solutions using FARFAR or SWM.

A combination of FARFAR and SWM can be helpful for achieving high accuracy models. Initial FARFAR modeling may provide hypotheses for regions of the global structure that should be proximal; these proximal helices and junction residues can be modeled in more detail in separate smaller SWM runs. Alternatively, if a multiway junction or tertiary contact is essential for achieving the correct global fold, it is feasible to first generate models for this region with SWM, and then seed these elements into a FARFAR modeling round as rigid bodies to generate more accurate global models.

During modeling runs, a choice needs to be made for whether to allow regions drawn from templates to be optimized away from their starting structures. Usually, it is best to keep these regions fixed so as to preserve the structural information from the starting template; however, we typically allow for minimization of nucleotides near the edges of these regions to allow for small changes that can propagate and "relax" newly built regions into

more energetically favored configurations (see Appendix for example command-lines and explanations). In some cases, including the walk-through examples below, we also carry out pre-optimization of template segments through the `rna_minimize` command; this step automatically uses coordinate constraints to the structure's initial coordinates for its first round of minimization, ensuring that any large clashes are relieved without major global changes to the RNA structure (see Appendix for example command-lines).

2.7 Model selection

For both FARFAR and SWM algorithms, hundreds to thousands of independent trajectories that stochastically build the new regions are carried out. With either sampling strategy, the presence of numerous independently built models with low RMSD to the top-scoring model suggests sufficient sampling. A plot of Rosetta score versus the RMSD to the top scoring model can be used to quickly assess the conformational space accessed over all modeling runs. For any given Rosetta score cutoff (say, a cutoff for the top 10 models), observing multiple structures close in RMSD to the top structure indicates support for the best scoring model. In addition to score v. RMSD plots, inspection of the top 10 or 20 models in 3D molecular viewers like Pymol allows for visual assessment of whether any subset of these models have converged to the same global fold, a hallmark of convergence of the modeling and typically a good sign of accuracy (Cheng, Chou, Kladwang, et al., 2015; Kappel et al., 2018; Shortle, Simons, & Baker, 1998). The resulting collection of top-scoring models are then the candidates for the predicted structure. More quantitatively, the mutual RMSD between the top scoring models can help assess convergence to the native structure, with lower mutual RMSD indicating more accurate models (Kappel et al., 2018).

3. Worked examples

This section briefly describes two worked examples of Rosetta RNA homology modeling, following the steps outlined above and in Fig. 2. For both examples, we test both alternatives for the *de novo* modeling step (SWM vs. FARFAR) to illustrate the rationale and outcomes involved in that choice. Readers wishing to work through these examples may also be interested in command-lines and a publicly available repository of computer files; see Appendix.

3.1 The adenine riboswitch

Riboswitches fold to form selective, tight pockets to bind their target ligands. A classic example is the purine riboswitch (Fig. 3), where a Watson-Crick base pair between the ligand and nucleotide 74, which is always a pyrimidine, defines the binding selectivity of the riboswitch. As a first example of Rosetta homology modeling, we model the *V. vulnificus* adenine riboswitch structure, comparing modeling using the FARFAR and SWM approaches.

The first steps in the method are to select and optimize an appropriate template structure. For this test, we select the *B. subtilis* guanine riboswitch (PDB ID 1Y27) (Serganov et al., 2004) and we run the structure through the ERRASER-Phenix pipeline. In the third step, we estimate the secondary structure for the target adenine riboswitch, which has only small deviations from the template's secondary structure (Fig. 3). In the fourth step of the method, we identify homologous regions in the template and target to determine which regions of the template structure to use. Here, there are sequence changes in RNA helices, and while these can affect the ensemble of favored structures for each helix (Yesselman et al., 2018), those changes are typically overwhelmed by a well-conserved tertiary context, as we see here. We excise the residues whose mutations fall in loops or that break or form Watson-Crick pairs (as in P2), marked as mismatches in Fig. 3. In the fifth step, we thread the target sequence onto the remaining backbone.

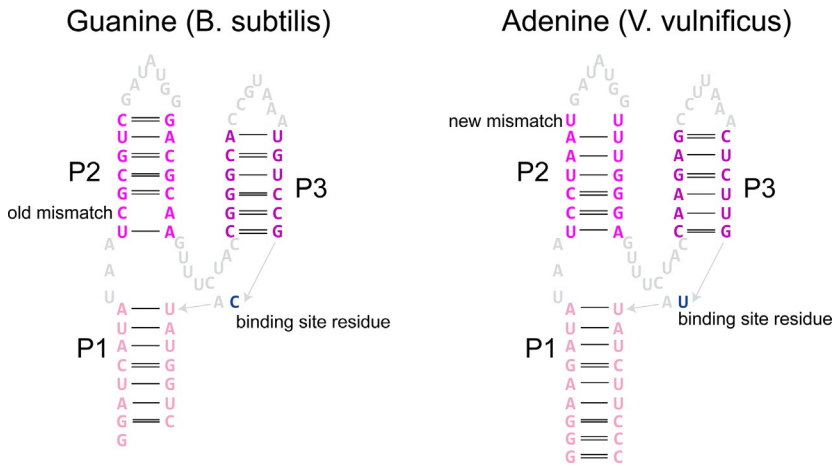

Fig. 3 The secondary structure and sequence of two purine riboswitches are highly similar, with sequence deviations outside of helices marked in blue.

Now we can move to the sixth step of the method, as outlined in Fig. 2: modeling the remaining residues in the target structure. We compare two simulation approaches for completing the modeling. We perform a simulation with SWM to build in the mutated residues, followed by a simulation with FARFAR just to extend the P1 helix. In the second approach, we perform a simulation where we model in both the mutated residues and the P1 helix extension with FARFAR. For the ligand in each simulation case, we place a new adenine in a perfect Watson-Crick orientation with U74, mapped from a C in the template guanine riboswitch structure. (While SWM can explicitly remodel ligand binding modes, this may not be necessary for an initial simulation or for a simple binding mode like a single Watson-Crick base pair.) For each type of simulation, we observed convergence by comparing the top 10 models (as ranked by Rosetta energy), and we chose the top model as our best structure (Fig. 4). We note that in

