Published in final edited form as:

J Phys Chem B. 2018 May 31; 122(21): 5327–5335. doi:10.1021/acs.jpcb.7b10102.

Hierarchical Assembly of RNA Three-Dimensional Structures Based on Loop Templates

Xiaojun Xu^{1,2}, Shi-Jie Chen^{2,*}

¹Institute of Bioinformatics and Medical Engineering, Jiangsu University of Technology, Changzhou, Jiangsu 213001, China

²Department of Physics, Department of Biochemistry, and Informatics Institute, University of Missouri, Columbia, MO 65211, USA

Abstract

The current RNA structure prediction methods cannot keep up the pace of the rapidly increasing number of sequences and the emerging new functions of RNAs. Template-based RNA three-dimensional structure prediction methods are restricted by the limited number of known RNA structures and traditional motif-based search for the templates does not always lead to successful results. Here we report a new template search and assembly algorithm, hierarchical loop template-assembly method (VfoldLA). The method searches for templates for single strand loop/junctions instead of the whole motifs, which often renders no available templates, or short fragments (several nucleotides), which requires long computational time to assemble and refine. The VfoldLA method has the advantage of accounting for local and nonlocal inter-loop interactions. Benchmark tests indicate that this new method can provide low-resolution predictions for RNA conformations at different levels of structural complexities. Furthermore, the VfoldLA-predicted conformations may also serve as reliable putative models for further structure prediction and refinements. VfoldLA is accessible at “<http://rna.physics.missouri.edu/vfoldLA>”.

Graphical Abstract

* Author to whom correspondence should be addressed; chenshi@missouri.edu.

Supporting Information

The Supporting Information (PDF) with additional figures and tables is available free of charge.

The authors declare no competing financial interest.

Introduction

RNA three-dimensional (3D) structures are critical for RNA cellular functions^{1–3} such as regulation of gene expression. For example, RNA riboswitch,^{4–8} a regulatory segment of a messenger RNA molecule that binds a small molecule, can control transcription or translation through conformational changes. However, compared with RNA primary structure (sequence)^{9–11} and secondary (two-dimensional, 2D) structure^{12–15} determination, experimental determination of RNA 3D structures at atomic precision is much more expensive and laborious. This has prompted the development of computational methods to predict RNA 3D structures from their nucleotide sequences.

Computational models for RNA 3D structure prediction^{16–22} fall into two general categories: physics-based and knowledge-based methods. Because of the enormous degrees of freedom of an RNA 3D structure, most physics-based algorithms use coarse-grained RNA conformations to reduce the computational complexity and to enhance the conformational sampling.^{23–34} Such systems may be modeled with knowledge-based coarse-grained potentials derived from the known structures. Combined with discrete molecular dynamics (DMD) or other similar approaches, coarse-grained models can predict large RNA structures. For example, iFoldRNA^{30,31} uses three-bead model to predict RNA 3D structures and folding dynamics by applying DMD simulations with base-pairing and base-stacking energy parameters. Among the 153 test RNAs, 84% of the predicted structures have an RMSD of $< 4 \text{ \AA}$ to the native RNA structures. SimRNA^{33,34} uses Monte Carlo sampling for (coarse-grained) conformations and a statistical potential for scoring the different structures, and graph theory-based method (RAG)^{35,36} samples global helix topology (represented as graphs) according to a knowledge-based statistical potential derived from the bending and torsion measures of internal loops and the radii of gyration of the known RNA structures.

Knowledge-based structure sampling methods,^{37–50} construct RNA 3D structures using templates ranging from nucleotide fragments to structural motifs. Guided by a knowledge-base energy function that encodes base stacking and pairing interactions, FARNA/
FARFAR^{37–39} models RNA structures by assembling short fragments (1–3 nucleotides)

extracted from a single crystal structure. The method can predict structures with atomic ($< 3.0 \text{ \AA}$) resolution for short ($< 30\text{nt}$) RNAs. To treat large RNAs, most methods require 2D structures as input in order to reduce the conformational search space.

The recently developed Vfold3D is a template-based coarse-grained structure prediction model. For a given 2D structure, Vfold3D^{40–42} divides the structure into different motifs. Helical regions are modeled as A-form helices. The other (non-helix) motifs, such as hairpin loops and multi-branched junctions, are modeled based on motif-based templates extracted from the known RNA structures. Vfold3D predicts 3D structures through the assembly of the structures of non-helix motifs and A-form helices. With a similar general approach, RNAComposer^{43,44} treats secondary structural motifs in a more detailed way. For example, the model allows slightly distorted helical conformations. Furthermore, the model considers motifs with tertiary contacts and with the CYANA structure calculation program,⁵¹ can generate 3D structure elements not included in the RNA FRABASE database.^{52,53} Based on the machine translation principle, RNAComposer enables prediction of large RNA 3D structures with relatively high quality. Similarly, MC-sym⁴⁵ builds all-atom structures using 3D nucleotide cyclic motif (NCM) fragments, which are not only extracted from known RNA 3D structures, but also built on the fly if necessary. MC-sym has been quite successful for structure prediction for a great variety of RNAs, especially for short RNAs with 2D structures as input.

Here, we present a new approach, VfoldLA, for predicting RNA conformations for the purpose of structure prediction. The method is based on a new template search method. Unlike the previous models which search for templates based on the whole motif or piece-wise fragments, VfoldLA searches for templates for single strands of loops/junctions. For example, for a two-way junction (internal loop) which is composed of two loop strands, the model searches templates for each strand separately rather than for the whole two-way junction motif together or for short fragments in each strand. VfoldLA has a much higher success rate to find appropriate templates than the previous motif-based approaches in Vfold3D and also has much higher computational efficiency than fragment-based approaches. Similar to the fragment-based methods, the loop/junction-based search strategy has the potential to give new structures for the different motifs.

