ISSN 1674-1056
http://cpb.iphy.ac.cn
http://iopscience.iop.org/cpb
中国物理 B
Chinese Physics
中国物理 Chinese Physics B

A novel knowledge-based potential for RNA 3D structure evaluation

Yi Yang(杨毅), Qi Gu(辜琦), Ben-Gong Zhang(张本龚), Ya-Zhou Shi(时亚洲), Zhi-Gang Shao(邵志刚)
Citation: Chin. Phys. B . 2018, 27(3): 038701. **doi:** 10.1088/1674-1056/27/3/038701

Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Wavelet optimization for applying continuous wavelet transform to maternal electrocardiogram component enhancing

Qiong Yu(于琼), Qun Guan(管群), Ping Li(李萍), Tie-Bing Liu(刘铁兵), Jun-Feng Si(司峻峰), Ying Zhao(肇莹), Hong-Xing Liu(刘红星), Yuan-Qing Wang(王元庆)
Chin. Phys. B . 2017, 26(11): 118702. **doi:** 10.1088/1674-1056/26/11/118702

Novel Fourier-based iterative reconstruction for sparse fan projection using alternating direction total variation minimization

Zhao Jin(金朝), Han-Ming Zhang(张瀚铭), Bin Yan(闫镔), Lei Li(李磊), Lin-Yuan Wang(王林元), Ai-Long Cai(蔡爱龙)
Chin. Phys. B . 2016, 25(3): 038701. **doi:** 10.1088/1674-1056/25/3/038701

Flexibility of nucleic acids: From DNA to RNA

Lei Bao(鲍磊), Xi Zhang(张曦), Lei Jin(金雷), Zhi-Jie Tan(谭志杰)
Chin. Phys. B . 2016, 25(1): 018703. **doi:** 10.1088/1674-1056/25/1/018703

Knowledge-based potentials in bioinformatics: From a physicist's viewpoint

Zheng Wei-Mou
Chin. Phys. B . 2015, 24(12): 128701. **doi:** 10.1088/1674-1056/24/12/128701

Theoretical studies on sRNA-mediated regulation in bacteria

Chang Xiao-Xue, Xu Liu-Fang, Shi Hua-Lin
Chin. Phys. B . 2015, 24(12): 128703. **doi:** 10.1088/1674-1056/24/12/128703

A novel knowledge-based potential for RNA 3D structure evaluation*

Yi Yang(杨毅)[1,2], Qi Gu(辜琦)[2], Ben-Gong Zhang(张本龚)[3], Ya-Zhou Shi(时亚洲)[2,3,†], and Zhi-Gang Shao(邵志刚)[1,‡]

[1] *Guangdong Provincial Key Laboratory of Quantum Engineering and Quantum Materials, SPTE, South China Normal University, Guangzhou* 510006, *China*

[2] *Center for Theoretical Physics and Key Laboratory of Artificial Micro & Nano-structures of Ministry of Education, School of Physics and Technology, Wuhan University, Wuhan* 430072, *China*

[3] *Research Center of Nonlinear Science, School of Mathematics and Computer Science, Wuhan Textile University, Wuhan* 430200, *China*

(Received 18 October 2017; revised manuscript received 5 December 2017; published online 13 February 2018)

Ribonucleic acids (RNAs) play a vital role in biology, and knowledge of their three-dimensional (3D) structure is required to understand their biological functions. Recently structural prediction methods have been developed to address this issue, but a series of RNA 3D structures are generally predicted by most existing methods. Therefore, the evaluation of the predicted structures is generally indispensable. Although several methods have been proposed to assess RNA 3D structures, the existing methods are not precise enough. In this work, a new all-atom knowledge-based potential is developed for more accurately evaluating RNA 3D structures. The potential not only includes local and nonlocal interactions but also fully considers the specificity of each RNA by introducing a retraining mechanism. Based on extensive test sets generated from independent methods, the proposed potential correctly distinguished the native state and ranked near-native conformations to effectively select the best. Furthermore, the proposed potential precisely captured RNA structural features such as base-stacking and base-pairing. Comparisons with existing potential methods show that the proposed potential is very reliable and accurate in RNA 3D structure evaluation.