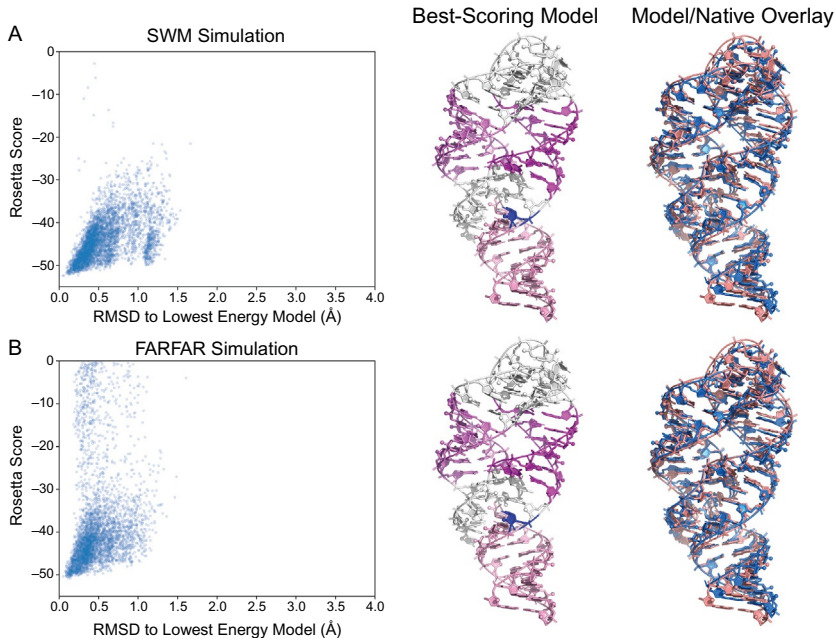

Fig. 4 On the left, plots comparing model score to the RMSD of the lowest energy model for SWM simulations (A) and FARFAR simulations (B) show convergence to the lowest energy structure. In the middle column are the lowest energy models from the SWM (top) and FARFAR (bottom) simulations. In the right column, the top scoring structures (salmon) are overlaid on the crystal structure of the target (blue, PDB ID: 4TZY) (Zhang & Ferre-D'Amare, 2014), which was not used in homology modeling.

both cases, the mutual RMSD between the top 10 structures is <0.5 Å, indicating strong convergence and suggesting that the lowest energy model will likely be close to the actual structure of this adenine riboswitch. In the seventh step, we therefore choose the lowest energy model from each run as the final model.

In this case, the experimental structure of the adenine riboswitch is known, and so the accuracy of the Rosetta homology models can be assessed. Compared to a crystal structure of this RNA (PDB ID: 4TZY) (Zhang & Ferre-D'Amare, 2014), the numerical values of the RMSD of the homology model to the experimental structure are around 1.4 Å (Fig. 4, Table 1). This RMSD is not quite atomic accuracy (which would be sub-Angstrom) and is greater than the deviation between crystal structures of the same molecule solved by independent laboratories (e.g., <0.5 Å RMSD between the adenine riboswitch structures 4TZY (Zhang & Ferre-D'Amare, 2014) and 1Y26 (Serganov et al., 2004)). Nevertheless, the modeling confirms that the guanine and adenine riboswitch aptamers can adopt the same fold around distinct purines, and the deviation is much smaller than the length scale of nucleotides (~6 Å from one nucleotide to the next). Indeed, not much flexibility had to be modeled; there was little chance with either SWM or FARFAR of the riboswitches adopting a fold with substantial deviations. Remodeling the binding site across similar binding modes is a task to which either SWM or FARFAR is comfortably amenable.

3.2 SAM I/IV riboswitch

The prior example was an illustration of what can be done for a ligand binding RNA with a well-conserved global architecture with small sequence changes in helices, loops, and the ligand binding site that contribute to selectivity. Modeling within a single family of folds, as above, is a very different scale of challenge from modeling one fold family based on another, which

Table 1 Accuracy achieved in 200 CPU-hour simulations, using either SWM for mutations and FARFAR for helix extension or using exclusively FARFAR, achieve high-resolution models of the adenine riboswitch. RMSD (in Å) is to the actual crystal structure of the adenine riboswitch (PDB ID: 4TZY) (Zhang & Ferre-D'Amare, 2014).

Simulation	RMSD (low-E)	RMSD (best of 5)	Best RMSD	# structures
SWM	1.485	1.477	1.404	5000
FARFAR	1.479	1.478	1.393	4960

can involve global "rewiring" of the strands connecting otherwise homologous core structural elements. We illustrate the case of cross-family modeling with the SAM I/IV riboswitch.

As the first step of homology modeling, we must identify a candidate template structure. The SAM I/IV riboswitch was proposed as an RNA-puzzle (Miao et al., 2017) at a time when multiple structures of the SAM I riboswitch had already been solved by crystallography, but no riboswitches of the SAM IV or SAM I/IV class had been solved (Schroeder, Daldrop, & Lilley, 2011). The most conserved region of the SAM I/IV riboswitch appeared highly homologous to the SAM I riboswitch, and the literature proposed homologies at the family level (Weinberg et al., 2008) (Fig. 5). Taking the next two steps, we optimize the structure for SAM I riboswitch (2YGH) (Schroeder et al., 2011) with ERRASER-Phenix, and we draw out the secondary structure of SAM I/IV (Fig. 5). Then, we isolate template regions homologous to the target structure. Specifically, the red region of Fig. 5, which contains the ligand binding site as well as three of the four helices comprising the core junctional architecture of the riboswitch, is highly

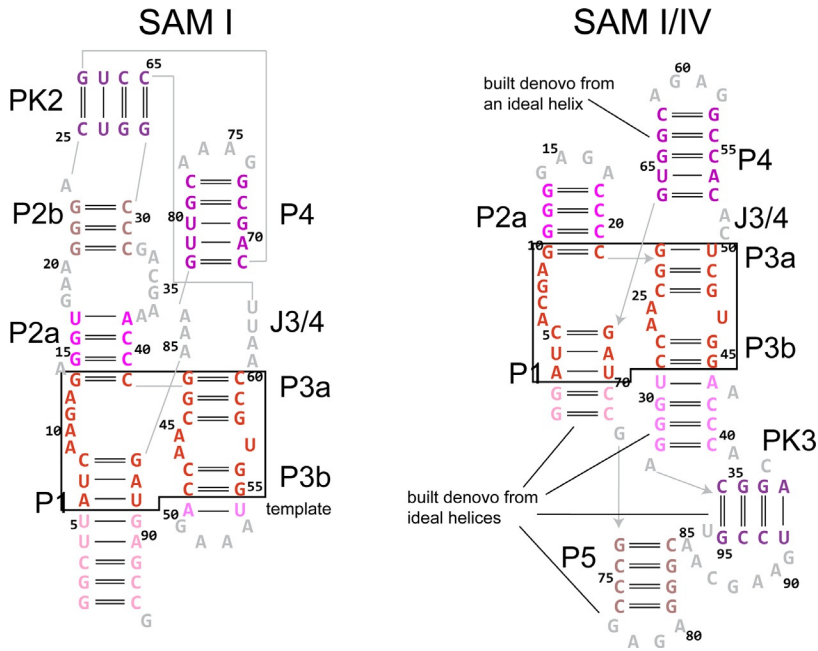

Fig. 5 The secondary structure and sequence of SAM I and SAM I/IV riboswitches are highly dissimilar, except for a well-conserved ligand binding site (red).

conserved. To model the SAM I/IV riboswitch "puzzle," we take this conserved binding site exactly, excising it from the SAM I riboswitch structure and threading on the target sequence, the fifth step of our homology modeling procedure (Fig. 2).

