OPEN ACCESS Freely available online

PLOS ONE

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.

Citation: Liu L, Chen S-J (2012) Coarse-Grained Prediction of RNA Loop Structures. PLoS ONE 7(11): e48460. doi:10.1371/journal.pone.0048460

Editor: Ying Xu, University of Georgia, United States of America

Received May 19, 2012; **Accepted** September 26, 2012; **Published** November 8, 2012

Copyright: © 2012 Liu, Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This research was supported by National Institutes of Health (NIH) grant GM063732 and National Science Foundation (NSF) grants MCB0920067 and MCB0920411. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: chenshi@missouri.edu

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 UUCG (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

Table 1. The number of dinucleotides with torsion angles (θ, η).

		η		
θ		g^+	t	g^-
	g^+	43	27	21
	t	46	99	21
	g^-	19	101	43

doi:10.1371/journal.pone.0048460.t001

[37]) and Monte Carlo simulation methods and polymer statistical mechanics methods. Molecular dynamics simulations have provided much insights into the atomic details of intraloop interactions and their contributions to the loop stabilities [1,38–40]. The polymer statistical mechanical models often employ low-resolution (coarse-grained) conformational models in order to capture the complete conformational ensemble. Along this line there have been different ways to construct the low-resolution RNA structures. For example, with a knowledge-based potential, Jonikas et al. [8] developed a structure filter model (*NAST*) where nucleotides are represented by the C_3^* atoms. In another coarse-grained model where nucleotide are represented by the P and C_3^* atoms, Keating and Pyle [41] developed a semi-automated approach to build RNA structures with a directed rotameric search strategy. Furthermore, in an attempt to develop a high-resolution RNA model (HiRE-RNA), Pasquali and Derreumaux [11] used six to seven beads for each nucleotide (one bead for the phosphate P, four beads for the sugar O_5^*, C_5^*, C_4^*, C_1, respectively, and one bead for a pyrimidine base and two beads for a purine base).

Our RNA folding model is based on a virtual bond-based RNA conformational model (called "Vfold" model; Fig. 5) [42]. The Vfold model uses two virtual bonds ($C_4^*-P-C_4^*$) for each nucleotide and samples RNA conformations through self-avoiding walks in a diamond lattice. It provides an effective tool to sample RNA conformations and to evaluate the conformational entropy. The model has shown a high promise in predicting the 2D and 3D structures and the folding stabilities from the sequence [14,42–48]. However, the Vfold model does not account for the sequence-dependent conformational propensity of the loop, namely, the model assumes that for any given loop, all the loop conformations generated in the model have the same energy. Such a simplification could cause inaccuracy in the prediction of loop stability and structure [36,49,50]. Physically, the sequence-dependence of loop stability arises from the local interactions, which affect the loop flexibility and hence conformational propensity [50], as well as the nonlocal interactions between the different nucleotides. In this study, we develop a method to extract a set of virtual bond-based (coarse grained) statistical potentials (scoring functions) from a set of non-redundant RNA structures such as the RNA09 database [51] and the Leontis database (http://rna.bgsu.edu/nrlist/). Specifically, we aim to derive a set of dinucleotide statistical potentials $u_{ij}(\theta, \eta)$ 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.

Table 2. The numbers of sequences with successfully predicted loop/junction structures for (I) all the 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*.

	I		II		III	
	$\#_{SP_{fold}}$	$\#_{SP_{native}}$	$\#_{SP_{fold}}$	$\#_{SP_{native}}$	$\#_{SP_{fold}}$	$\#_{SP_{native}}$
Top-1	7364	6992	6555	6222	1070	1001
Top-2	7686	7411	6819	6563	1083	1022
Top-3	7796	7715	6909	6826	1101	1060
Top-4	7956	7798	7043	6895	1102	1061
Top-5	8006	7857	7089	6949	1102	1065
TOTAL-#	8452		7459		1119	

For each dataset, the numbers in columns are calculated with the SP_{fold} ("$\#_{SP_{fold}}$", "true" potential parameters) and SP_{native} ("$\#_{SP_{native}}$", "extracted" potential parameters), respectively. The potentials are obtained from the RNA09 database. The "Top-#" in the first column means that the "correct" structure is in the top-# lowest-potential conformations.

doi:10.1371/journal.pone.0048460.t002

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 ($\theta = 150^\circ$, $\eta = 225^\circ$). 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 types and numbers of loops and junctions in (*A*) the RNA09 dataset, (*B*) the Capriotti's dataset and (*C*) the *TEST-III* dataset, and the number of the correct predictions with the "true" potentials SP_{fold}.

	RNA09		Capriotti's		*TEST-III*	
Types	**Total #**	**Correct #**	**Total #**	**Correct #**	**Total #**	**Correct #**
Hairpin	72	72	37	37	408	395
Internal/bulge	25	25	14	14	166	165
Pseudoknot	14	14	5	4	148	147
Multibranched	35	35	15	15	329	297
Junction	6	6	1	0	68	66
TOTAL	152	152	72	70	1119	1070

doi:10.1371/journal.pone.0048460.t003

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				
	$\#_{SP_{fold}}$**(A)**	$\#_{SP_{fold}}$**(B)**	$\#_{SP_{native}}$**(A)**	$\#_{SP_{native}}$**(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			
TEST-II				
	$\#_{SP_{fold}}$**(A)**	$\#_{SP_{fold}}$**(B)**	$\#_{SP_{native}}$**(A)**	$\#_{SP_{native}}$**(B)**
Top-1	6555	6567	6222	6332
Top-2	6819	6861	6563	6657
Top-3	6909	6961	6826	6791
Top-4	7043	7055	6895	6947
Top-5	7089	7083	6949	6976
TOTAL-#	7459			
TEST-III				
	$\#_{SP_{fold}}$**(A)**	$\#_{SP_{fold}}$**(B)**	$\#_{SP_{native}}$**(A)**	$\#_{SP_{native}}$**(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			

For each dataset, the numbers in columns "A" are calculated from SP_{fold} ("$\#_{SP_{fold}}$", "true" potential parameters) and SP_{native} ("$\#_{SP_{native}}$", "extracted" potential parameters), obtained from the RNA09 database, and the numbers in columns "B" are calculated from SP_{fold} ("$\#_{SP_{fold}}$") and SP_{native} ("$\#_{SP_{native}}$"), 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

Quasi-chemical approximation-based potentials

The 152 RNA loops/junctions within our training dataset consist of a total number of $N_{total} = 572$ nucleotides, with the numbers of each type of the nucleotide $(\rho_A,\rho_C,\rho_G,\rho_U) = (217,105,121,129)$ and the corresponding mole fraction of each nucleotide $(\chi_A,\chi_C,\chi_G,\chi_U) = (0.38,0.18,0.21,0.23)$.