The new model here has the advantage of accounting for interactions within the loop/junction strand. The template-based methods for RNA 3D structure prediction are severely limited by the low success rate of finding a proper template for a given motif. The advantage of this new approach is its ability to significantly augment the possibility of finding a proper template. The improvement is achieved by focusing on the search for the template for each loop strand instead of for the whole structural motif. However, such identified loop templates may lose information about inter-loop interactions, resulting in less accurate predictions.

We note that if the specific whole motif-based templates are available and can be successfully found, it is possible that the original whole motif-based approach can give better predictions than the current loop strand template-based method. However, the merit of the current new method is not for more accurate structure predictions than all the existing

methods for all the tested cases. Instead, the merit here is for its ability to predict a set of low-resolution structures based on the limited available templates. Indeed, as shown in the present study, benchmark tests indicate that the new method can successfully build template-based low-resolution conformations for structure prediction. Although for certain test cases, the current loop-assembled structures might not be as accurate as those predicted by other structure prediction models, the VfoldLA-generated structures may serve as useful scaffolds for future structure refinement studies.

Algorithm and methods

Helices and four types of loops.

An RNA 2D structure can be divided into helices and single-strand loops. Helices are canonical double-stranded regions and can be connected by single strand loops. According to different helix-loop-helix connection mode, we classify four types of loops, as shown in Fig. 1: “helix2” loop for a single strand that connects two helices; “hairpin” loop, which can be viewed as a special type of “helix2” loop with two overlapping helices, “Tail5” and “tail3” loops for segments of unpaired nucleotides at the 5’ and the 3’ ends, respectively.

Loop template database.

The loop template database is built from 2557 RNA PDB structures. We extract the 3D templates for the four types of loops with different loop length L :

1. For a given RNA 3D structure, we first extract the A-form helices. We allow only canonical base pairs (A-U, G-C, and G-U) in a helix. Noncanonical base pairs will break a otherwise continuous helix into separate helices.
2. We identify all the loops and their connected helices. In order to account for the helix-loop structural correlations, we include the terminal base pair(s) of the connected helix (helices) in the loop conformation (see the red oval shape in Fig. 1).
3. We remove the redundant loop templates. Two loop structures are redundant if they belong to the same loop type, have the same sequence, and their heavy-atom root mean square deviation (RMSD) is within 1.5Å.
4. We construct a template database for all the non-redundant loop structures. Fig. 1 shows the statistics for the current template database. See Supplementary Table S1 for a more completed database.

Sequence similarity-based loop score.

Given a query loop sequence, VfoldLA scores the loop sequence similarity according to the following rule: The score s_i for nucleotide position i is equal to 0 for perfect match, 1 for purine/pyrimidine-type match, and 2 for no match. The scoring scheme considers the similarity of nucleotides, assuming that less changes in base substitution result in less changes in loop structure, including the local and nonlocal interactions. For example, $s^{C \rightarrow U} = 1$ for a nucleotide C substituted by U, and $s^{C \rightarrow A/G} = 2$ for a nucleotide C (pyrimidine)

replaced with A or G (purine). We rank the loop templates according to the loop score defined as the sum over all the nucleotides in the loop: $S^{\text{loop}} = \sum_i s_i$.

Loop template directed 3D structure assembly.

For a given 2D structure, VfoldLA predicts the 3D structures from the loop-based templates. We use the case of class II preQ1 riboswitch, shown in Fig. 2, as an example to illustrate the structure prediction method.

1. To identify the structure motifs (helices and the different types of the loops) from the given 2D structure. For example, the 2D structure shown in Fig. 2A has four helices and eight loops: L_1 and L_6 are hairpin loops, L_8 is a tail3 loop, and rest are helix2 loops.
2. To select $N-1$ helix2 loops for helix orientations. The overall global fold can be determined by the helix orientations, and helix orientations can be determined by the helix2 loops. Specifically, an N -helix structure can be determined by $N-1$ helix2 loops. For a fixed N -helix orientation as determined by the $N-1$ helix2 loops, as shown in Fig. 2B, the orientations of the four helices can be sampled by three helix2 loops (L_2 , L_3 , and L_5). By assembling the 3D structures of the loops based on loop template database and the helices, we can determine the helix configurations from the template structures of the $N-1$ helix2 loops.

There may exist multiple ways to select the $N-1$ helix2 loops for an N -helix configuration. To configure the helices, VfoldLA follows the 5' to 3' sequential order to select the $N-1$ helix2 loops. If no template is found for one or more helix2 loops, VfoldLA uses alternative selection rules to assemble the helices. The search for the templates may involve base substitutions. For the base conformation, VfoldLA uses the three-atom alignment (shown in Supplementary Figure S1) to keep the base orientations in the templates.

3. To assemble helices based on the loop templates. For each of the $N-1$ helix2 loops, VfoldLA scans the template database based on the loop type and the loop length L , and ranks the loop templates according to the aforementioned sequence similarity-based scores. The helices are then configured based on the optimal loop templates. In the helix assembly process, structures with steric clash between helices and loops are excluded. To check steric clashes in the assembled structures, we allow at most three loop-helix atom pairs with a distance $> 1.6\text{\AA}$ for each loop strand.
4. To configure the rest helix2 loop conformations by connecting the helices. We allow a $5\text{\AA}$ heavy-atom RMSD of the helix terminal base pairs when selecting templates for the rest helix2 loops. Specifically, the optimal templates are chosen based on the loop type, loop size, sequence-based loop score, and the criteria of within $5\text{\AA}$ -RMSD-cutoff. For example, the L_4 and L_7 helix2 loops in Fig. 2(C). L_4 is a loop connecting two terminal base pairs: A-U and A-U. We use the heavy atoms of those four nucleotides to calculate the RMSD for each templates to find the optimal template with the best score (if RMSD is within $5\text{\AA}$ cutoff). In the

selection of the templates, loop structures that have steric clash with any of the pre-existing atoms would be excluded. If none of the loop templates can satisfy the loop closing criteria, the specific helix orientation would be removed.