Keywords: RNA, 3D structure evaluation, knowledge-based potential

PACS: 87.14.gn, 87.15.bg, 87.10.Vg **DOI:** 10.1088/1674-1056/27/3/038701

1. Introduction

In living systems, RNAs perform essential roles including transmission of genetic information, regulation of gene expression, and catalysis of biochemical reactions.[1–3] Understanding and utilizing these functions require comprehensive knowledge of RNA 3D structures. However, obtaining high-resolution RNA 3D structures through experimental methods such as x-ray crystallography or NMR is a challenging task.[4] Some alternative computational methods for RNA 3D structure prediction have been developed.[5–20] However, for a given RNA sequence, the existing methods usually generate a set of possible near-native conformations rather a best one.[5,7,8,14,17,21] For example, the recent coarse-grained model presented by Tan's group can predict a serial of 3D structures (or decoys) for a 25-nt RNA hairpin with a mean RMSD of 2.5 Å in spite of the minimum 1.0-Å RMSD.[13] Therefore, the selection of the best structures from an ensemble of conformations is a vital but challenging task.[5,22] Meanwhile, it has occurred in the evaluation of protein tertiary structure[23–25] and protein–RNA/DNA complex.[26,27]

In recent years, several potentials have been proposed for RNA 3D structure evaluation; e.g., the nucleic acid simulation tool (NAST),[28] the Rosetta,[8,29] the ribonucleic acids statistical potential (RASP),[30,31] the RNA coarse-grained and all-atom knowledge-based (RNA KB) potentials,[32] and 3dRNAscore.[33] The NAST can generate, cluster, and rank RNA tertiary structures using an RNA-specific knowledge-based potential in a coarse-grained molecular dynamics (MD) engine; however, it needs a secondary structure, tertiary contact information, and some experimental data. Instead, Rosetta, also known as FARNA/FARFAR, is successful in predicting and evaluating RNA 3D structures just from sequences, but only applies to small RNAs. RASP is another full-atom potential that explicitly embraces base-pairing and base-stacking interactions and is able to discriminate conformations between near-native and misfolded RNA conformations including even several non-canonical base pairs. However, this potential is derived from only 85 RNA structures and the precision could be further improved by involving more diverse RNA structures.[34] RNA KB potential, which is derived from the diverse RNAs of six different families, effectively assesses types of RNA conformations and can be used as a force field in molecular dynamics. Nonetheless, the above potentials only include interactions between two atoms in different nucleotides and could ignore the fluctuations of local structures.[35,36] Very recently, a novel full-atom statis-

*Project supported by the National Science Foundation of China (Grants Nos. 11605125, 11105054, 11274124, and 11401448).

†Corresponding author. E-mail: yzshi@wtu.edu.cn

‡Corresponding author. E-mail: zgshao@scnu.edu.cn

© 2018 Chinese Physical Society and IOP Publishing Ltd

tical potential 3dRNAscore developed by Xiao's group performed better than the above potentials in identifying RNA native structures from a pool of decoys by combing the distance-dependent energies with a new dihedral-dependent energy. However, the 3dRNA score as well as the other existing potentials consider only the universality of the limited experimental structures in a training set[37] and ignore the distinctions among the isolation of individuals, which can be significant in RNA folding.[38,39]

Here, we propose a new all-atom knowledge-based potential to accurately assess RNA tertiary structures. The proposed potential has the following features: (i) it is derived from a set of 380 non-redundant RNA structures that embraces most of the common RNAs and motifs, (ii) it considers the local interactions between two atoms within one nucleotide by adding a new energy term that efficiently represents the local geometrical features of RNA structures and depicts the flexibility of RNAs,[40,41] and (iii) introduces a retraining process for each RNA to identify the individual variation of different RNAs where the potential can be optimized based on several low-energy conformations from initial scoring. Moreover, the proposed potential is validated by different test sets widely used to analyze the performance of knowledge-based potential in recent works. Our results show that the proposed potential effectively evaluates RNA 3D structures and selects the native structure from a pool of RNA conformations, ranking near-native structures to identify the best ones, and capturing base-base interactions. Furthermore, the proposed potential is compared to existing methods such as RASP, RNA KB, 3dRNAscore, and Rosetta to show its feasibility in the assessment of RNA 3D structures.

2. Methods and materials

2.1. Knowledge-based potential

The thermodynamic hypothesis proposed by Anfinsen is that the native state tends to have the lowest free energy.[42] To evaluate an RNA 3D structure, the proposed potential considers two energy terms. The total energy u_{total} of a structure is given by

$$u_{\text{total}} = u_{\text{inter}} + \varepsilon u_{\text{intra}}, \tag{1}$$

where u_{inter} is the energy between two atoms of different nucleotides, and u_{intra} is the energy between two atoms in one nucleotide, which can preclude the rigid bond stretching and angle bending interactions. The ε in Eq. (1) is the weight of u_{intra} with respect to u_{inter}.

