

# 3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures

Jian Wang<sup>†</sup>, Yunjie Zhao<sup>†</sup>, Chunyan Zhu and Yi Xiao\*

Biomolecular Physics and Modeling Group, Department of Physics and Key Laboratory of Molecular Biophysics of the Ministry of Education, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China

Received March 06, 2014; Revised February 04, 2015; Accepted February 06, 2015

## ABSTRACT

**Model evaluation is a necessary step for better prediction and design of 3D RNA structures. For proteins, this has been widely studied and the knowledge-based statistical potential has been proved to be one of effective ways to solve this problem. Currently, a few knowledge-based statistical potentials have also been proposed to evaluate predicted models of RNA tertiary structures. The benchmark tests showed that they can identify the native structures effectively but further improvements are needed to identify near-native structures and those with non-canonical base pairs. Here, we present a novel knowledge-based potential, 3dRNAscore, which combines distance-dependent and dihedral-dependent energies. The benchmarks on different testing datasets all show that 3dRNAscore are more efficient than existing evaluation methods in recognizing native state from a pool of near-native states of RNAs as well as in ranking near-native states of RNA models.**

## INTRODUCTION

RNA molecules play different biological roles besides messengers between DNA and protein (1,2), e.g. regulatory functions (3). Like proteins, the 3D structural information is needed for better understanding of the functions of RNAs. Since the number of available RNA experimental structures is very limited at present, several computational methods have been proposed for structural modeling or RNA tertiary structures prediction (4–13). These methods usually generate a large set of candidates that need to be evaluated.

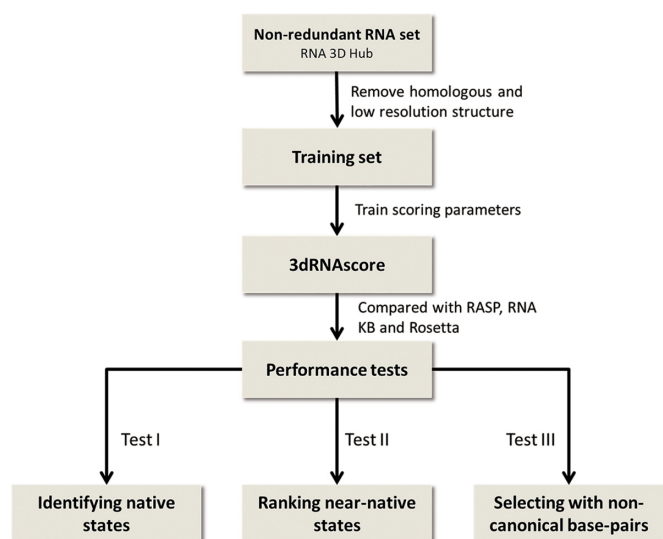
Knowledge-based statistical potential has been proved to be a powerful approach for evaluating models of protein tertiary structures (14–16). Currently, some knowledge-based potentials have also been proposed to evaluate models of RNA tertiary structures (17–21). For examples, the Ribonucleic Acids Statistical Potential (RASP) developed by

Capriotti *et al.* (17,19) and the coarse-grained and all-atom RNA KB potentials by Bernauer *et al.* (18). The all-atom version of RASP, RASP-ALL, is a distance-dependent statistical potential with 23 clustered atom types and is trained on a non-redundant training set (randstr) generated by MODELLER (22). The RNA KB potentials are also distance-dependent statistical potentials. In the coarse-grained version of RNA KB potential, five atoms (P, C4' in backbone, and C2, C4, C6 in base) are selected to represent the nucleotide. In the all-atom version of RNA KB potential, unlike RASP in which the atom types are clustered, the atom types in different nucleotides are considered to be different, so totally 85 atom types are considered rather than 23 atom types used in RASP-ALL. Furthermore, The RNA KB potentials used a Dirichlet process mixture model to obtain the distance distributions instead of bin counting (23). The fully differentiable feature also makes it possible for molecular dynamics simulations. The benchmark tests showed that the RASP and KB potentials could identify the native state structures effectively (17,18). However, further improvements are needed to rank near-native structures and pick out the structure closest to the native state from near-native structures including those with non-canonical base pairs, which is important in the prediction of RNA tertiary structures. Besides, there are other statistical potentials for evaluating RNA tertiary structures embedded in the RNA tertiary structure prediction programs and they have been compared in (17). For example, a full atom RNA potential (FARFAR, fragment assembly of RNA with full-atom refinement) available within the ROSETTA suite was successfully used for the de novo prediction and design of non-canonical RNA 3D structures (6,11). This full-atom potential contains weak carbon hydrogen bonding and solvation terms, as well as a complete description for potential hydrogen bonds between bases and backbone oxygen atoms.

In this work, we introduce a novel all-heavy-atom knowledge-based statistical potential, 3dRNAscore, to evaluate the 3D structure of RNA. Unlike the aforementioned two knowledge-based potentials that utilize the distances between atoms, a new energy contribution based on backbone torsion angle (dihedral) is involved in 3dR-

\*To whom correspondence should be addressed. Tel: +86 27 87556652; Fax: +86 27 87542219; Email: yxiao@mail.hust.edu.cn

<sup>†</sup>These authors contributed equally to the paper as first authors.



**Figure 1.** Flow charts of the building steps of 3dRNAscore.

NAscore. The dihedral-based energy can describe the flexibility of RNA molecules more efficiently (24,25). Furthermore, we also consider the RNA stacking interactions in adjacent bases in the calculation of the distance-based energy of 3dRNAscore. It turns out that 3dRNAscore performs better than RASP and KB potentials in identifying RNA native structures from a pool of structural decoys as well as ranking a tremendous amount of near-native RNA tertiary structures.

## MATERIALS AND METHODS

The steps for building 3dRNAscore (Figure 1) are as follows. First, we design the functional form of 3dRNAscore from Boltzmann distribution, which contains two energy terms: the distance-dependent energy and the backbone dihedral-dependent energy. Second, in order to train the parameters in the scoring function, we select a training set of non-redundant RNA tertiary structures in which the structures having high similarity and those having similar submotifs with the structures in the test sets are removed. Using the training set, we determine the parameters in the scoring function. Third, we use the test sets to test the performance of 3dRNAscore. Here, we select three existing test sets. We use different metrics to compare the performance of 3dRNAscore with other scoring methods. The detail of each building step of 3dRNAscore is described in the following figure.

### Distance-dependent energy

Our knowledge-based potential 3dRNAscore is composed of two parts: the first part is based on distance between any two non-bonded heavy atoms located at different residues in the molecule, and the second part is based on the backbone torsion angles. Among them, the distance-dependent part is constructed as a mean force potential (26) derived from Boltzmann distribution.

Samudrala (27) pointed out that three assumptions underlie the mean force potential scoring method. The first

assumption is that the total free energy of a molecule relative to some reference state,  $\Delta G$ , can be expressed as a sum of the relative free energy  $\Delta G(R)$  of a number of individual contributions, where  $R$  represents the value of a 'reaction coordinate'. The reaction coordinate may be any convenient measure of properties of the molecule. When using the distance  $d$  between atoms  $i$  and  $j$  of types  $a$  and  $b$  as the reaction coordinate, the assumption can be expressed as:

$$\Delta G = \Delta G(\{d_{ab}^{ij}\}) = \sum_{ij} \Delta G_{ij}(d_{ab}^{ij}) \quad (1)$$

The second assumption is that the relative free energy can be deduced from the inverse of Boltzmann's law:

$$\Delta G(d) = -k_B T \ln \frac{f_{ab}^{\text{OBS}}(d)}{f_{ab}^{\text{REF}}(d)} \quad (2)$$

Substituting Equation (2) into Equation (1), we have:

$$\Delta G = -k_B T \sum_{ij} \ln \frac{f_{ab}^{\text{OBS}}(d_{ab}^{ij})}{f_{ab}^{\text{REF}}(d_{ab}^{ij})} \quad (3)$$

where  $T$  is the absolute temperature and is set to 298 K,  $k_B$  is the Boltzmann's constant,  $f_{ab}^{\text{OBS}}(d_{ab}^{ij})$  is the observed probability of the distance of  $d_{ab}^{ij}$  between two atoms of types  $a$  and  $b$  in native RNA structures and  $f_{ab}^{\text{REF}}(d_{ab}^{ij})$  is the probability of the distance between two atoms of  $a$  and  $b$  in reference state structures.

The third assumption is the thermodynamic hypothesis (28): the lowest free energy conformation represents the native state. The Equation (3) thus is useful for addressing the issue of ranking near-native structures.

