Chapter 1

A Method to Predict the 3D Structure of an RNA Scaffold

Xiaojun Xu and Shi-Jie Chen

Abstract

The ever increasing discoveries of noncoding RNA functions draw a strong demand for RNA structure determination from the sequence. In recently years, computational studies for RNA structures, at both the two-dimensional and the three-dimensional levels, led to several highly promising new developments. In this chapter, we describe a recently developed RNA structure prediction method based on the virtual bond-based coarse-grained folding model (Vfold). The main emphasis in the Vfold method is placed on the loop entropy calculations, the treatment of noncanonical (mismatch) interactions and the 3D structure assembly from motif-based template library. As case studies, we use the glycine riboswitch and the G310-U376 domain of MLV RNA to illustrate the Vfold-based prediction of RNA 3D structures from the sequences.

Key words Partition function, Loop entropy, Mismatched stacks, 2D structure motif, Structure assembly

1 Introduction

To perform crucial cellular functions, RNA molecules fold up to form compact three-dimensional (3D) structures [1–5]. The RNA structure determination by experiments alone cannot keep up the pace with the ever increasing number of RNA sequences and new functions. The gap between the number of known RNA 3D structures and the number of biologically significant RNA sequences underscores more than ever the request for accurate computational models for RNA structure prediction.

An RNA structure can be described at the two-dimensional (2D) and three-dimensional (3D) levels. A 2D structure is defined as the sum of all the base-base pairs in the structure, including long-range base pairs in tertiary folds. Computational prediction of RNA 2D structures falls into two categories [6–10]: sequence comparison (alignment) analysis and free energy-based modeling. In general, sequence comparison-based methods can give more reliable predictions than free energy-based methods, but it depends on the availability of homologous sequences and often cannot directly provide information about the alternative structures.

Luc Ponchon (ed.), *RNA Scaffolds: Methods and Protocols*, Methods in Molecular Biology, vol. 1316,
DOI 10.1007/978-1-4939-2730-2_1, © Springer Science+Business Media New York 2015

For the free energy-based modeling, a key problem is to determine the helix stabilities and loop free energies. The free energy parameters for a helix stem can be calculated from the Turner experimental data [11], but the loop free energy requires a model.

The recently developed Vfold model is a statistical mechanics-based RNA folding model. The model relies on a coarse-grained (virtual bond) representation of RNA structures [12–14]. Compared with other free energy-based RNA 2D structure prediction models, such as Mfold [15] and RNAstructure [16], the Vfold model computes loop entropy parameters from explicit conformational sampling. Furthermore, by enumerating all the possible (sequence-dependent) intra-loop mismatches, the Vfold model partially accounts for the sequence-dependence of the loop free energy. Through application to a broad range of experimental and biological problems, the Vfold-based predictions have shown to be able to provide novel insights for RNA mechanisms, such as pseudoknot-involved conformational switch between bistable secondary structures [17], microRNA–gene target interactions [18], and RNA–RNA kissing dimerization in viral replication [19, 20].

Knowing RNA 2D structures alone is often not sufficient to understand RNA function. We also need RNA 3D structure information in order to understand the interactions between RNA and other molecules and RNA functions [21–24]. One way to predict RNA 3D structure is to combine a coarse-grained RNA structure model with the knowledge-based force field and fold the RNA through discrete molecular dynamics (DMD) simulations [25–28]. Due to the limitation of conformational sampling, this method would be most suitable for short RNAs or large RNAs with auxiliary constraints from experimental data. Based on the assumption that 3D structure is more conserved and can be recognized by the alignment of sequences and structure motifs, (3D structure) template-based modeling has become a promising method in RNA 3D structure predictions [29–31]. The template-based methods build RNA 3D structures using known structures ranging from fragments of 1–3 nucleotides to larger structural motifs. One of the common limitations for the template (structure assembly) approaches is the completeness of the fragment library. The lack of reliable structural motifs for many loops and junctions greatly hampers the success of accurate 3D structure prediction.