5. To configure the rest non-helix2 loops. After configuring the helices and the helix2 loops, we add the structures for the remaining hairpin loops and the single strand tail loops based on the templates and the helix 3D configuration, following the same criteria as for the loop closing. For example, as shown in Fig. 2D, the two hairpin loops (L_1 and L_6) and the tail3 loop L_8 share a base pair with their respective helices. When configuring such loops, we align the common base pairs to the pre-assembled helices.
6. To refine the overall structure. We apply Replica-Exchange Molecular Dynamics (MD) simulation to refine the assembled structures. In the structure assembly process, loop strand crossing (into the helical regions) could possibly occur for unusual cases. Manual local structure adjustment may be applied to disentangle the strands as a pre-simulation treatment.
7. To score the predicted structure. The sequence score of a 3D structure S^{seq} is the sum over the scores for the loops $\sum S^{\text{loop}}$. To account for the inter-loop interactions, we introduce the MD score S^{MD} , which is the total potentials from the simulations. We rank all predicted structures by the total score $S^{\text{tot}} = S^{\text{seq}} \times w + S^{\text{MD}}$. Here, we set the weighting factor $w = 10$.

For a given sequence, the current version of the VfoldLA model sets the maximum number of the output helix orientations (global folds) to be 100. For each helix configuration, the different loop conformations usually do not cause significant changes to the global fold. Therefore, for each given set of helix configuration, VfoldLA predicts one all-atom (loop) structure with the optimal score. We use an RMSD cutoff of 5 Å to cluster the predicted ranked structures. The centroid structures of the top five clusters are considered as the predicted structures for the given 2D structure.

Results and discussion

Benchmark test against other models.

FARNA/FARFAR^{37–39} and MC-Fold |MC-Sym pipeline⁴⁵ are designed to predict RNA 3D structures with near atomic resolutions. We use the cases from the previously published paper⁴⁰ to test our predictions against the FARNA/FARFAR, MC-Fold |MC-Sym pipeline, and the Vfold3D^{40,41} model. As shown in Fig. 3, with the 2D structures (shown in Supplementary Table S2) extracted from native structures as input, VfoldLA achieves similar performance compared with the above three models. However, the prediction of top ranked cluster center may not always be the best prediction (with minimum RMSD) among the top five cluster centers, indicating the limitation of the sequence similarity-based scoring function used in the VfoldLA model.

To further test the model, we compare VfoldLA with RNAComposer,⁴³ SimRNA,^{33,34} and Vfold3D⁴¹ servers, for 22 RNAs as shown in Table. 1. The comparisons are based on the

same 2D structures provided as input. The corresponding 2D structures are extracted from native 3D structures (see Supplementary Table S3 for details). To avoid using (motif or loop) templates from the native structures, we exclude all the native loop and motif structures from the template database for the VfoldLA and Vfold3D predictions. For RNAComposer, we upload the native structures to exclude templates from the native structures (in batch mode). The test RNAs involve a variety of structural motifs: six stem-loop structures (containing hairpin, internal, or bulge loops), four 3-way, one 4-way, and one 5-way multi-branched structures, five pseudoknotted structures, and five tertiary structures (containing crossing base pairs).

We consider the centroid structures of the top five clusters as the VfoldLA predictions. We use the first five output structures as the predicted structures for Vfold3D. For SimRNA, we use the three clusters given by the default output. For the RNAComposer predictions, we set the maximum number of the 3D models for each secondary structure to be five. The cluster center with minimum RMSD is selected as the predicted structures for each model. The qualities of the predicted structures are evaluated by the RNAAssess⁵⁴ server, with heavy-atom RMSDs, INF (interaction network fidelity) for all the base pairing and stacking interactions, and the clash scores.

As shown in Table. 1, our VfoldLA model gives near atomic predictions for the six stem-loop structures. The average RMSD for the stem-loop structure predictions is 4.5 Å. For the multi-branched structures, five of the six tested cases have low-resolution (< 10.0 Å) predictions, while only one of them (PDB: 2mhi) has a large RMSD (12.6 Å), indicating that no proper loop templates exist in the current template database for at least one of the "helix2" loops. The average RMSD for the predictions of the tested multi-branched structures is 8.5 Å. For the pseudoknotted structures, our model can also achieve near atomic predictions (average RMSD is 4.9 Å for the five tested cases). Differed from our previous model, Vfold3D, our new model can also deal with RNAs with tertiary contacts (crossing base pairs). Because of the base pair crossing, besides the local motifs (hairpin and multi-branched loops), a 2D structure may also contain tertiary motifs, such as hairpin-hairpin, 2-way-hairpin kissing structures (see Supplementary Figure S2 for illustration). Currently, there are a limited number of such motifs in the known RNA 3D structures. As a result, most template-based methods, such as MC-sym and Vfold3D, can not deal with RNAs with tertiary contacts. For the five tested tertiary motifs (see Supplementary Table S3 for detailed 2D structures), VfoldLA can obtain low-resolution (< 10.0 Å) predictions, except for one (pdb: 4jf2) case, which has the RMSD of 12.6 Å. The average RMSD for the five tested tertiary structure predictions is 7.7 Å. Overall, our new model can reach low-resolution (6.4 Å in average) predictions with only 2 cases of high RMSDs (> 10.0 Å) out of 22 cases.