For the next step, we need to plan out how to model new regions of the target SAM I/V riboswitch *de novo* that are not shared with the SAM I template. The two families appear quite different outside the core (Fig. 5). Compared to the template SAM I riboswitch, in the SAM I/IV riboswitch there is a 16-nt deletion around P2, a 10-nt insertion in P3, the J3/4 loop truncated by 6 nt, and P1 is shortened by 6 nt but with an additional 4 bp helix P5 added (Fig. 5). We first summarize the regions that remain for *de novo* modeling after using the SAM I binding site homology, and we then outline the specific commands and Rosetta utilities that can be employed to carry out the modeling.

Compared to the SAM I structure, in the SAM I/IV riboswitch, P2a needs be truncated and then a GNRA tetraloop needs to be added to the remaining four Watson-Crick base pairs; this addition can be completed with FARFAR or SWM modeling. For a change in J3/4, the problem amounts to solving for the best orientation of two nucleotides (the CA loop). P4 is a 5 bp helix capped with a GNRA tetraloop and thus involves well-known RNA structures. That segment could be threaded from P4 in the template, but given its conventional structure, here we separately build it and let it be placed relative to the core as a fixed body in the next stage of *de novo* sampling of J3/4. These transformations occur on the "top" side of the template (Fig. 5) and are strictly independent of the other required modifications, described next.

On the "bottom" side of the template are the P3b and P5 modifications, which must be considered jointly and are the bulk of the challenge. P3b is extended by multiple base pairs and terminates in a loop implicated in a pseudoknot (PK3). At the same time, P5 is appended to the 3′ end of P1, terminating in a seven-residue loop, four bases in PK3, and a final nucleotide. Solving the joint fold of L3b, PK3, P5, and the 3′ loop is the major challenge.

We may solve these two challenges using either stepwise Monte Carlo or FARFAR. For the user, the requirements for running each method are broadly similar, including specification of a FASTA file (see Appendix) and a PDB file of input RNA coordinates representing the template region, renumbered to correspond to the SAM I/IV system (sequence in Appendix). Additionally, a FARFAR run requires the specification of an explicit secondary structure in "dot-bracket" notation. In FARFAR, the secondary structure may be used to impose energetic restraints to encourage base pair formation, or to guide the generation of fixed helical inputs. In contrast,

SWM currently requires the additional provision of fixed helical input structures. The secondary structure files and fixed helical inputs for use in FARFAR and SWM modeling are included in the Appendix.

With the sequence, secondary structure, and initial structure files on hand, we run simulations using two approaches, each with either FARFAR or SWM. In the first approach, we conduct a nearly naïve simulation, starting from the correct binding site but providing the rest of the structure simply as helices. In the second approach, we conduct strategic simulations which subdivided the problem into separate, more tractable modeling challenges—we first truncate P2a and complete the tetraloops capping P2a, P4, and P5. We then include the best-scoring models made for each of these sub-problems as rigid bodies in a subsequent simulation that tackles the P3b and P5 modifications. These strategic simulations reduce the scope of the problem from a problem that is very large for the SWM method (27 nucleotides to be built, seven input pieces of RNA) to something more manageable (15 nucleotides, 6 input pieces of RNA). We expected that the latter simulations would perform better because they can focus more computational power for the harder parts of the problem.

We used equal computational time for each case. The complexity of the naïve SWM challenge required simulations of five thousand Monte Carlo cycles. These simulations required on average 88,491 s, or almost 25 h, for each generated structure. Accordingly, we limited ourselves to 6000 CPU-hours to the other simulation conditions as well (Table 2). FARFAR completed an order of magnitude more total trajectories in the same amount of CPU time than SWM, for either style of problem specification. The top 10 structures from each simulation showed adequate convergence, with numerous independently modeled structures within 4 Å of the top scoring model, and with the top 10 structures having mutual RMSD <4 Å for the FARFAR simulations. As expected from this convergence level (Kappel et al., 2018), the models were also accurate compared to the actual

Table 2 6000 CPU-hour simulations with diverse starting assumptions achieve high-resolution models of the SAM I/IV riboswitch.

Simulation	RMSD (low-E)	RMSD (best of 5)	Best RMSD	# structures
SWM_naïve	10.914	3.137	3.137	222
FARFAR_naïve	5.050	4.344	3.495	12,966
SWM_strategic	4.077	3.932	2.931	691
FARFAR_strategic	5.379	4.195	3.106	17,755

experimental structure of the SAM I/IV riboswitch aptamer (PDB ID: 4L81) (Trausch et al., 2014), with the correct global fold of the RNA even in the newly modeled regions outside the ligand-binding core (Fig. 6). The best RMSD to native for the top five lowest scoring structures fell under 5 Å (Fig. 6). The RMSDs of the lowest energy models to the native structure