The energies between the two atoms in Eq. (1) are knowledge-based potentials derived from a Boltzmann distribution or Bayesian formulations,[43] and the energy between two atoms with type i and type j is

$$u_{i,j}(r) = -RT \ln \left[\frac{N_{i,j}^{\text{OBS}}(r) N_{\text{total}}^{\text{OBS}}}{N^{\text{OBS}}(r) N_{i,j}^{\text{OBS}}} \right], \tag{2}$$

where R and T are the Boltzmann constant and the temperature in Kelvin, respectively. $N_{i,j}^{\text{OBS}}(r)$ is the observed number of an atom pair (i,j) at a distance r in experimental RNA structures in PDB (http://www.rcsb.org/pdb/home/home.do). $N^{\text{OBS}}(r)$ is the count of observed contact between all pairs of atom types at a distance r. $N_{i,j}^{\text{OBS}}$ is the number of atom pairs of type i and type j in the entire distance region. $N_{\text{total}}^{\text{OBS}}$ is the total number of contacts between all pairs of atom types summed over the entire distance r, which is the total count. For u_{inter}, atoms i and j are in different nucleotides, while for u_{intra} they are located in the same nucleotide. The proposed potential utilizes all 85 atom types in the four nucleotides: 22 atom types in adenine, 20 atom types in cytosine, 23 atom types in guanine, and 20 atom types in uracil (Fig. S1). More detailed information regarding the proposed potential can be found in the Supplementary Material.

2.2. Determination of parameters

The parameters of the proposed potential were calculated based on the statistical analysis of known RNA 3D structures in the PDB/NDB database. First, we gathered 1369 non-redundant structures including most of the common RNAs (e.g., mRNA, rRNA, tRNA, ribozyme, and riboswitch) and various RNA complexes where only RNAs were taken into account. Second, we discarded the structures with sequence identities greater than 80% using the BLASTN program with default options[44] and removed low-quality structures with a resolution ≥ 3.5 Å. Finally, 380 structures with only normal heavy atoms containing various typical motifs (e.g., hairpin/internal/bulge loops, double/triple helices, junctions, and non-Waston–Crick base pairs) were selected as the training set and the PDB ID for each of selected structures is listed in Table S1. Based on the careful selection of the training set, the distance-dependent parameters of the potential can be obtained as a matrix with a bin width of 0.3 Å and a cut-off distance of 20 Å as used in similar recent potentials.[30,33]

In structure ranking, the initial potential obtained above can be further optimized for decoys of an RNA by a retraining mechanism (Fig. S5). In the retraining process of each RNA, based on the energies calculated by the initial potential, structures with the lowest energies are used as a new training set to further obtain an RNA-specific potential, and energies of all conformations can be calculated again by the retrained potential (Fig. S5(a)). Note that the number of conformations with low energies is generally taken as 10, where the potential has the highest precision (Fig. S5(b)).

2.3. Test sets

To test the performance of the proposed potential, we used two different decoy sets widely used in recent works to analyze the performance of knowledge-based potential.[8,29–33] Test set I, which has been used by RASP and 3dRNAscore, consisted of decoys generated from 85 native structures (only with normal heavy atoms) by MODELLER[45] with a set of Gaussian restraints for dihedral angles and bond stretches and can be downloaded from http://melolab.org/supmat/RNApot/Home.html. Test set II contained three parts which were a set of decoys generated by three separate methods recently used to assess the scoring functions of KB, Rosetta, and 3dRNAscore. The first part of Test set II contained five distinct RNA decoy sets generated by MD by using a position restraint potential on each heavy atom to constrain the motions of the RNA, named MD decoys.[46,47] The second part of Test set II included decoys of 15 RNAs generated by the normal-mode (NM) perturbation method,[48] named NM decoys. These two decoys can be downloaded from http://csb.stanford.edu/rna/. The third part of Test set II, named FARNA decoys, included 3D structures of 17 RNAs with many non-native base pairs blindly predicted by Rosetta 3.1 from sequences downloadable from https://daslab.stanfold.edu/resources. It should be noted that a very small portion of tested RNAs existed in the training set (Sup. Table 1) that could not affect the reliability of the proposed potential.