For a given 2D structure, the Vfold-based 3D structure prediction method searches for the appropriate template for each loop/junction in the structure, and assembles the 3D template structures into a scaffold for further structure refinement. In comparison with other template-based (structure assembly) methods such as FARNA/FARFAR [29] and MC-Sym [31], which sample structures from small fragments of known RNA structures, the Vfold-based method uses motif-based instead of fragment-based templates.

2 Algorithms

2.1 RNA Motif-Based Loop Entropy

Using two virtual bonds per nucleotide to represent the backbone conformation, the Vfold model samples fluctuations of loops/junction conformations in 3D space through conformational enumeration (*see* Fig. 1). By calculating the probability of loop formation, it gives the conformational entropy parameters for the formation of the different types of loops such as hairpin, bulge, internal, pseudoknot loops. The model has the advantage of accounting for chain connectivity, exclude volume and the completeness of conformational ensemble.

1. Enumerate all the possible virtual bond backbone conformations for a given chain length (*see* **Note 1**) and count the total number Ω_{coil} of the conformations.
2. From the conformational ensemble above, identify the loop conformations according to the loop closure condition. For example, for hairpin loops, the two ends of a loop conformation should be fitted to an A-form base pair. Count the total number Ω_{loop} of loop conformations.
3. Calculate the loop entropy $\Delta S_{\text{loop}} = k_B \ln\left(\Omega_{\text{loop}} / \Omega_{\text{coil}}\right)$. Here, k_B is the Boltzmann constant.

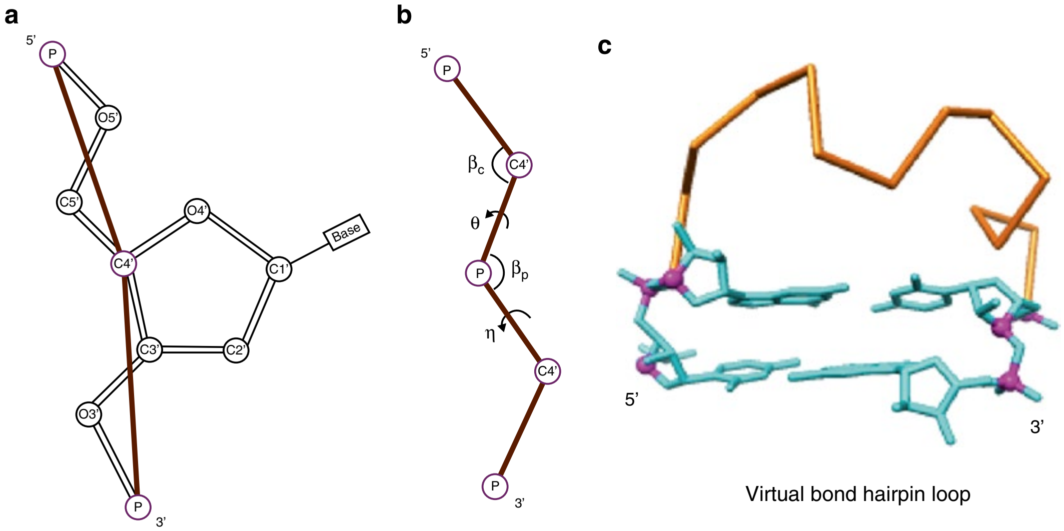

Fig. 1 Vfold computes loop entropies by sampling virtual bond conformations in 3D space. (**a**) Virtual bond representation: two bonds (P-C4′ and C4′-P) per nucleotide. (**b**) The bond angles (β_c, β_p) and the torsional angles (θ, η) for the virtual bonds. Vfold enumerates RNA backbone conformations on a diamond lattice with bond length of 3.9 Å, bond angle of ~109.5° and three equiprobable torsional angles (60°, 180°, 300°). (**c**) A virtual bond backbone conformation of a hairpin loop, with two ends fitted to the base pair structure in an A-form helix

Table 1
RNA motif-based template library

Motif name	Number of templates
Hairpin loops	2,366
Internal/bulge loops	3,260
3-way junctions	820
4-way junctions	506
5-way junctions	222
6-way junctions	49
7-way junctions	61
H-type pseudoknots	56

4. Vfold computations lead to pre-tabulated entropy parameters for hairpin loops [12], internal/bulge loops [12], H-type pseudoknots with/without inter-helix junction [32, 33] and hairpin-hairpin kissing motifs [19].