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 (θ, η), we calculate the numbers $\rho_{ij}^{obs}(\theta, \eta)$ of the dinucleotides conformations and the total observed number $N(\theta, \eta) = \sum_{i,j} \rho_{ij}^{obs}(\theta, \eta)$. According to Eq. 8, we then calculate the expected number $\rho_{ij}^{exp}(\theta, \eta)$ of each type of dinucleotides (i,j) with pseudo-torsion angles (θ, η). If no pseudo-torsion angles (θ,η) is observed for dinucleotide (i,j), we assign an unfavorable potential $u_{ij}(\theta, \eta) = +2.0k_BT$ 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^{-3}, 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 EM64T 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 SP_{fold} are less uniform than the "extracted" potentials (Fig. 2), suggesting that the "true" potentials SP_{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 (SP_{native} or SP_{fold}) 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

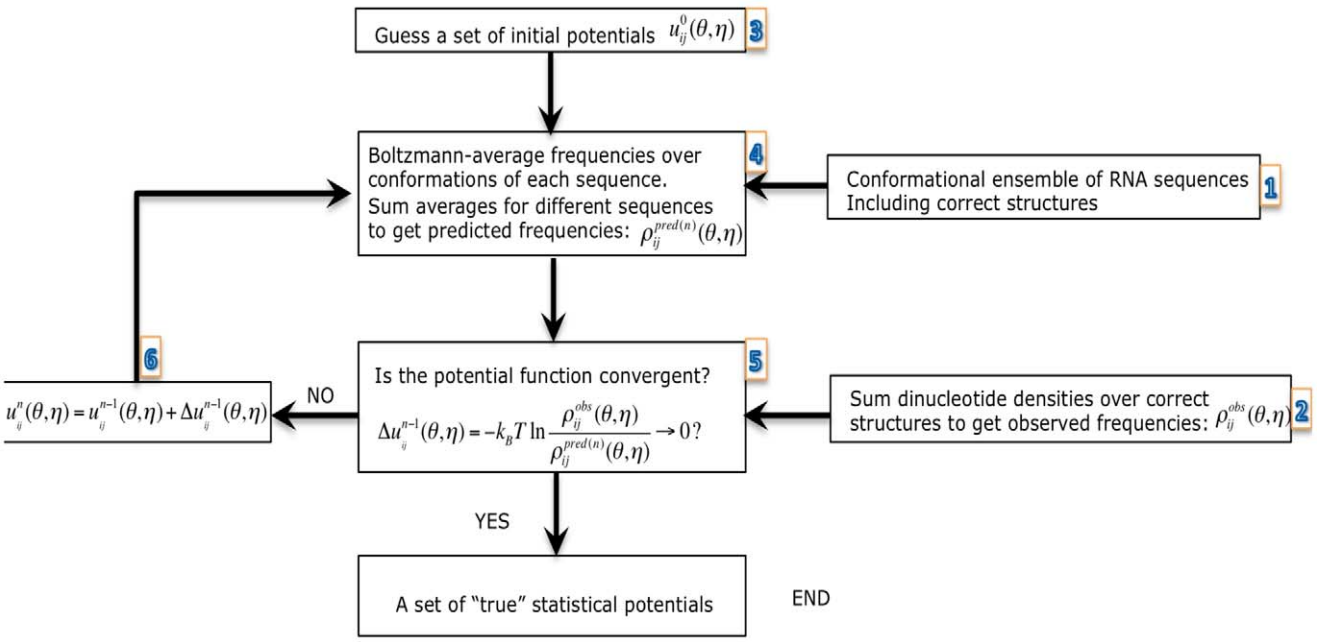

Figure 1. Flow chart of the iterative approach.
doi:10.1371/journal.pone.0048460.g001

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_{native} and SP_{fold} are derived based on the frequency $\rho_{ij}(\theta, \eta)$ (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_{native} succeeded in finding 130 coarse-grained "correct" (lowest potential) structures out of 152 loops and junctions. In contrast, SP_{fold} can give successful predictions for all the 152 the coarse-grained "correct" conformations (Table 3A). SP_{fold} is more reliable than SP_{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 Å. 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_{native} and SP_{fold} potentials can successfully find out 62 and 70 coarse-grained "correct" conformations, respectively, out of the 72 test cases (Table 3B). The result indicates that the "true" potential parameters SP_{fold} are more reliable than the directly extracted potentials SP_{native} (success rate 97% vs 86%). One of the two failed predictions for SP_{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 (*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 (*TEST-II*), and 934 of these 1609 RNAs have high-resolution structures ($\leq$3.0 Å), which contain 1119 RNA loops and junctions (*TEST-III*).

Our test results show that the numbers of the correct predictions with the statistical potentials SP_{fold} and SP_{native} are 7364 (success rate = 87%) v.s. 6992 (success rate = 83%) for *TEST-I*, 6553 (success rate = 88%) v.s. 6222 (success rate = 83%) for *TEST-II*, and 1070 (success rate = 95%) v.s. 1001 (success rate = 89%) for *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_{fold} v.s. 7857 with SP_{native} for *TEST-I*, 7089 v.s. 6949 for *TEST-II* and 1102 v.s. 1065 for *TEST-III* (Table 2). Fig. 11 shows the minimal-RMSD structure, the 9-th potential structure computed with SP_{native}, and the predicted

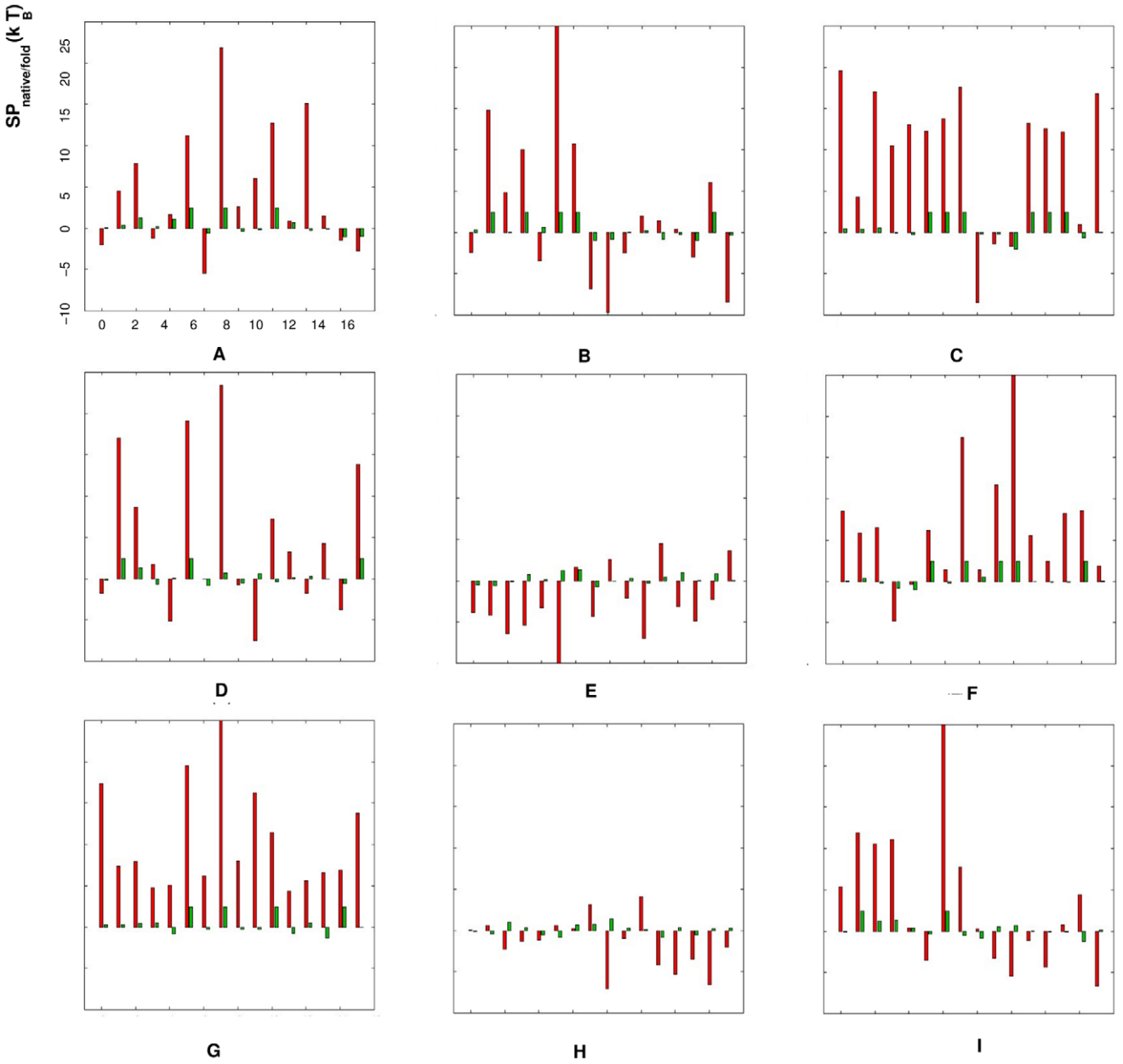

Figure 2. Comparison between the ''extracted'' statistical potentials SP_{native} and the ''true'' potentials SP_{fold}. The figures from (A) to (I) stand for the torsion angles (θ, η =): (A) (g^+, g^+), (B) (g^+, t), (C) (g^+, g^-), (D) (t, g^+), (E) (t, t), (F) (t, g^-), (G) (g^-, g^+), (H) (g^-, t) and (I) (g^-, g^-), respectively. In each figure, the red bars represent the statistical potentials SP_{fold}, the green bars represent the statistical potentials SP_{native}, both of which are obtained 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.g002