Among the four models, Vfold3D gives the lowest RMSD for the 5 cases in the totally 12 (stem-loop and multi-way) structures. However, due to the limited availability of the current motif-based templates, Vfold3D cannot predict structures with pseudoknotted and other tertiary motifs. SimRNA gives comparable accurate predictions (7.3 Å on average for the tested structures). More importantly, it shows similar performance for the different structure types, indicating that the energy functions used to guide the MC simulations are likely not motif type-biased. RNAComposer achieves the lowest RMSD value for the 8 cases in the

totally 22 tested RNAs. The average RMSD is 6.6 Å. Among them, there are four cases with high-resolution (RMSD < 2.0) predictions. This can be explained by the use of known near native fragments for interior and multi-loop motifs, leading to near exact matches. Instead of the standard A-form helices, the use of the templates for helices with slight sequence-dependent distortions might also contribute to the improved accuracy achieved by the RNAComposer. The accurate modeling of helices may be important because a small difference between the A-form helix and the realistic (distorted) RNA helix could result in a notable structural difference in the global fold.⁴² By excluding all the templates from the native structures, our model VfoldLA can also reach comparable predictions (as the RNAComposer) with 4 lowest RMSD values predictions out of 22 tested cases.

With the same 2D structures as input, the INF value shown in Table. 1, which evaluates the predictions of the interaction network, is in the range of (0.54–0.64) on average, indicating that all the four models may capture the similar interaction patterns. However, the RNAComposer predictions have much lower clash values (51.0 on average) than other three models (around 150.0 on average). Therefore, all the predicted structures, especially from SimRNA, Vfold3D, and VfoldLA, need further structural refinement.

Computational efficiency.

For 3D structure prediction methods based on Molecule dynamics and Monte Carlo simulations, such as SimRNA, the conformational sampling may take a long computational time. For motif template-based 3D structure prediction method, such as RNAComposer and Vfold3D, the calculation takes a shorter computational time. VfoldLA is loop template directed 3D structure assembly/prediction model, which dissects whole motifs (such as multi-way junctions) into loops. The computational time of VfoldLA varies, depending on the availability of the loop templates. For the tested RNAs, the computational time ranges from minutes to hours, see Supplementary Table S4 for details.

Class II preQ1 riboswitch: 4jf2.—As shown in Table. 1, VfoldLA can predict structures of the different types. If homologous loop templates are available in the database, the model can likely offer reliable predictions for the structure. If no homologous templates exist in the database, the model might also predict near native global folds, as illustrated by the case below for class II preQ1 riboswitch.

As shown in Fig. 4A, the 77-nt class II preQ1 riboswitch consists of four helices, eight loops, and the long-range base pairing helix H₃. With the native structure (PDB: 4jf2) removed from the template database, the VfoldLA predicted top-ranked structure has RMSD of 12.6 Å to the PDB structure. With the same 2D structure as input, the SimRNA prediction shows similar accuracy (11.9 Å) as the VfoldLA model. In contrast, RNAComposer achieves much better prediction with an RMSD of 6.0 Å. As we can see in Fig. 4B, VfoldLA correctly predicts the main global fold (the orientation of helices). The primary difference between the VfoldLA prediction and the PDB structure comes from the orientation of helix H₃ and the resultant placement of helix H₄. As shown in the inset of Fig. 4, the selected template of the L₃ loop, which connects helices H₂ and H₃, causes the suboptimal prediction for the H₃ and H₄ orientations.

Predictions for RNA Puzzles.

RNA Puzzles is a CASP-like blind test and critical assessment of RNA structure prediction algorithms.^{55–57} For the challenges #15 and #17, we participated the *RNA Puzzles* competition by using VfoldLA and Vfold3D,⁴¹ which generates RNA 3D structures based on motif-based templates, to predict the structures. The detailed prediction results can be found on the *RNA Puzzles* server (<http://ahsoka.u-strasbg.fr/rnapuzzles>). To further test the capability of the VfoldLA model, we here also predict the structure for the released challenge #3 of the *RNA Puzzles*. The sequences and 2D structures used for the three puzzles are listed in Supplementary Table S5. The purpose here is to analyze the structures, generated by VfoldLA for *RNA Puzzles* predictions, to illustrate the advantage and limitations of this new model.

Challenge #3: Glycine riboswitch.—Since our model needs sequence and 2D structure as input, we extract the base pairing information from the native structure (PDB: 3owz, Glycine riboswitch), and use the sequence and the 2D structure as the input to VfoldLA. To test the prediction, we remove all the similar structures with PDB IDs 3owz, 3owi, 3oww, 3ow0, 3owb, 3owd, 3owe, 3owm, and 3owj. As shown in Fig. 5(A-2), and Supplementary Figure S4(A), the input 2D structure has 84 nucleotides, containing five helices and nine loops. The predicted all-atom structure, shown in Fig. 5(A-1), has an RMSD of 9.7 Å to the native structure (PDB: 3owz), with the main difference in the 3-way junction structure. The loop templates used for the best prediction are listed in Supplementary Figure S3.

Our original Vfold3D-predicted structure for challenge #3 has an RMSD of 7.2 Å and is more accurate than the VfoldLA prediction. The difference between the original Vfold3D and the current VfoldLA predictions mainly comes from the use of the different templates. For example, for the three-way junction (3wj) in the glycine riboswitch, which consists of three loop strands, Vfold3D uses the motif-based templates, where each template is a 3wj structure. VfoldLA, however, predicts the 3wj structure through the assembly of three templates, each for a loop strand. Therefore, VfoldLA may lose the information about inter-strand interactions.

Challenge #17: Pistol ribozyme.—Starting from the 62-nt sequence, we first predicted the 2D structure (as shown in Fig. 5B-2) using Vfold2D,⁴¹ and used the predicted 2D structure as the input for VfoldLA. As shown in Fig. 5(B-2) and Supplementary Table S5, the predicted 2D structure contains four helices and seven loops, including crossing base paired helices. The predicted structure, shown in Fig. 5(B-1), and Supplementary Figure S4(B), has an RMSD of 9.4 Å to the native structure (PDB: 5k7c). The predicted structure has native-like global fold except for the H₄ helix and some misplaced loop parts. The hairpin loop shown in cyan in the sequence is actually not included in the crystal structure. The loop templates used for the best prediction are listed in Supplementary Figure S3.

