Coarse-Grained Prediction of RNA Loop Structures

Liang Liu, Shi-Jie Chen*
Department of Physics and Department of Biochemistry, University of Missouri, Columbia, Missouri, United States of America

Abstract

One of the key issues in the theoretical prediction of RNA folding is the prediction of loop structure from the sequence. RNA loop free energies are dependent on the loop sequence content. However, most current models account only for the loop length-dependence. The previously developed “Vfold” model (a coarse-grained RNA folding model) provides an effective method to generate the complete ensemble of coarse-grained RNA loop and junction conformations. However, due to the lack of sequence-dependent scoring parameters, the method is unable to identify the native and near-native structures from the sequence. In this study, using a previously developed iterative method for extracting the knowledge-based potential parameters from the known structures, we derive a set of dinucleotide-based statistical potentials for RNA loops and junctions. A unique advantage of the approach is its ability to go beyond the the (known) native structures by accounting for the full free energy landscape, including all the nonnative folds. The benchmark tests indicate that for given loop/junction sequences, the statistical potentials enable successful predictions for the coarse-grained 3D structures from the complete conformational ensemble generated by the Vfold model. The predicted coarse-grained structures can provide useful initial folds for further detailed structural refinement.

Introduction

The ability to predict RNA 3D structure is critical for understanding RNA functions. Recent developments in de novo prediction of RNA 3D structures have led to highly promising results [1–18] (for review, see [19–24]). In particular, several de novo structure prediction methods have been developed based on the knowledge-based energy function. For example, Dima and co-workers [25] extracted the base-pair stacking parameters from RNA native structures and the extracted parameters agree with the experimental data [26]. Wu et al. [27] explored the correlation between RNA secondary structural motifs and their thermodynamic stability to derive energy parameters for base-pair stackings, and free segments such as hairpin loops, internal loops and bulge loops. Bernauer and co-workers [15] further extracted a set of distance-dependent energy parameters between any two bases, irrespective of the locations of the bases (in a helix or a loop). Das and Baker [5,9] obtained energy function – made available in the Rosetta software package – based on the base orientations and interactions. Parisien and Major [7] predicted 3D structures with a pipeline of two computer programs: MC-Fold and MC-Sym, developed based on the nucleotide cyclic motifs (NCMs). All of these methods have provided valuable insights into the correlation between loop sequence and their stability. Especially, these methods are particularly useful for selecting the most probable conformation from an ensemble of near-native structures.

A key issue in the predictions of RNA stability is how to compute the loop free energy. For hairpins and RNA secondary structures in general, the nearest neighbor model, which assumes that the total free energy is an additive sum of the free energy of each elements (base-pair stacking, loop), and other models [6,26,28–32] has enabled successful predictions for RNA structures and stabilities. In most of the existing models, loop stability is often assumed to depend on loop size, the identity of the closing base pair, the interaction of the first mismatch with the closing base pair, and an additional stabilization term for loops with GA or UU first mismatches [33–35]. Further detailed sequence-dependence of the loop stability has been ignored. Experimental results suggest that loop stability may be sensitive to the sequence context inside the loop. For example, for unusually stable RNA hairpin loops, Dale et al. [36] performed optical melting studies for a series of hairpins. The study led to a set of different stability parameters for different loop sequences, such as GNRA and UUGC (where N is any nucleotides and R is a purine) tetraloops, hexaloops with UU first mismatches, and hairpin loop of iron responsive element, GAGUGC, all of which are significantly more stable than other hairpin loops of the same length.

The prediction of sequence-dependent loop free energy requires a model that goes beyond simple fitting with the experimentally measured empirical parameters. The formation of the intraloop base pairs and stacks would cause significant restriction of the loop conformational space and the loop entropy. It is practically impossible to exhaustively measure the loop free energy for all the different possible intraloop contacts for the different sequences and loop lengths. Therefore, evaluation of loop free energy with an ensemble of possible intraloop base pairing/stacking interactions cannot be achieved by experiment alone. We also need a computational model. The main purpose of the present study is to develop a statistical potential model that enables predictions of loop and junction three-dimensional structures from the sequence.

In general, there are two classes of physics-based models for RNA structure prediction: molecular dynamics (For review, see
database and the Leontis database (http://rna.bgsu.edu/). Based on the coarse-grained statistical potentials (scoring functions) from this study, we develop a method to extract a set of virtual bond-pairs, which could cause inaccuracy in the prediction of loop stability generated in the model as a function of the (4×4) types of the dinucleotides base i and j and the backbone pseudo-torsion angles (θ, η) of the dinucleotide conformation. We use the RNA09 database [51], the Leontis database, the Capriotti’s database [52], and the PDB database [53,54], respectively, to test the extracted potential functions. We note that the dinucleotide in this work refers to the two continuous nucleotides within the same loop. The goal is for a given RNA loop/junction sequence, to identify the lowest-RMSD structures from the Vfold-generated complete conformational ensemble.

### Results

**Pseudo-torsion angles (θ, η)**

The θ and η values of dinucleotides in the 152 RNA loops/junctions are plotted as a 2D scatter plot (Fig. 6), where each point represents the θ-η coordinates for a dinucleotide. In contrast to the analysis in the previous studies [55,56], here we find a distinct category of dinucleotides conformation around (θ = 150°, η = 225°). This region was once considered as the helical region, because the dinucleotides with θ-η coordinates located in this group are most likely found within the RNA helix. However, as we only count the dinucleotides in RNA loops and junctions, the plot shows that the loop/junction residues can also have the tendency to have the helix-like conformation. This observation provides a rational in the next step for building the 3D all-atom structures by adding the helical residues back to the coarse-grained backbone model.

### Table 3. The numbers and correct predictions with the “true” potentials SPfold.

| Types                  | RNA09 Total # | Correct # | Capriotti’s Total # | Correct # | TEST-III Total # | Correct # |
|------------------------|--------------|-----------|---------------------|-----------|------------------|-----------|
| Hairpin                | 72           | 37        | 14                  | 166       | 395              | 150       |
| Internal/bulge         | 25           | 14        | 14                  | 166       | 165              | 165       |
| Pseudoknot             | 14           | 5         | 14                  | 148       | 147              | 147       |
| Multibranched          | 35           | 15        | 15                  | 329       | 297              | 297       |
| Junction               | 6            | 1         | 0                   | 68        | 66               | 66        |
| **TOTAL**              | **152**      | **72**    | **72**              | **1070**  | **1070**         | **1070**  |

doi:10.1371/journal.pone.0048460.0003
Quasi-chemical approximation-based potentials

The 152 RNA loops/junctions within our training dataset consist of a total number of \( N_{\text{total}} = 572 \) nucleotides, with the numbers of each type of the nucleotide \((\rho_{A4,PC}, \rho_{G,PU}) = (217,105,121,129)\) and the corresponding mole fraction of each nucleotide \((\rho_{A4,PC}, \rho_{G,PU}) = (0.38, 0.18, 0.21, 0.23)\).