2.2 RNA Motif-Based Template Library

The (3D structure) template library was built from 2,621 PDB structures (*see* **Note 2**), including RNA-involved complexes. It contains 3D templates for hairpin loops, internal/bulge loops, H-type pseudoknots, and multibranched junctions.

1. For a given RNA 3D structure, extract the A-form helices. From the information of helices and base pairs, the corresponding 2D structure is determined.
2. Identify all the non-helix 2D structure motifs for the given 3D structure.
3. Remove the redundant templates for those with root mean square deviation (RMSD) ≤1.5 Å for the same motif, same size, and identical sequence.
4. Collect all the nonredundant motif structures to construct a template library. Table 1 shows the statistics for the current template library.

3 Methods

To predict RNA 3D structures, Vfold first predicts the 2D structures from the sequence. Using the 2D structures as constraint, the model then predicts the corresponding 3D structures.

3.1 RNA 2D Structure Prediction from the Sequence

The key of the free energy-based RNA 2D structure prediction is the enthalpy and entropy parameters used to evaluate the stability of sampled structures. The enthalpy and entropy for the canonical and mismatched base stacks are calculated from Turner's experimental data. The loop entropies are from the Vfold pre-tabulated parameters.

1. Enumerate all the possible base pair arrangements (2D structures), including H-type pseudoknots with/without interhelix loops and non-pseudoknotted secondary structures, for a given RNA sequence (*see* **Note 3**).
2. For each helix, calculate its free energy ΔG_{helix} (*see* **Note 4**) as a sum of the free energy ΔG_{stack} of each constituent base stack: $\Delta G_{\text{helix}} = \sum_{\text{stack}} \Delta G_{\text{stack}}$ based on the nearest-neighbor model, where ΔG_{stack} is determined from Turner's experimental parameters[11] (*see* **Note 5**).
3. Enumerate all the possible intra-loop mismatches and compute the loop free energy for each given set of intra-loop mismatches (*see* **Note 6**) (*see* Fig. 2 for a hairpin loop for illustration). The loop free energy ΔG_{loop} is calculated (*see* **Note 4**) from the loop partition function Q_{loop}, the Boltzmann sum over all the possible arrangements of intra-loop mismatched base stacks:

$$\Delta G_{\text{loop}} = -k_{\text{B}}\text{T} \ln Q_{\text{loop}}, \quad Q_{\text{loop}} = \sum_{\text{mismatches}} \text{e}^{-\left(\Delta G_{\text{mm}} - T\Delta S_{\text{loop}}\right)/k_{\text{B}}T}$$