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 first 20 lowest-potential conformations for the three databases. The comparisons between the predicted results with SP_{fold} and SP_{native} supports our conclusions in the previously two benchmark tests that SP_{fold} is more reliable than SP_{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_{fold} and SP_{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 ($\Delta u_{ij}^n(\theta, \eta) = 0.001$ 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 $\Delta u_{ij}^n(\theta, \eta)$ do not have strong influence on the accuracy of the structure predictions.

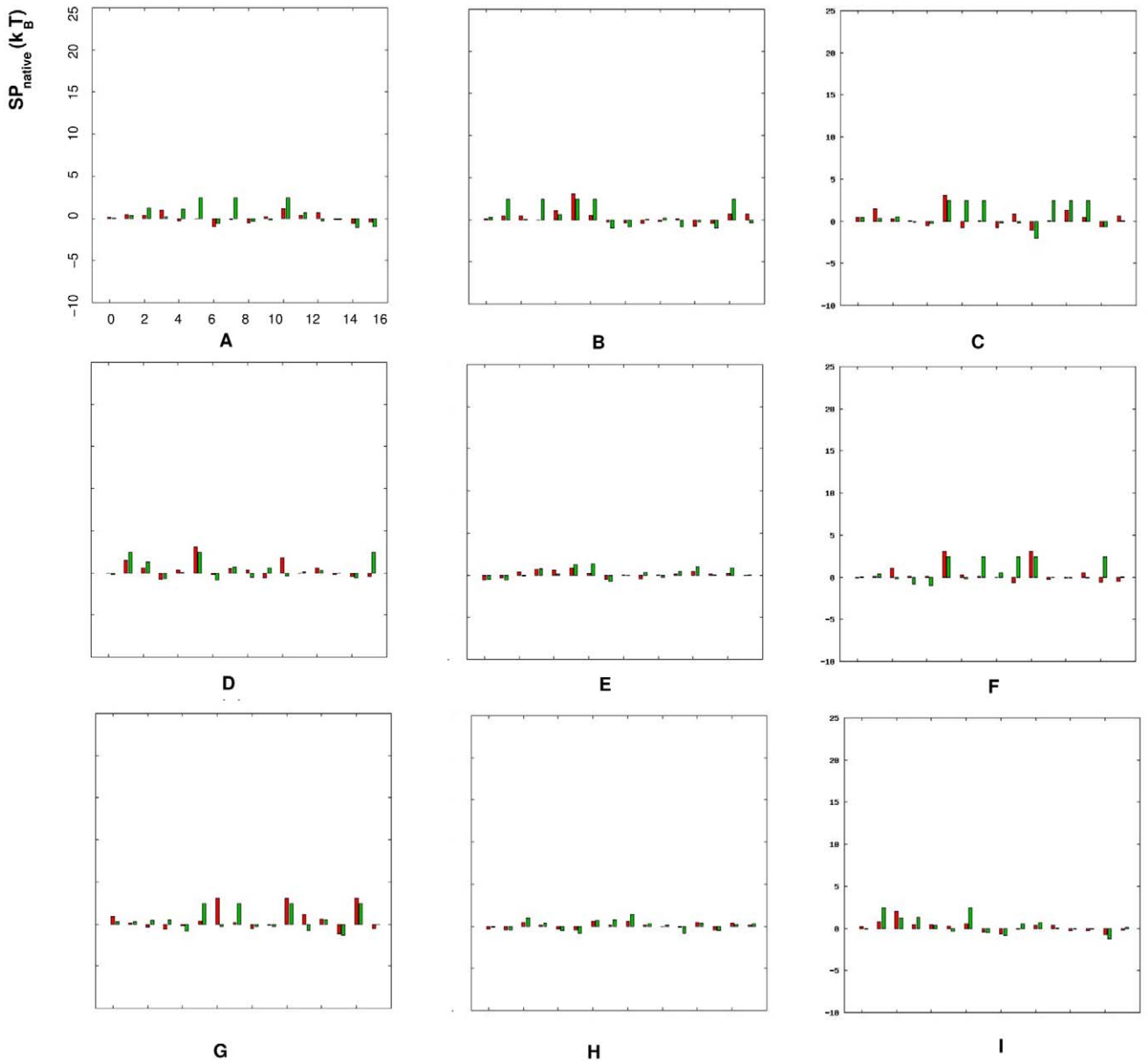

Figure 3. Comparison between the ''extracted'' statistical potentials SP_{native} from the RNA09 dataset and the Leontis dataset, respectively. The figures from (A) to (I) stand for the torsion angles (θ, η =): (A) (g^+, g^+), (B) (g^+, t), (C) (g^+, g^-), (D) (t, g^+), (E) (t, t), (F) (t, g^-), (G) (g^-, g^+), (H) (g^-, t) and (I) (g^-, g^-), respectively. In each figure, the red bars represent the statistical potentials SP_{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

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_{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 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, 255 internal/bulge loops, 184 pseudoknot loops, 1265 multibranched loops and 119 junctions. The correct (top-1) predictions with SP_{fold} for each type of RNA loops and junctions are 1274 (success rate = 90.6%), 245 (success rate = 96.1%), 178 (success rate = 96.7%), 1162 (success rate = 91.9%) and 110