2.4. Enrichment score

Generally, the enrichment score (ES) is often used to describe the performance of a scoring function in identifying the best RNA structures.[32,49] The ES is defined as

$$\mathrm{ES} = \frac{|E_{\mathrm{top10\%}} \bigcap R_{\mathrm{top10\%}}|}{0.1 \times 0.1 \times N_{\mathrm{decoys}}}, \tag{3}$$

where $E_{\mathrm{top10\%}}$ is the number of structures with energies in the lowest 10% of the energy range. Since the root mean square deviation (RMSD) depicts the global geometry differences between two structures and the deformation index (DI) is particularly designed for describing hydrogen-bonding networks in RNA (see more detail in Supplementary Material), $R_{\mathrm{top10\%}}$ is the number of structures with an RMSD or DI in the lowest 10% for RMSD-based or DI-based ESs. $|E_{\mathrm{top10\%}} \cap R_{\mathrm{top10\%}}|$ is the intersection of $E_{\mathrm{top10\%}}$ and $R_{\mathrm{top10\%}}$. If the relationship between the energy scores and RMSD or DI is completely linear, ES equals 10, suggesting a perfect score. If the relationship is random, ES equals 1, corresponding to a poor score.

2.5. Optimization of parameter ε

Owing to the different sample sizes for the two potentials in Eq. (1), the coupling parameter ε is extremely important in the combination of the two energy terms.[50] To obtain a proper value of ε, we selected five RNAs (PDB ID: 1KKA, 1MSY, 1ZIH, 255D, and 437D) with different sizes and representative motifs including hairpins, bulges, pseudoknots, and junctions. For each of the five RNAs, we used the MC-Fold/MC-Sym pipeline to generate a series of 3D structures[7] that were scored by the proposed potential with different values of ε. As shown in Fig. S2, when ε is equal to 0.15, the calculated average ES (RMSD) of the five decoy sets was the maximum value. Therefore, an ε value of 0.15 was used for all the tested RNAs in this study.

3. Results and discussion

The proposed knowledge-based potential can be used to select the native state and rank near-native structures to identify the best ones to capture RNA structural features. Compared with existing potentials, the proposed potential performs much better in RNA 3D structures assessment.

3.1. Selecting native structure from decoys

To test the performance of the proposed potential, we first use our potential to select the native structure from decoys based on the thermodynamic hypothesis.[42] For decoys in Test sets I and II and the corresponding native structures, we use the proposed potential with initial parameters to calculate the energy for each conformation and select the conformation with the lowest energy as a native structure. As shown in Fig. 1 and Figs. S3 and S4, the proposed potential accurately selected the native structures from even the poor decoys. For example, our potential identified 83 native structures out of 85 RNAs in Test set I and 34 native structures from 39 RNAs in Test set II.

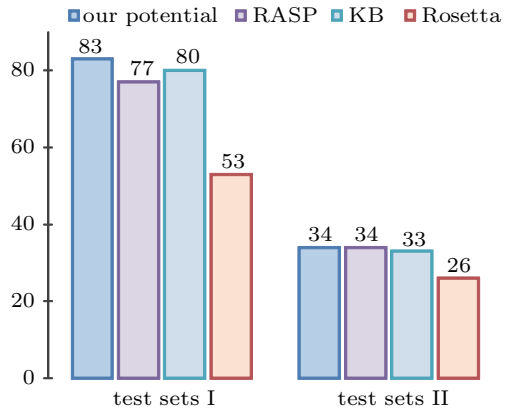


Fig. 1. (color online) Counts of native states identified correctly from Test sets I and II by RASP, KB, Rosetta, and our proposed potential.

We also compared the proposed potential against well-established existing potentials such as RASP, KB, and Rosetta in identifying the native structures from different decoys. We employed the RASP, KB, and Rosetta potentials with corresponding programs or software to select native structures for

the two test sets, and the results are shown in Fig. 1. For Test set I, 77, 80, and 53 out of 85 native structures were identified by the three potentials, respectively, which are all less than that identified by our potential. Although the proposed potential was as good as the RASP for Test set II in selecting native structures, it performed slightly better than the other two potentials.

It is noted that several native structures, such as 1ESY, 1KKA, and 1QWA in the test sets, could not be correctly selected by either the proposed potential or the other existing potentials as shown in Fig. 1 and Figs. S3 and S4. One possible reason could be that these structures have their own special features, such as complex ligand binding or structure stretching.[51–53] In addition, the number of RNAs in the training set was limited and resulted in over-training.[34]

3.2. Ranking near-native RNA structures and selecting the best conformations

The identification of the best structures is a vital task in RNA tertiary structure prediction.[8,13] For example, although a recent new method using evolutionary restraints of RNAs was proposed to predict RNA 3D structures, it still needs to recognize the best conformations from the predicted structures using a knowledge-based potential such as 3dRNAscore.[21] However, it is still challenging and essential for the existing predictive methods to accurately recognize the best conformations from a series of predicted conformations. Thus, the precise identification of the best RNA conformations from a set of near-native structures is an essential purpose for the scoring function. To assess the quality of our potential, we employed the proposed potential to discriminate and rank the near-native conformations of RNA.