From the dataset of the coarse-grained “correct” structures (Table 1), for each pair of the nucleotides \( i = 1,2,3,4 \) and \( j = 1,2,3,4 \) and pseudo-torsion angles \((\theta, \eta)\), we calculate the numbers \( \rho_{ij}^\text{obs}(\theta, \eta) \) of the dinucleotides conformations and the total observed number \( N(\theta, \eta) = \sum_{i,j} \rho_{ij}^\text{obs}(\theta, \eta) \). According to Eq. 8, we then calculate the expected number \( \rho_{ij}^\text{exp}(\theta, \eta) \) of each type of dinucleotides \((i,j)\) with pseudo-torsion angles \((\theta, \eta)\). If no pseudo-torsion angles \((\theta, \eta)\) is observed for dinucleotide \((i,j)\), we assign an unfavorable potential \( w_{ij}(\theta, \eta) = +2.0 k_B T \) kcal/mol. From Eq. 7, we compute the potentials as a \( 4 \times 4 \times 3 \times 3 \) tensor.

Iterative approach-derived potentials

The convergence speed of the iteration depends on the selection of convergence criteria. For instance, if the convergence threshold parameter in Eq. 13 is set to \( 10^{-5} \), the iterative process would converge after around 3000 iterations. In contrast, if the convergence threshold parameter is set to \( 10^{-2} \), the iterative process would converge after 340 iterations. On an Intel(R) Xeon(R) CPU 5150 @ 2.66 GHz on Dell E6440 cluster system, the 3000-cycle iteration process took about 30 minutes and a 300-step iteration took less than 6 minutes.

Comparison between the two sets of the derived potential parameters shows that the “true” potentials \( \mathcal{SP}_{\text{fold}} \) are less uniform than the “extracted” potentials (Fig. 2), suggesting that the “true” potentials \( \mathcal{SP}_{\text{fold}} \) are more sensitive to the sequences and conformations of the dinucleotide and thus have the ability to discriminate the “correct” conformation from an ensemble of conformations for a given RNA loop/junction sequence.

Tests on training datasets

We test the statistical potentials for the accuracy in loop/junction structure prediction for a large number of sequences. For a given sequence, we generate the full ensemble of the virtual bond loop/junction conformations using our Vfold model. Each conformation is then scored by the total statistical potential, which is evaluated as the sum of the statistical potentials \( \mathcal{SP}_{\text{fold}} \) or \( \mathcal{SP}_{\text{native}} \) for the dinucleotide pairs in the conformation. The conformation with the lowest value of the total statistical potential is identified as the predicted native structure. If the predicted native structure is the same as the coarse-grained “correct” structure (i.e., the virtual bond structure that is closest to the PDB

### Table 4. The numbers of sequences of the successfully predicted loop/junction structures for (I) all 8452 RNA loops/junctions in TEST-I (II) the 7459 RNA loops and junctions in TEST-II and (III) the 1119 RNA loops and junctions in TEST-III.

| TEST-I | \#\( \mathcal{SP}_{\text{fold}} \)(A) | \#\( \mathcal{SP}_{\text{fold}} \)(B) | \#\( \mathcal{SP}_{\text{native}} \)(A) | \#\( \mathcal{SP}_{\text{native}} \)(B) |
|--------|-----------------|-----------------|-----------------|-----------------|
| Top-1  | 7364            | 7369            | 6992            | 7102            |
| Top-2  | 7686            | 7686            | 7411            | 7500            |
| Top-3  | 7796            | 7862            | 7715            | 7665            |
| Top-4  | 7956            | 7969            | 7798            | 7844            |
| Top-5  | 8006            | 8007            | 7857            | 7878            |
| TOTAL-#| 8452            | 8452            | 8452            | 8452            |

| TEST-II | \#\( \mathcal{SP}_{\text{fold}} \)(A) | \#\( \mathcal{SP}_{\text{fold}} \)(B) | \#\( \mathcal{SP}_{\text{native}} \)(A) | \#\( \mathcal{SP}_{\text{native}} \)(B) |
|---------|-----------------|-----------------|-----------------|-----------------|
| Top-1   | 6555            | 6567            | 6222            | 6332            |
| Top-2   | 6819            | 6861            | 6563            | 6557            |
| Top-3   | 6909            | 6961            | 6826            | 6791            |
| Top-4   | 7043            | 7055            | 6895            | 6947            |
| Top-5   | 7089            | 7083            | 6949            | 6976            |
| TOTAL-# | 7459            | 7459            | 7459            | 7459            |

| TEST-III | \#\( \mathcal{SP}_{\text{fold}} \)(A) | \#\( \mathcal{SP}_{\text{fold}} \)(B) | \#\( \mathcal{SP}_{\text{native}} \)(A) | \#\( \mathcal{SP}_{\text{native}} \)(B) |
|---------|-----------------|-----------------|-----------------|-----------------|
| Top-1   | 1070            | 1077            | 1001            | 1016            |
| Top-2   | 1083            | 1084            | 1022            | 1053            |
| Top-3   | 1101            | 1087            | 1060            | 1055            |
| Top-4   | 1102            | 1092            | 1061            | 1058            |
| Top-5   | 1102            | 1098            | 1065            | 1058            |
| TOTAL-# | 1119            | 1119            | 1119            | 1119            |

For each dataset, the numbers in columns "A" are calculated from \( \mathcal{SP}_{\text{fold}} \) \("true" potential parameters) and \( \mathcal{SP}_{\text{native}} \) \("extracted" potential parameters), obtained from the RNA09 database, and the numbers in columns "B" are calculated from \( \mathcal{SP}_{\text{fold}} \) \("SPfold"") and \( \mathcal{SP}_{\text{native}} \) \("SPnative"), obtained from the Leontis’ database respectively. The “Top-#” in the first column means that the “correct” structure is in the top # lowest-potential conformations.

doi:10.1371/journal.pone.0048460.t004
structure, evaluated by RMSD), the prediction is successful for the sequence; otherwise, the prediction fails.

We first apply the two sets of statistical potentials to the RNA loops and junctions in the training dataset. Such a test for structure prediction is nontrivial because the SP\textsubscript{native} and SP\textsubscript{fold} are derived based on the frequency \( r_{ij}(h,g) \) (see Eqs. 7 and 10) instead of the structure. As described above, 152 loops and junctions are constructed from the 262 RNA structures in the RNA09 dataset, including 72 RNA hairpin loops, 25 internal/bulge loops, 14 pseudoknot loops, 35 multibranched loops and 6 junctions (the free segments other than the above 4 types) (Table 3A). Here we note that the two loops in an internal loops are counted separately in our calculation, as our model does not specify specific types of loops or junctions. This rule is also applied to the pseudoknot loops and multibranched loops.