The probability  $f_{ab}^{\text{OBS}}(d_{ab}^{ij})$  could be evaluated with the number of occurrence observed in experimental structures:

$$f_{ab}^{\text{OBS}}(d_{ab}^{ij}) = \frac{N_{ab}(d_{ab}^{ij})}{\sum_d N_{ab}(d_{ab}^{ij})} = \frac{N_{ab}(d_{ab}^{ij})}{N_{ab}} \quad (4)$$

The probability  $f_{ab}^{\text{REF}}(d_{ab}^{ij})$  could not be compiled from experimental structures directly. It relies on which reference state we choose. The RASP (17) and KB (18) potentials used averaged (RAPDF) (27) and quasi-chemical (KBP) (29) approximation reference states, respectively (30). These reference states consider covalent-bond constraints between atoms when counting the number of atom pair distances and usually need larger datasets of experimental RNA structures to obtain accurate potentials. In the average reference state,  $f_{ab}^{\text{REF}}(d_{ab}^{ij})$  is evaluated by ignoring the type of atoms:

$$f_{ab}^{\text{REF}}(d_{ab}^{ij}) = \frac{\sum_{ab} N_{ab}(d_{ab}^{ij})}{\sum_{ab} \sum_d N_{ab}(d_{ab}^{ij})} = \frac{N(d_{ab}^{ij})}{N} \quad (5)$$

**Table 1.** Atom types in 3dRNAscore potential

A	P OP1 OP2 O5' C5' C4' O4' C3' O3' C2' O2' C1' N9 C8 N7 C5 C6 N6 N1 C2 N3 C4
U	P OP1 OP2 O5' C5' C4' O4' C3' O3' C2' O2' C1' N1 C2 O2 N3 C4 O4 C5 C6
C	P OP1 OP2 O5' C5' C4' O4' C3' O3' C2' O2' C1' N1 C2 O2 N3 C4 N4 C5 C6
G	P OP1 OP2 O5' C5' C4' O4' C3' O3' C2' O2' C1' N9 C8 N7 C5 C6 O6 N1 C2 N2 N3 C4

Equation (3) could then be expanded by substituting Equations (4) and (5) to it:

$$\Delta G_{ij}(d_{ab}^{ij}) = -k_B T \ln \frac{f_{ab}^{\text{OBS}}(d_{ab}^{ij})}{f_{ab}^{\text{REF}}(d_{ab}^{ij})} = -k_B T \ln \frac{N_{ab}(d_{ab}^{ij})N}{N(d_{ab}^{ij})N_{ab}} \quad (6)$$

where  $N_{ab}(d_{ab}^{ij})$  is the counts of the occurrence of the distance of  $d$  between two atoms of types  $a$  and  $b$ .  $N$  is the total counts.  $N(d_{ab}^{ij})$  is the counts of the occurrence of the distance  $d$  regardless of atom types.  $N_{ab}$  is the counts of the occurrence of atom pairs of types  $a$  and  $b$  in whole distance region.

In general, the atom pair in which the two atoms belong to two adjacent nucleotides along the sequence would not be considered. This is because that there may be bonding interactions between two atoms in adjacent nucleotides. The distance-dependent statistical potential just considers non-bonding interactions. However, atoms in adjacent bases do not have bonding interactions, and there exist base stacking interactions between them. Therefore, unlike other statistical potentials, the atom pair in two adjacent bases are considered in 3dRNAscore. Furthermore, the maximum value of the distance  $d$  in the process of statistics, namely the cut-off, is taken as 20 Å in 3dRNAscore like other scoring methods.

Our all-heavy-atom distance-dependent potential utilizes all the 85 atom types (Table 1) in the four nucleotides: adenine (A), cytosine (C), guanine (G) and uracil (U). We number these 85 atom types from 1 to 85. For each atom-pair from 1–1 to 85–85, we count the distance distribution in a discrete space with a bin width of 0.15 Å. We use a matrix to represent the distance distribution information that each row denotes an atom-pair type and each column the counts of occurrences of the distance in the corresponding bin. So the data in the  $i$ th row and  $j$ th column of the matrix represents the counts of occurrences that the  $i$ th atom-pair has appeared at a distance of  $j$  in native RNA structure. Finally, we deposit the matrix into a parameter file that will be used in the scoring process.

### Selection of the bin width

The probability  $f_{ij}^{\text{OBS}}(d)$  and  $f_{ij}^{\text{REF}}(d)$  are stored as histograms with bin width of  $\Delta d$ . The size of the bin has a great influence on the probability distribution. Once the bin width is oversized, the probability  $f_{ij}^{\text{OBS}}(d)$  and  $f_{ij}^{\text{REF}}(d)$  stored would be truly rough. When the bin width is undersized, there may be none or little samples located in certain

bins, leading to an inappropriate and artificial discontinuity of the probability distribution. The size of the bin width should be compatible with the total counts of the samples  $N$ . With the increasement of the number of the samples, the size of the bin should be as small as possible, the samples located in the bins should be sufficient, and the samples of the bins should be quite a very small part of the total samples.

Sippl (26) used a bin width of 1 Å, Samudrala (27) used a bin width of 1 Å and then carried out spline fitting, Capriotti (RASP) also used a bin width of 1 Å, Bernauer (KB) used a Dirichlet process mixture model, which leads to analytically differentiable potential functions, rather than fixed binning and spline fitting.

According to Scott's work (31) in 1979, the most appropriate bin width, which provides the most efficient and unbiased estimation of the population distribution, is achieved when:

$$W = 3.49\sigma N^{-\frac{1}{3}} \quad (7)$$

Here,  $W$  means the bin width,  $\sigma$  means the standard deviation, and  $N$  means the number of samples.

We then extracted information from experimental structure, and then got the total samples  $N$  and calculated the standard deviation  $\sigma$ . Thus, we found that the most appropriate value of the bin width for 3dRNAscore is 0.3 Å. Bin width of 0.3 Å is small relative to those aforementioned, which is due to the large amount of samples in training set.

### The problem of sparse data

In mathematical statistics, a larger size of samples yields a more accurate population distribution. A too small size of samples is likely to result in severe uncertainty for population distribution. The accuracy of the distance distribution of a specific pair of atom types thus heavily depends on the size of the total observations on this pair of atom types. That is to say, the  $N_{ij}$  in Equation (4) should not be far too small.

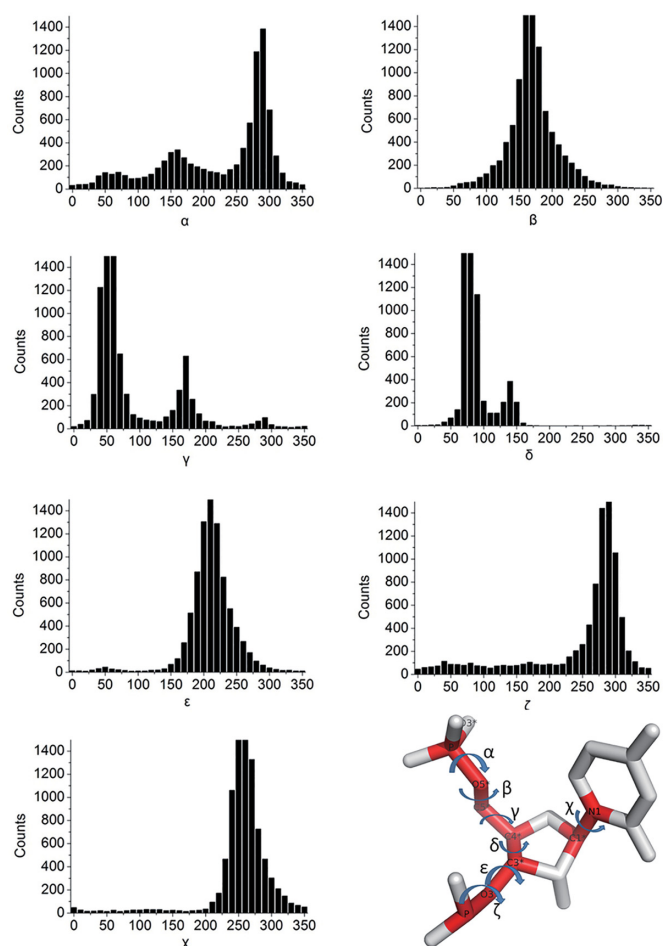
In 1990, Sippl developed a method to address the problem of small data sets (26). He approximated the genuine frequency  $g_{ij}(r)$  by the sum of the total densities  $f(r)$  and the statistical frequencies  $f_{ij}(r)$ :

$$g_{ij}(d) \approx \frac{1}{1+m\sigma} f(d) + \frac{m\sigma}{1+m\sigma} f_{ij}(d) \quad (8)$$

where  $m$  is actually  $N_{ij}$ ,  $\sigma$  represents a custom constant. In the limit  $m \rightarrow \infty$ ,  $g_{ij}(r)$  converges to  $f_{ij}(r)$ . In the limit  $m \rightarrow 0$ ,  $g_{ij}(r)$  converges to  $f(r)$ . Equation (8) ensures that the genuine frequency distribution  $g_{ij}(r)$  resembles the total distribution  $f(r)$  when encountering the situation of sparse data.