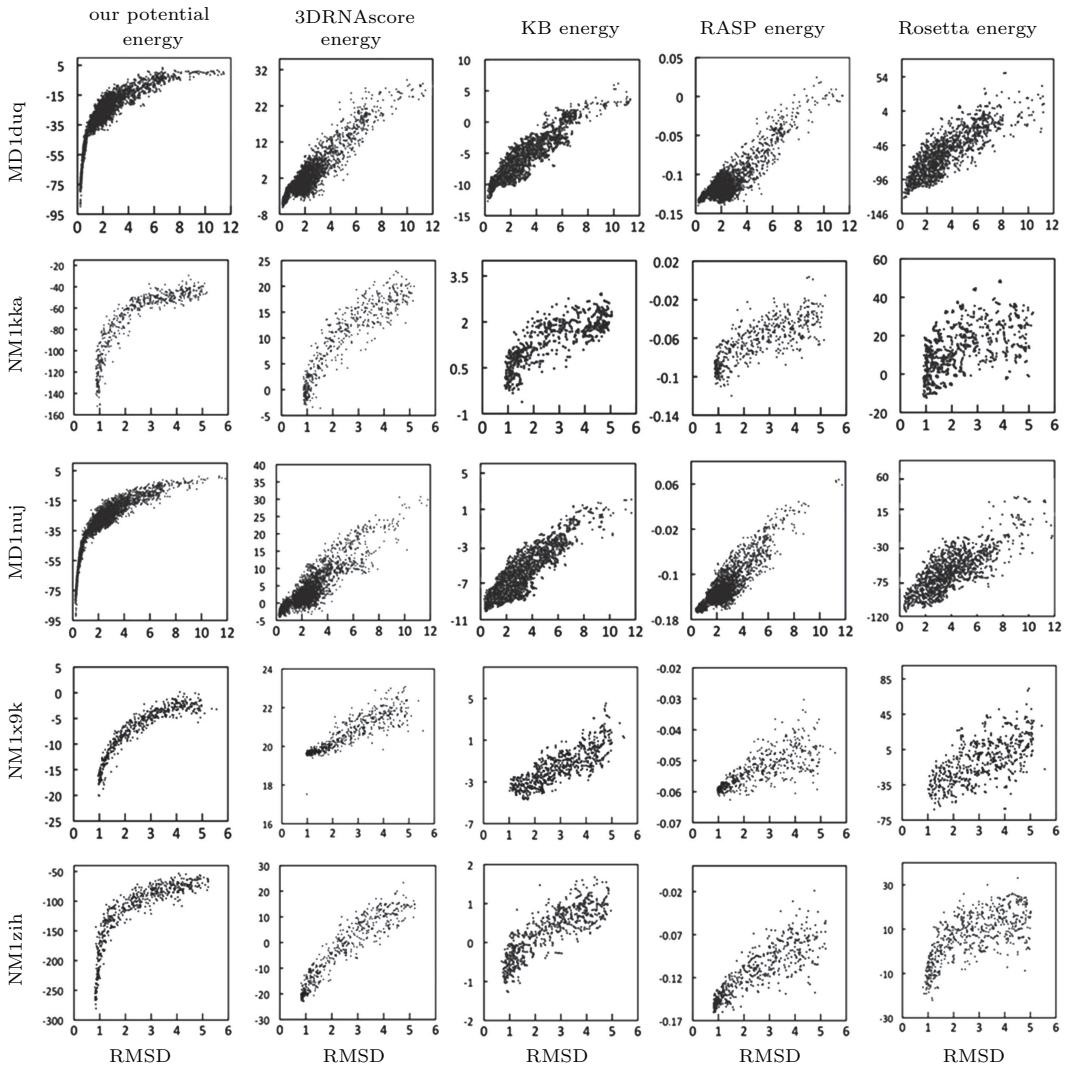


Fig. 2. (color online) Energy as a function of RMSD for five typical MD/NM decoys (rows). The proposed potential, 3dRNAscore, KB, RASP, and Rosetta are shown in turns (columns).

For each RNA in Test set II, we first calculated the energies of all decoys excluding the native structure by using the proposed potential[30,33] and ranked the decoys based on energy (Fig. S5). Then, selecting 10 decoys with the lowest energies as a new training set, the potential was trained again to consider the characteristics of the RNA[37,39] This retraining

process step proved to be effective for improving the accuracy of RNA structure evaluation. Finally, we further ranked the decoys of the RNA again based on the energies scored by the retrained potential (Figs. S6 and S7) and assumed that the lower the energy, the better the structure.