We found that SP\textsubscript{native} succeeded in finding 130 coarse-grained “correct” (lowest potential) structures out of 152 loops and junctions. In contrast, SP\textsubscript{fold} can give successful predictions for all the 152 the coarse-grained “correct” conformations (Table 3A). SP\textsubscript{fold} is more reliable than SP\textsubscript{native} in structure prediction with success rate 100% vs 85.5%.

Test on the Capriotti’s dataset

To rigorously test the reliability of the two sets of potential parameters, we need to perform the test on loops and junctions outside the training dataset. We will perform tests for several such test sets. We first choose a dataset collected by Capriotti, et al. [52], consisting of 85 structures with length \( \geq 20 \) nucleotides and solved at resolution better than 3.5 A\textsubscript{˚}. The 3DNA software determined a total of 72 loops/junctions with lengths ranging from 3 nt to 8 nt, excluding the 5’/3’-terminal dangling regions. The 72 loops/junctions can be classified into 37 hairpin loops, 14 internal/bulge loops, 5 pseudoknot loops, 1 junction and 15 multibranched loops (Table 3B).

Our results indicate that SP\textsubscript{native} and SP\textsubscript{fold} potentials can successfully find out 62 and 70 coarse-grained “correct” conformation, respectively, out of the 72 test cases (Table 3B). The result indicates that the “true” potential parameters SP\textsubscript{fold} are more reliable than the directly extracted potentials SP\textsubscript{native} (success rate 97% vs 86%). One of the two failed predictions for SP\textsubscript{native} is for a pseudoknot loop, which has special loop structure due to the tertiary interactions (base triplets) with the helices. Another failed prediction is a junction located in a large RNA molecule and the junction structure is determined by not only its sequence content but also the surrounding structural environment.

Test on the PDB dataset

The January 2012 version of PDB database contains 2227 structures that contain at least one strand of RNA sequence. These structures range from hairpin-loop structures to RNA-protein complexes or RNA-DNA hybrids. We found 8452 loops and junctions in the 2227 PDB structures (\textit{TEST-I}). All the loops and junctions (excluding the 3’/5’-terminal dangling regions) have lengths from 3 nt to 8 nt. Within these 2227 RNA molecules, 1609 have structures determined by X-ray crystallography, which contain 7459 RNA loops and junctions (\textit{TEST-II}), and 934 of these 1609 RNAs have high-resolution structures (\( \leq 3.0 \) A\textsubscript{˚}), which contain 1119 RNA loops and junctions (\textit{TEST-III}).

Our test results show that the numbers of the correct predictions with the statistical potentials SP\textsubscript{fold} and SP\textsubscript{native} are 7364 (success rate = 87%) vs. 6992 (success rate = 83%) for \textit{TEST-I}, 6553 (success rate = 88%) vs. 6222 (success rate = 83%) for \textit{TEST-II}, and 1070 (success rate = 95%) vs. 1001 (success rate = 89%) for \textit{TEST-III}, respectively. Fig. 10 shows illustrations for two of the results (a hairpin loop from PDB structure 1IVS and an internal loop from PDB structure 1JJ2).

If we include the top five lowest-potential conformations, the numbers of successfully determined loops and junctions are increased to 8006 with SP\textsubscript{fold} v.s. 7857 with SP\textsubscript{native} for \textit{TEST-I}, 7089 v.s. 6949 for \textit{TEST-II} and 1102 v.s. 1065 for \textit{TEST-III} (Table 2). Fig. 11 shows the minimal-RMSD structure, the 9-th potential structure computed with SP\textsubscript{native}, and the predicted
structure with the lowest-potential, for multibranched loop from PDB structure 3CCM.

Moreover, we test the predictions of the first 20 lowest-potential conformations for the loops and junctions in all the three databases TEST-I, TEST-II and TEST-III as well as the influence of the convergence threshold parameter on the accuracy of the structure prediction. Fig. 12 shows the numbers of successfully determined loops and junctions in the three databases. The comparisons between the predicted results with SP\text{fold} and SP\text{native} supports our conclusions in the previously two benchmark tests that SP\text{fold} is more reliable than SP\text{native}. The loops/junctions that we failed to predict mostly involve tertiary interactions with helices or other cofactors such as protein and DNA.

The predicted results using the statistical potentials SP\text{fold} and SP\text{native}, extracted from the Leontis dataset also support the above conclusions (Fig. 9 and Table 4).

Fig. 13 shows the sensitivity of the convergence threshold parameter \( \Delta \mu_{ij}(\theta, \eta) \) of 0.001 v.s. 0.01 to the RNA structural predictions. The comparisons of the numbers of the successfully determined loops and junctions in the three databases show that the convergence threshold parameters \( \Delta \mu_{ij}(\theta, \eta) \) do not have strong influence on the accuracy of the structure predictions.
Furthermore, we categorize the sequences in the dataset TEST-III and the predicted results according to the types of RNA loops and junctions. The 1119 RNA loops and junctions can be grouped into 408 hairpin loops, 166 internal/bulge loops, 148 pseudoknot loops, 329 multibranched loops and 68 junctions. The correct predictions with \(SP_{\text{fold}}\) for each type of RNA loops and junctions are 395, 165, 147, 297, and 66, respectively (Table 3C). For junctions, multibranched loops, pseudoknots, internal/bulge loops and five of the hairpin loops, the failed predictions are due to the interactions beyond the dinucleotide context, such as the loop-helix tertiary interactions and RNA-protein interactions. The possible reason for the other six failed predictions for hairpin loops is that these hairpin loops are not closed by canonical base pairs. These loops are closed by non-canonical pairs, such as AA or AC, which may lead to different hairpin loop structures.

For the TEST-I (the 8452 RNA structures in PDB) and TEST-II (the 7459 x-ray structures in TEST-I) databases, we randomly selected 3229 RNA loops/junctions from the database TEST-I and categorize them according to the types of RNA loops and junctions. The 3229 RNA loops and junctions contain 1406 hairpin loops, 253 internal/bulge loops, 184 pseudoknot loops, 1265 multibranched loops and 119 junctions. The correct (top-1) predictions with \(SP_{\text{fold}}\) for each type of RNA loop junctions are 1274 (success rate = 90.6%), 245 (success rate = 96.1%), 178 (success rate = 96.7%), 1162 (success rate = 91.9%) and 110 (success rate = 96.4%).