In order to avoid sparse data, we checked all the training parameters. There are totally about  $5.25 \times 10^8$  atom pairs in the dataset of 317 structures. Hence, there are averagely  $\sim 1090$  samples ( $5.25 \times 10^8 / (85 \times 85 \times 20/0.3)$ ) located in each distance bin. The atom distances are ignored when we are counting the total samples for each nucleotide pair type (one of  $85 \times 85$  types). There are averagely 39 030 samples for each nucleotide pair type. Thus, both the average number of samples in each bin ( $\sim 1090$ ) and the average number of samples in each nucleotide pair type ( $\sim 39\ 030$ )





**Figure 2.** The frequency distributions of seven RNA torsion angles ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$  for the nucleotide backbone, and  $\chi$  for the base).

are large enough to avoid the sparse data problem. Furthermore, we do not use the statistics in the 0–3 Å regions where sparse data most likely appear. Instead, we utilize a penalty to substitute it. That is to say, once the distance between two atoms is <3 Å, it will give a penalty to the total energy score. In physical sense, the penalty corresponds to the Van der Waals exclusive force between two atoms. These analyses show that 3dRNAscore has no need of using the method proposed by Sippl to address the issue of sparse data.

### Dihedral-dependent energy

We used not only distance-dependent potential, but also a dihedral-dependent potential, involving seven RNA dihedral angles ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$  for the nucleotide backbone, and  $\chi$  for the base). First, we calculated their statistical distribution over the training set (Figure 2).

Once we get their statistical distributions, just like the distance-dependent potential, it can be assumed that all of these angles obey Boltzmann statistical distribution, and once again the mean force potential (26) is used:

$$\Delta G_i(\theta_a^i) = -k_B T \ln \frac{f_a^{\text{OBS}}(\theta_a^i)}{f_a^{\text{REF}}(\theta_a^i)} \quad (9)$$

where  $T$  is the absolute temperature that is set to 298 K,  $k_B$  is the Boltzmann's constant,  $f_a^{\text{OBS}}(\theta_a^i)$  is the observed probability of angle  $i$  (one of  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$  and  $\chi$ ) of  $\theta$  degree (0–360) and type  $a$  in the database of experimental RNA structures and  $f_a^{\text{REF}}(\theta_a^i)$  is the expected probability of arbitrary angle  $i$  of  $\theta$  degree and type  $a$  in reference state.

Then, we could also get a total score  $\Delta G_{\text{torsion}}$  in the same way as the distance-dependent potential energy:

$$f_i^{\text{OBS}}(\theta) = \frac{N_i(\theta)}{\sum_{\theta} N_i(\theta)} = \frac{N_i(\theta)}{N_i} \quad (10)$$

$$f_i^{\text{REF}}(\theta) = \frac{\sum_i N_i(\theta)}{\sum_i \sum_{\theta} N_i(\theta)} = \frac{N(\theta)}{N} \quad (11)$$

Just like the distance-dependent energy, we also use a matrix to store all the dihedral distribution information. Each row denotes a dihedral-pair type and each column the counts of occurrences of the dihedral value in the corresponding bin with a bin width of 4.5° according to eq.(7). So the matrix has 7 × 7 rows and 80 columns.

### Combination of the two energy terms

In 3dRNAscore the two energy terms are combined together to get the final total energy:

$$\Delta G_{\text{total}} = \Delta G_{\text{distance}} + \omega \Delta G_{\text{dihedral}} \quad (12)$$

where  $\Delta G_{\text{total}}$  is the total energy,  $\Delta G_{\text{distance}}$  is the distance-dependent energy,  $\Delta G_{\text{dihedral}}$  is the dihedral-dependent energy, and  $\omega$  is the weight.

To get an appropriate value for  $\omega$ , we adopt a statistical optimization method, which is based on the coarse-grained force field parameterization protocol in recent work by Leonarski *et al.* (32). First, we picked out four RNAs of different types from the training set. These four RNAs (PDB ID: 28SP, 119X, 1KPZ, 1J1U) are representative. 28SP is the structure of the most conserved internal loop in SRP RNA. It includes a hairpin loop and an internal loop, and the length is 28nt. 119X is the structure of a model branchpoint-U2 snRNA duplex. It is a duplex structure containing two bulge. It consists of 26 nucleotides. 1KPZ is the structure of a luteoviral P1–P2 frameshifting mRNA pseudoknot with the length of 33nt. 1J1U is a tRNA containing a multi-branch loop and it comprises 77 nucleotides. We use 3dRNA (12) to predict 1000 models for each RNA. We then calculated each model's DI and the energy score. After that, we use the gradient descent method to maximize the ES. The final optimized  $\omega$  is 3.68.

### Training set

We have trained 3dRNAscore over the RNA 3D Hub non-redundant RNA set. RNA 3D Hub hosts non-redundant set of RNA-containing 3D structures extracted from experimental RNA structures according to the methodology described in Chapter 13 in ref. (33). Leontis' group summarized two categories of structural redundancy existed in PDB/NDB database (34,35), which are redundancy within

a given PDB file and redundancy in PDB/NDB database. They figured out ways to get rid of these redundancies, and then clustered all structures into 749 classes (Release 1.32, 2013–10–12) based on redundancies between structures. About how they get rid of these redundancies, please see Chapter 13 in (33). For each class, they chose a structure to represent it and assigned a unique and stable id to it. The dataset can be accessed at <http://rna.bgsu.edu/nrlist/oldsite.html>.

To construct the training set, we first gathered all the representative structures of 749 classes of RNA 3D Hub non-redundant RNA set. RNA 3D Hub ensured that sequence identity between any two sequences is <95%. By the way, RASP also used a sequence identity of 95% and the coverage of 80% to reduce redundancy in training set. RNA KB potential used a lower sequence identity of 80% to delete homologies in training set. We then used the blastn (36) program to discard all the RNAs with sequence identity >80% and coverage greater than 80%. After that, we carried out 3D structure alignment by the ARTS program (37) to discard all the RNAs where the coverage of aligned part was >80%. However, there exist recurrent motifs, e.g. the sarcin motifs and the GNRA motifs, in RNA 3D structures, and they often have quite different sequences and are hardly found by sequence alignment. These recurrent motifs may also affect the training of a scoring function and should remove this kind of redundancy. Based on RNA 3D Motif Atlas (38), which is a collection of RNA 3D motifs, we can find out the motifs contained in each structure in the training set. Then, RNAs having the motifs also contained in other RNAs in training set or in test sets were removed from the training set. These steps ensure that structures in the training set and test sets have no sequence and structural overlap. Finally, structures having low quality (resolution > 3.5 Å) are removed. Thus, there are 317 structures remained in the training set, which do not share any structures or motifs that are homologous to those in the test sets and have high structure quality (resolution < 3.5 Å).

### Test sets

We tested our knowledge-based potential using three different decoy sets available at present. Test set I is a randstr decoy set (17), which is generated by MODELLER (39) with a set of Gaussian restraints for dihedral angles and atom distances from 85 native structures. This is the largest decoy dataset in the benchmark. It can be downloaded from <http://melolab.org/supmat.html>.

Test set II consists of the decoys built by Bernauer group (18) and FARNA decoys (11). The former is generated by position restrained dynamics, REMD simulation and normal-mode perturbation method (40,41). In the REMD simulation, 1ns REMD simulations are performed for each RNA structures. The temperature is roughly distributed from 285 to 592 K for 50 different temperatures. In the normal-mode perturbation approach, the structures possess stereochemically correct bond lengths and angles but without correct contacts (42). These methods can generate decoy sets with RMSD ranging from 0 to 10 Å. It can be downloaded from <http://csb.stanford.edu/rna>. The FARNA

decoys used in this study consist of lots of near-native tertiary models (11).

Test set III is the FARFAR decoy set which was generated by RNA modeling with ROSETTA-3.1 (11). The FARFAR decoy dataset consists of five lowest energy clusters of tertiary structures with non-canonical base pairs for each of 32 motifs. The FARNA and FARFAR decoys can be downloaded from <http://www.stanford.edu/~rhiju/data.html>.