As shown in Fig. 2 and Figs. S6 and S7, for each set of decoys from different RNAs, the energy scores generally declined with a decrease of RMSD (or DI) and such a reduction became steeper near small values of RMSD, which suggests that the proposed potential was very effective at ranking near-native structures and selecting the best.

In addition, we benchmarked the performance of 3dRNAscore, RASP, KB, Rosetta, and the proposed potential in selecting the best structures on Test set II using both RMSD-based ES and DI-based ES. As shown in Fig. 3 and Table S2, the overall results of Test set II by the proposed potential (ES (RMSD) = 5.6 and ES (DI) = 5.5) were better than those from the KB (ES (RMSD) = 3.7 and ES (DI) = 3.7), RASP (ES (RMSD) = 3.6, and ES (DI) = 3.8), 3dRNAscore (ES (RMSD) = 4.5 and ES (DI) = 4.6), and Rosetta (ES (RMSD) = 2.7 and ES (DI) = 2.7), which is also suggested by the p-values less than 0.01 from the non-parametric and distribution-free Kolmogorov–Smirnov (KS) test (Table S3(a) and S3(b)).

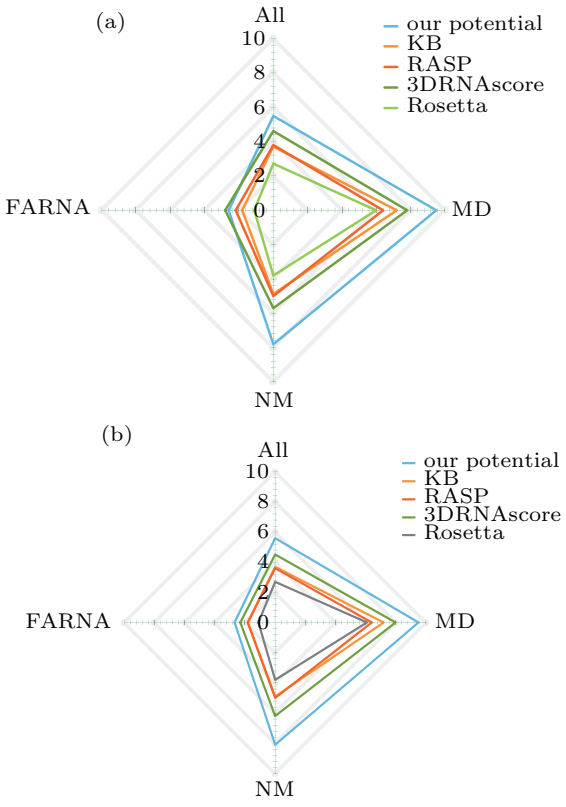


Fig. 3. (color online) Comparison between average values of RMSD-based ES (a) or DI-based ES (b) calculated from Rosetta, 3dRNAscore, RASP, KB, and our potential with retraining mechanism on Test set II: all decoys, FARNA decoys, NM decoys, and MD decoys.

Figure 3 and Table S2 also show that for FARNA decoys, the ES values calculated by the proposed potential (ES (RMSD) = 2.7 and ES (DI) = 2.6) as well as the others (ES $\leq$ 2.8) were smaller than those of the other decoys. This may be attributed to the fact that the FARNA decoy set contained many tertiary structures that included many non-canonical base pairs.[8] The result also suggests that all the existing potentials' evaluation of structures predicted from blind prediction experiments (e.g., RNA-Puzzles) needs to be improved.[5] In future studies, the proposed potential should be further tested in more real RNA prediction cases.

3.3. Captured base-base interactions

Generally, base-pairing and base-stacking interactions provide a strong force in stabilizing RNA 3D structures.[17,54] Thus, the capacity to precisely capture base-pairing/stacking interactions is one of the significant criteria for determining the quality of potentials. We further analyzed the energies between the specific atom pairs involved in base-pairing and base-stacking such as N1 of adenine and N3 of uracil in the proposed potential (Fig. 4(a)) which were deduced from a Boltzmann distribution or Bayesian formulations (Eq. (1)) based on our training set.[42,43] As shown in Fig. 4(a), there were two visible wells: one at 3.0 Å and the other at 5.1 Å, which accurately represent the distances between corresponding atoms in pairing and stacking bases, respectively,[55,56] as shown in Fig. 4(b). We also calculated the energies between any two bases using the proposed potential for RNA (PDB ID: 434D) because RNA has a typical helical structure and represents many RNAs.[8,57] The secondary and 3D structure of RNA are shown in Fig. 4(d). As shown in Fig. 4(c), the lower the energy, the stronger the base-base interaction: (i) energies between the nucleotides 1–14, 2–13, 3–12, 4–11, 6–9, and 5–6 are the lowest energies, which exactly represents the four types of Watson–Crick base pairs in 434D (Fig. 4(d)), and (ii) energies between bases 3–12, 4–11, 6–9, and 5–6 are stronger than those between bases 1–14 and 2–13, which represents an interaction of cytosine and guanine stronger than that of adenine and uracil and is in line with previous works.[8,57] Above all, the proposed potential effectively captured the interaction of base-pairing and base-stacking.