Figure 3. Comparison between the “extracted” statistical potentials \(SP_{\text{native}}\) from the RNA09 dataset and the Leontis dataset, respectively. The figures from (A) to (I) stand for the torsion angles \((\theta,\eta)=\): (A) \((g^+,g^-)\), (B) \((g^+,\theta)\), (C) \((g^+,g^+)\), (D) \((t,\theta)\), (E) \((t,t)\), (F) \((t,g^+)\), (G) \((g^-,g^-)\), (H) \((g^-,\theta)\) and (I) \((g^-,g^-)\), respectively. In each figure, the red bars represent the statistical potentials \(SP_{\text{native}}\) extracted from the Leontis dataset, the green bars represent the ones from the RNA09 dataset, and the x-axis stands for the dinucleotides with different nucleotides \((i,j)\) AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG and UU from 1 to 16.

doi:10.1371/journal.pone.0048460.g003
The March 17, 2012 version of the Leontis dataset contains 642 RNA sequences with structures of resolution better than 4.0 Å. The 3DNA software identified a total of 435 RNA loops and junctions with lengths ranging from 3 nt to 8 nt, excluding the 5'/-3'-terminal dangling regions. Our calculations show that the difference between the diamond lattice-represented structures and the PDB structures varies for the different loop lengths and sequence contents with RMSD from 0.74 Å to 3.93 Å (Fig. 8), and the mean and standard deviation of RMSD values are 1.37 Å and 0.29 Å, respectively. We use such coarse-grained structures to calculate the observed dinucleotide frequencies, to extract SP\text{native} and to search for the “true” potential functions SP\text{fold}.

Figs. 3 and 4 show the comparison between the potentials SP\text{native}/SP\text{fold} extracted from the dataset RNA09 and from the Leontis dataset. We apply the “true” potential SP\text{fold} extracted from the Leontis dataset to predict the loop/junction structures in three testing datasets: TEST-I, TEST-II and TEST-III, constructed from the January 2012 version of PDB database. Our test results (Table 4) show a success rates of 87% (7369 out of 8452) for TEST-I, 88% (6567 out of 7459) for TEST-II, and 96% (1077 out of 1119) for TEST-III, respectively.
Moreover, the predictions of the top 20 lowest-potential conformations for the loops and junctions in all the three test datasets, TEST-I, TEST-II and TEST-III, are shown in Fig. 9, with comparisons with the predictions based on the SPfold from the RNA09 dataset. The comparisons for the three test datasets (Table 4 and Fig. 9) show that the two “true” statistical potentials SPfold extracted from the RNA09 dataset and the Leontis dataset lead to similar success rates in structure prediction and the predictions are not sensitive to the choice of the specific training set.

**Discussion**

Motivated by the biological significance to predict sequence-dependent loop and junction structures, we have developed a knowledge-based scoring functions/potentials to predict the structures of RNA loops and junctions. We use a coarse-grained conformational model (the virtual bond model) to sample RNA loop/junction conformations. From the known RNA structures, we extract a set of sequence-dependent dinucleotide-based statistical potentials using two methods. In the first method, the statistical potentials (SPnative) are derived from the distributions of the dinucleotide conformations in the known (native) structures. In the second method, the statistical potentials (SPfold) are derived based on the folding stability of the native structures (against all the other nonnative folds).

For a given sequence, the extracted statistical potentials enable ranking of the different conformations with the top ranked (the lowest-potential) structure as the predicted native structure. Extensive tests indicate that the statistical potentials can successfully predict the native structure for a large class of loop and junction sequences and that SPfold consistently outperforms SPnative in structure prediction. Our test results also indicate that our results are not sensitive to the choice of the specific training dataset (Fig. 9 and Table 4).

The present approach has several advantages. First, the extraction of the statistical potential SPfold is based on the sampling of the complete conformational ensemble, including the native and all the nonnative folds. Second, because we consider all the nonnative folds in the derivation of the statistical potential, our statistical potentials can be used to predict folding from the sequence. The build-up strategy for RNA loop structure in this study is a de novo approach. Compared with other 3D loop structure prediction models, such as ModeRNA [22] and RLooM [50], our loop/junction structure prediction method is based on the complete ensemble of the (coarse-grained) conformations and does not rely on the information of template structures or homologous RNA structures. The only input information for the prediction is the sequence. Therefore, the model can predict the low-resolution structure from the sequence if no known homol-
ogous conformations can be found in the PDB. With the predicted low-resolution scaffold, one may further predict the all-atom structures using the all-atom potential methods that are derived based on near-native structures [5,15]. The present coarse-grained model offers a useful complement to the other all-atom based models.

We have performed extensive benchmark tests using the different databases. However, a direct comparison between our model other methods is not very straightforward. This is because our model is a coarse-grained model while other models mainly focus on the all-atom structures. Furthermore, our model aims to fold a low-resolution structure from the sequence without using any input information such as homologous templates, while other models mainly focus on the prediction of the native structures from near-native folds. Future development of our method, which may give all-atom structures from the low-resolution folds, would make direct comparison between our model and other models possible.

Applications of the present statistical potentials to the Vfold structure prediction model [14] may provide an effective strategy for better structure prediction. Despite the success, there are several limitations of our model. First, the current study is based on the coarse-grained Vfold model, which only provides a low-resolution approximation for the all-atom structure of RNA loops and junctions. Ultimately an all atom-based model is required to treat the detailed tertiary interactions. Future development of the model should address the issue to build all-atom RNA structures based on the low-resolution (Vfold-generated) representations. Second, the statistical potentials are derived for dinucleotide conformations. In realistic loop and junction structures, the long-range sequence effects within the loop, such as intra-loop
interactions, can influence the loop structure. The present model cannot explicitly account for such long-range intraloop interactions, which occur frequently in loops and between the different loops. Following the same procedure outlined above, one may develop a trinucleotide or higher order many-body statistical potentials to address this issue. Third, in the current form of the model, we use the same set of statistical potentials for the different types of loops/junctions. More refined potentials according to RNA loop/junction types may lead to further improvement of the accuracy of model.

Materials and Methods

Vfold model

We use the Vfold model to generate the full conformation ensemble of a given RNA loop or junction sequence. The Vfold model is a virtual bond-based RNA folding model [42,45,48], which is developed based on the two observations: the C=O torsion in the nucleotide backbone of RNA tend to adopt the trans (t) rotational isometric state (Fig. 5A) and the $P-O_2-C_3-C_4$ bonds and the $C_4-C_1-O_2-P$ bonds in the nucleotide backbone are approximately planar [57]. Therefore, the nucleotide backbone conformations can be reduced into two effective virtual bonds $P-C_4$ and $C_4-P$ [58–60]. The length of each backbone virtual bond is about 3.9 Å. The virtual bonds show rotamer-like configurations gauche$^+$ (g$^+$), trans (t) and gauche$^-$ (g$^-$) (Fig. 5C) [55,56,61]. Such virtual bond conformations can be well represented by the bonds in a diamond lattice. Therefore, we can use self-avoiding walks on a diamond lattice to enumerate the conformations of RNA sequences. This approach promises proper treatment of the excluded volume effect between the different atoms and the complete sampling of the conformational ensemble.