### Metrics of measuring RNA structures

To compare any two RNA structures quantitatively, we should make use of some metrics to assess their tertiary structures. The most commonly used metric is RMSD (root mean square deviation). RMSD depicts the global geometry differences between two RNA 3D structures but it is usually difficult to describe the hydrogen bond networks of RNA molecules. Hence, some metrics accounting for hydrogen-bonding networks intramolecular in RNA have been proposed to assess RNA structures. One of the commonly used metrics specifically devised for RNA is DI (deformation index) proposed by Parisien (43). The DI is defined as

$$DI(A, B) = \frac{RMSD(A, B)}{INF(A, B)} \quad (13)$$

where RMSD reflects the geometry discrepancy and INF reflects the topology discrepancy.  $INF(A, B)$  is the interaction network fidelity between two structures A and B. Base-pairing and base-stacking are the two major interactions in RNA. These two interactions constitute the interaction network of RNA. Suppose  $S_r$  is the set of interactions in reference structure and  $S_m$  is the set of interactions in modeled structure, then the true positives TP is defined as  $TP = S_r \cap S_m$ , the false positives FP is defined as  $FP = S_m - S_r$ , and the false negatives FN is defined as  $FN = S_r - S_m$ . In Parisien's paper, INF is defined as MCC, that is to say

$$INF(A, B) = MCC(A, B) \quad (14)$$

where MCC is the Matthews correlation coefficient. They used the approximate definition of MCC (44,45):

$$MCC \approx \sqrt{PPV \times STY} \quad (15)$$

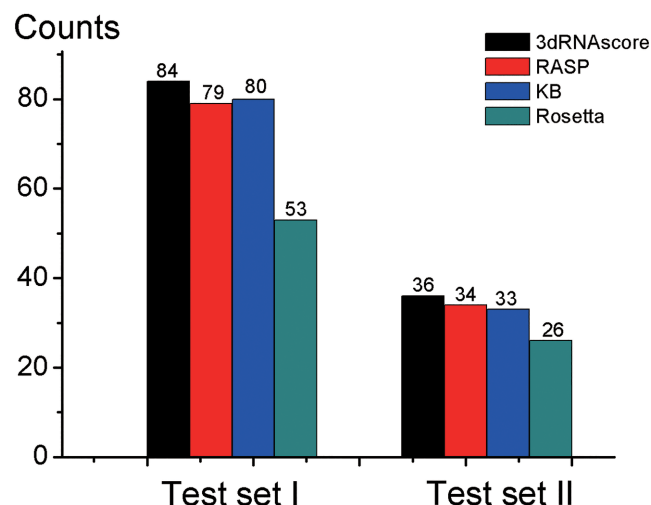
where  $PPV(\text{precision}) = \frac{TP}{TP+FP}$  and  $STY(\text{sensitivity}) = \frac{TP}{TP+FN}$ . For comparison, here we also use this approximation.

In this work, we use both RMSD and DI to measure how well a RNA model recapitulates the corresponding experimental structure in the benchmark of the performance of 3dRNAscore and other scoring functions.

## RESULTS

### Identifying native RNA structures

An important function of scoring methods is to identify the native-like tertiary structure of target RNA in a pool of structural decoys correctly. To compare such ability of 3dRNAscore with three existing RNA knowledge-based potentials: RASP (17), KB (18) and Rosetta (FARFAR) (6), for the sake of fairness, we use all-heavy-atom representation



**Figure 3.** Counts of native states identified correctly from test set I and test set II by 3dRNA score, RASP, KB and Rosetta, respectively.

for all the potentials and evaluate them over the same decoy sets.

Two tests are done on test set I and test set II, respectively (see Figure 3). When using 3dRNA score, 84 out of 85 native structures are identified in test set I, and 36 out of 39 in test set II. For RASP, 79 out of 85 in test set I and 34 out of 39 in test set II. KB potentials could identify 80 out of 85 native structures in test set I, and 33 out of 39 in test set II. For Rosetta, the former is 53 out of 85 and the latter is 26 out of 39. These results show that 3dRNA score has a better performance than other three methods on identifying native RNA structures.

It is noted that Capriotti's group has verified that RASP has a better performance than NAST (46) and the molecular force field energy function: AMBER pseudo-energies (47).

### Ranking near-native RNA structures

Another significant function of scoring function is to rank near-native structures reasonably. In the RNA structure prediction, a natural question is how to affirm that a predicted structure is closer than others to the native state. For RNA, a structure could be geometrically and topologically close to another structure. Hence, both RMSD and DI are used in this work to evaluate structure discrepancy. Mentioned earlier, RMSD measures similarity of two RNA structures from the aspect of geometry and DI is from both geometry and topology.

To describe the performance of a scoring function, the ES (enrichment score (18,48)) is employed here and it is defined as

$$ES = \frac{|E_{\text{top}10\%} \cap R_{\text{top}10\%}|}{0.1 \times 0.1 \times N_{\text{decoys}}} \quad (16)$$

where  $E_{\text{top}10\%}$  is the number of structures with energies (scores given by scoring function) in the lowest 10% of the energy range. For RMSD-based ES,  $R_{\text{top}10\%}$  is the number of structures with RMSD in the lowest 10%. For DI-based

ES,  $R_{\text{top}10\%}$  is the number of structures with DI in the lowest 10%.  $|E_{\text{top}10\%} \cap R_{\text{top}10\%}|$  is the intersection of  $E_{\text{top}10\%}$  and  $R_{\text{top}10\%}$ . If the relationship between the scores and RMSD or DI is completely linear, then ES is equal to 10. If the relationship is random, ES is equal to 1, so

$$ES = \begin{cases} 10, & \text{perfect scoring} \\ 1, & \text{perfectly random} \\ < 1, & \text{bad scoring} \end{cases} \quad (17)$$

We have benchmarked the performance of 3dRNA score, RASP, KB and Rosetta in ranking near-native structures on test set II, using both RMSD and DI metrics. Table 2 shows that when we are using ES of DI, 3dRNA score ( $ES = 4.5$ ) outperforms other three scoring methods (RASP( $ES = 3.8$ ), KB( $ES = 3.7$ ) and Rosetta( $ES = 2.7$ )) on the overall average level. It also outperforms them on the REMD decoys, normal mode decoys and FARNAs decoys of test set II. The result is the same when the test set II is divided into two parts: NMR and X-ray. When we are using ES of RMSD, the result is similar. These results suggest that 3dRNA score is better than other methods when it's used to rank near-native structures. A more detailed table than Table 2 and energy-RMSD (DI, INF) plots are provided in the supplementary data.

### Selecting correct structure containing non-canonical base pairs.

Non-canonical base pairs occur frequently in RNA tertiary folds and functional motifs (49–51). The selection of near-native conformations containing non-canonical base pairs is an inevitable and intractable problem in current RNA 3D structure prediction (6). Test set III is used to test the ability of 3dRNA score in selecting correct structures containing non-canonical base pairs. Test set III (the FARFAR decoy set) consists of five lowest energy clusters of tertiary structures with non-canonical base pairs for each of 32 motifs. We calculated the energy score with RASP, KB, ROSETTAMin and 3dRNA score, respectively. The lowest energy models according to each potential are then selected and compared with the lowest DI models. The results are presented in Table 4. 3dRNA score, RASP, KB and Rosetta are able to identify the lowest DI models of 10, 9, 8 and 2 out of the 32 RNA motifs from the FARFAR decoy set, respectively (see Figure 4 (A)). 3dRNA score can pick out models with lower DI than those by RASP, KB and Rosetta for 8, 12 and 21 out of 32 motifs and models with DI identical to those by RASP, KB and Rosetta for 22, 16 and 1 out of 32 motifs, respectively (see Figure 4 (B)). Moreover, 3dRNA can pick out models with DI greater than those by RASP, KB and Rosetta for 2, 4 and 10 out of 32 motifs, respectively.

### Contribution of the dihedral-dependent potential

Most statistical potentials for RNA structures are based on distance distribution of intermolecular atom pairs or residue pairs. 3dRNA score incorporated an additional energy contribution based on the distribution of backbone dihedrals (torsion angles) in the light of molecular force field. This will be discussed in the 'Discussion' section.