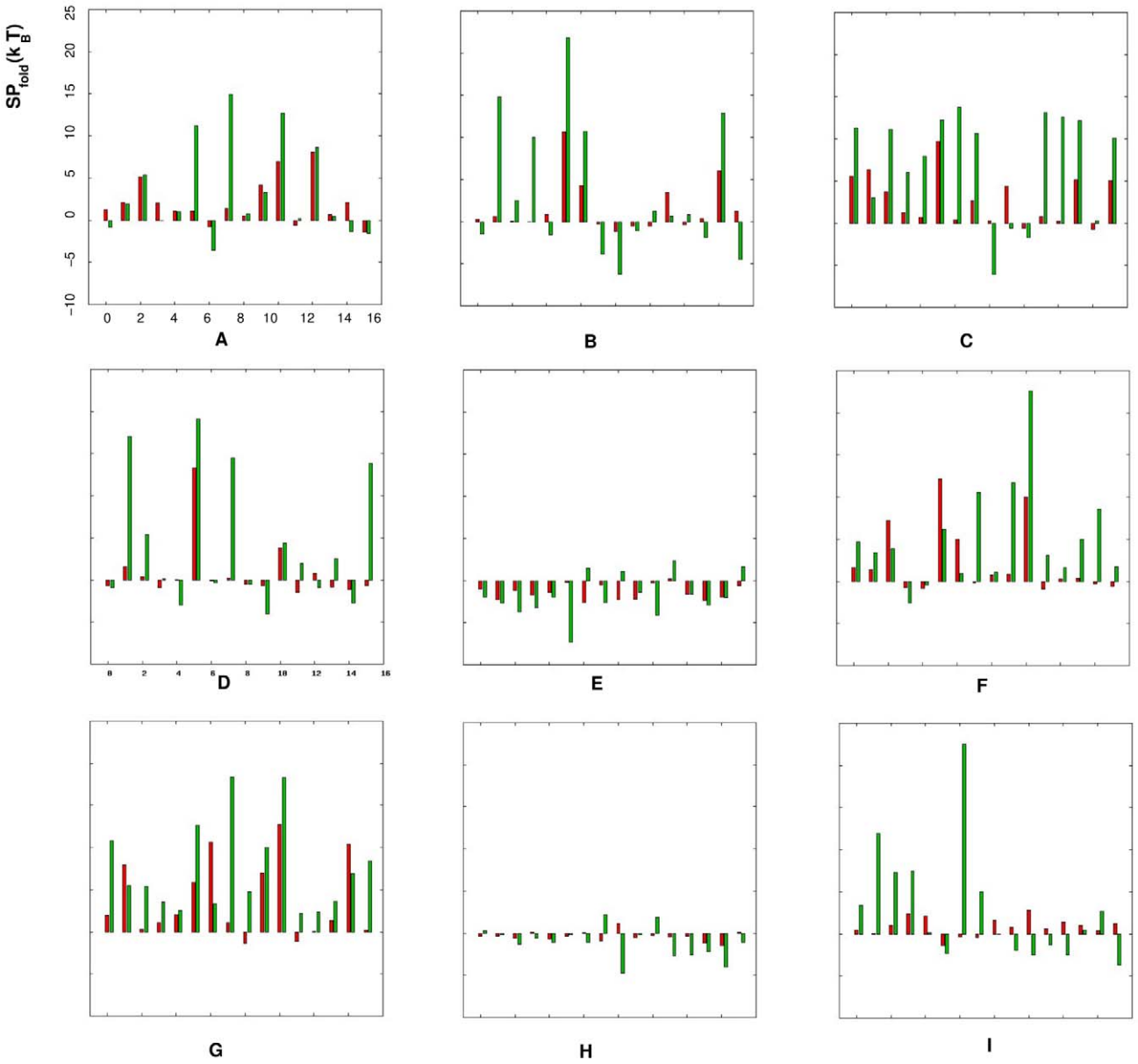

Figure 4. Comparison between the "true" statistical potentials SP_{fold} obtained from the RNA09 dataset and the Leontis dataset, respectively. The figures from (A) to (I) stand for the torsion angles (θ, η =): (A) (g^+, g^+), (B) (g^+, t), (C) (g^+, g^-), (D) (t, g^+), (E) (t, t), (F) (t, g^-), (G) (g^-, g^+), (H) (g^-, t) and (I) (g^-, g^-), respectively. In each figure, the red bars represent the statistical potentials SP_{fold} 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.g004

(success rate = 92.4%), respectively. The total number of the successful predictions is 2969 (success rate = 91.9%).

Statistical potentials and the Leontis dataset

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 $\mathbf{SP}_{native}$ and to search for the "true" potential functions $\mathbf{SP}_{fold}$.

Figs. 3 and 4 show the comparison between the potentials $\mathbf{SP}_{native}/\mathbf{SP}_{fold}$ extracted from the dataset RNA09 and from the Leontis dataset. We apply the "true" potential $\mathbf{SP}_{fold}$ 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.

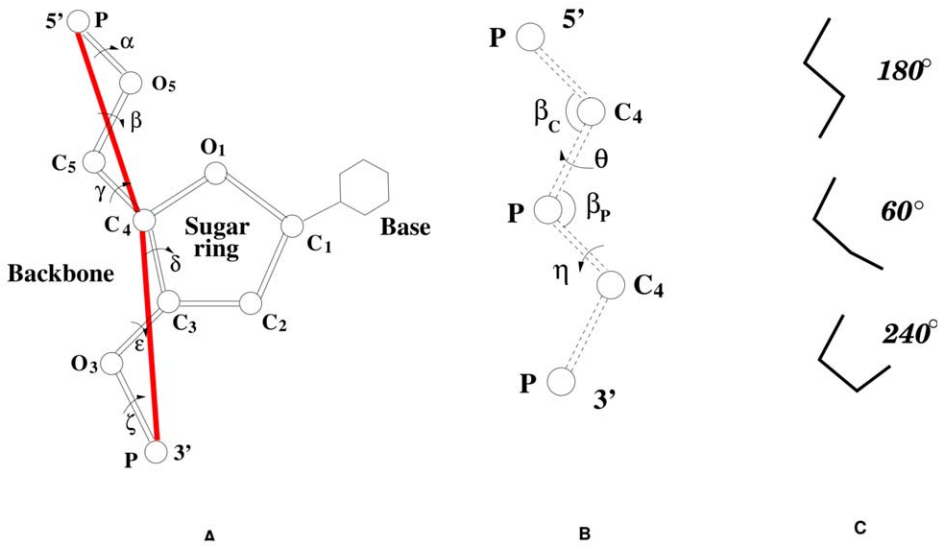

Figure 5. The pseudo-torsional angles. (*A*) The virtual bond scheme for RNA nucleotides. (*B*) The bond angles (β_C, β_P) and the pseudo-torsional angles (θ, η) of the virtual bonds. (*C*) The three preferred rotamer-like configurations of the virtual bonds: t,g^+ and g^-, with the torsional angles equal to 180°, 60° and 300°, respectively.
doi:10.1371/journal.pone.0048460.g005

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 SP_{fold} from the RNA09 dataset. The comparisons for the three test datasets (Table 4 and Fig. 9) show that the two "true" statistical potentials SP_{fold} 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.

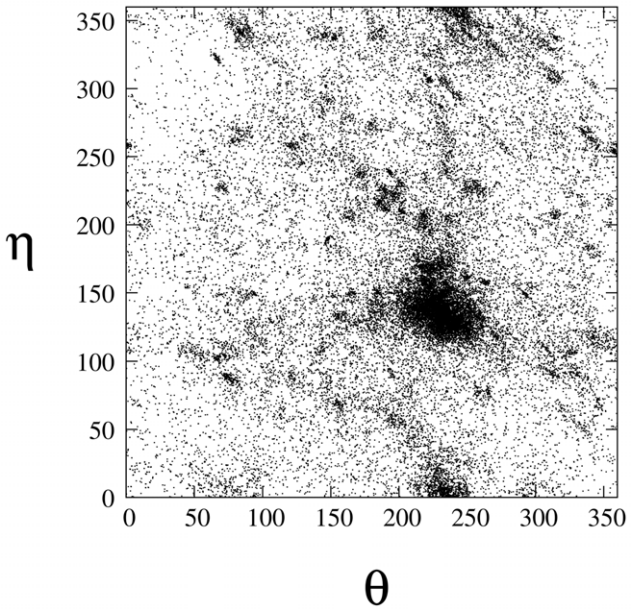

Figure 6. The distribution of the θ and η values for the different types of dinucleotides in RNA loops and junctions. The RNA loops and junctions are selected from the RNA09 dataset and the angles are calculated based on the PDB structures.
doi:10.1371/journal.pone.0048460.g006

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 (SP_{native}) are derived from the distributions of the dinucleotide conformations in the known (native) structures. In the second method, the statistical potentials (SP_{fold}) 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 SP_{fold} consistently outperforms SP_{native} 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 SP_{fold} 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-