The loop structures are enumerated on a diamond lattice through exhaustive self-avoiding walks. If the first atom ($P$) is fixed at position $\mathbf{X}_0$, then the coordinates ($\mathbf{X}_N$) of $N$-th atom ($P$ or $C_4$) can be calculated with

$$
\mathbf{X}_N = \mathbf{X}_0 + \sum_{i=1}^{N} l_i \mathbf{b}_i
$$

(1)

where, $l_i = 3.9$ Å is the bond length of virtual bond $\mathbf{b}_i$ and $\mathbf{b}_i$ is the unit vector of virtual bond $\mathbf{b}_i$. Also, the relation between the ($i-1$)-th virtual bond ($\mathbf{b}_{i-1}$) and the $i$-th virtual bond ($\mathbf{b}_i$) is:

$$
\mathbf{b}_i = T(\beta_i, \theta_i) \mathbf{b}_{i-1}
$$

(2)

where, $\beta_i$ and $\theta_i$ are the related bond angles $\beta_i = 120^\circ$ and pseudo-torsion angle $\theta_i = g^+ (60^\circ)$, $g^- (300^\circ)$ or $t (180^\circ)$ (Fig. 5C). The matrix $T$ is defined as

$$
T(\beta_i, \theta_i) = \begin{bmatrix}
1 & 0 & 0 \\
0 & \cos \beta_i & -\sin \beta_i \\
0 & \sin \beta_i & \cos \beta_i
\end{bmatrix} \times
\begin{bmatrix}
\cos \theta_i & -\sin \theta_i & 0 \\
\sin \theta_i & \cos \theta_i & 0 \\
0 & 0 & 1
\end{bmatrix}
$$

(3)
Therefore, if the unit vector of the first virtual bond \(P_{C_3/C_4}\) is \(\vec{b}_1\), we can obtain the unit vector of the \(i\)-th virtual bond \(P_{C_3/C_4}\) with

\[
\vec{b}_i = \prod_{j=2}^{i} T(\beta_j, \theta_j) \vec{b}_1
\]  

(4)

By considering excluded volume effect, i.e., two atoms cannot occupy the same site on a diamond lattice, we can generate the atomic coordinates using equations 1, 2, 3 and 4 and the conformation ensemble of RNA loops; see Ref. [42] for the detailed calculations. Here we give an example for illustration. In Fig. 5, following the 5'→3' direction, if the first atom \(P\) is fixed at \(X_0 = \text{col}(0, 0, 3.9)\) (Å), the second atom \(C_3/C_4\) is fixed at \(\vec{x}_1 = \text{col}(0, 0, 0)\) (Å), then, \(\vec{b}_1 = \text{col}(0, 0, -1)\). According to Eq. 1, the coordinate of the third atom \(P\) can be computed:

\[
\vec{X}_2 = \vec{X}_0 + l_1 \vec{b}_1 + l_2 T(\pi - \beta_P, \eta) \vec{b}_1
\]

where \(l_1 = l_2 = 3.9\) Å are the bond lengths, \(\beta_P = 120°\) is the bond angle and \(\eta\) is the torsion angle (we select 180° for illustration). According to Eq. 3, the matrix is computed as:

\[
T(60°, 180°) = \begin{pmatrix}
-1 & 0 & 0 \\
0 & -1/2 & -\sqrt{3}/2 \\
0 & -\sqrt{3}/2 & 1/2
\end{pmatrix}
\]

Therefore, the coordinate of the third atom \(P\) is \(\text{col}(0, \sqrt{3}/2, -1/2)\) (3.9 Å). Following the procedure and considering the excluded volume between any two atoms, we can compute the coordinates for other atoms and generate the conformations for a given RNA loop.

**Statistical potential**

Assuming an equilibrium Boltzmann distribution for the structures in the PDB [53,54] and NDB [62] database, one can extract the interaction potentials:

\[
u(r) = -k_B T \ln \frac{\rho(r)}{\rho^*}\]

(5)

**Figure 11.** Comparison between the virtual bond-based PDB structure and top-ranked predicted structures with statistical potentials \(\text{SP}_{\text{fold}}\). The loop is a multibranched loop from PDB structure 3CCM. In the figure, the structure shown in brown \(P\) and yellow \(C_3\) stands for the PDB structure, the one shown in brown \(P\) and pink \(C_4\) represents the minimal-RMSD structure (RMSD = 2.42 Å), and the one shown in brown \(P\) and lightgray \(C_4\) represents the predicted structure with lowest potentials (RMSD = 2.87 Å), computed with statistical potentials \(\text{SP}_{\text{fold}}\). The minimal-RMSD structure has the 9-th lowest potential. The atomic structures are illustrated with Pymol software (http://www.pymol.org/).

**Figure 12.** The success rate of coarse-grained correct loop/junction structure predictions for (A) all the 8452 RNA loops/junctions in \(\text{TEST-I}\), (B) the 7459 RNA loops and junctions in \(\text{TEST-II}\), and (C) the 1119 RNA loops and junctions in \(\text{TEST-III}\). The “Top-#” in x-axis means that the “correct” structure is in the top # lowest-potential conformations. In each figure, • and □ represent the success rate with \(\text{SP}_{\text{fold}}\) and \(\text{SP}_{\text{native}}\), respectively.

**Table 1.** Comparison of the success rate of coarse-grained correct loop/junction structure predictions for all the 8452 RNA loops/junctions in \(\text{TEST-I}\), the 7459 RNA loops and junctions in \(\text{TEST-II}\), and the 1119 RNA loops and junctions in \(\text{TEST-III}\). The “Top-#” in x-axis means that the “correct” structure is in the top # lowest-potential conformations. In each figure, • and □ represent the success rate with \(\text{SP}_{\text{fold}}\) and \(\text{SP}_{\text{native}}\), respectively.

**Figure 11.** Comparison between the virtual bond-based PDB structure and top-ranked predicted structures with statistical potentials \(\text{SP}_{\text{fold}}\). The loop is a multibranched loop from PDB structure 3CCM. In the figure, the structure shown in brown \(P\) and yellow \(C_3\) stands for the PDB structure, the one shown in brown \(P\) and pink \(C_4\) represents the minimal-RMSD structure (RMSD = 2.42 Å), and the one shown in brown \(P\) and lightgray \(C_4\) represents the predicted structure with lowest potentials (RMSD = 2.87 Å), computed with statistical potentials \(\text{SP}_{\text{fold}}\). The minimal-RMSD structure has the 9-th lowest potential. The atomic structures are illustrated with Pymol software (http://www.pymol.org/).

**Figure 12.** The success rate of coarse-grained correct loop/junction structure predictions for (A) all the 8452 RNA loops/junctions in \(\text{TEST-I}\), (B) the 7459 RNA loops and junctions in \(\text{TEST-II}\), and (C) the 1119 RNA loops and junctions in \(\text{TEST-III}\). The “Top-#” in x-axis means that the “correct” structure is in the top # lowest-potential conformations. In each figure, • and □ represent the success rate with \(\text{SP}_{\text{fold}}\) and \(\text{SP}_{\text{native}}\), respectively.
Here, $k_B$ is the Boltzmann constant, $T$ is the absolute temperature, $\rho(r)$ is the observed density, and $\rho'(r)$ is the density in a “reference” state where no interactions occur.