Challenge #15: Hammerhead ribozyme.—The challenge #15 is an RNA-RNA complex, consisting of sequences of 48 and 20 nts, respectively. We used VfoldCPX⁵⁸ to predict the 2D structure. The structure contains five helices and ten loops, including crossing base pairs (Fig. 5C-2). Using the 2D structure as the input, the VfoldLA predicted the 3D

structure (Fig. 5C-1, and Supplementary Figure S4(C)) with an RMSD of 8.5 Å to the native structure (PDB: 5di4). The loop templates used for the best prediction are listed in Supplementary Figure S3. Even though the H₁ helix in the predicted 2D structure does not exist in the crystal structure, VfoldLA can still predict the near-native global fold.

Conclusions

We have developed a loop template assembly method for RNA 3D structure prediction. Tests of the model show that the model can give low-resolution predictions for 3D structures: the average RMSDs for the predicted stem-loop, N-way (N = 3), pseudoknotted, and tertiary structures are 4.5, 8.5, 4.9, and 7.7 Å, respectively. The improved accuracy in the structure prediction can be attributed to the implicit consideration of the local (intra-loop) interactions within each loops. For the fragment assembly based structure prediction methods, such as FARNA/FARFAR,^{37–39} the smaller fragments account for local (short-range) intra-loop interactions but ignore the long-range (intra- and inter- loop) interactions. As a result, those methods need sufficient sampling of the conformational space with proper scoring functions to achieve better predictions. For motif-based assembly methods, such as MC-sym,⁴⁵ Vfold3D,^{40,41} and RNAComposer,⁴³ the use of the whole motifs, instead of small fragments, can account for not only the local intra-loop contacts, but also the inter-loop interactions within motifs. With the increase in the diversity of the motif template library, we may expect improved predictions from motif-based template methods.

RNA solution structures, especially multibranched junctions, can show multiple helix orientations, and a crystal structure may represent only one of the several structures in a flexible ensemble. Moreover, interactions with other molecules, such as proteins, may result in large structural changes of the RNA. Therefore, predicting multiple RNA conformations is important for understanding RNA folding and functions, which often involve multiple conformational switches. The VfoldLA model developed here, which is based on the template database extracted from known structures of RNA-only and RNA-involved complex systems, can give multiple putative RNA structures. Therefore, the VfoldLA model may represent a step toward accurate prediction of RNA structural ensemble. Furthermore, with the continuous expansion of the database of known structures, the VfoldLA may become increasingly reliable.

Our loop template-assembly based method (VfoldLA) can efficiently explores the global arrangements of A-form helices guided by the known homologous loop structures. By assembling loop templates, the VfoldLA method attempts to strike the balance between the conservation of structural motifs and the computational efficiency. The predicted low-resolution structures may serve as starting points for other physics-based predictions of high-resolution structures. This multiscale (from low- to high-resolution) approach can potentially offer a useful way for accurate prediction of RNA structures, especially for structures with long-range interactions such as inter-loop kissing base pairs. Moreover, as shown in Supplementary Figures S5–S9, the assembled structures might have same/similar sequence scores but large RMSDs between them, indicating that the sequence similarity-based scoring function may not always correlate with the 3D structural similarity. By accounting for the interactions between loops and helices, the total score, which is the combination of sequence

and MD energy scores, gives better correlations between score and structural similarity (RMSD) than the sequence score alone.

In general, VfoldLA can predict a set of all-atom structures from any 2D structure, with the condition that homologous loop structures exist in the template database. However, due to the large number of combinations for the different templates of the different loops, the current model can only handle RNAs containing small number of helices (< 8 helices) in order to avoid excessively long computational times. To enhance the computational efficiency, further development of the model could use a step-wise approach: prediction of the all-atom structures for each motif, followed by the motif assembly to obtain the whole structure. This hybrid approach may combine the advantages of the motif- and loop/fragment-based assembly methods and can be computationally efficient.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

This work was supported by the National Institutes of Health through grant R01-GM063732.

References

1. Clancy S RNA functions. Nat. Edu 2008, 1, 102.
2. Eddy SR Non-coding RNA genes and the modern RNA world. Nat. Rev. Genet 2001, 2, 919–929. [PubMed: 11733745]
3. He L; Hannon GJ MicroRNAs: small RNAs with a big role in gene regulation. Nat. Rev. Genet 2004, 5, 522–531. [PubMed: 15211354]
4. Nudler E; Mironov AS The riboswitch control of bacterial metabolism. Trends Biochem. Sci 2004, 29, 11–17. [PubMed: 14729327]
5. Cheah MT; Wachter A; Sudarsan N; Breaker RR Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. Nature 2007, 447, 497–500. [PubMed: 17468745]
6. Bocobza S; Adato A; Mandel T; Shapira M; Nudler E; Aharoni A Riboswitch-dependent gene regulation and its evolution in the plant kingdom. Genes Dev 2007, 21, 2874–2879. [PubMed: 18006684]
7. Blount KF; Breaker RR Riboswitches as antibacterial drug targets. Nat. Biotechnol 2006, 24, 1558–1564. [PubMed: 17160062]
8. Chen S-J; Burke D Computational methods for understanding riboswitches; Methods Enzymol 553, Academic Press 2015.
9. Chu Y; Corey DR RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic. Acid. Ther 2012, 22, 271–274. [PubMed: 22830413]
10. Siu H; Zhu Y; Jin L; Xiong M Implication of next-generation sequencing on association studies. BMC Genomics 2011, 12, 322. [PubMed: 21682891]
11. Wang Z; Gerstein M; Snyder M RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet 2009, 10, 57–63. [PubMed: 19015660]
12. Felden B RNA structure: experimental analysis. Curr. Opin. Microbiol 2007, 10, 286–291. [PubMed: 17532253]
13. Xu Z; Culver G RNA structure experimental analysis—chemical modification. Methods Enzymol 2013, 530, 363–380. [PubMed: 24034333]
14. Ishitani R; Yokoyama S; Nureki O Structure, dynamics, and function of RNA modification enzymes. Curr. Opin. Struct. Biol 2008, 18, 330–339. [PubMed: 18539024]