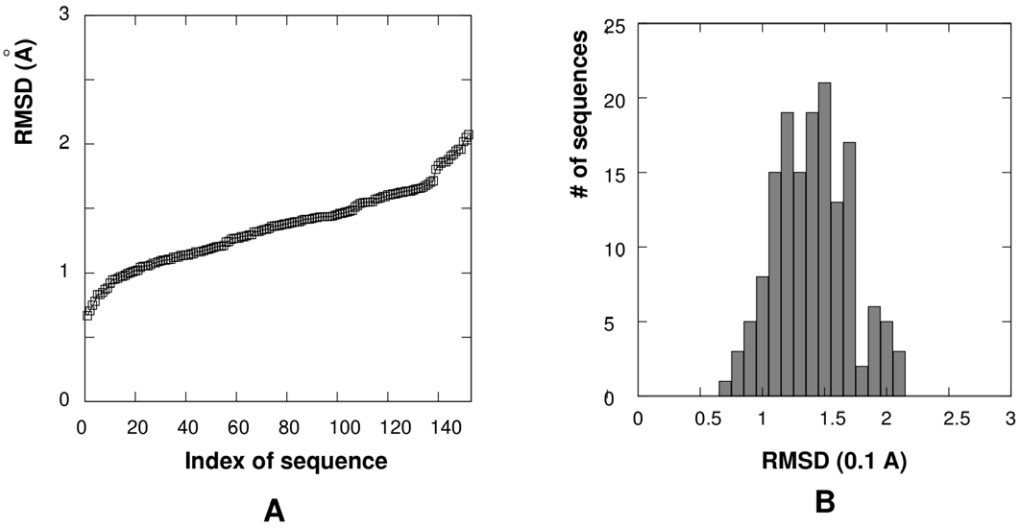

Figure 7. (*A*) The RMSD between the PDB structures and the diamond lattice-represented structures for RNA loops and junctions in the RNA09 dataset. The x-axis represents the index of RNA loops/junctions in the RNA09 dataset. (*B*) The number of RNA loops/junctions within each RMSD-value bin (0.1 Å). The mean and standard deviation of RMSD values for the 152 RNA loops/junctions are 1.35 Å and 0.30 Å, respectively.
doi:10.1371/journal.pone.0048460.g007

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

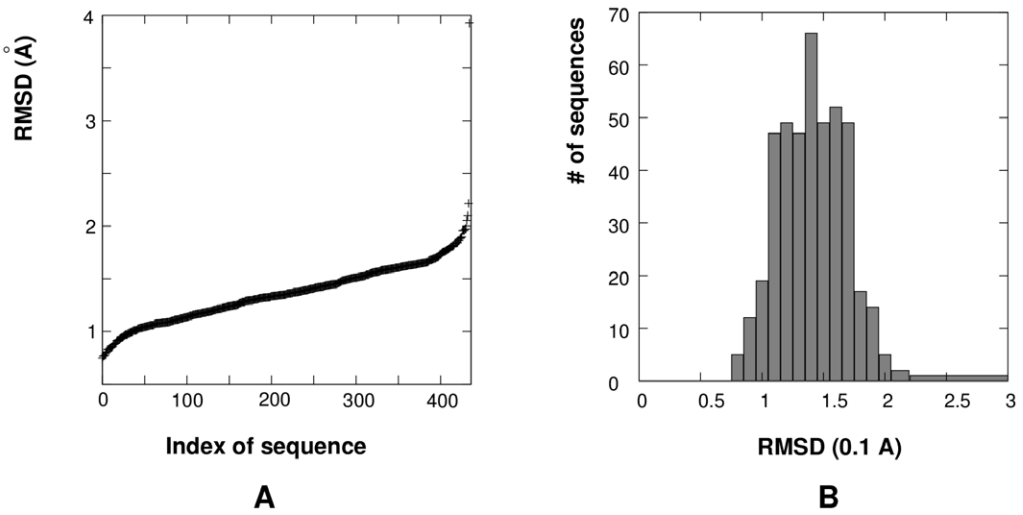

Figure 8. (*A*) The structural differences (RMSD values) between PDB structures and diamond lattice-represented structures for the RNA loops/junctions within the Leontis dataset. The x-axis represents the index of RNA loops/junctions in the Leontis dataset. (*B*) The number of RNA loops/junctions within each RMSD-value bin (0.1 Å). The mean and standard deviation of RMSD values for the 435 RNA loops/junctions are 1.37 Å and 0.29 Å, respectively.
doi:10.1371/journal.pone.0048460.g008

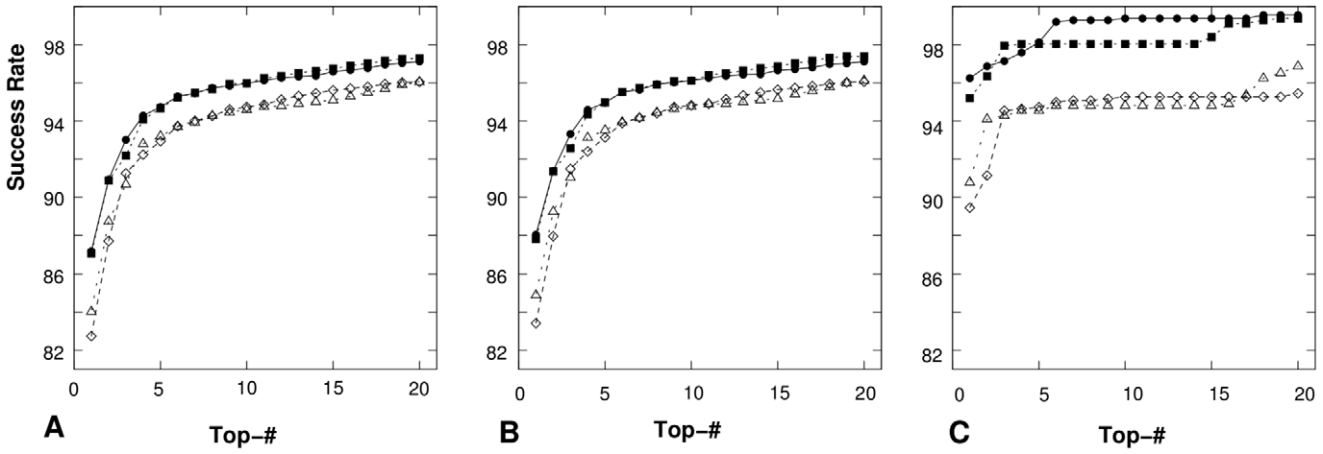

Figure 9. The success rate for the prediction of the coarse-grained correct loop/junction structures for (A) all the 8452 RNA loops and junctions in *TEST-I*, (B) the 7459 RNA loops and junctions in *TEST-II* and (C) the 1119 RNA loops and junctions in *TEST-III*. The "Top-#" in x-axis means that the "correct" structure is in the top # lowest-potential conformations. In each figure, "•" and "■" represent the success rate with SP_{fold} extracted from the Leontis' database and the RNA09 dataset, respectively; while, "Δ" and "◇" represent the success rate with SP_{native} extracted from the Leontis' database and the RNA09 dataset, respectively.
doi:10.1371/journal.pone.0048460.g009

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 $P-O_5^*-C_5^*-C_4^*$ bonds and the $C_4^*-C_3^*-O_3^*-P$ 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.