The above approach can give continuous distance-dependent pairwise potentials [15,63–66] or discrete base pairs/stacks potentials [27,67]. In this work, we extract the potentials for the different dinucleotide conformations described by the pseudo-torsion angles $\theta$ ($P - C_4 - P - C_4$) and $\eta$ ($C_4 - P - C_4 - P$) and the different dinucleotide sequences (Fig. 5). The potential $u_{ij}(\theta, \eta)$ is extracted based on the following formula:

$$
u_{ij}(\theta, \eta) = -k_B T \ln \frac{\rho_{ij}^{\text{obs}}(\theta, \eta)}{\rho_{ij}^{\text{exp}}(\theta, \eta)}$$ (6)

where $i$ and $j$ denote the types of the nucleotides (bases) of the two consecutive nucleotides within the same loop and $(\theta, \eta)$ are the pseudo-torsion angles of the dinucleotide.

Reference state

The proper choice of the reference state is a key issue in deriving the statistical potentials from the known structures. As pointed out by Thomas and Dill [68,69], an accurate ideal reference state is not achievable. Different approximations such as the quasi-chemical approximation [64] have been used to model the reference state. Furthermore, to circumvent the reference state problem, an iterative method for successive refinement of the statistical potentials was developed and was shown to give reliable scoring functions for protein folding [68,69] and protein-ligand interactions [63]. Here, for comparison, we employ the above two approaches to extract two different sets of RNA knowledge-based potentials and apply the extracted potentials to RNA loop and junction structure prediction.

Quasi-chemical approximation-based approach. In the quasi-chemical approximation [64] we use an “expected state” as the “reference state”. Such an approximation leads to the Boltzmann relation (Eq. 6):

$$
u_{ij}(\theta, \eta) = -k_B T \ln \frac{\rho_{ij}^{\text{obs}}(\theta, \eta)}{\rho_{ij}^{\text{exp}}(\theta, \eta)}$$ (7)

where $\rho_{ij}^{\text{obs}}(\theta, \eta)$ and $\rho_{ij}^{\text{exp}}(\theta, \eta)$ are the observed number and the expected number, respectively, of the dinucleotides $(i, j)$ with pseudo-torsion angles $(\theta, \eta)$ in the entire training dataset. $\rho_{ij}^{\text{exp}}(\theta, \eta)$ is computed from the following formulas:

$$ho_{ij}^{\text{exp}}(\theta, \eta) = N(\theta, \eta) x_i x_j$$ (8)

where $N(\theta, \eta)$ is the total observed number of dinucleotides with pseudo-torsion angles $\theta$ and $\eta$ (calculated with a composition-independent scale, meaning each nucleotide is independent to other nucleotides in the state (Eq. 8)) [70], and $x_i$ and $x_j$ are the mole fraction of nucleotide $i$ and $j$ of the sequences in the entire training dataset, respectively [70].

The quasi-chemical approximation-based statistical potentials are the “extracted” energy-like parameters derived from the observed density in the database of native structures [68]. We denote such extracted statistical potentials as $\text{SP}_{\text{native}}$. The $\text{SP}_{\text{native}}$ parameters are not the “true” potentials extracted from the criteria of identifying the “correct”/native structure from an ensemble of the nonnative conformations. We denote the potentials that can account for the nonnative conformations as the “true” statistical potentials $\text{SP}_{\text{fold}}$.

Convergent iterative approach. To extract the “true” energy $\text{SP}_{\text{fold}}$, Thomas and Dill developed an iterative approach based on the folding stability of the native structure [68,69]. The method has the advantage of accounting for the effect of the distribution of the whole conformational ensemble, including both the native and the nonnative states, for a given sequence. The overall strategy of the iterative approach is to train a set of potential parameters iteratively until the collection of the native and the nonnative states, for a given sequence. The overall strategy of the iterative approach is to train a set of potential parameters iteratively until the collection of the native and the nonnative states, for a given sequence.

In our calculation, the iterative process starts with a set of initial values for the potentials, $u_{ij}^{(0)}(\theta, \eta)$. The superscript $(n)$ denotes the $n$-th iterative step. We use the quasi-chemical approximation-based statistical potentials $\text{SP}_{\text{native}}$ as the initial input.

At each step, we compute the total potential/score for each conformation of loop/junction by summing over the potential of all the dinucleotides contained in the loop/junction:

![Figure 13. The sensitivity of the structure prediction to the change of the convergence threshold parameter. Δuij(\theta, \eta) = 0.001 v.s. 0.01. The tests are based on the predictions of the RNA loops and junctions in all the three databases: (A) TEST-I, (B) TEST-II and (C) TEST-III. In each figure, ● and □ represent the success rate of the SPfold-based structure prediction with Δuij(\theta, \eta) = 0.001 and 0.01, respectively. doi:10.1371/journal.pone.0048460.g013](image-url)
\[ E^{(n)} = \sum_{\text{dimers}} u^{(n)}_{ij}(\theta, \eta) \]

Then the predicted frequencies (numbers) of the different dinucleotide conformations are computed from the Boltzmann average over all the conformations:

\[ \rho^{\text{pred}}_{ij}(\theta, \eta) = \frac{1}{S} \sum_{s} \frac{f^{(s)}_{ij}(\theta, \eta) e^{-E^{(s)}(ij)/k_BT}}{\sum_{l} e^{-E^{(s)}(ij)/k_BT}} \]

where \( E^{(s)}(i, j) \) is the potential/score for the \( l \)-th conformation of the \( s \)-th RNA loop/junction sequence, computed with the potentials \( u^{(n)}_{ij}(\theta, \eta) \) at the \( n \)-th step (Eq. 9), \( n_{ij}(\theta, \eta) \) is the number of dimucleotide \((i, j)\) with pseudo-torsion angles \((\theta, \eta)\) in the \( l \)-th conformation of the \( s \)-th RNA loop/junction, \( L \) is the number of conformations of the \( s \)-th RNA loop/junction in the training dataset, and \( S \) is the number of RNA loop/junction sequences in the training dataset.

In each step, we also calculate the difference between the observed and predicted dinucleotide frequencies:

\[ \Delta u^{(n)}_{ij}(\theta, \eta) = \frac{\rho^{\text{pred}}_{ij}(\theta, \eta) - \rho^{\text{obs}}_{ij}(\theta, \eta)}{\rho^{\text{pred}}_{ij}(\theta, \eta)} \]

where \( \rho^{\text{pred}}_{ij}(\theta, \eta) \) is calculated by Eq. 10 and \( \rho^{\text{obs}}_{ij}(\theta, \eta) \) is the dinucleotide frequency observed from the “correct” (native) structures in the training dataset.