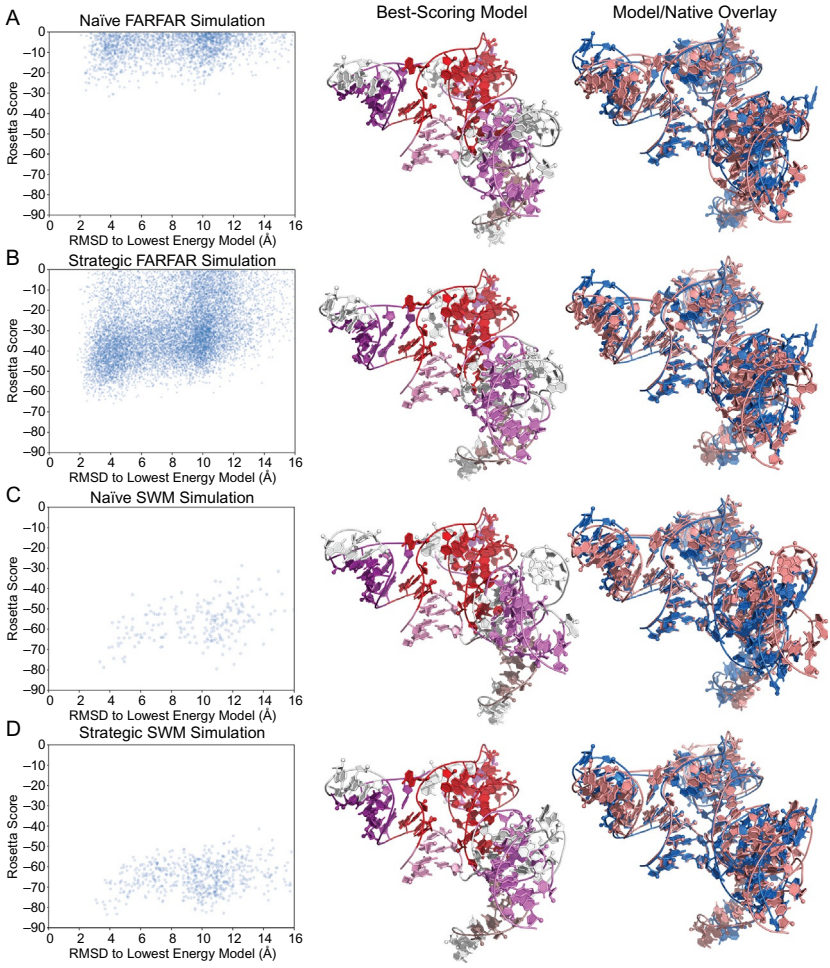

Fig. 6 Rosetta score in the default rna_res_level_energy4.wts scoring function, plotted against heavyatom RMSD to the lowest energy generated structure, comparing (A) SWM_naive, (B) FARFAR_naive, (C) SWM_strategic, and (D) FARFAR_strategic (blue). Top scoring structures for each simulation setting are depicted in the middle column with coloring corresponding to the secondary structure diagram (Fig. 4). In the right column, the top scoring structures (salmon) are overlaid on the crystal structure of the target (PDB ID: 4L81) (Trausch et al., 2014) which was not used in homology modeling (blue).

suggest that both SWM and FARFAR can achieve acceptable convergence and acceptable accuracy on this challenge, with FARFAR's ability to complete more trajectories making up for any shortcomings relative to the "high-resolution" SWM method.

4. Summary

Rosetta provides useful tools for RNA homology modeling, and their performance is well-suited to common challenges arising in current RNA structural biology. In the future, more of the steps discussed above for effective homology modeling will be automated. Very large or complex structures, for which homologous regions are limited or for which sequence/structure similarity to prior solved structures is hard to parse, currently remain inaccessible. With the acceleration of experimental methods such as cryo-EM and multidimensional chemical mapping, additional data may be employed to further accelerate these particularly challenging cases (Kappel et al., 2018; Tian & Das, 2016).

Appendix

Installation and setup

Documentation for installing Rosetta can be found here: https://www.rosettacommons.org/demos/latest/tutorials/install_build/install_build

After installing Rosetta, to install the python scripts distributed with Rosetta as `rna_tools`, execute the shell script `Rosetta/tools/rna_tools/INSTALL` (or add it to `.bashrc`).

In the next two Appendix sections, we will include command lines used for homology modeling simulations of the adenine riboswitch and SAM I/IV riboswitch. The command lines were run with Rosetta 3.10.

For each simulation step, we have included input files, output files, and command lines in the Github repository here:

https://github.com/everyday847/rosetta_rna_homology_modeling_examples

Note that when running the command lines in the following sections, it might be necessary to append "`.macosclangrelease`" or "`.linuxgccrelease`" to each executable name. Typing the executable as written below (for instance, `rna_thread`) and tab completing will show the name of the executable on your machine upon installation.

Adenine riboswitch simulation files and command lines

- The template structure used is the guanine riboswitch PDB structure 1Y27 (Serganov et al., 2004). For simplicity, we remove the ligand, water molecules, and metal ions.
- Threading the adenine riboswitch sequence onto the guanine riboswitch template. Template structures are passed in with the `-s` flag, and `-seq` indicates the sequence to place onto the template:

```
rna_thread -s 1y27_start_culled.pdb -seq ggaagauauaauccuaauga
uaugguuugggaguuucuaccaagagccuuaaacucuugauuaucuuc
```

- Minimizing the threaded template to relieve any clashes. The `-score:weights` flag here specifies the current best performing RNA scoring function, `rna_res_level_energy4.wts`. The `-restore_talaris_behavior` flag ensures that the score function exactly reproduces the default settings from when the weights were first optimized.

```
rna_minimize -s threaded.pdb -score:weights
stepwise/rna/rna_res_level_energy4.wts -
restore_talaris_behavior
```

- Since the threading application by default changes residue numbering to chain A, numbered sequentially from 1, at this point we restore standard purine riboswitch numbering. In this case, since the target structure has already been deposited with a particular numbering scheme, we can further confirm that our numbering matches up.

```
renumber_pdb_in_place.py threaded_minimize.pdb X:14-81
```

- Mismatched residues, including part of the P1 stem, are deleted from `threaded_minimize.pdb` via text editor: specifically, residues X:14–15, X:26, X:31, X:39, X:44, X:62, X:74, and X:81 are removed.
- FASTA file which defines the sequence for the complete adenine riboswitch system to be modeled, which includes a longer P1 stem than is present in the template. Note here that this FASTA file sequencing numbering must be consistent with the threaded template structure numbering after the steps above.

```
>4tzy_target.pdb X:13-83
gggaagauauaauccuaaugauaugguuugggaguuucuaccaagagccuuaaacucuug
auuaucuuccc
```