**Table 2.** Comparison of the performance of 3dRNA score, KB potential and Rosetta methods in test set II

Decoy	RNA	Length	Method	Enrichment score (RMSD)				Enrichment score (DI)			
				3dRNA score	KB	RASP	Rosetta	3dRNA score	KB	RASP	Rosetta
Position restrained dynamics and REMD (A)	1duq	26	X-ray	8.5	7.6	7.6	7.1	8.3	7.5	7.6	7.0
	1f27	30	X-ray	8.3	7.9	6.6	6.2	8.1	7.8	6.6	6.2
	1msy	27	X-ray	7.5	5.7	5.7	3.6	7.6	6.0	5.6	3.5
	1nuj	24	X-ray	7.7	7.3	5.2	6.9	7.4	7.2	5.2	6.7
	434d	14	X-ray	8.0	7.7	7.0	6.8	7.6	7.7	6.9	6.8
Normal modes (B)	1duq	26	X-ray	7.5	7.0	5.7	3.8	6.9	7.0	5.7	3.5
	1esy	19	NMR	4.9	5.4	4.5	5.6	6.5	5.5	4.7	5.7
	1f27	30	X-ray	5.9	5.8	3.7	2.6	6.7	5.8	3.7	2.5
	1i9v	76	X-ray	6.1	2.6	5.3	3.0	5.5	2.7	5.1	3.0
	1kka	17	NMR	5.7	4.6	4.1	4.6	6.9	4.3	4.1	4.8
	1msy	27	X-ray	5.9	5.6	2.2	4.6	6.7	5.4	2.7	4.5
	1nuj	24	X-ray	7.1	7.4	5.9	2.4	4.5	7.0	5.9	2.2
	1qwa	21	NMR	3.5	3.2	2.0	3.8	3.7	3.3	2.2	3.9
	1x9k	62	X-ray	6.7	1.6	5.2	3.0	3.1	1.9	5.2	2.8
	1xjr	46	X-ray	7.7	5.4	7.9	2.2	6.2	5.0	7.9	2.5
FARNA (C)	1ykq	19	X-ray	4.6	3.4	3.5	2.8	5.0	3.3	3.8	2.5
	1zih	12	NMR	7.7	5.4	5.7	6.6	6.9	5.3	5.7	6.5
	28sp	28	NMR	5.7	4.0	6.5	1.8	5.7	4.5	6.7	2.9
	2f88	34	NMR	6.8	5.4	4.9	4.4	4.1	5.4	4.7	4.2
	434d	14	X-ray	7.7	7.4	7.4	5.2	6.9	7.4	7.6	5.2
	1a4d	41	NMR	2.4	3.8	2.0	0.8	2.1	4.0	2.0	0.7
	1esl	28	X-ray	2.0	1.5	1.6	1.3	2.4	1.4	1.6	1.4
	1dqf	19	X-ray	4.2	1.8	2.8	1.0	3.6	2.0	3.2	1.0
	1esy	19	NMR	3.2	3.7	4.8	1.2	3.8	3.5	4.2	1.1
	1i9x	26	X-ray	3.4	1.3	2.2	1.5	4.0	1.5	3.6	1.6
Average values	1j6s	24	X-ray	0.2	1.4	0.2	0.6	1.6	1.7	0.8	0.5
	1kd5	22	X-ray	3.2	0.3	0.8	0.2	2.4	0.5	1.6	0.3
	1kka	17	NMR	1.4	1.2	0.6	0.6	2.0	0.8	0.8	0.6
	1l2x	27	X-ray	0.6	3.2	1.0	1.8	1.5	3.2	1.2	1.8
	1mhk	32	X-ray	1.6	1.2	1.6	1.0	1.4	1.2	1.4	1.2
	1q9a	27	X-ray	2.6	0.5	0.8	0.8	3.2	0.5	1.2	0.8
	1qwa	21	NMR	2.2	1.2	0.4	1.0	1.4	1.3	0.6	1.0
	1xjr	46	X-ray	2.4	2.0	2.4	1.2	3.2	1.9	3.4	1.0
	1zih	12	NMR	4.8	5.0	4.8	2	6.8	5.5	7.2	1.9
	255d	24	X-ray	2.4	0.7	0.6	1.3	2.0	0.7	0.6	1.1
Average values	283d	24	X-ray	1.4	0.8	0.8	0.7	1.8	0.8	1.0	0.7
	28sp	28	NMR	2.8	1.5	3.0	1.7	3.8	1.2	4.2	1.8
	2a43	26	X-ray	2.2	2.0	1.4	0.6	3.2	1.8	2.0	0.6
	2f88	34	NMR	3.2	1.3	2.2	1.3	3.6	1.3	1.6	1.2
	(A)			8.0	7.2	6.4	6.1	7.8	7.2	6.4	6.0
Average values	(B)			6.2	4.9	5.0	3.8	5.7	4.9	5.0	3.8
	(C)			2.3	1.8	1.8	1.1	2.8	1.8	2.2	1.1
	X-ray			4.8	3.8	3.7	2.8	4.7	3.8	3.9	2.7
	NMR			4.2	3.5	2.7	2.7	4.4	3.5	3.7	2.8
	All			4.5	3.7	3.6	2.7	4.6	3.7	3.8	2.7

To verify the contribution of the dihedral-based energy, we tested the performance of single distance-dependent energy, single dihedral-dependent energy and combined total energy according to Equation (12). The results are shown in Table 3. On average, single distance-dependent energy performs better than single dihedral-dependent energy but in some case (5 out of 40) the later is better than the former, especially four of the five (1a4d, 1j6s, 1l2x, and 28sp) is for FARNA dataset that consists of many near-native tertiary models (11) and they are just the cases where single distance-dependent energy performs worth than the best results of other methods (see Table 2). Furthermore, the average ES given by single distance-based energy is 4.37 while that given by the combined energy is 4.57. These results indicate that including dihedral-dependent energy can further improves accuracy of the distance-dependent energy on average.

### Relationship between 3dRNA score and physical interactions in RNA

Capriotti and coworkers have found that all-atom version of RASP, RASP-ALL, could well capture base-pairing and base-stacking interactions in RNA. We observed the same phenomenon using 3dRNA score. It seems that base-pairing and base-stacking interactions are implicitly contained in all-atom distance-dependent statistical potential. For example, Figure 5 depicts the distance distribution of the atom pair between N9 of adenine and N1 of uracil. Three apparent peaks appear on the distance distribution. The first peak is at the distance of 4.65 Å, and it stems from the base-stacking interaction between adjacent residues along the nucleotide residues chain. The second peak is at the distance of 7.05 Å, and it's from the indirect interaction between the  $i$ th residue and the  $(i + 2)$ th residue. The third peak is at the distance of 8.7 Å, and it results from the base-pairing interaction between adenine and uracil.



**Table 3.** Comparison of performance of distance-dependent energy, dihedral-dependent energy, and total energy of 3dRNAscore in test set II

Decoy	RNA	Length	Method	Enrichment score (DI)		
				Distance term	Dihedral term	Total
Position restrained dynamics and REMD (A)	1duq	26	X-ray	8.2	7.5	8.3
	1f27	30	X-ray	7.9	5.7	8.1
	1msy	27	X-ray	7.4	6.7	7.6
	1nuj	24	X-ray	7.3	7.5	7.4
	434d	14	X-ray	7.5	7.0	7.6
Normal modes (B)	1duq	26	X-ray	6.3	4.8	6.9
	1esy	19	NMR	6.5	5.2	6.5
	1f27	30	X-ray	6.7	4.3	6.7
	1i9v	76	X-ray	5.5	4.6	5.5
	1kka	17	NMR	6.9	2.6	6.9
	1msy	27	X-ray	6.3	4.7	6.7
	1nuj	24	X-ray	4.5	1.6	4.5
	1qwa	21	NMR	3.5	2.2	3.7
	1×9k	62	X-ray	3.1	2.5	3.1
	1xjr	46	X-ray	5.8	4.8	6.2
	1ykq	19	X-ray	5.2	5.0	5.0
	1zih	12	NMR	5.9	3.4	6.9
	28sp	28	NMR	6.1	5.4	5.7
	2f88	34	NMR	3.9	3.2	4.1
	434d	14	X-ray	6.5	5.2	6.9
	1a4d	41	NMR	1.8	2.4	2.1
	1csl	28	X-ray	2.2	1.8	2.4
	1dqf	19	X-ray	3.6	2.0	3.6
	1esy	19	NMR	3.8	2.4	3.8
FARNA (C)	1i9x	26	X-ray	4.2	1.6	4.0
	1j6s	24	X-ray	0.6	2.2	1.6
	1kd5	22	X-ray	2.4	1.6	2.4
	1kka	17	NMR	1.7	0.8	2.0
	1l2x	27	X-ray	0.4	1.4	1.5
	1mhk	32	X-ray	1.2	0.6	1.4
	1q9a	27	X-ray	3.0	1.6	3.2
	1qwa	21	NMR	1.2	0.6	1.4
	1xjr	46	X-ray	3.0	1.8	3.2
	1zih	12	NMR	6.8	2.4	6.8
	255d	24	X-ray	1.8	0.8	2.0
	283d	24	X-ray	1.7	0.6	1.8
	28sp	28	NMR	3.6	4.2	3.8
	2a43	26	X-ray	2.8	1.4	3.2
	2f88	34	NMR	3.5	2.1	3.6
Average values	(A)			7.66	6.88	7.80
	(B)			5.51	3.97	5.69
	(C)			2.59	1.70	2.83
	X-ray			4.43	3.43	4.65
	NMR			4.25	2.84	4.41
	All			4.37	3.24	4.57

Figure 6 shows the base-stacking and base-pairing energies analyzed by 3dRNAscore in RNA 1AFX, which has a hairpin 3D structure. Figure 6(B) suggests that stacking energies between the bases from the fourth to eighth nucleotides are lower than the other base-stacking energies, which just meet the 2D structure: (((...))). Figure 6(C) suggests that the base-pairing energies between the nucleotides 1–12, 2–11, 3–10 and 4–9 are the four lowest energies, which exactly represent the four Watson–Crick base-pairing in 1AFX.