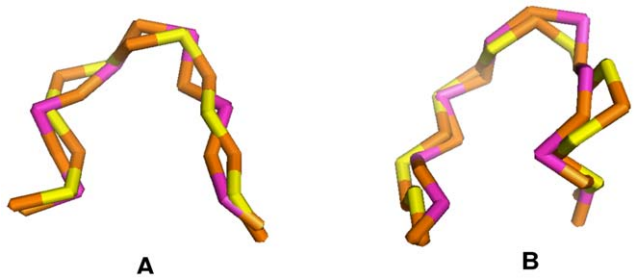

Figure 10. Comparison between the virtual bond-based PDB structure and correctly predicted structure with statistical potentials SP_{fold}**.** The loops are (A) a hairpin loop from PDB structure 1IVS and (B) an internal loop from PDB structure 1JJ2, respectively. In both figures, the structures shown in brown (P) and yellow (C_4^*) stand for the PDB structure, and the ones shown in brown (P) and purple (C_4^*) represent the correctly predicted structure, which have the lowest potential and the minimal RMSD. The RMSD values are (A) 1.52 Å and (B) 1.96 Å, respectively. The atomic structures are illustrated with Pymol software (http://www.pymol.org/).
doi:10.1371/journal.pone.0048460.g010

The loop structures are enumerated on a diamond lattice through exhaustive self-avoiding walks. If the first atom (P) is fixed at position $\overrightarrow{X_0}$, then the coordinates ($\overrightarrow{X_N}$) of N-th atom (P or C_4^*) can be calculated with

$$\overrightarrow{X_N} = \overrightarrow{X_0} + \sum_{i=1}^{N} l_i \overrightarrow{b_i} \quad (1)$$

where, $l_i = 3.9\text{A}$ is the bond length of virtual bond b_i and $\overrightarrow{b_i}$ is the unit vector of virtual bond b_i. Also, the relation between the $(i-1)$-th virtual bond (b_{i-1}) and the i-th virtual bond (b_i) is:

$$\overrightarrow{b_i} = T(\beta_i, \theta_i) \cdot \overrightarrow{b_{i-1}} \quad (2)$$

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

$$T(\beta_i, \theta_i) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\beta_i & -\sin\beta_i \\ 0 & \sin\beta_i & \cos\beta_i \end{bmatrix} \times \begin{bmatrix} \cos\theta_i & -\sin\theta_i & 0 \\ \sin\theta_i & \cos\theta_i & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

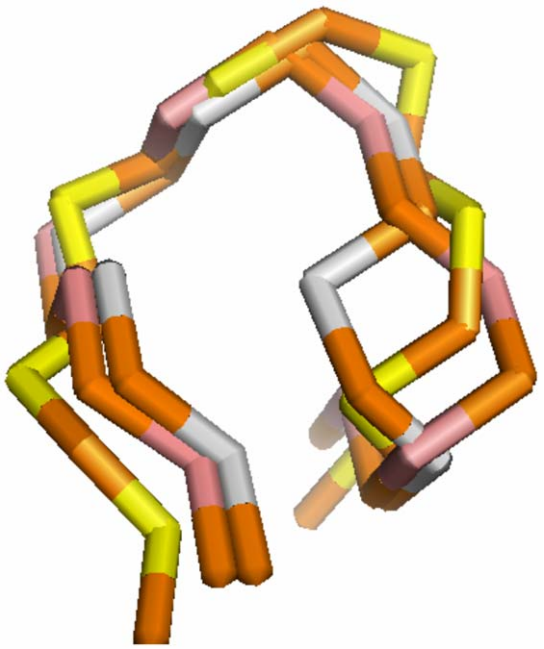

Figure 11. Comparison between the virtual bond-based PDB structure and top-ranked predicted structures with statistical potentials SP_{fold}**.** The loop is a multibranched loop from PDB structure 3CCM. In the figure, the structure shown in brown (P) and yellow (C_4^*) stands for the PDB structure, 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 SP_{fold}. The minimal-RMSD structure has the 9-th lowest potential. The atomic structures are illustrated with Pymol software (http://www.pymol.org/).
doi:10.1371/journal.pone.0048460.g011

Therefore, if the unit vector of the first virtual bond ($P-C_4^*$) is $\overrightarrow{b_1}$, we can obtain the unit vector of the i-th virtual bond ($P-C_4^*$ or C_4^*-P) with

$$\overrightarrow{b_i} = \prod_{j=2}^{i} T(\beta_j, \theta_j)\cdot\overrightarrow{b_1} \tag{4}$$

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 $5'\rightarrow 3'$ direction, if the first atom P is fixed at $\overrightarrow{X_0}$ = col (0, 0, 3.9) (Å), the second atom C_4^* is fixed at $\vec{x}_1$ = col (0, 0, 0) (Å), then, $\overrightarrow{b_1}$ = col (0, 0, −1). According to Eq. 1, the coordinate of the third atom P can be computed:

$$\overrightarrow{X_2} = \overrightarrow{X_0} + l_1\overrightarrow{b_1} + l_2 T(\pi-\beta_P, \eta)\cdot\overrightarrow{b_1}$$

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

$$T(60^\circ, 180^\circ) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1/2 & -\sqrt{3}/2 \\ 0 & -\sqrt{3}/2 & 1/2 \end{bmatrix}$$

Therefore, the coordinate of the third atom P is 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:

$$u(r) = -k_B T \ln\frac{\rho(r)}{\rho^*(r)} \tag{5}$$

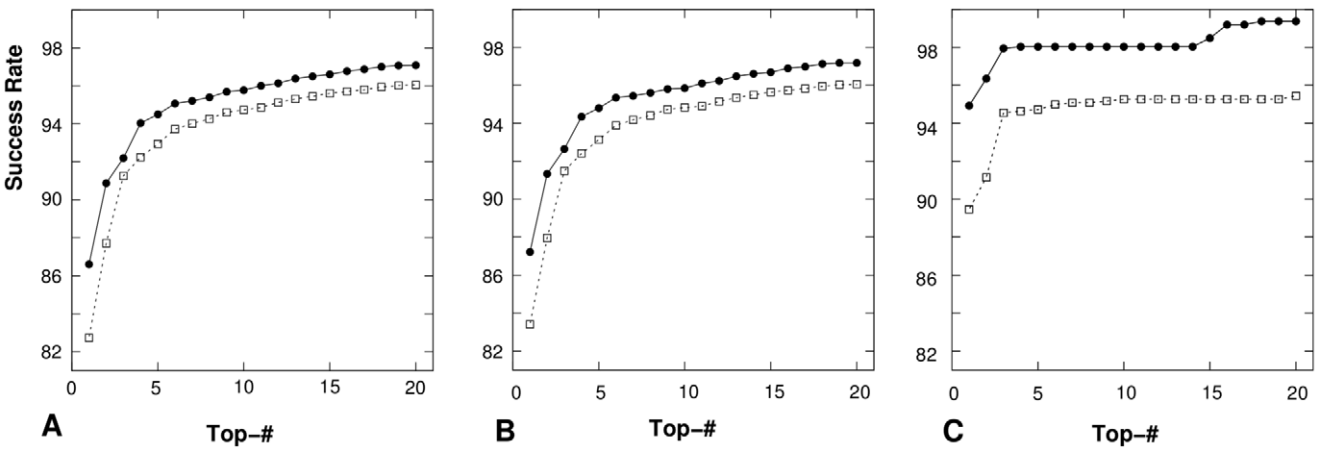

Figure 12. The success rate of coarse-grained correct loop/junction structure predictions for (A) all the 8452 RNA loops/junctions in *TEST-I*, (B) the 7459 RNA loops and junctions in *TEST-II*, and (C) the 1119 RNA loops and junctions in *TEST-III*. The "Top-#" in x-axis means that the "correct" structure is in the top # lowest-potential conformations. In each figure, • and □ represent the success rate with SP_{fold} and SP_{native}, respectively.
doi:10.1371/journal.pone.0048460.g012