- Running FARFAR simulation for building remaining residues, left out from the template. The flags used here are as follows. The `-s` flag specifies the starting structure, which here is the threaded template structure. The `-native` flag specifies the native structure, only used for scoring

the models for later RMSD analysis. The `-minimize_rna` flag indicates that the modeling will include energy minimization steps. As before, the `-score:weights` flag specifies the current best performing RNA scoring function, and the `-restore_talaris_behavior` flag includes global scoring corrections necessary for that score function. With the `-extra_minimize_res` flag, we can allow residues in the template structure that are near *de novo* modeled residues to resample and minimize their energies. Finally, `-use_legacy_job_distributor` is necessary for cluster execution using the `rosetta_submit.py` script, as described in the Github repository.

```
rna_denovo -s threaded_minimize_culled.pdb -native
4tzy_target.pdb -fasta target.fasta -minimize_rna true -
nstruct 50 -score:weights
stepwise/rna/rna_res_level_energy4.wts -
restore_talaris_behavior -use_legacy_job_distributor -
extra_minimize_res X:25 X:27 X:30 X:32 X:38 X:40 X:42 X:45
X:61 X:63 X:73 X:75 -out:file:silent farna_rebuild.out
```

- Running SWM simulation. The flags here are analogous to those in the FARFAR simulation above, with the addition of the `-cycles` flag which indicates the number of stepwise Monte Carlo cycles to perform.

```
stepwise -s threaded_minimize_culled.pdb -native
4tzy_target.pdb -fasta target.fasta -cycles 50 -nstruct 5 -
score:weights stepwise/rna/rna_res_level_energy4.wts -
restore_talaris_behavior
-use_legacy_stepwise_job_distributor -extra_min_res X:25
X:27 X:30 X:32 X:38 X:40 X:42 X:45 X:61 X:63 X:73 X:75 -
out:file:silent swm_rebuild.out
```

Each of the resulting 500 models are extracted to individual PDB files using `extract_lowscore_decoys.py swm_rebuild.out 500` and subsequently finished (the P1 helix is extended) by FARFAR, generating 10 models each:

```
rna_denovo -s swm_rebuild.out.1.pdb -native 4tzy_target.pdb
-fasta target.fasta -minimize_rna true -nstruct 10 -
score:weights stepwise/rna/rna_res_level_energy4.wts -
restore_talaris_behavior -use_legacy_job_distributor -
extra_minimize_res X:25 X:27 X:30 X:32 X:38 X:40 X:42 X:45
X:61 X:63 X:73 X:75 -out:file:silent
threaded_minimize_stepwise_finished_by_farfar.out
```

- Retrieving top models. The `extract_lowscore_decoys` utility can be used to retrieve the top ten scoring models from each simulation:

```
extract_lowscore_decoys.py
threaded_minimize_stepwise_finished_by_farfar.out 10
extract_lowscore_decoys.py farna_rebuild.out 10
```

SAM I/IV simulation files and command lines

Many of the flags used in the command lines below are explained in more detail above in the adenine riboswitch appendix.

- The template structure for this example is the SAM I riboswitch structure (PDB ID: 2YGH) (Schroeder et al., 2011). For both FARFAR and SWM simulations, we input as a PDB file the section of the template that is homologous to the target, threaded with the target sequence. Below is the FASTA file corresponding to the inputted homology structure. As described above, this PDB file must be renumbered; for this, we use the standard numbering for the SAM I/IV riboswitch family. Command lines for these steps are analogous to the adenine riboswitch example above.

```
>start.pdb A:1-10 A:21-28 A:45-50 A:67-71
ggaucacgagcggcaaccggugcugaucc
```

- FASTA file for complete SAM I/IV riboswitch system to be modeled:

```
>sam_I/IV A:1-96
ggaucacgaggggagaccccggcaaccugggacggacacccaaggugcucacaccggag
acgguggauccggcccgagagggcaacgaaguccgu
```

- PDB files for non-template helices. For modeling helices of SAM I/IV that are not taken from the template structure, we can include PDB files corresponding to idealized A-form helices. These PDBs may be generated automatically by passing the corresponding sequence to rna_helix.py and renumbering (example below). Below, we list FASTA files for all the non-template helices used as inputs for FARFAR and SWM modeling.

```
rna_helix.py -o HELIX2.pdb -seq ggg ccc
renumber_pdb_in_place.py HELIX2.pdb A:11-13 A:18-20
```

All FASTA files for the input helix PDB files:

```
>HELIX2.pdb A:11-13 A:18-20
gggccc
>HELIX4.pdb A:29 A:44
ua
```

```
>HELIX5.pdb A:30-32 A:40-42
gggccc
>HELIX6.pdb A:53-57 A:62-66
caccgcggug
>HELIX7.pdb A:73-76 A:81-84
gcccgggc
>HELIX8.pdb A:34-37 A:92-95
cggauccg
```

- Secondary structure file for the SAM I/IV riboswitch in dot-bracket notation:

```
(((((....((((....))))(((..((((((.[[[[..))).))).))))..
(((((....))))))))))).((((....))))........]]]].
```

  Note: FARFAR supports traditional "dot-bracket" notation as well as the use of square, curly, and angular brackets for first through third order pseudoknots, and matching letters for higher order pseudoknots.

- Running naïve FARFAR simulation. In addition to the flags used above in the adenine riboswitch example, we use the flag `-allow_complex_loop_graph` to allow accurate scoring for the pseudoknotted structure being modeled, and we include `-superimpose_over_all` to compute RMSDs by aligning over all residues in this extensive modeling challenge:

```
rna_denovo -s starting_puzzle8_chunk.pdb HELIX2.pdb
HELIX4.pdb HELIX5.pdb HELIX6.pdb HELIX7.pdb HELIX8.pdb
-native 4L81.pdb -fasta target.fasta -save_times
-allow_complex_loop_graph true -superimpose_over_all
-cycles 100000 -nstruct 50 -use_legacy_job_distributor true
-score:weights stepwise/rna/rna_res_level_energy4.wts
-restore_talaris_behavior -minimize_rna true
-out:file:silent farna_rebuild.out
```

- Running naïve SWM simulation. Here and in later command lines, `-motif_mode` ensures that nucleotides on the ends of input structures may be energy-minimized along with newly built residues during simulation:

```
stepwise -s starting_puzzle8_chunk.pdb HELIX2.pdb
HELIX4.pdb HELIX5.pdb HELIX6.pdb HELIX7.pdb HELIX8.pdb -
native 4L81.pdb -fasta target.fasta -save_times -motif_mode
-allow_complex_loop_graph true -superimpose_over_all
```

```
-cycles 5000 -nstruct 5
-use_legacy_stepwise_job_distributor true
-score:weights stepwise/rna/rna_res_level_energy4.wts
-restore_talaris_behavior -out:file:silent swm_rebuild.out
```

- Generating intermediate inputs for strategic simulations:

  The template-P2 combination:

```
target_1.fasta:
>starting_puzzle8_chunk.pdb A:1-28 A:45-50 A:67-71
ggaucacgagggggagaccccggcaaccgugcugaucc
stepwise -s starting_puzzle8_chunk.pdb HELIX2.pdb -native
4L81.pdb -fasta target_1.fasta -save_times -motif_mode
-allow_complex_loop_graph true -superimpose_over_all
-cycles 200 -nstruct 5000 -
use_legacy_stepwise_job_distributor true
-score:weights stepwise/rna/rna_res_level_energy4.wts
-restore_talaris_behavior
-out:file:silent swm_rebuild1.out
```

  Finishing the P4 tetraloop:

```
rna_denovo -s HELIX6.pdb -working_res A:53-66 -native 4L81.pdb
-fasta target.fasta -save_times -motif_mode
-allow_complex_loop_graph true -superimpose_over_all
-cycles 10000 -nstruct 5000 -use_legacy_job_distributor true
-score:weights stepwise/rna/rna_res_level_energy4.wts
-restore_talaris_behavior -minimize_rna true
-out:file:silent farna_rebuild2.out
```

  Finishing the P5 tetraloop:

```
 rna_denovo -s HELIX7.pdb -working_res A:73-84 -native 4L81.pdb
-fasta rna_puzzle_8.fasta -save_times -motif_mode
-allow_complex_loop_graph true -superimpose_over_all
-cycles 10000 -nstruct 5000 -use_legacy_job_distributor true
-score:weights stepwise/rna/rna_res_level_energy4.wts
-restore_talaris_behavior -minimize_rna true
-out:file:silent farna_rebuild3.out
```

  Extracting the top-scoring model from each simulation to seed the final simulation:

```
 extract_lowscore_decoys.py swm_rebuild1.out 1
 extract_lowscore_decoys.py farna_rebuild2.out 1
 extract_lowscore_decoys.py farna_rebuild3.out 1
```

- Running strategic FARFAR simulation:

```
rna_denovo -s farna_rebuild3.out.1.pdb farna_rebuild2.out.1.
pdb swm_rebuild1.out.1.pdb HELIX4.pdb HELIX5.pdb HELIX8.pdb
-native 4L81.pdb -fasta target.fasta
-secstruct_file secstruct -save_times
-allow_complex_loop_graph true -superimpose_over_all
-cycles 100000 -nstruct 5000 -use_legacy_job_distributor true
-score:weights stepwise/rna/rna_res_level_energy4.wts
-restore_talaris_behavior -minimize_rna true
-out:file:silent farna_rebuild_final.out
```

- Running strategic SWM simulation:

```
stepwise -s farna_rebuild3.out.1.pdb farna_rebuild2.out.1.pdb
swm_rebuild1.out.1.pdb HELIX4.pdb HELIX5.pdb HELIX8.pdb
-native 4L81.pdb -fasta target.fasta -save_times -motif_mode
-allow_complex_loop_graph true
-superimpose_over_all -cycles 2000 -nstruct 5000
-use_legacy_stepwise_job_distributor true -score:weights step-
wise/rna/rna_res_level_energy4.wts
-restore_talaris_behavior -out:file:silent swm_rebuild_final.
out
```

- Obtaining low-energy models. To choose a subset of these models, we can use the following commands:

```
extract_lowscore_decoys.py farna_rebuild.out 10
extract_lowscore_decoys.py swm_rebuild.out 10
extract_lowscore_decoys.py farna_rebuild_final.out 10
extract_lowscore_decoys.py swm_rebuild_final.out 10
```

References

Akiyama, B. M., Laurence, H. M., Massey, A. R., Costantino, D. A., Xie, X., Yang, Y., et al. (2016). Zika virus produces noncoding RNAs using a multi-pseudoknot structure that confounds a cellular exonuclease. *Science*, *354*(6316), 1148–1152. https://doi.org/10.1126/science.aah3963.

Ames, T. D., & Breaker, R. R. (2014). Bacterial aptamers that selectively bind glutamine. *RNA Biology*, *8*(1), 82–89. https://doi.org/10.4161/rna.8.1.13864.

Barrick, J. E., & Breaker, R. R. (2007). The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biology*, *8*(11), R239. https://doi.org/10.1186/gb-2007-8-11-r239.

Chan, C. W., Chetnani, B., & Mondragon, A. (2013). Structure and function of the T-loop structural motif in noncoding RNAs. *Wiley Interdisciplinary Reviews RNA*, *4*(5), 507–522. https://doi.org/10.1002/wrna.1175.

Chao, J. A., & Williamson, J. R. (2004). Joint X-ray and NMR refinement of the yeast L30e-mRNA complex. *Structure*, *12*(7), 1165–1176. https://doi.org/10.1016/j.str.2004.04.023.

Chapman, E. G., Costantino, D. A., Rabe, J. L., Moon, S. L., Wilusz, J., Nix, J. C., et al. (2014). The structural basis of pathogenic subgenomic flavivirus RNA (sfRNA) production. *Science*, *344*(6181), 307–310. https://doi.org/10.1126/science.1250897.

Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., et al. (2010). MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D, Biological Crystallography*, *66*(Pt 1), 12–21. https://doi.org/10.1107/S0907444909042073.

Cheng, C. Y., Chou, F. C., & Das, R. (2015). Modeling complex RNA tertiary folds with Rosetta. *Methods in Enzymology*, *553*, 35–64. https://doi.org/10.1016/bs.mie.2014.10.051.

Cheng, C. Y., Chou, F. C., Kladwang, W., Tian, S., Cordero, P., & Das, R. (2015). Consistent global structures of complex RNA states through multidimensional chemical mapping. *eLife*, *4*,. e07600https://doi.org/10.7554/eLife.07600.

Cheng, C. Y., Kladwang, W., Yesselman, J. D., & Das, R. (2017). RNA structure inference through chemical mapping after accidental or intentional mutations. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(37), 9876–9881. https://doi.org/10.1073/pnas.1619897114.

Chou, F. C., Echols, N., Terwilliger, T. C., & Das, R. (2016). RNA structure refinement using the ERRASER-phenix pipeline. *Methods in Molecular Biology*, *1320*, 269–282. https://doi.org/10.1007/978-1-4939-2763-0_17.

Chou, F. C., Sripakdeevong, P., Dibrov, S. M., Hermann, T., & Das, R. (2013). Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nature Methods*, *10*(1), 74–76. https://doi.org/10.1038/nmeth.2262.