3.4. u_{intra} versus retraining mechanism

Based on the carefully selected training set shown in Figs. 1–3, the proposed potential with the combination of the new energy term u_{intra} and the employment of a retraining mechanism performed better than the existing methods in assessing RNA structures for Test sets I and II.

To clarify the contributions of the two improvements, we further performed two additional tests for each RNA decoy in Test set II using our potential without the u_{intra} or the retraining mechanism (Table sup). As shown in Fig. 5, the two improvements made positive contributions to the overall

improvements, which is also suggested by the p-values less than 0.01 from the non-parametric and distribution-free KS test (Table S5), and the inclusion of the retraining mechanism made a very strong contribution (Fig. 5). As shown in Table S5, the proposed potential with only U_{inter} performed slightly better than the other potentials, which possibly contributed to the success of our training set in including extensive non-redundant RNAs and embracing most motifs. It is noted that although the local interactions U_{intra} introduced into the proposed potential only slightly improved the results (Table S5), it would be more significant in the evaluation of complex RNAs including many local tertiary interactions.[33] Although the addition of the new term slightly improved the performance of our potential in some cases (Table S5), it is still particularly necessary for large RNAs with complex structures because the new term represents the local geometrical features and flexibility of RNA structures[40,41] Overall, the specificity of RNA molecules and the interactions between two atoms within one nucleotide can be necessary.

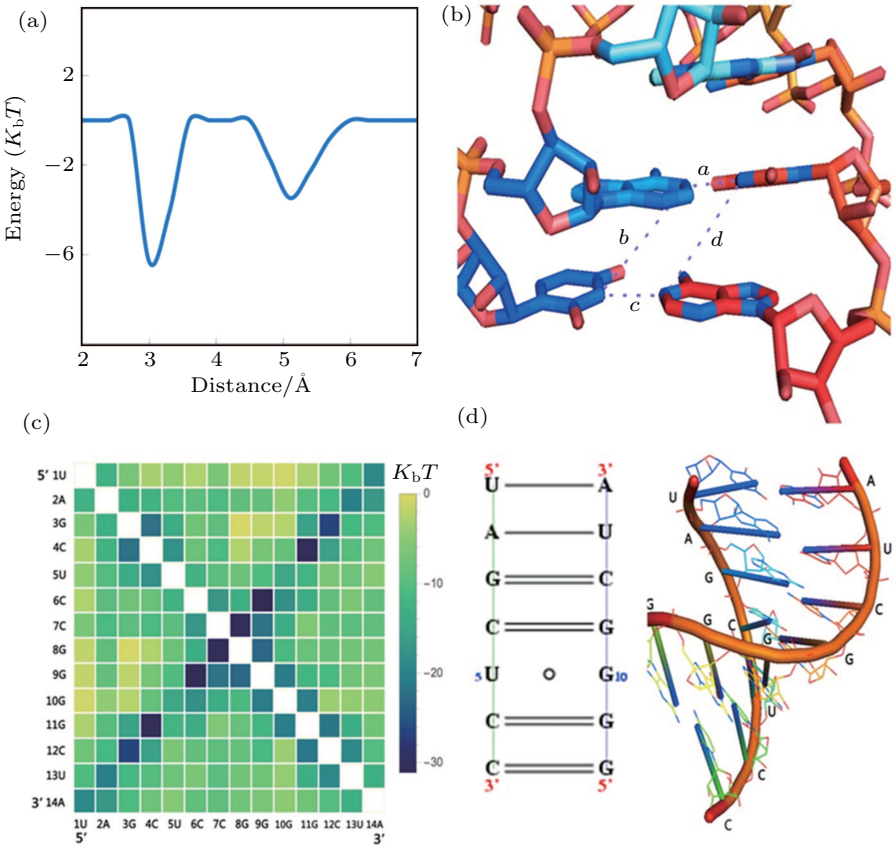


Fig. 4. (color online) (a) Energy distribution of distance between N1 of adenine and N3 of uracil. (b) Diagram of representative distance between N1 of adenine and N3 of uracil. (c) Base-stacking and base-pairing energies between each possible base-pair calculated by the proposed potential for 434D. (d) Secondary structure and 3D structure of 434D.