15. Butcher SE; Pyle AM The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Acc. Chem. Res* 2011, 44, 1302–1311. [PubMed: 21899297]
16. Laing C; Schlick T Computational approaches to RNA structure prediction, analysis, and design. *Curr. Opin. Struct. Biol* 2011, 21, 306–318. [PubMed: 21514143]
17. Sim AY; Minary P; Levitt M Modeling nucleic acids. *Curr. Opin. Struct. Biol* 2012, 22, 273–278. [PubMed: 22538125]
18. Rother K; Rother M; Boniecki M; Puton T; Bujnicki JM RNA and protein 3D structure modeling: similarities and differences. *J. Mol. Model* 2011, 17, 2325–2336. [PubMed: 21258831]
19. Shapiro BA; Yingling YG; Kasprzak W; Bindewald E Bridging the gap in RNA structure prediction. *Curr. Opin. Struct. Biol* 2007, 17, 157–165. [PubMed: 17383172]
20. Magnus M; Matelska D; Lach G; Chojnowski G; Boniecki MJ; Purta E; Dawson W; Dunin-Horkawicz S; Bujnicki JM Computational modeling of RNA 3D structures, with the aid of experimental restraints. *RNA Biol* 2014, 11, 522–536. [PubMed: 24785264]
21. Chen S-J RNA folding: Conformational statistics, folding kinetics, and ion electrostatics. *Annu. Rev. Biophys* 2008, 37, 197–214. [PubMed: 18573079]
22. Sun LZ; Zhang D; Chen S-J Theory and modeling of RNA structure and interactions with metal ions and small molecules. *Annu. Rev. Biophys* 2017, 46, 227–246. [PubMed: 28301768]
23. Bernauer J; Huang X; Sim AY; Levitt M Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA* 2011, 17, 1066–1075. [PubMed: 21521828]
24. Wang R; Alexander RW; VanLoock M; Vladimirov S; Bukhtiyarov Y; Harvey SC; Cooperman BS Three-dimensional placement of the conserved 530 loop of 16S rRNA and of its neighboring components in the 30S subunit. *J. Mol. Biol* 1999, 286, 521–540. [PubMed: 9973568]
25. Tan RKZ; Petrov AS; Harvey SC YUP: A molecular simulation program for coarse-grained and multi-scaled models. *J. Chem. Theory Comput* 2006, 2, 529–540. [PubMed: 22844233]
26. Jonikas MA; Radmer RJ; Laederach A; Das R; Pearlman S; Herschlag D; Altman RB Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA* 2009, 15, 189–199. [PubMed: 19144906]
27. Shi YZ; Wang FH; Wu YY; Tan Z-J A coarse-grained model with implicit salt for RNAs: Predicting 3D structure, stability and salt effect. *J. Chem. Phys* 2014, 141, 105102. [PubMed: 25217954]
28. Xia Z; Gardner DP; Gutell RR; Ren P Coarse-grained model for simulation of RNA three-dimensional structures. *J. Phys. Chem. B* 2010, 114, 13497–13506. [PubMed: 20883011]
29. Pasquali S; Derreumaux P HiRE-RNA: A high resolution coarse-grained energy model for RNA. *J. Phys. Chem. B* 2010, 114, 11957–11966. [PubMed: 20795690]
30. Ding F; Sharma S; Chalasani P; Demidov VV; Broude NE; Dokholyan NV Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA* 2008, 14, 1164–1173. [PubMed: 18456842]
31. Sharma S; Ding F; Dokholyan NV iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics* 2008, 24, 1951–1952. [PubMed: 18579566]
32. Kerpedjiev P; Siederdisen CH; Hofacker IV Predicting RNA 3D structure using a coarse-grain helix-centered model. *RNA* 2015, 21, 1110–1121. [PubMed: 25904133]
33. Boniecki MJ; Lach G; Dawson WK; Tomala K; Lukasz P; Soltysinski T; Rother KM; Bujnicki JM SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res* 2016, 44, e63. [PubMed: 26687716]
34. Magnus M; Boniecki MJ; Dawson W; Bujnicki JM SimRNAweb: a web server for RNA 3D structure modeling with optional restraints. *Nucleic Acids Res* 2016, 44, W315–W319. [PubMed: 27095203]
35. Izzo JA; Kim N; Elmetwaly S; Schlick T RAG: an update to the RNA-As-Graphs resource. *BMC Bioinformatics* 2011, 12, 219. [PubMed: 21627789]
36. Kim N; Laing C; Elmetwaly S; Jung S; Curuksu J; Schlick T Graph-based sampling for approximating global helical topologies of RNA. *Proc. Natl. Acad. Sci. U. S. A* 2014, 111, 4079–4084. [PubMed: 24591615]