Correll, C. C., Beneken, J., Plantinga, M. J., Lubbers, M., & Chen, Y. (2003). The common and the distinctive features of the bulged-G motif based on a 1.04 a resolution RNA structure. *Nucleic Acids Research*, *31*(23), 6806–6818. https://doi.org/10.1093/nar/gkg908.

Cruz, J. A., & Westhof, E. (2011). Sequence-based identification of 3D structural modules in RNA with RMDetect. *Nature Methods*, *8*(6), 513–521. https://doi.org/10.1038/nmeth.1603.

Das, R., Karanicolas, J., & Baker, D. (2010). Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature Methods*, *7*(4), 291–294. https://doi.org/10.1038/nmeth.1433.

Dirks, R. M., & Pierce, N. A. (2003). A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry*, *24*(13), 1664–1677. https://doi.org/10.1002/jcc.10296.

Ditzler, M. A., Otyepka, M., Sponer, J., & Walter, N. G. (2010). Molecular dynamics and quantum mechanics of RNA: Conformational and chemical change we can believe in. *Accounts of Chemical Research*, *43*(1), 40–47. https://doi.org/10.1021/ar900093g.

Do, C. B., Woods, D. A., & Batzoglou, S. (2006). CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, *22*(14), e90–e98. https://doi.org/10.1093/bioinformatics/btl246.

Einvik, C., Nielsen, H., Westhof, E., Michel, F., & Johansen, S. (1998). Group I-like ribozymes with a novel core organization perform obligate sequential hydrolytic cleavages at two processing sites. *RNA*, *4*(5), 530–541.

Gesteland, R. F., Cech, T. R., & Atkins, J. F. (2005). *The RNA world*. New York: Cold Spring Harbor Press.

Jaeger, L., Michel, F., & Westhof, E. (1994). Involvement of a GNRA tetraloop in long-range RNA tertiary interactions. *Journal of Molecular Biology*, *236*(5), 1271–1276. https://doi.org/10.1016/0022-2836(94)90055-8.

Jaeger, L., Verzemnieks, E. J., & Geary, C. (2009). The UA_handle: A versatile submotif in stable RNA architectures. *Nucleic Acids Research*, *37*(1), 215–230. https://doi.org/10.1093/nar/gkn911.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., et al. (2018). Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research*, *46*(D1), D335–D342. https://doi.org/10.1093/nar/gkx1038.

Kappel, K., Liu, S., Larsen, K. P., Skiniotis, G., Puglisi, E. V., Puglisi, J. D., et al. (2018). De novo computational RNA modeling into cryo-EM maps of large ribonucleoprotein complexes. *Nature Methods*, *15*(11), 947–954. https://doi.org/10.1038/s41592-018-0172-2.

Keating, K. S., & Pyle, A. M. (2012). RCrane: Semi-automated RNA model building. *Acta Crystallographica. Section D, Biological Crystallography*, *68*(Pt 8), 985–995. https://doi.org/10.1107/S0907444912018549.

Kladwang, W., VanLang, C. C., Cordero, P., & Das, R. (2011). Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry*, *50*(37), 8049–8056. https://doi.org/10.1021/bi200524n.

Laing, C., & Schlick, T. (2011). Computational approaches to RNA structure prediction, analysis, and design. *Current Opinion in Structural Biology*, *21*(3), 306–318. https://doi.org/10.1016/j.sbi.2011.03.015.

Leontis, N. B., & Zirbel, C. L. (2012). Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In *RNA 3D structure analysis and prediction* (pp. 281–298). .

Lipchock, S. V., & Strobel, S. A. (2008). A relaxed active site after exon ligation by the group I intron. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(15), 5699–5704. https://doi.org/10.1073/pnas.0712016105.

Lorenz, R., Bernhart, S. H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., et al. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, *6*, 26. https://doi.org/10.1186/1748-7188-6-26.

Meyer, M., Nielsen, H., Olieric, V., Roblin, P., Johansen, S. D., Westhof, E., et al. (2014). Speciation of a group I intron into a lariat capping ribozyme. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(21), 7659–7664. https://doi.org/10.1073/pnas.1322248111.

Miao, Z., Adamiak, R. W., Antczak, M., Batey, R. T., Becka, A. J., Biesiada, M., et al. (2017). RNA-puzzles round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA*, *23*(5), 655–672. https://doi.org/10.1261/rna.060368.116.

Miao, Z., Adamiak, R. W., Blanchet, M. F., Boniecki, M., Bujnicki, J. M., Chen, S. J., et al. (2015). RNA-puzzles round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA*, *21*(6), 1066–1084. https://doi.org/10.1261/rna.049502.114.

Miao, Z., & Westhof, E. (2017). RNA structure: Advances and assessment of 3D structure prediction. *Annual Review of Biophysics*, *46*, 483–503. https://doi.org/10.1146/annurev-biophys-070816-034125.

Mirihana Arachchilage, G., Sherlock, M. E., Weinberg, Z., & Breaker, R. R. (2018). SAM-VI RNAs selectively bind S-adenosylmethionine and exhibit similarities to SAM-III riboswitches. *RNA Biology*, *15*(3), 371–378. https://doi.org/10.1080/15476286.2017.1399232.

Montange, R. K., & Batey, R. T. (2006). Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. *Nature*, *441*(7097), 1172–1175. https://doi.org/10.1038/nature04819.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2018). Critical assessment of methods of protein structure prediction (CASP)-round XII. *Proteins*, *86*(Suppl 1), 7–15. https://doi.org/10.1002/prot.25415.

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*(22), 2933–2935. https://doi.org/10.1093/bioinformatics/btt509.

Peselis, A., & Serganov, A. (2012). Structural insights into ligand binding and gene expression control by an adenosylcobalamin riboswitch. *Nature Structural & Molecular Biology*, *19*(11), 1182–1184. https://doi.org/10.1038/nsmb.2405.

Petrov, A. I., Zirbel, C. L., & Leontis, N. B. (2013). Automated classification of RNA 3D motifs and the RNA 3D motif atlas. *RNA*, *19*(10), 1327–1340. https://doi.org/10.1261/rna.039438.113.

Piatkowski, P., Kasprzak, J. M., Kumar, D., Magnus, M., Chojnowski, G., & Bujnicki, J. M. (2016). RNA 3D structure modeling by combination of template-based method ModeRNA, template-free folding with SimRNA, and refinement with QRNAS. *Methods in Molecular Biology*, *1490*, 217–235. https://doi.org/10.1007/978-1-4939-6433-8_14.