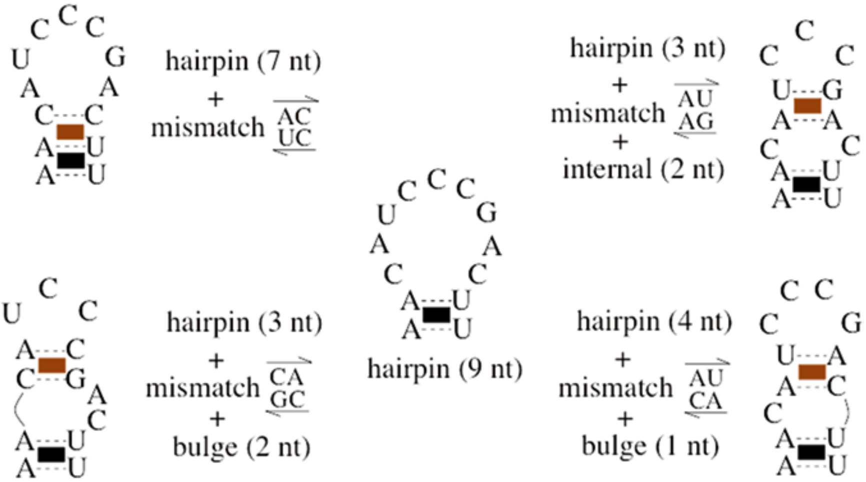

Fig. 2 Ensemble of a 9-nt hairpin loop closed by an A–U base pair, containing five different arrangements of mismatched base stacks within the loop

Here ΔG_{mm} is total free energies of the mismatched base stacks and ΔS_{loop} is the loop entropy for the given intra-loop mismatch constraints.

4. Assign the free energy for each sampled 2D structure: $\Delta G_{\mathrm{s}} = \Delta G_{\mathrm{helix}} + \Delta G_{\mathrm{loop}}$.
5. Calculate the total partition function as the sum over all the possible (2D) structures:

$$Q_{\mathrm{tot}} = \sum_{\mathrm{structures}} \mathrm{e}^{-\Delta G_{\mathrm{s}}/k_{\mathrm{B}}T}$$

6. Following the similar procedure as above for the total partition function, compute the conditional partition function Q_{ij} for all the 2D structures with nucleotides i and j base paired.
7. Calculate the probability of forming the (i, j) base pair: $p_{ij} = Q_{ij} / Q_{\mathrm{tot}}$.
8. From the base pairing probability for all the possible (i, j) pairs, extract the predicted most probable (*see* **Note 7**) as well as alternative structures.

We use the glycine riboswitch (PDB: 3owi) as an example to show how Vfold predicts the 2D structure. Given the 84-nt RNA sequence of the glycine riboswitch (*see* **Note 8**), Vfold calculates the base pairing probabilities *pij* for all the possible base pairs. The predicted most probable 2D structure (*see* **Note 7**) can be predicted from *pij*, as shown in Fig. 3. It should be noted that depending on the sequence, the Vfold model predicts all the stable structures, including the most probable (most stable) structure as well as the alternative (metastable) structures. Therefore, it is recommended to also find out the possible alternative structures from the base pairing probabilities.

3.2 RNA 3D Structure Prediction for a Given 2D Structure

The Vfold model predicts the 3D structure from a 2D structure by assembling motif-specific structural templates. Currently, due to the limited structural template database, Vfold can only predict the 3D structures with hairpin loops, internal/bulge loops, multi-branched junctions and pseudoknots.

1. Identify the structure motifs (such as hairpin loop, internal loop, pseudoknot loop, and three-way junction) from the given 2D structure.
2. Build the virtual bond 3D structure for helices according to the A-form helix template.
3. For each non-helix motif, search for the best templates from the template library. The search criteria are based on the size (first) and sequence (second) matches (*see* **Note 9**).
4. From the (all-atom) templates found in the previous step (*see* **Notes 10** and **11**), build the virtual bond 3D structures of each motifs.

i – j	pij
1 C–G 84	0.89
2 U–A 83	0.95
3 C–G 70	0.01
3 C–G 82	0.95
4 U–A 69	0.01
4 U–A 81	0.95
5 G–A 69	0.04
5 G–C 80	0.93
6 G–C 68	0.05
6 G–A 79	0.90
7 A–U 67	0.07
8 G–C 22	0.03
8 G–C 66	0.07
8 G–C 68	0.09
9 A–U 21	0.03
9 A–U 65	0.07
9 A–U 67	0.09
10 G–A 20	0.02
10 G–C 26	0.01
10 G–C 64	0.07
10 G–C 66	0.09
...	