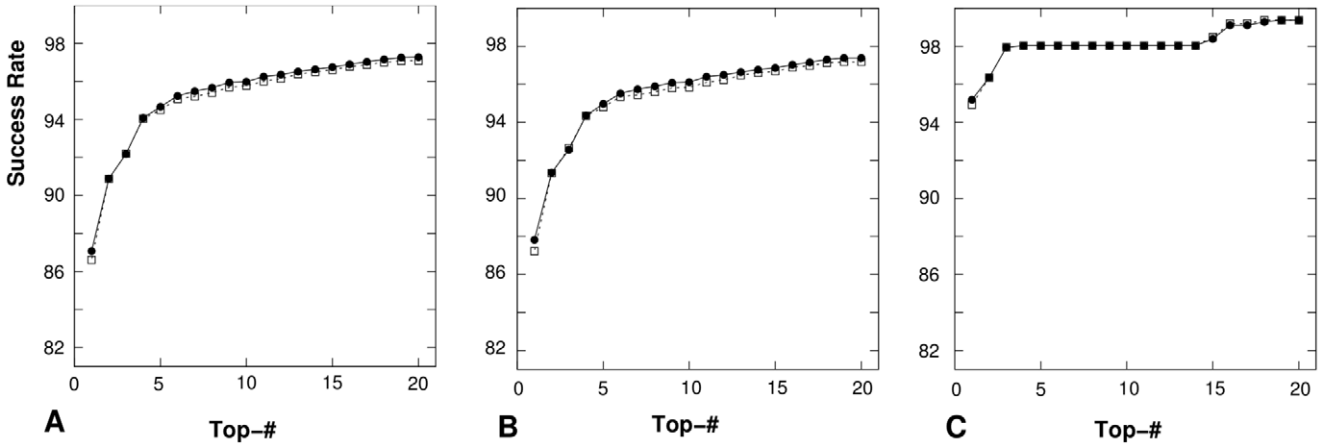

Figure 13. The sensitivity of the structure prediction to the change of the convergence threshold parameter: $\Delta u_{ij}^{n}(\theta, \eta) = 0.001$ v.s. 0.01. The tests are based on the predictions of the RNA loops and junctions in all the three databases: (A) *TEST-I*, (B) *TEST-II* and (C) *TEST-III*. In each figure, ● and □ represent the success rate of the SP_{fold}-based structure prediction with $\Delta u_{ij}^{n}(\theta, \eta) = 0.001$ and 0.01, respectively.
doi:10.1371/journal.pone.0048460.g013

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 θ ($P-C_4^*-P-C_4$) and η ($C_4^*-P-C_4-P$) and the different dinucleotide sequences (Fig. 5). The potential $u_{ij}(\theta, \eta)$ is extracted based on the following formula:

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

where i and j denote the types of the nucleotides (bases) of the two consecutive nucleotides within the same loop and (θ, η) 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):

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

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

$$\rho_{ij}^{exp}(\theta, \eta) = N(\theta, \eta)\chi_i\chi_j \quad (8)$$

where $N(\theta,\eta)$ is the total observed number of dinucleotides with pseudo-torsion angles θ and η (calculated with a composition-independent scale, meaning each nucleotide is independent to other nucleotides in the state (Eq. 8)) [70], and χ_i and χ_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 SP_{native}. The SP_{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 (SP_{fold}).

Convergent iterative approach. To extract the "true" energy (SP_{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 structures in the training dataset and the full conformational ensemble (native and nonnative) lead to the same frequencies of the different dinucleotide conformations [63,68,69].

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 SP_{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:

$$E^{(n)} = \sum_{\text{dinucleotides}} u_{ij}^{(n)}(\theta, \eta) \quad (9)$$

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

$$\rho_{ij}^{\text{pred}(n)}(\theta, \eta) = \sum_{s}^{S} \frac{\sum_{l}^{L} n_{ij}(\theta,\eta) \cdot e^{-E^{(n)}(s,l)/k_B T}}{\sum_{l}^{L} e^{-E^{(n)}(s,l)/k_B T}} \quad (10)$$

where $E^{(n)}(s, l)$ is the potential/score for the l-th conformation of the s-th RNA loop/junction sequence, computed with the potentials $u_{ij}^{(n)}(\theta, \eta)$ at the n-th step (Eq. 9), $n_{ij}(\theta, \eta)$ is the number of dinucleotide (i, j) with pseudo-torsion angles (θ, η) 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:

$$\Delta u_{ij}^{(n)}(\theta, \eta) = -k_B T \ln \frac{\rho_{ij}^{\text{obs}}(\theta, \eta)}{\rho_{ij}^{\text{pred}(n)}(\theta, \eta)} \quad (11)$$

where $\rho_{ij}^{\text{pred}(n)}(\theta, \eta)$ is calculated by Eq. 10 and $\rho_{ij}^{\text{obs}}(\theta, \eta)$ 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 ($\Delta u_{ij}^{(n)}(\theta, \eta) \gg 0$). For each n-th step, we refine the potentials by accounting for the differences between the "observed" and the "predicted" frequencies of the dinucleotide conformations (see the parameter $\Delta u_{ij}^{(n)}(\theta,\eta)$ in Eq. 11):

$$u_{ij}^{(n)}(\theta, \eta) = u_{ij}^{(n-1)}(\theta, \eta) + \Delta u_{ij}^{(n-1)}(\theta, \eta) \quad (12)$$

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

Flowchart

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

1. Prepare the training dataset of the native structures. Download RNA structures from PDB database and extract the loops and junctions. For each RNA loop/junction in the training set, generate the full conformational ensemble of RNA backbone by self-avoiding exhaustive walks on a diamond lattice (Vfold model). The conformation ensemble is used in the iterative calculation.
2. Calculate the observed dinucleotide frequencies $\rho_{ij}^{\text{obs}}(\theta, \eta)$ by summing the dinucleotide densities observed in the training dataset.
3. Choose a set of initially guessed potentials. In this work, we start with the "extracted" quasi-chemical approximation-based potentials ($\mathbf{SP}_{\text{native}}$).
4. Calculate the potential/score for each generated conformation of each RNA loop/junction sequence by using Eq. 9 with iterative potentials $u_{ij}^{(n-1)}(\theta, \eta)$. Then calculate the weighted predicted frequencies $\rho_{ij}^{\text{pred}(n)}(\theta, \eta)$ of each dinucleotide (i, j) with pseudo-torsion angles (θ, η), by employing Eq. 10.
5. Calculate the convergence parameter $\Delta u_{ij}^{(n-1)}(\theta, \eta)$ according to Eq. 11. If the following convergence condition is satisfied:

$$|\Delta u_{ij}^{(n-1)}(\theta, \eta)| \leq C_0 \quad (13)$$

with a preset small value of the threshold parameter C_0, say, 10^{-3}, then skip to the final step and the "true" potentials $\mathbf{SP}_{\text{fold}}$ are obtained; otherwise, continue to the next step.

6. Adjust the current potentials $u_{ij}^{(n-1)}(\theta, \eta)$ by adding the correction $\Delta u_{ij}^{(n-1)}(\theta, \eta)$ (Eq. 12), and obtain a new set of potentials $u_{ij}^{(n)}(\theta, \eta)$. Move to the next ($n$-th) iterative step and return to Step 4 for the next cycle.

Preparation of the training dataset

We use the RNA 2009 database (RNA09; http://Kinemage.biochem.duke.edu/databases/rnadb.phb, an updated version of previous 2005 database (RNA05) [51]) to extract the $\mathbf{SP}_{\text{native}}$ potential and use the database as the training dataset to search for the "true" potential $\mathbf{SP}_{\text{fold}}$. The RNA09 dataset consists of 262 RNA sequences with the experimentally determined structures of resolution ≤ 3.0 Å.

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 several automated tools to detect the base pairs in RNA structures. In this work, we use the 3DNA software package (http://3dna.rutgers.edu/home) [73] to search for all the possible base pairs from the RNA structures.