In general, the initially guessed potential parameters are not equal to the “true” potentials and thus there are differences between the observed dinucleotide frequencies and the predicted dinucleotide frequencies (\( \Delta u^{(n)}_{ij}(\theta, \eta) \)). For each \( n \)-step, we refine the potentials by accounting for the differences between the “observed” and the “predicted” frequencies of the dinucleotide conformations (see the parameter \( \Delta u^{(n)}_{ij}(\theta, \eta) \) in Eq. 11):

\[ u^{(n)}_{ij}(\theta, \eta) = u^{(n-1)}_{ij}(\theta, \eta) + \Delta u^{(n-1)}_{ij}(\theta, \eta) \]

We repeat the above iterative process until \( |\Delta u^{(n)}_{ij}(\theta, \eta)| \) approaches 0 (smaller than a threshold number, e.g. \( 10^{-3} \)). The final set of potentials \( u^{(n)}_{ij}(\theta, \eta) \) is the “true” potentials \( \text{SP}_{\text{old}} \).

Flowchart

The iterative method is summarized as follows (Fig. 1):

1. Prepare the training dataset of the native structures. Download RNA sequences with the experimentally determined structures of both the canonical (Watson-Crick and wobbles) base pairs and non-canonical base pairs (normally classified as tertiary structures) from Kinemage borinophysics.duke.edu/databases/rnadb.phb, an updated version of previous 2005 database (RNA05) [51] to extract the \( \text{SP}_{\text{native}} \) potential and use the database as the training dataset to search for the “true” potential \( \text{SP}_{\text{old}} \). The RNA09 dataset consists of 262 RNA sequences with the experimentally determined structures of resolution \( \leq 3.0 \) Å.

2. RNA loops and junctions are constructed by all the unpaired nucleotides in RNA structure. We detect the loops and junctions by removing the nucleotides involved in the base pairs. Furthermore, we remove all the loop structures with modified bases or non-RNA atoms and molecules. Advances in the knowledge of the structures of both the canonical (Watson-Crick and wobbles) base pairs and non-canonical base pairs (normally classified as tertiary interactions) [71,72] enable the development of a number of automated tools to detect the base pairs in RNA structures. In this work, we use the 3DN software package (http://3dna.rutgers.edu/home) [73] to search for all the possible base pairs from the RNA structures.

3. In addition, the time consumption to enumerate the loop conformations on a diamond lattice increases exponentially with the length of the loop [45]. Therefore, we only choose loops/junctions with length \( \leq 8 \) nt in the dataset due to the long computer time for the exhaustive enumeration of the loop conformations for longer loops [45]. In this study, we use the algorithm reported in Ref. [74] to search for the best fits on diamond lattice for the RNA loops/junctions.

Larger loops often involve tertiary interactions with other subunits of RNA structures, which are not accounted for in this study. As a result, we found 152 loops/junctions with the length ranging from 3 nt to 8 nt in the RNA09 dataset. These loop and junction structures are used as the training dataset in our iterative approach.

To further test the influence of the use of the different training dataset on the statistical potentials and the predictive power of the statistical potentials, we also use another non-redundant high-resolution RNA structure dataset, collected by Leontis’ lab (http://rna.bgsu.edu/nrlist/).

Conformational ensembles of RNA loops/junctions

We use the Vifold model to generate the full conformational ensemble of a given RNA loop or junction sequence. In the Vifold-
generated loop/junction conformational ensemble, because each of the two pseudo-torsion angles (θ and η) of a dinucleotide occupies three torsional states in a diamond lattice, there exist 4 × 4 × 3 × 3 parameters for the potentials for the four types of each nucleotide (A, C, G and U) and the three possible rotamer-like states for each torsional angle (and hence 3 × 3 types of the pseudo-torsion angle pairs θ and η).

For each PDB structure of the loop/junction, we find the minimum-RMSD fit of the virtual bond conformation on the diamond lattice. We call such a coarse-grained (correct) structure the “coarse-grained correct structure”; see Step 1 in Fig. 1. Our calculations show that the differences between the diamond-lattice-represented structures and the PDB structures varies for the different loop lengths and sequence contents with RMSD in the range from 0.67 Å to 2.07 Å (Fig. 7), the mean and standard deviation of RMSD values are 1.35 Å and 0.30 Å, respectively.

We will use the coarse-grained correct structure to calculate the observed dinucleotide frequencies, from which $S_{\text{native}}$ and $S_{\text{fold}}$ are extracted.

### Acknowledgments

Most of the numerical calculations involved in this research were performed on the HPC resources at the University of Missouri Bioinformatics Consortium (UMBC).

### Author Contributions

Conceived and designed the experiments: LL SJC. Performed the experiments: LL. Analyzed the data: LL SJC. Wrote the paper: LL SJC.

### References

1. Ding F, Sharma S, Chalasani P, Demidov V, Broude N, et al. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. RNA 14: 1164–1173.

2. Zhang J, Dundas J, Lin M, Chen R, Wang W, et al. (2009) Prediction of geometrically feasible three-dimensional structures of pseudoknotted RNA through free energy estimation. RNA 15: 2240–2263.

3. Gherghe C, Leonard C, Ding F, Dohohany N, Weeks K (2009) Native-like RNA tertiary structures using a sequence-encoded cleavage agent and refinement by discrete molecular dynamics. J Am Chem Soc 131: 2541–2546.

4. Hajduk C, Ding F, Dohohany N, Weeks K (2010) On the significance of an RNA tertiary structure prediction. RNA 16: 1340–1349.

5. Das R, Baker D (2007) Automated de novo prediction of native-like RNA tertiary structures using 3D discrete molecular dynamics. J Mol Biol 371: 769–783.

6. Das R, Karamicloj J, Baker D (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. Nat Methods 7: 291–294.

7. Yang S, Parisien M, Major F, Roux B (2010) RNA structure determination using SAXS data. J Phys Chem B 114: 10039–10040.

8. Pasquale S, DeRremaux P (2010) HRE:RNA: a high resolution coarse-grained energy model. J Phys Chem B 114: 11957–11966.

9. Xia Z, Gardner D, Gutell R, Ren P (2010) Coarse-grained model for simulation of RNA three-dimensional structures. J Phys Chem B 114: 13987–13996.

10. Flores S, Altman R (2010) Turning limited experimental information into 3D models of RNA. RNA 16: 1755–1766.

11. Cao S, Chen SJ (2011) Physics-based de novo prediction of RNA 3D structures. J Phys Chem B 115: 4216–4226.

12. Bernard J, Huang X, Sun A, Levi M (2011) Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. RNA 17: 1066–1075.

13. Strpăk, Mere傲 P, Khudawang W, Das R (2011) An enumerative stepwise algorithm enables atomic-accuracy RNA loop modeling. Proc Natl Acad Sci U S A 108: 20575–20578.

14. Parisien M, Major F (2012) Determining RNA three-dimensional structures using low resolution data. J Struct Biol: in press.