Popenda, M., Szachniuk, M., Antczak, M., Purzycka, K. J., Lukasiak, P., Bartol, N., et al. (2012). Automated 3D structure composition for large RNAs. *Nucleic Acids Research*, *40*(14), e112. https://doi.org/10.1093/nar/gks339.

Read, R. J., Adams, P. D., Arendall, W. B., 3rd, Brunger, A. T., Emsley, P., Joosten, R. P., et al. (2011). A new generation of crystallographic validation tools for the protein data bank. *Structure*, *19*(10), 1395–1412. https://doi.org/10.1016/j.str.2011.08.006.

Ren, A., Xue, Y., Peselis, A., Serganov, A., Al-Hashimi, H. M., & Patel, D. J. (2015). Structural and dynamic basis for low-affinity, high-selectivity binding of L-glutamine by the glutamine riboswitch. *Cell Reports*, *13*(9), 1800–1813. https://doi.org/10.1016/j.celrep.2015.10.062.

Reuter, J. S., & Mathews, D. H. (2010). RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, *11*, 129. https://doi.org/10.1186/1471-2105-11-129.

Rother, K., Rother, M., Boniecki, M., Puton, T., & Bujnicki, J. M. (2011). RNA and protein 3D structure modeling: Similarities and differences. *Journal of Molecular Modeling*, *17*(9), 2325–2336. https://doi.org/10.1007/s00894-010-0951-x.

Sarver, M., Zirbel, C. L., Stombaugh, J., Mokdad, A., & Leontis, N. B. (2008). FR3D: Finding local and composite recurrent structural motifs in RNA 3D structures. *Journal of Mathematical Biology*, *56*(1–2), 215–252. https://doi.org/10.1007/s00285-007-0110-x.

Schroeder, K. T., Daldrop, P., & Lilley, D. M. (2011). RNA tertiary interactions in a riboswitch stabilize the structure of a kink turn. *Structure*, *19*(9), 1233–1240. https://doi.org/10.1016/j.str.2011.07.003.

Serganov, A., & Patel, D. J. (2012). Molecular recognition and function of riboswitches. *Current Opinion in Structural Biology*, *22*(3), 279–286. https://doi.org/10.1016/j.sbi.2012.04.005.

Serganov, A., Yuan, Y. R., Pikovskaya, O., Polonskaia, A., Malinina, L., Phan, A. T., et al. (2004). Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chemistry & Biology*, *11*(12), 1729–1741. https://doi.org/10.1016/j.chembiol.2004.11.018.

Shortle, D., Simons, K. T., & Baker, D. (1998). Clustering of low-energy conformations near the native structures of small proteins. *Proceedings of the National Academy of Sciences*, *95*(19), 11158–11162. https://doi.org/10.1073/pnas.95.19.11158.

Sim, A. Y., Minary, P., & Levitt, M. (2012). Modeling nucleic acids. *Current Opinion in Structural Biology*, *22*(3), 273–278. https://doi.org/10.1016/j.sbi.2012.03.012.

Tian, S., & Das, R. (2016). RNA structure through multidimensional chemical mapping. *Quarterly Reviews of Biophysics, 49*, e7. https://doi.org/10.1017/S0033583516000020.

Toor, N., Keating, K. S., Taylor, S. D., & Pyle, A. M. (2008). Crystal structure of a self-spliced group II intron. *Science*, *320*(5872), 77–82. https://doi.org/10.1126/science.1153803.

Trausch, J. J., Xu, Z., Edwards, A. L., Reyes, F. E., Ross, P. E., Knight, R., et al. (2014). Structural basis for diversity in the SAM clan of riboswitches. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(18), 6624–6629. https://doi.org/10.1073/pnas.1312918111.

Vicens, Q., Mondragon, E., & Batey, R. T. (2011). Molecular sensing by the aptamer domain of the FMN riboswitch: A general model for ligand binding by conformational selection. *Nucleic Acids Research*, *39*(19), 8586–8598. https://doi.org/10.1093/nar/gkr565.

Wang, X., Kapral, G., Murray, L., Richardson, D., Richardson, J., & Snoeyink, J. (2008). RNABC: Forward kinematics to reduce all-atom steric clashes in RNA backbone. *Journal of Mathematical Biology*, *56*(1–2), 253–278. https://doi.org/10.1007/s00285-007-0082-x.

Watkins, A. M., Geniesse, C., Kladwang, W., Zakrevsky, P., Jaeger, L., & Das, R. (2018). Blind prediction of noncanonical RNA structure at atomic accuracy. *Science Advances*, *4*(5), eaar5316. https://doi.org/10.1126/sciadv.aar5316.

Weinberg, Z., Lunse, C. E., Corbino, K. A., Ames, T. D., Nelson, J. W., Roth, A., et al. (2017). Detection of 224 candidate structured RNAs by comparative analysis of specific subsets of intergenic regions. *Nucleic Acids Research*, *45*(18), 10811–10823. https://doi.org/10.1093/nar/gkx699.

Weinberg, Z., Regulski, E. E., Hammond, M. C., Barrick, J. E., Yao, Z., Ruzzo, W. L., et al. (2008). The aptamer core of SAM-IV riboswitches mimics the ligand-binding site of SAM-I riboswitches. *RNA*, *14*(5), 822–828. https://doi.org/10.1261/rna.988608.

Westhof, E., Masquida, B., & Jossinet, F. (2011). Predicting and modeling RNA architecture. *Cold Spring Harbor Perspectives in Biology*, *3*(2). a003632. https://doi.org/10.1101/cshperspect.a003632.

Wu, L., Chai, D., Fraser, M. E., & Zimmerly, S. (2012). Structural variation and uniformity among tetraloop-receptor interactions and other loop-helix interactions in RNA crystal structures. *PLoS One*, 7(11). e49225https://doi.org/10.1371/journal.pone.0049225.

Xu, X., & Chen, S. J. (2018). Hierarchical assembly of RNA three-dimensional structures based on loop templates. *The Journal of Physical Chemistry. B*, *122*(21), 5327–5335. https://doi.org/10.1021/acs.jpcb.7b10102.

Yesselman, J. D., Denny, S. K., Bisaria, N., Herschlag, D., Greenleaf, W. J., & Das, R. (2018). RNA tertiary structure energetics predicted by an ensemble model of the RNA double helix. *BioRxiv*, 341107. https://doi.org/10.1101/341107.

Zhang, J., & Ferre-D'Amare, A. R. (2014). Dramatic improvement of crystals of large RNAs by cation replacement and dehydration. *Structure*, *22*(9), 1363–1371. https://doi.org/10.1016/j.str.2014.07.011.