### Choosing a structure model in a robust way

Evaluation function is used to select proper structure models. Here, we give an example of how to choose a structure model in a robust way during RNA tertiary structure prediction. We firstly use the RNA 3D structure building program 3dRNA (12) to predict 1000 models for a duplex

RNA (PDB id: 1FQZ). The 1000 models can be clustered into groups based on their INF values and with a threshold. Table 5 and Figure 7 are results for a threshold of 0.18. Each group is then scored by 3dRNAscore. The structure with the lowest energy score is selected as the representative structure of its group. All of these representative structures constitute the candidates of the native structure.

Table 5 lists the seven largest groups. Figure 7 shows that Group I includes the model with the lowest DI. Although the overall energy scores of Group II are lower than Group I, the model with the lowest DI in Group I can still be picked up through clustering. Furthermore, most groups exhibit a funnel shape after clustering, which is conducive to 3dRNAscore that the model with the lowest energy score is just the one with the minimum DI.



**Table 4.** The ranking results of Rosetta, RASP-ALL, KB and 3dRNA score methods on the FARFAR decoy set

Motif name	DI of the lowest energy model				Minimum DI model
	ROSETTAmin	RASP-ALL	RNA KB	3dRNA score	
G-A base pair	4.026	1.408	2.455	1.408	1.408
Fragment with G/G and G/A pairs, SRP helix VI	13.910	6.073	6.073	6.073	5.267
Helix with A/C base pairs	2.414	3.262	5.410	5.866	2.414
Four-way junction, HCV IRES	21.236	21.236	21.526	20.067	15.040
Loop 8, A-type Ribonuclease P	8.421	4.290	4.290	4.290	1.615
Helix with U/C base pairs	6.012	3.136	3.136	3.136	3.136
Curved helix with G/A and A/A base pairs	1.062	1.929	3.565	1.929	0.998
Pre-catalytic conformation, hammerhead ribozyme	30.097	12.830	12.830	12.830	12.830
Loop E motif, 5S RNA	2.404	2.561	2.561	2.561	1.986
UUCG tetraloop	1.284	1.278	1.403	1.278	1.278
Rev response element high affinity site	7.802	5.826	4.227	5.826	4.227
Fragment with A/C pairs, SRP helix VI	3.930	2.424	2.424	2.424	2.424
Signal recognition particle Domain IV	5.516	3.224	3.224	3.224	1.338
Bulged G motif, sarcin/ricin loop	1.902	7.066	7.838	7.066	1.588
Tertiary interaction, hammerhead ribozyme	25.215	24.647	24.647	24.647	19.453
GAGA tetraloop from sarcin/ricin loop	1.075	0.956	0.956	0.956	0.956
Pentaloop from conserved region of SARS genome	4.147	4.145	4.147	4.145	1.068
L2/L3 tertiary interaction, purine riboswitch	16.371	19.160	19.160	19.160	16.371
L3, thiamine pyrophosphate riboswitch	6.076	2.262	2.262	2.262	2.262
Kink-turn motif from SAM-I riboswitch	1.592	13.086	12.107	12.107	1.454
Active site, hammerhead ribozyme	20.394	23.413	22.927	22.927	17.947
P1/L3, SAM-II riboswitch	14.885	18.103	10.478	18.174	10.478
J4/5 from P4-P6 domain, Tetrahymena ribozyme	2.504	2.919	2.919	2.919	2.199
Stem C internal loop, L1 ligase	3.479	3.479	3.226	3.479	3.226
J5/5a hinge, P4-P6 domain, Tetr. ribozyme	25.063	23.920	28.689	23.920	20.910
Three-way junction, purine riboswitch	13.693	13.734	10.616	10.616	10.387
J4a/4b region, metal-sensing riboswitch	6.237	5.012	6.549	5.012	5.012
Kink-turn motif	15.0735	23.997	28.729	16.805	11.695
Tetraloop/helix interaction, L1 ligase crystal	1.398	0.948	1.398	0.948	0.948
Hook-turn motif	7.598	16.240	15.354	1.985	1.985
Tetraloop/receptor, P4-P6 domain, Tetr. ribozyme	16.803	11.112	4.477	4.477	3.818
Pseudoknot, domain III, CPV IRES	10.097	7.063	7.771	4.377	4.274

**Table 5.** Seven largest groups clustered from 1000 models predicted by 3dRNA for 1FQZ

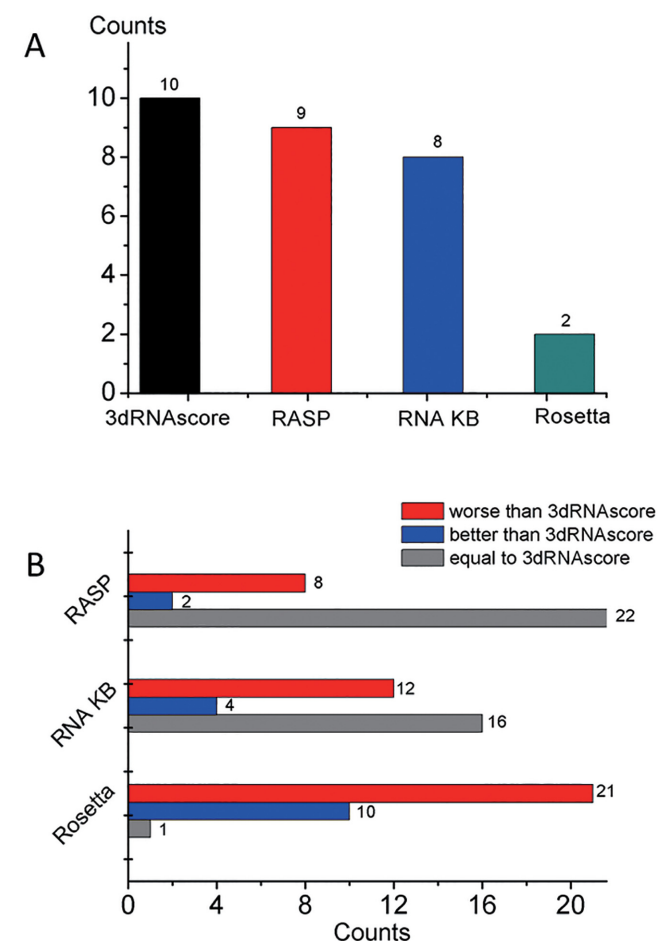
	Group I	Group II	Group III	Group IV	Group V	Group VI	Group VII	Other
Size	190	39	56	40	19	38	21	597
Center DI	3.738	5.234	18.727	4.632	23.136	4.173	6.724	13.741
Minimum DI	1.429	4.373	13.270	2.713	9.758	2.437	5.129	2.612
DI with minimum energy	1.429	4.487	16.132	2.713	26.532	2.437	5.384	9.537

## DISCUSSION

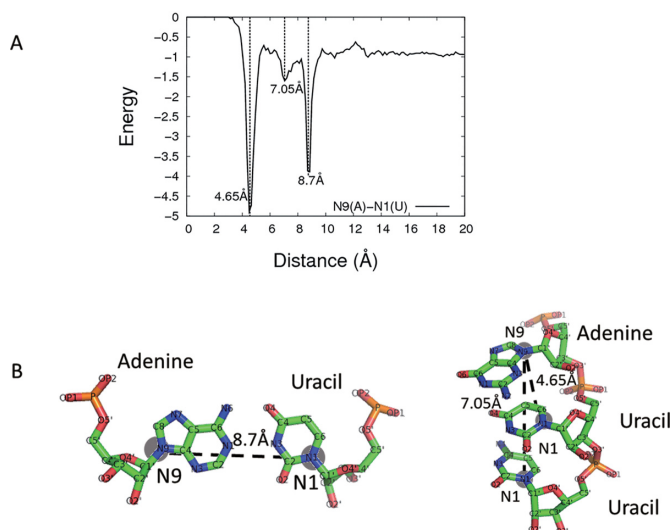
The native structure tends to have the lowest free energy according to the thermodynamic hypothesis proposed by Anfinsen (28). So precisely to say, we need to use free energy to evaluate a structure. For classical molecular force fields, it is easy to calculate the enthalpy of a structure but it is very time consuming to calculate its entropy. In contrast to force fields, statistical potentials extracted from experimen-

tal data of known RNA structures include both enthalpy and entropy information. Although they are not directly equal to free energy, they in principle correlate with the latter. Furthermore, the calculation is easy and fast. So, statistical potentials are more often used for RNA and protein structure scoring than molecular force fields.