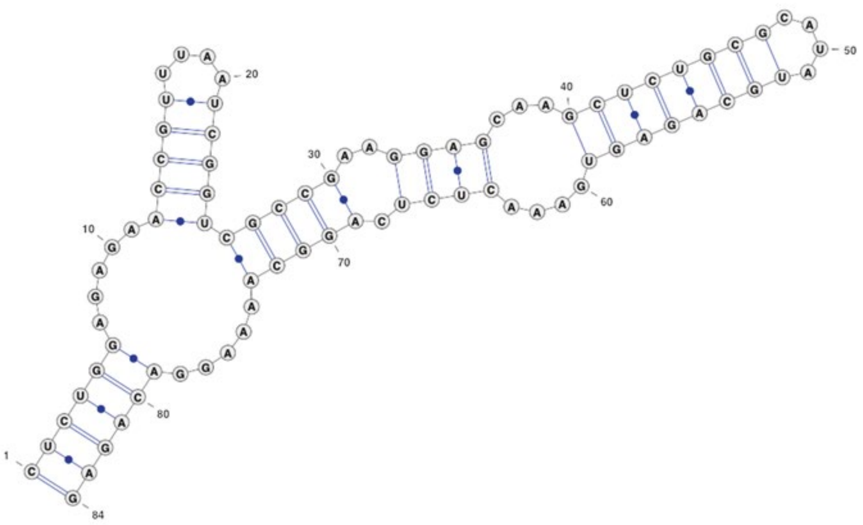

Fig. 3 *Left*: Partial list for the predicted base pairing probabilities for the glycine riboswitch. The first two columns are the position and the nucleotide type of residue *i*; the fourth and fifth columns are the nucleotide type and position for residue *j*; the last column is the probability of forming the *i*, *j* base pair. *Right*: predicted most probable 2D structure derived from the base pairs with probability p_{ij} around 0.90 (e.g., 0.89, 0.95)

5. Assemble the virtual bond 3D structure of the motifs to construct the 3D scaffold of the whole RNA.
6. Add bases to the virtual bond backbone according to the templates for base configurations (*see* **Note 12**).
7. Refine the 3D structure by AMBER energy minimization (*see* **Note 13**).

We use the glycine riboswitch (PDB: 3owi) case for the illustration of the 3D structure prediction. For the purpose of 3D structure prediction, we list only the canonical base pairs in the 2D structure and treat noncanonical base pairs as part of the loop/junction structure (*see* Fig. 3). The predicted 2D structure shows a three-way junction, two internal loops, two hairpin loops and five helices. If the motif structure in the glycine riboswitch is included in the template library, as shown in Fig. 4a, the RMSD between the predicted structure and the experimentally determined structure is 6.3 Å. This RMSD is smaller than the previous prediction of 7.24 Å, shown in Fig. 4b, which excluded the native templates in the template library.

It should be noted that even with the native templates included in the template library, the predicted structure still shows a non-zero RMSD with the PDB structure. The small difference between the A-form helix and the real (slightly distorted) RNA helix (*see* **Note 14**) could results in a notable structural difference in the global fold.