15. Bida J, Maher III L (2012) Improved prediction of RNA tertiary structure with geometrical and energetic criteria. RNA 18: 385–393.

16. Capotosti E, Marti-Renom M (2008) Computational RNA structure prediction. Curr Opin Struct Biol 18: 157–165.

17. Capotosti E, Marti-Renom M (2008) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci U S A 101: 7287.

18. Mathews D, Turner D (2006) Prediction of RNA secondary structure by free energy minimization. Curr Opin Struct Biol 16: 270–278.

19. Serra M, Barnes T, Betschart K, Gutierrez M, Sprouse K, et al. (1997) Improved parameter sets for the prediction of RNA hairpin stability. Biochemistry 36: 4844–4853.

20. Giese M, Betschart K, Dale T, Riley C, Rowan C, et al. (1998) Stability of RNA hairpins closed by wobble base pairs. Biochemistry 37: 1094–1100.

21. Mathews D, Dinney M, Childs J, Schroeder S, Zuker M, et al. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci U S A 101: 7287.

22. Mathews D, Turner D (2006) Prediction of RNA secondary structure by free energy minimization. Curr Opin Struct Biol 16: 270–278.

23. Cross J, Blanchet M, Boniecki M, Bujnicki J, Chen S, et al. (2012) RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. RNA 18: 1358–1363.

24. Mathews D, Dinney M, Childs J, Schroeder S, Zuker M, et al. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci U S A 101: 7287.

25. Dale T, Smith R, Serra M (2000) A test of the model to predict unusually stable RNA hairpin loop stability. RNA 6: 608–613.

26. Schlick T, Collepardo-Guevara R, Halvorsen L, Jung S, Xiao X (2011) Biomolecular mod e1ing and simulation: a field coming of age. Q Rev Biophys 44: 191.

27. Hyon C, Thirumalai D (2007) Mechanical unfolding of RNA: from hairpins to structures with internal multiloops. Biophys J 92: 731–743.

28. Villa A, Stock G (2006) What NMR relaxation can tell us about the internal motion of an RNA hairpin: a molecular dynamics simulation study. J Chem Theory Comput 2: 1228–1236.

29. Swin E, Rhee Y, Nakatsui B, Pande V (2003) Insights into nucleic acid conformational dynamics from massively parallel stochastic simulations. Biophys J 85: 790–803.

30. Feiltinger K, Pyle A (2010) Semiautomated model building for RNA crystallography using a directed rotameric approach. Proc Natl Acad Sci U S A 107: 8177.

31. Cao S, Chen SJ (2005) Predicting RNA folding thermodynamics with a reduced chain representation model. RNA 11: 1884–1897.

32. Cao S, Chen SJ (2006) Predicting RNA pseudoknot folding thermodynamics. Nucleic Acids Res 34: 2634.

33. Cao S, Chen SJ (2009) Predicting structures and stabilities for type pseudoknots with interhelical loops. RNA 15: 696–706.

34. Li L, Chen SJ (2010) Computing the conformational entropy for RNA folds. J Chem Phys 132: 235104.

35. Cao S, Giedroc D, Chen SJ (2010) Predicting loop-helix tertiary structural contacts in RNA pseudoknots. RNA 16: 339–352.

36. Cao S, Chen SJ (2012) Predicting kissing interactions in microRNA-target complex and assessment of microRNA activity. Nucleic Acids Res 40: 4681–4690.

37. Chen SJ (2008) RNA folding: conformational statistics, folding kinetics, and ion electro statics. Annu Rev Biophys 37: 197.

38. Vecenie C, Morrow C, Zyra A, Serra M (2006) Sequence dependence of the stability of RNA hairpin molecules with six nucleotide loops. Biochemistry 45: 1460–1467.
50. Schudoma C, May P, Nikiforova V, Walther D (2010) Sequence–structure relationships in RNA loops: establishing the basis for loop homology modeling. Nucleic Acids Res 38: 970–980.

51. Murray L, Arensdall W, Richardson D, Richardson J (2003) RNA backbone is rotamerics. Proc Natl Acad Sci U S A 100: 13904.

52. Capriotti E, Norambuena T, Marti-Renom M, Melo F (2011) All-atom knowledge-based potential for RNA structure prediction and assessment. Bioinformatics 27: 1606.

53. Bernstein F, Koetzle T, Williams G, Meyer E, Brice M, et al. (1977) The protein data bank: A computer-based archival file for macromolecular structures. J Mol Biol 112: 535–542.

54. Berman H, Olson W, Beveridge D, Westbrook J, Gelbin A, et al. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys J 63: 751.

55. Duarte C, Pyle A (1998) Stepping through an RNA structure: a novel approach to conformational analysis1. J Mol Biol 284: 1465–1478.

56. Wadley L, Keating K, Duarte C, Pyle A (2007) Evaluating and learning from RNA pseudotorsional space: quantitative validation of a reduced representation for RNA structure. J Mol Biol 372: 942–957.

57. Bloomfield V, Cothrs D, Tinoco I (2000) Nucleic acids: structures, properties, and functions. Univ Science Books.

58. Olson W, Flory P (1972) Spatial configurations of polynucleotide chains. I. steric interactions in polyribo nucleic acids. Biopolymers 11: 1–23.

59. Olson WK (1980) Configuration statistical of polynucleotide chains: A single virtual bond model. Macromolecules 13: 721–728.

60. Flory P, et al. (1969) Statistical mechanics of chain molecules. Biopolymers 8: 699–700.

61. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, et al. (2000) The protein data bank. Nucleic Acids Res 28: 235–242.

62. Huang S, Zou X (2006) An iterative knowledge-based scoring function to predict protein–ligand interactions: I. derivation of interaction potentials. J Comput Chem 27: 1866–1875.

63. Lu H, Skolnick J (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins: Struct, Funct, Bioinf 44: 223–232.

64. Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11: 2714–2726.

65. Zhao H, Yang Y, Zhou Y (2011) Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. Nucleic Acids Res 39: 3017–3025.

66. Sharma S, Ding F, Dukholyan N (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. Bioinformatics 24: 1951–1952.

67. Thomas P, Dill K (1996) Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 257: 457–469.

68. Thomas P, Dill K (1996) An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci U S A 93: 11628–11633.

69. Skolnick J, Kolinski A, Ortiz A (2000) Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins: Struct, Funct, Bioinf 38: 3–16.

70. Leontis N, Westhof E (2001) Geometric nomenclature and classification of RNA base pairs. RNA 7: 512.

71. Leontis N, Westhof E (2002) Motif prediction in ribosomal RNAs: lessons and prospects for automated motif prediction in homologous RNA molecules. Biochimie 84: 961–973.

72. Lu X, Olson W (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res 31: 5108.

73. Ferro D, Hermans J (1977) A different best rigid-body molecular fit routine. Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography 33: 345–347.