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 ≤ 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 Vfold model to generate the full conformation ensemble of a given RNA loop or junction sequence. In the Vfold-

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 $4 \times 4 \times 3 \times 3$ 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 as 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 **A** and 0.30 Å, respectively.. We will use the coarse-grained correct structure to calculate the observed dinucleotide frequencies, from which $\mathbf{SP}_{\text{native}}$ and $\mathbf{SP}_{\text{fold}}$ 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 SCJ.

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: 2248–2263.
3. Gherghe C, Leonard C, Ding F, Dokholyan 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. Hajdin C, Ding F, Dokholyan 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 struc22 tures. Proc Natl Acad Sci U S A 104: 14664.
6. Shapiro B, Yingling Y, Kasprzak W, Bindewald E (2007) Bridging the gap in RNA structure prediction. Curr Opin Struct Biol 17: 157–165.
7. Parisien M, Major F (2008) The MC-fold and MC-Sym pipeline infers RNA structure from sequence data. Nature 452: 51–55.
8. Jonikas M, Radmer R, Laederach A, Das R, Pearlman S, et al. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA 15: 189–199.
9. Das R, Karanicolas J, Baker D (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. Nat Methods 7: 291–294.
10. Yang S, Parisien M, Major F, Roux B (2010) RNA structure determination using SAXS data. J Phys Chem B 114: 10039–10048.
11. Pasquali S, Derreumaux P (2010) HiRE-RNA: a high resolution coarse-grained energy model for RNA. J Phys Chem B 114: 11957–11966.
12. Xia Z, Gardner D, Gutell R, Ren P (2010) Coarse-grained model for simulation of RNA three-dimensional structures. J Phys Chem B 114: 13497–13506.
13. Flores S, Altman R (2010) Turning limited experimental information into 3D models of RNA. RNA 16: 1769.
14. Cao S, Chen SJ (2011) Physics-based de novo prediction of RNA 3D structures. J Phys Chem B 115: 4216–4226.
15. Bernauer J, Huang X, Sim A, Levitt M (2011) Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. RNA 17: 1066–1075.
16. Sripakdeevong P, Kladwang W, Das R (2011) An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling. Proc Natl Acad Sci U S A 108: 20573–20578.
17. Parisien M, Major F (2012) Determining RNA three-dimensional structures using low resolution data. J Struct Biol: in press.
18. Bida J, Maher III L (2012) Improved prediction of RNA tertiary structure with insights into native state dynamics. RNA 18: 385–393.
19. Capriotti E, Marti-Renom M (2008) Computational RNA structure prediction. Current Bioinformatics 3: 32–45.
20. Laing C, Schlick T (2010) Computational approaches to 3D modeling of RNA. J Phys: Condens Matter 22: 283101.
21. Laing C, Schlick T (2011) Computational approaches to RNA structure prediction, analysis, and design. Curr Opin Struct Biol 21: 306–318.
22. Rother K, Rother M, Boniecki M, Puton T, Bujnicki J (2011) RNA and protein 3D structure modeling: similarities and differences. Journal of molecular modeling : 1–12.
23. Cruz 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 8: 1–16.
24. Leontis N,Westhof E (2012) RNA 3D Structure Analysis and Prediction. Springer-Verlag.
25. Dima R, Hyeon C, Thirumalai D (2005) Extracting stacking interaction parameters for RNA from the data set of native structures. J Mol Biol 347: 53–69.
26. Mathews D, Sabina J, Zuker M, Turner D (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure1. J Mol Biol 288: 911–940.
27. Wu J, Gardner D, Ozer S, Gutell R, Ren P (2009) Correlation of RNA secondary structure statistics with thermodynamic stability and applications to folding. J Mol Biol 391: 769–783.
28. Nussinov R, Jacobson A (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc Natl Acad Sci U S A 77: 6309.
29. Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. Science 244: 48–52.
30. Hofacker I, Fontana W, Stadler P, Bonhoeffer L, Tacker M, et al. (1994) Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie/ Chemical Monthly 125: 167–188.
31. Hofacker I (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31: 3429–3431.
32. Mathews D, Turner D (2006) Prediction of RNA secondary structure by free energy minimization. Curr Opin Struct Biol 16: 270–278.
33. Serra M, Barnes T, Betschart K, Gutierrez M, Sprouse K, et al. (1997) Improved param eters for the prediction of RNA hairpin stability. Biochemistry 36: 4844–4851.
34. 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.
35. Mathews D, Disney 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.
36. Dale T, Smith R, Serra M (2000) A test of the model to predict unusually stable RNA hairpin loop stability. RNA 6: 608–615.
37. Schlick T, Collepardo-Guevara R, Halvorsen L, Jung S, Xiao X (2011) Biomolecular mod eling and simulation: a field coming of age. Q Rev Biophys 44: 191.
38. Hyeon C, Thirumalai D (2007) Mechanical unfolding of RNA: from hairpins to structures with internal multiloops. Biophys J 92: 731–743.
39. 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.
40. Sorin E, Rhee Y, Nakatani B, Pande V (2003) Insights into nucleic acid conformational dynamics from massively parallel stochastic simulations. Biophys J 85: 790–803.
41. Keating K, Pyle A (2010) Semiautomated model building for RNA crystallography using a directed rotameric approach. Proc Natl Acad Sci U S A 107: 8177.
42. Cao S, Chen SJ (2005) Predicting RNA folding thermodynamics with a reduced chain representation model. RNA 11: 1884–1897.
43. Cao S, Chen SJ (2006) Predicting RNA pseudoknot folding thermodynamics. Nucleic Acids Res 34: 2634.
44. Cao S, Chen SJ (2009) Predicting structures and stabilities for h-type pseudoknots with interhelix loops. RNA 15: 696–706.
45. Liu L, Chen SJ (2010) Computing the conformational entropy for RNA folds. J Chem Phys 132: 235104.
46. Cao S, Giedroc D, Chen SJ (2010) Predicting loop–helix tertiary structural contacts in RNA pseudoknots. RNA 16: 538–552.
47. Cao S, Chen SJ (2012) Predicting kissing interactions in microRNA–target complex and assessment of microrna activity. Nucleic Acids Res 40: 4681–4690.
48. Chen SJ (2008) RNA folding: conformational statistics, folding kinetics, and ion electro statics. Annu Rev Biophys 37: 197.
49. Vecenie C, Morrow C, Zyra A, Serra M (2006) Sequence dependence of the stability of RNA hairpin molecules with six nucleotide loops. Biochemistry 45: 1400–1407.

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, Arendall W, Richardson D, Richardson J (2003) RNA backbone is rotameric. 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: 1086.
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, Crothers 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 polyribonucleotides: a virtual bond model. Biopolymers 11: 1–23.
59. Olson WK (1975) Configuration statistical of polynucleotide chains: A single virtual bond treatment. Macromolecules 8: 272–275.
60. Olson WK (1980) Configurational statistics of polynucleotide chains: An updated virtual bond model to treat effects of base stacking. Macromolecules 13: 721–728.
61. Flory P, et al. (1969) Statistical mechanics of chain molecules. Biopolymers 8: 699–700.
62. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, et al. (2000) The protein data bank. Nucleic Acids Res 28: 235–242.
63. 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.
64. Lu H, Skolnick J (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins: Struct, Funct, Bioinf 44: 223–232.
65. 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.
66. 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.
67. Sharma S, Ding F, Dokholyan N (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. Bioinformatics 24: 1951–1952.
68. Thomas P, Dill K (1996) Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 257: 457–469.
69. 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.
70. 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.
71. Leontis N, Westhof E (2001) Geometric nomenclature and classification of RNA base pairs. RNA 7: 499–512.
72. Leontis N, Stombaugh J, Westhof E (2002) Motif prediction in ribosomal RNAs lessons and prospects for automated motif prediction in homologous RNA molecules. Biochimie 84: 961–973.
73. 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.
74. 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.