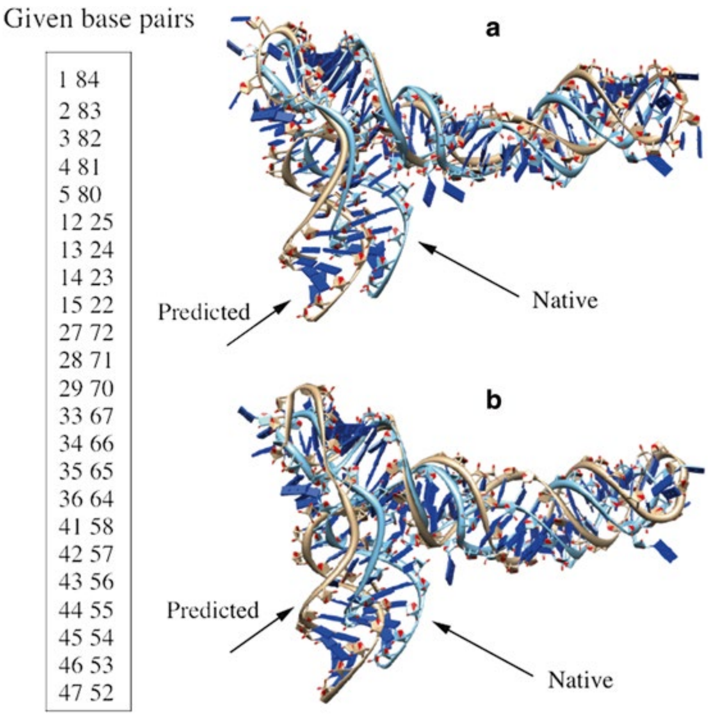

Fig. 4 Given the base pairs shown in the *left*, the predicted 3D structures (**a**) with and (**b**) without including templates from the native structure in the template library, respectively, and the comparisons with the PDB structures (PDB: 3owi) for the glycine riboswitch. The RMSDs are (**a**) 6.3 Å and (**b**) 7.2 Å , respectively

As another example, we predict the 3D structure for the G310-U376 domain of MLV RNA (*see* **Note 15**). Vfold correctly predicts the 2D structure (*see* **Note 16**) (*see* Fig. 5). If the native motif structure (PDB: 1s9s) is not included in the template library, we find a larger RMSD between the predicted and the PDB structure.

Because the template-based 3D structure prediction algorithm relies on the knowledge of the known structures, we can realistically expect continuous improvements in the quality of the structure prediction as more and more structures are solved.

4 Notes

1. A survey of the known structures suggests that the virtual bonds (P-C4′ and C4′-P) have bond length of ~3.9 Å, and have bond angles of (β_c , β_p) in the range of 90–120°.
2. The list of the 2,621 PDB structures used for constructing the template library includes all the PDB entries released before January of 2014. It includes RNA-involved complexes except RNA/DNA hybrids.