37. Das R; Baker D Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U. S. A* 2007, 104, 14664–14669. [PubMed: 17726102]
38. Das R; Kudaravalli M; Jonikas M; Laederach A; Fong R; Schwans JP; Baker D; Piccirilli JA; Altman RB; Herschlag D Structural inference of native and partially folded RNA by high-throughput contact mapping. *Proc. Natl. Acad. Sci. U. S. A* 2008, 105, 4144–4149. [PubMed: 18322008]
39. Das R; Karanicolas J; Baker D Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods* 2010, 7, 291–294. [PubMed: 20190761]
40. Cao S; Chen SJ Physics-based de novo prediction of RNA 3D structures. *J. Phys. Chem. B* 2011, 115, 4216–4226. [PubMed: 21413701]
41. Xu X; Zhao P; Chen S-J Vfold: a web server for RNA structure and folding thermodynamics prediction. *PLoS One* 2014, 9, e107504. [PubMed: 25215508]
42. Xu X; Chen S-J A method to predict the 3D structure of an RNA scaffold. *Methods Mol. Biol* 2015, 1316, 1–11. [PubMed: 25967048]
43. Popenda M; Szachniuk M; Antczak M; Purzycka KJ; Lukasiak P; Bartol N; Blazewicz J; Adamiak RW Automated 3D structure composition for large RNAs. *Nucleic Acids Res* 2012, 40, e112. [PubMed: 22539264]
44. Purzycka KJ; Popenda M; Szachniuk M; Antczak M; Lukasiak P; Blazewicz J; Adamiak RW Automated 3D RNA structure prediction using the RNAComposer method for riboswitches. *Methods Enzymol* 2015, 553, 3–34. [PubMed: 25726459]
45. Parisien M; Major F The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 2008, 452, 51–55. [PubMed: 18322526]
46. Bida JP; Maher LJ 3rd. Improved prediction of RNA tertiary structure with insights into native state dynamics. *RNA* 2012, 18, 385–393. [PubMed: 22279150]
47. Yamasaki S; Nakamura S; Fukui K Prospects for tertiary structure prediction of RNA based on secondary structure information. *J. Chem. Inf. Model* 2012, 52, 557–567. [PubMed: 22239168]
48. Zhang J; Dundas J; Lin M; Chen R; Wang W; Liang J Prediction of geometrically feasible three-dimensional structures of pseudoknotted RNA through free energy estimation. *RNA* 2009, 15, 2248–2263. [PubMed: 19864433]
49. Rother M; Rother K; Puton T; Bujnicki JM ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res* 2011, 39, 4007–4022. [PubMed: 21300639]
50. Reinharz V; Major F; Waldispuhl J Towards 3D structure prediction of large RNA molecules: an integer programming framework to insert local 3D motifs in RNA secondary structure. *Bioinformatics* 2012, 28, i207–i214. [PubMed: 22689763]
51. Guntert P; Mumenthaler C; Wüthrich K Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol* 1997, 273, 283–298. [PubMed: 9367762]
52. Popenda M; Blazewicz M; Szachniuk M; Adamiak RW RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures. *Nucleic Acids Res* 2008, 36, D386–D391. [PubMed: 17921499]
53. Popenda M; Szachniuk M; Blazewicz M; Wasik S; Burke EK; Blazewicz J; Adamiak RW RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics* 2010, 11, 231. [PubMed: 20459631]
54. Lukasiak P; Antczak M; Ratajczak T; Szachniuk M; Popenda M; Adamiak RW; Blazewicz J RNAssess—a web server for quality assessment of RNA 3D structures. *Nucleic Acids Res* 2015, 43, W502–W506. [PubMed: 26068469]
55. Cruz JA; Blanchet MF; Boniecki M; Bujnicki JM; Chen S-J; Cao S; Das R; Ding F; Dokholyan NV; Flores SC; et al. RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA* 2012, 18, 610–625. [PubMed: 22361291]
56. Miao Z; Adamiak RW; Blanchet MF; Boniecki M; Bujnicki JM; Chen S-J; Cheng C; Chojnowski G; Chou FC; Cordero P; et al. RNA-Puzzles round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA* 2015, 21, 1066–1084. [PubMed: 25883046]

57. Miao Z; Adamiak RW; Antczak M; Batey RT; Becka AJ; Biesiada M; Boniecki MJ; Bujnicki JM; Chen S-J; Cheng CY; et al. RNA-Puzzles round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA* 2017, 23, 655–672. [PubMed: 28138060]
58. Xu X; Chen S-J VfoldCPX server: Predicting RNA-RNA complex structure and stability. *PLoS One* 2016, 11, e0163454. [PubMed: 27657918]

Type	Symbol	L	Templates	Type	Symbol	L	Templates
Helix2		0	88	Tail5		1	60
		1	437			2	136
		2	865			3	118
		3	869			4	109
		4	746			5	61
Hairpin		3	51	Tail3		1	81
		4	281			2	143
		5	253			3	158
		6	167			4	150
		7	271			5	68

Figure 1:

RNA loop template database. Cylinders represent RNA helices. The red oval shape in each helix denotes the terminal base pair attached to the single strand loops (shown in red). A larger database can be found in Supplementary Table S1.

Figure 2:

Loop template-based RNA 3D structure prediction. (A) Motif extraction: helices and loops are extracted from a 2D structure. (B) The configuration of N helices can be configured by the $N-1$ helix2 loops. Each helix is modeled with A-form helix structure. Helix configurations can be sampled by assembling the loops from the different templates. For a given configuration of the helix system, the remaining (C) helix2 and (D) hairpin, tail5, and tail3 loops are configured based on the corresponding templates, as selected based on the loop score (first) and RMSD (second) criteria. If none of templates satisfies this criteria, we exclude this helix configuration. The assembled whole structures are ranked by their overall scores. The 2D structure in (A) has four helices and eight loops (hairpin: L_1 , and L_6 ; tail3: L_8 ; helix2: L_2 , L_3 , L_4 , L_5 , and L_7). The coordinates of the four helices can be sampled by the L_2 , L_3 , and L_5 helix2 loops. For each helix configuration shown in (B), L_4 , and L_7 template structures are selected to ensure the loop connections between the (fixed) helices.

Figure 3: (A) Comparison of the RMSDs between VfoldLA, Vfold3D,⁴⁰ and the model from Das and Baker (2007).³⁷ (B) and (C) Comparison of the RMSDs between VfoldLA and Vfold3D⁴⁰ and the model from Parisien and Major (2008).⁴⁵ The data are extracted from the published paper.⁴⁰ The sequences and 2D structures listed in Supplementary Table S2 are used for the VfoldLA structure predictions.