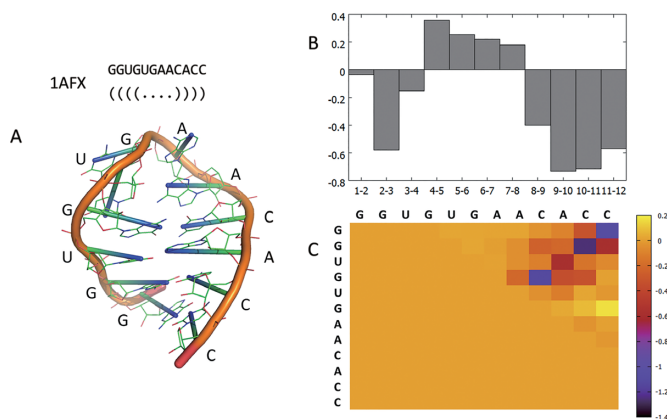
A significant difference between 3dRNA score and other RNA statistical potentials is the combination of the conventional distance-dependent energy with a dihedral-



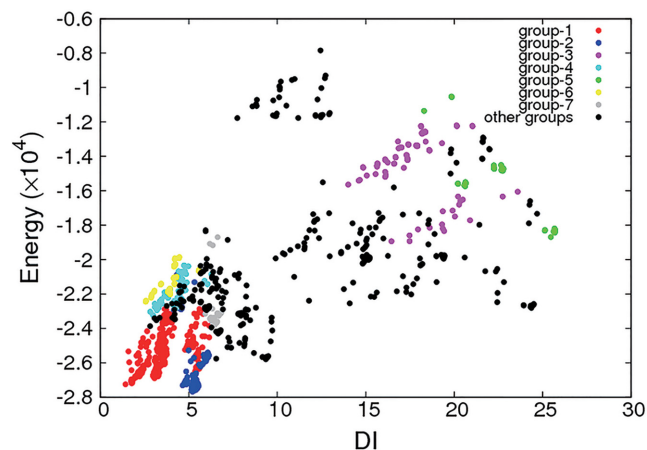
**Figure 4.** Results on test III. (A) The counts of models having the lowest energy selected by 3dRNA score, RASP, RNA KB and Rosetta, respectively. (B) The red part represents the counts of models that are worse than those selected by 3dRNA score. Blue represents the counts of models that are better than those selected by 3dRNA score. Gray represents the counts of models that are the same as those selected by 3dRNA score.



**Figure 5.** (A) Energy distribution of the distance between N9 of adenine and N1 of uracil. (B) Diagram of the three representative distance between N9 of adenine and N1 of uracil.



**Figure 6.** (A) Sequence, secondary structure and 3D structure of 1AFX. (B) Base-stacking energies between adjacent two nucleotides in 1AFX calculated by 3dRNA score. '1-2' means base-stacking energy between the first and the second nucleotide, '2-3' means the second to the third ... and so on. The lower the energy, the better the base-stacking. (C) Base-pairing energies between each possible base-pair in 1AFX calculated by 3dRNA score. The lower the energy, the better the base-pairing.



**Figure 7.** Energy versus DI plot of the groups of the 1000 prediction models of 1FQZ.

dependent energy. Thus, in principle, 3dRNA score can take account of both overall shape and base pairs. The results above indicate a better performance of this combined potential. It is known that molecular force fields are the basis of molecular dynamics simulation to study conformations of biomolecules. Many force fields in use today consist of four components: bond stretching, angle bending, the rotation of bonds and non-bonded interactions. The distance-dependent statistical potential corresponds to non-bonded interactions part of a force field while the dihedral-dependent potential corresponds to bond rotation part. The latter can efficiently describe the flexibility of RNA molecules, which is one of the major features of RNA structures that are different from that of protein structures. Originally, we planned to design four different statistical potentials corresponding to the four components of a force field. However, we found that the potentials corresponding to bond stretching and angle bending have no significant effect because all of the decoy sets don't involve any varia-

tion of bond length and bond angle in the process of their generations.

RNA backbone is rotameric (52,53), and this may help recognize native-like models. Murray and coworkers (52) processed the backbone torsion angle distributions of an 8636-residues RNA database with quality-filtering techniques like resolution, crystallographic B factor and all-atom steric clashes. With noise levels greatly reduced, clear signal appears for the underlying angle preferences. It suggests that native-like models have obvious preference for certain specific backbone torsion angle distribution. The dihedral-dependent energy of 3dRNAscore utilizes this preference to identify the native-like model from decoys where the dihedral distributions of RNA structures deviate from the normal dihedral distributions.

The Rosetta RNA scoring function (FARFAR) (6) is more detailed and precise than most other RNA knowledge-based potentials. The RNA energy function used in FARFAR includes a term weakly favoring compactness (proportional to radius-of-gyration), a term to penalize steric clashes within molecules and other terms that are specially designed for RNA interactions (6,11). On account of these precise energy terms contained in FARFAR, its scoring performance depends largely on the quality of the decoy sets. The less good scoring performance of FARFAR than that of other three scoring methods in the test above may be attributed to its unfitness to these decoys sets.

Test set I is generated by Gaussians on distances and torsions and Test set II is generated by molecular dynamics in a relative short period of time, hence decoys in these two datasets have low diversity and are distributed around a local minimum free energy state (in these two sets, the local minimum free energy state is just the native state) in the sense of free energy landscape. Results in these two sets show that 3dRNAscore is quite qualified for identifying local minimum free energy state and ranking structures around the local minimum. Test set III is a more real-world decoy set which is generated by RNA modelling with FARFAR (6), so decoys in test set III have a very high diversity, which is owed to the Monte Carlo algorithm adopted by FARFAR. Results in test set III show that 3dRNAscore's performance in ranking structures widely distributed in free energy landscape is not so good, but still better than other existing statistical potentials.

Other limitations exist for knowledge-based statistical potentials at present. The limited number of RNA tertiary structures of the training set is still the major problem for developing knowledge-based potentials (54–56). Almost all the knowledge-based potentials face the possibility of over-training problem because the parameters depend on the limited number of structures in training datasets (57). We have tried to use a non-redundant RNA tertiary structure set as the training set and then remove all the structures that are similar to certain structures in test sets. More non-redundant structures would improve the accuracy of knowledge-based potentials (58,59). Furthermore, RNAs are very sensitive to electrostatic interactions because of the negative charges in the phosphate groups of the backbone (60–62). RNA may be unable to form the functional folds in the absence of positive ions. The knowledge-based potentials only implicitly consider this effect by counting the

experimental structures. We will study these problems in future.

## CONCLUSION

In this paper, we have developed a novel RNA knowledge-based potential for identifying native RNA structures and ranking predicted structures and functional structural motifs with non-canonical base pairs. We used a non-redundant RNA training set to train the parameters by combining the distances of paired atoms and torsion angles to construct our statistical potential. The benchmark tests show that our method could identify not only appropriate tertiary folds, but also tertiary motifs with the non-canonical base pairs. Although some limitations, 3dRNAscore performs consistently better than existing methods in both two cases.

## AVAILABILITY

<http://biophy.hust.edu.cn/download.html>.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGMENT

Computations presented in this paper were carried out using the High Performance Computing Center experimental testbed in SCTS/CGCL.

## FUNDING

National High Technology Research and Development Program of China [2012AA020402]; National Science Foundation of China (NSFC) [11374113]. Funding for open access charge: National High Technology Research and Development Program of China [2012AA020402]; NSFC [11374113].

*Conflict of interest statement.* None declared.

## REFERENCE

- Guttman, M. and Rinn, J.L. (2012) Modular regulatory principles of large non-coding RNAs. *Nature*, **482**, 339–346.
- Montange, R.K. and Batey, R.T. (2008) Riboswitches: emerging themes in RNA structure and function. *Annu. Rev. Biophys.*, **37**, 117–133.
- Dethoff, E.A., Chugh, J., Mustoe, A.M. and Al-Hashimi, H.M. (2012) Functional complexity and regulation through RNA dynamics. *Nature*, **482**, 322–330.
- Laing, C. and Schlick, T. (2011) Computational approaches to RNA structure prediction, analysis, and design. *Curr. Opin. Struct. Biol.*, **21**, 306–318.
- Cao, S. and Chen, S.J. (2011) Physics-based de novo prediction of RNA 3D structures. *J. Phys. Chem. B*, **115**, 4216–4226.
- Das, R., Karanicolas, J. and Baker, D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods*, **7**, 291–294.
- Frellsen, J., Moltke, I., Thiim, M., Mardia, K.V., Ferkinghoff-Borg, J. and Hamelryck, T. (2009) A probabilistic model of RNA conformational space. *PLoS Comput. Biol.*, **5**, e1000406.
- Sharma, S., Ding, F. and Dokholyan, N.V. (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, **24**, 1951–1952.