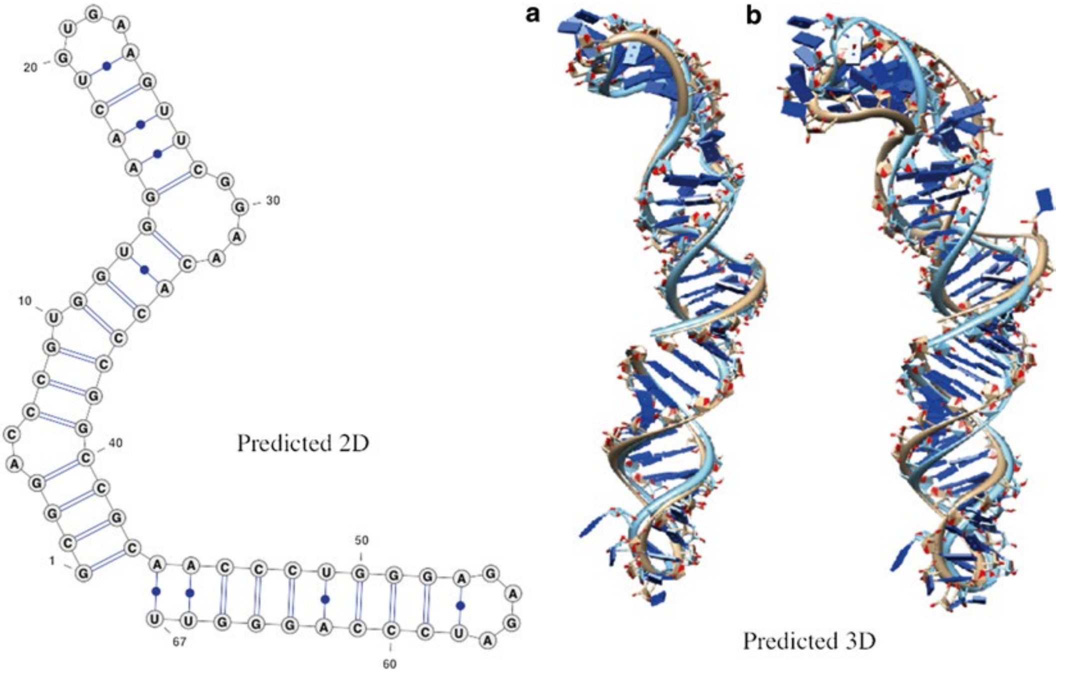

Fig. 5 The prediction of the G310-U376 domain of MLV RNA. The RMSDs between the predictions (**a**) with and (**b**) without templates from the native structure and the experimentally determined all-atom structure (PDB: 1s9s) are 3.1 Å and 7.0 Å, respectively. Vfold correctly predicts its 2D structure, shown in the *left*

3. The computational time scales with the chain length N as $O(N^6)$ and the memory scales as $O(N^2)$.

4. The thermodynamic parameters for the different base stacks are from the experimental data (Turner parameters). We use Vfold derived loop entropies to evaluate the loop free energy.

5. The nearest-neighbor model of RNA structure assumes that the stability of a base pair depends only on its adjacent bases, which could be either base-paired (a stacking contribution) or unpaired (a mismatch contribution).

6. A mismatched base stack is formed by a canonical base pair (A–U, G–C, or G–U) and a noncanonical base pair.

   Consecutive noncanonical base pairs are considered to be unstable and are not accounted for in the intra-loop mismatch rearrangements.

7. The predicted most probable 2D structure is formed by base pairs of the largest base pairing probabilities.

8. The sequence of the glycine riboswitch is: 5′CUCUGGAGA GAACCGUUUAAUCGGUCGCCGAAGGAG-CAAGCU CUGCGCAUAUGCAGAGUGAAACUCUCAGGCAAAA GGACAGAG3′.

9. To find the best template for a loop, Vfold screens the template library according to the loop size (first criteria) and the sequence (second criteria) matches. If necessary, this may involve sequence replacement in order to match the sequences in the template library. Vfold defines the sequence distance $H = \sum_i h^i$ to find the optimal templates. Here, *hi* is the hamming distance between nucleotide *i* in the selected template and the corresponding nucleotide in the target sequence through the following substitution cycles: A → G → C → U, C → U → A → G, G → A → U → C, U → C → G → A.
10. If there is no templates available for a motif, no 3D structures will be predicted.
11. There might have more than one optimal templates available for the same motif. This can lead to more than one predicted 3D structures.
12. For the nucleotides in helices, the base atoms are added following the A-form helix.
13. The all-atom energy (such as AMBER) minimization causes only small change in the RMSD of the structure.
14. Helices contain canonical base pairs only: A–U, G–C, and G–U base pairs. The RMSD between a helix of known RNA structures and the standard A-form helix is <1.2 Å.
15. The sequence of the G310-U376 domain of MLV RNA is: 5 ′ G C G G A C C C G U G G U G G A A C U G U G AAGU-UCGGAACACCCGGCCGCAACCCUGGGA GAGAUCCCAGGGUU3′.
16. The base pairs used to predict the 3D structures of MLV RNA: (1 43), (2 42), (3 41), (4 40), (7 39), (8 38), (9 37), (11 36), (12 35), (13 34), (14 33), (15 28), (16 27), (17 26), (18 25), (19 24), (44 67), (45 66), (46 65), (47 64), (48 63), (49 62), (50 61), (51 60), (52 59), (53 58).

Acknowledgment

This research was supported by NIH grant GM063732 and NSF grant MCB0920411.