Figure 4:

(A) The 2D structure of 4jf2, which contains four helices and eight loops (details shown in Fig. 2). (B) Excluding the native structure from the template database, our model predicts the all-atom structure with RMSD of 12.6 Å. Even though our predicted structure has similar global fold (helix configuration) with the native structure, the orientations of helices H₃ and H₄ differ significantly from the native ones, which may be caused by the improper template of the L₃ loop that connects helix H₂ and H₃ (shown as the inset).

Figure 5:

The 2D and predicted 3D (in dark gray) structures for the *RNA Puzzles* #3 (A), #17 (B), and #15 (C) challenges. We exclude all the templates from the PDB structures 3owz, 3owi, 3oww, 3ow0, 3owb, 3owd, 3owe, 3owm, and 3owj to predict 3D structures for the #3 challenge. The predicted all-atom structure (with RMSD of 9.7 Å) captures the global fold of the native structure (shown in orange), with the major difference in the 3-way junction structure. For the challenges #15 and #17, we participated the *RNA Puzzles* competition^{55,56} by using VfoldLA to predict the structures. As shown in (B) and (C), the heavy atom RMSD to the native structures are 9.4, and 8.5 Å, respectively. The hairpin loop in cyan for the #17 challenge is not included in the crystal structure. We use the magenta circles and arrows to highlight the major differences between the predicted and native structures. The templates used for all loops are shown in Supplementary Figure S3.

Table 1:

Comparison of the RNAComposer, SimRNA, Vfold3D, and VfoldLA predictions. We use the online tools of RNAComposer, SimRNA, and Vfold3D to predict the structures, with the same 2D structures as the input. We use the RNAssess⁵⁴ server to assess the quality of the predicted structures. The score includes the Interactions Network Fidelity (INF) for all the base pairing and stacking interactions. The RMSDs (in Å) are calculated over all the heavy atoms. For VfoldLA predictions, we also list the ranking of the minimum RMSD cluster center within the predicted top clusters in the “RMSD” column. Here “n/a” indicates no predictions. By dissect the motifs into loops, VfoldLA achieves low-resolution predictions for RNAs with no homologous motif templates.

PDB	Size (nt)	Motif	RNAComposer			SimRNA			Vfold3D			VfoldLA		
			RMSD	INF	Clash	RMSD	INF	Clash	RMSD	INF	Clash	RMSD	INF	Clash
2qux	25	stem-loop	1.9	0.88	42.1	4.1	0.57	85.3	4.4	0.84	126.5	3.9 (1/5)	0.83	121.1
1xsh	27	stem-loop	2.8	0.75	23.1	3.2	0.55	169.4	3.5	0.69	142.5	3.1 (1/5)	0.61	136.7
2lun	28	stem-loop	5.8	0.59	40.0	5.9	0.60	150.3	4.4	0.61	138.3	4.6 (1/5)	0.65	170.2
2kpv	34	stem-loop	4.2	0.63	31.9	2.8	0.68	105.6	3.3	0.62	185.6	3.5 (1/5)	0.67	160.8
1so3	47	stem-loop	3.1	0.85	37.6	3.9	0.59	153.2	3.6	0.69	97.5	5.1 (5/5)	0.62	120.5
4c7o	48	stem-loop	8.3	0.81	50.3	7.3	0.62	95.9	7.0	0.72	140.8	6.7 (1/5)	0.66	121.3
2lu0	49	3-way	6.8	0.61	46.8	7.6	0.61	166.1	4.3	0.69	181.7	8.7 (2/5)	0.62	142.1
2mhi	53	3-way	17.1	0.47	54.8	15.8	0.60	155.9	15.8	0.45	204.1	12.6 (2/5)	0.51	163.4
1kuq	57	3-way	3.5	0.63	42.3	8.3	0.56	156.8	3.2	0.65	160.9	5.5 (1/5)	0.49	192.2
1hc8	58	3-way	5.3	0.54	82.6	7.4	0.44	182.8	4.3	0.46	166.0	6.6 (1/5)	0.43	186.2
3knh	75	4-way	5.2	0.55	50.5	12.9	0.49	165.7	6.1	0.48	121.3	7.9 (1/5)	0.53	151.9
3j5y	84	5-way	22.5	0.58	66.8	19.0	0.45	136.8	6.8	0.57	133.9	9.8 (1/5)	0.59	114.6
437d	26	pk	1.3	0.72	33.9	3.7	0.48	140.5	n/a			4.2 (1/5)	0.49	122.4
2ap5	28	pk	5.0	0.47	51.7	4.0	0.52	248.9				6.0 (1/5)	0.32	158.9
2a43	26	pk	0.9	0.75	60.4	3.6	0.47	160.6				4.9 (1/5)	0.30	162.5
1yg4	28	pk	4.8	0.53	55.0	4.1	0.47	171.6				4.7 (1/5)	0.39	221.3
2tpk	36	pk	4.7	0.60	44.1	3.5	0.51	183.6				4.9 (2/3)	0.68	149.2
1e8o	50	tertiary	10.4	0.59	64.7	4.7	0.62	127.4				4.9 (3/5)	0.68	128.6
2miy	59	tertiary	12.9	0.45	69.8	9.3	0.53	137.4				8.9 (2/5)	0.43	208.3
1y26	71	tertiary	11.7	0.68	56.1	10.4	0.46	185.1				5.6 (1/5)	0.55	142.9
1vbx	73	tertiary	1.3	0.74	63.7	7.0	0.47	157.4				6.3 (2/5)	0.54	133.6
4jf2	77	tertiary	6.0	0.71	53.8	11.9	0.54	147.0				12.6 (1/5)	0.63	128.1
Average			6.6	0.64	51.0	7.3	0.54	153.8	5.6	0.62	149.9	6.4	0.56	151.7