9. Parisien, M. and Major, F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
10. Martinez, H.M. Jr, Maizel, J.V. and Shapiro, B.A. (2008) RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA. *J. Biomol. Struct. Dyn.*, **25**, 669–683.
11. Das, R. and Baker, D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 14664–14669.
12. Zhao, Y., Huang, Y., Gong, Z., Wang, Y., Man, J. and Xiao, Y. (2012) Automated and fast building of three-dimensional RNA structures. *Sci. Rep.*, **2**, 734.
13. Zhao, Y., Gong, Z. and Xiao, Y. (2011) Improvements of the hierarchical approach for predicting RNA tertiary structure. *J. Biomol. Struct. Dyn.*, **28**, 815–826.
14. Zhang, Y. (2008) Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.*, **18**, 342–348.
15. Zhou, H. and Skolnick, J. (2011) GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.*, **101**, 2043–2052.
16. Adhikari, A.N., Freed, K.F. and Sosnick, T.R. (2013) Simplified protein models: Predicting folding pathways and structure using amino acid sequences. *Phys. Rev. Lett.*, **111**, 28103.
17. Capriotti, E., Norambuena, T., Marti-Renom, M.A. and Melo, F. (2011) All-atom knowledge-based potential for RNA structure prediction and assessment. *Bioinformatics*, **27**, 1086–1093.
18. Bernauer, J., Huang, X., Sim, A.Y.L. and Levitt, M. (2011) Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA*, **17**, 1066–1075.
19. Norambuena, T., Cares, J.F., Capriotti, E. and Melo, F. (2013) WebRASP: a server for computing energy scores to assess the accuracy and stability of RNA 3D structures. *Bioinformatics*, **29**, 2649–2650.
20. Sim, A.Y.L., Schwander, O., Levitt, M. and Bernauer, J. (2012) evaluating mixture models for building rna knowledge-based potentials. *J. Bioinform. Comput. Biol.*, **10**, 1241010.
21. Olechnovic, K. and Venclovas, C. (2014) The use of interatomic contact areas to quantify discrepancies between RNA 3D models and reference structures. *Nucleic Acids Res.*, **42**, 5407–5415.
22. Melo, F., Nchez, R. and Sali, A. (2002) Statistical potentials for fold assessment. *Protein Sci.*, **11**, 430–448.
23. Neal, R.M. (2000) Markov chain sampling methods for Dirichlet process mixture models. *J. Comp. Graph. Stat.*, **9**, 249–265.
24. Holbrook, S.R. (2008) Structural Principles From Large RNAs. *Annu. Rev. Biophys.*, **37**, 445–464.
25. Reiter, N.J., Chan, C.W. and Mondragón, A. (2011) Emerging structural themes in large RNA molecules. *Curr. Opin. Struct. Biol.*, **21**, 319–326.
26. Sippl, M.J. (1990) Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, **213**, 859–883.
27. Samudrala, R. and Moulton, J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.*, **275**, 895–916.
28. Anfinsen, C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.
29. Lu, H. and Skolnick, J. (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins: Struct. Funct. Bioinform.*, **44**, 223–232.
30. Rykunov, D. and Fiser, A.A.S. (2007) Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. *Proteins: Struct. Funct. Bioinform.*, **67**, 559–568.
31. Scott, D.W. (1979) On optimal and data-based histograms. *Biometrika*, **66**, 605–610.
32. Leonarski, F., Trovato, F., Tozzini, V., Les, A. and Trylska, J. (2013) Evolutionary algorithm in the optimization of a coarse-grained force field. *J. Chem. Theory Comput.*, **9**, 4874–4889.
33. Leontis, N.B.Z.C. (2012) *Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking* Springer, Berlin, Heidelberg **27**, 281–298.
34. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank, *Nucleic Acids Res.*, **28**, 235–242.
35. Berman, H.M., Olson, W.K., Beveridge, D.L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S., Srinivasan, A.R. and Schneider, B. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751.
36. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
37. Dror, O., Nussinov, R. and Wolfson, H. (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, **21**, i47–i53.
38. Petrov, A.I., Zirbel, C.L. and Leontis, N.B. (2013) Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA*, **19**, 1327–1340.
39. Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
40. Huang, E.S., Subbiah, S., Tsai, J. and Levitt, M. (1996) Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. *J. Mol. Biol.*, **257**, 716–725.
41. Hansmann, U.H. and Okamoto, Y. (1999) New Monte Carlo algorithms for protein folding. *Curr. Opin. Struct. Biol.*, **9**, 177–183.
42. Summa, C.M. and Levitt, M. (2007) Near-native structure refinement using in vacuo energy minimization. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 3177–3182.
43. Parisien, M., Cruz, J.A., Westhof, E. and Major, F. (2009) New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, **15**, 1875–1885.
44. Gorodkin, J., Stricklin, S.L. and Stormo, G.D. (2001) Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res.*, **29**, 2135–2144.
45. Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA-Protein Struct. Mol.*, **405**, 442–451.
46. Jonikas, M.A., Radmer, R.J., Laederach, A., Das, R., Pearlman, S., Herschlag, D. and Altman, R.B. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, **15**, 189–199.
47. Wang, J., Cieplak, P. and Kollman, P.A. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, **21**, 1049–1074.
48. Tsai, J., Bonneau, R., Morozov, A.V., Kuhlman, B., Rohl, C.A. and Baker, D. (2003) An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins*, **53**, 76–87.
49. He, S., Hickman, A.B., Dyda, F., Johnson, N.P., Chandler, M. and Ton-Hoang, B. (2011) Reconstitution of a functional IS608 single-strand transpososome: role of non-canonical base pairing. *Nucleic Acids Res.*, **39**, 8503–8512.
50. Orioli, A., Pascali, C., Quartararo, J., Diebel, K.W., Praz, V., Romascano, D., Percudani, R., van Dyk, L.F., Hernandez, N., Teichmann, M. et al. (2011) Widespread occurrence of non-canonical transcription termination by human RNA polymerase III. *Nucleic Acids Res.*, **39**, 5499–5512.
51. Betel, D., Koppal, A., Agius, P., Sander, C. and Leslie, C. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.*, **11**, R90.
52. Murray, L.J.W., Arendall, W.B., Richardson, D.C. and Richardson, J.S. (2011) RNA backbone is rotameric. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 13904–13909.
53. Humphris-Narayanan, E. and Pyle, A.M. (2012) Discrete RNA libraries from pseudo-torsional space. *J. Mol. Biol.*, **421**, 6–26.
54. Ngan, S., Hung, L., Liu, T. and Samudrala, R. (2008) *Protein Structure Prediction*. Springer, pp. 243–281.
55. Berman, H.M., Bhat, T.N., Bourne, P.E., Feng, Z., Gilliland, G., Weissig, H. and Westbrook, J. (2000) The Protein Data Bank and the challenge of structural genomics. *Nat. Struct. Mol. Biol.*, **7**, 957–959.
56. Westbrook, J., Feng, Z., Chen, L., Yang, H. and Berman, H.M. (2003) The protein data bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
57. Zheng, S., Robertson, T.A. and Varani, G. (2007) A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. *FEBS J.*, **274**, 6378–6391.
58. Handl, J., Knowles, J. and Lovell, S.C. (2009) Artefacts and biases affecting the evaluation of scoring functions on decoy sets for protein structure prediction. *Bioinformatics*, **25**, 1271–1279.



59. Li, Y., Rata, I. and Jakobsson, E. (2011) Sampling multiple scoring functions can improve protein loop structure prediction accuracy. *J. Chem. Inf. Model.*, **51**, 1656–1666.
60. Chen, S. (2008) RNA folding: conformational statistics, folding kinetics, and ion electrostatics. *Annu. Rev. Biophys.*, **37**, 197.
61. Chin, K., Sharp, K.A., Honig, B. and Pyle, A.M. (1999) Calculating the electrostatic properties of RNA provides new insights into molecular interactions and function. *Nat. Struct. Mol. Biol.*, **6**, 1055–1061.
62. Draper, D.E. (2013) Folding of RNA tertiary structure: linkages between backbone phosphates, ions, and water. *Biopolymers*, **99**, 1105–1113.