References

1. Doudna JA, Cech TR (2002) The chemical repertoire of natural ribozymes. Nature 418: 222–228
2. Bachellerie JP, Cavaille J, Huttenhofer A (2002) The expanding snoRNA world. Biochimie 84:774–790
3. Gong C, Maquat LE (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3 UTRs via Alu elements. Nature 470:284–288
4. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136:215–233

5. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. Nat Genet 39:1278–1284
6. Gardner PP, Giegerich R (2004) A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics 5:140
7. Mathews DH, Moss WN, Turner DH (2010) Folding and finding RNA secondary structure. Cold Spring Harb Perspect Biol 2:a003665
8. Washietl S (2010) Sequence and structure analysis of noncoding RNAs. Methods Mol Biol 609:285–306
9. Machado-Lima A, del Portillo HA, Durham AM (2008) Computational methods in noncoding RNA research. J Math Biol 56:15–49
10. Mathews DH, Turner DH (2006) Prediction of RNA secondary structure by free energy minimization. Curr Opin Struct Biol 16: 270–278
11. Turner DH, Mathews DH (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res 38:D280–D282
12. Cao S, Chen S-J (2005) Predicting RNA folding thermodynamics with a reduced chain representation model. RNA 11:1884–1897
13. Chen S-J (2008) RNA folding: conformational statistics, folding kinetics, and ion electrostatics. Annu Rev Biophys 37:197–214
14. Xu X, Zhao P, Chen S-J (2014) Vfold: a web server for RNA structure and folding thermodynamics prediction. PLoS ONE 9(9): e107504
15. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406–3415
16. Bellaousov S, Reuter JS, Seetin MG, Methews DH (2013) RNAstructure: web servers for RNA secondary structure prediction and analysis. Nucleic Acids Res 41:W471–W474
17. Xu X, Chen S-J (2012) Kinetic mechanism of conformational switch between bistable RNA hairpins. J Am Chem Soc 134:12499–12507
18. Cao S, Chen S-J (2012) Predicting kissing interactions in microRNA-target complex and assessment of microRNA activity. Nucleic Acids Res 40:4681–4690
19. Cao S, Chen S-J (2011) Structure and stability of RNA/RNA kissing complex: with application of HIV dimerization initiation signal. RNA 17:2130–2143
20. Cao S, Xu X, Chen S-J (2014) Predicting structure and stability for RNA complexes with intermolecular loop- loop base pairing. RNA 20: 835–845
21. Shapiro BA, Yingling YG, Kasprzak W, Bindewald E (2007) Bridging the gap in RNA structure prediction. Curr Opin Struct Biol 17:157–165
22. Rother K, Rother M, Boniecki M, Puton T, Bujnicki JM (2011) RNA and protein 3D structure modeling: similarities and differences. J Mol Model 17:2325–2336
23. Laing C, Schlick T (2011) Computational approaches to RNA structure prediction, analysis, and design. Curr Opin Struct Biol 21: 306–318
24. Sim AY, Minary P, Levitt M (2012) Modeling nucleic acids. Curr Opin Struct Biol 22:1–6
25. Tan RK, Petrov AS, Harvey SC (2006) YUP: a molecular simulation program for coarse-grained and multi-scaled models. J Chem Theory Comput 2:529–540
26. Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D, Altman RB (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA 15:189–199
27. Sharma S, Ding F, Dokholyan NV (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. Bioinformatics 24: 1951–1952
28. Xia Z, Bell DR, Shi Y, Ren P (2013) RNA 3D structure prediction by using a coarse-grained model and experimental data. J Phys Chem B 117:3135–3144
29. Das R, Karanicolas J, Baker D (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. Nat Methods 7:291–294
30. Cao S, Chen S-J (2011) Physics-based de novo prediction of RNA 3D structures. J Phys Chem B 115:4216–4226
31. Parisien M, Major F (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature 452:51–55
32. Cao S, Chen S-J (2006) Predicting RNA psuedoknot folding thermodynamics. Nucleic Acids Res 34:2634–2652
33. Cao S, Chen S-J (2009) Predicting structures and stabilities for H-type pseudoknots with inter-helix loop. RNA 15:696–706