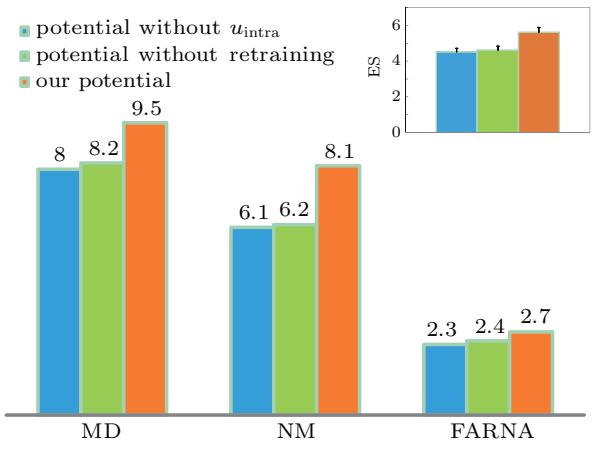


Fig. 5. (color online) Comparison between average values of RMSD-based ES calculated from the proposed potential, the potential without u_{intra}, and the potential without retraining mechanism on Test set II.

However, for some RNAs such as 1DQF, 1I9X, 1J6S, 1KD5, 1KKA, and 1MHK in the FARNA decoy set, the employment of the retraining process can result in a decrease of ES (Table S4). The possible reason is that the retraining process is based on the results generated from the initial scoring; however, for these RNAs with complex tertiary interactions (e.g., non-native hydrogen bonds), the initial scoring obtained by the proposed potential as well as the other existing potentials still needs improvement.

3.5. Reference state

Although the knowledge-based statistical potential has proven to be an efficient method for structural evaluation, the selection of an effective reference state is still an inherent limi-

tation for the knowledge-based scoring function. Selecting diverse reference states to score the same decoys can lead to different results.[58,59] Since an ideal reference state is not achievable, current statistical potentials generally construct reference states through randomizing disconnected atoms; e.g., the averaged reference state in 3dRNAscore and RASP and the quasi-chemical approximation reference state in KB, which ignore the diversity of various RNAs.[60] Therefore, on the basis of choosing the averaged reference state, we further propose a retraining mechanism in the proposed potential to distinguish the characteristic of different RNAs, which has proved to be very effective in assessing RNA 3D structures (Fig. 5).

4. Conclusions

In this work, we developed a novel all-atom knowledge-based potential to evaluate and analyze RNA 3D structures. First, the proposed potential correctly selected native structures from different poor decoys by combining the energies between two atoms within one nucleotide into the conventional statistical potential. Second, the proposed potential was effective in ranking near-native structures for various decoys and then selecting the best ones based on introducing a retraining mechanism to distinguish the characteristics of different RNAs. Third, the benchmark test with the existing potentials on extensive test sets showed that the proposed potential performed better than the others in RNA 3D structure evaluation. Finally, the proposed potential effectively and precisely captured the features of RNA structure, such as Watson–Crick base-pairing and base-stacking.

Despite this success, the proposed potential has some limitations. For example, the averaged reference state used by the proposed potential and other potentials has its limitations.[58] Although our proposed retraining process overcame the insufficiency to some degree, we believe that a unified potential should be proposed by including more detailed energies to solve the problem.[61–63] Furthermore, since RNAs are strongly charged polyanionic chains, there is a strong intra-chain Coulombic repulsion during RNA folding.[54] Therefore, RNA 3D structures can be highly sensitive to ion conditions.[64–68] However, the existing knowledge-based potentials including the proposed potential only implicitly consider this effect by counting the experimental structures and cannot be used to assess RNA structures in different ion conditions. Moreover, although the tests in this work show the reliable availability of our potential, more comprehensive tests on various different decoy sets, such as RNA conformations from RNA-Puzzles,[5] will be performed in the future, and then available software of the proposed potential will be further developed.

Acknowledgment

We are grateful to Professor Zhi-Jie Tan (Wuhan University) and Dr. Haiyou Deng (HZAU) for valuable discussions.

References

[1] Guttman M and Rinn J L 2002 *Nature* **482** 339
[2] Frank T D 2011 *Phys. Lett. A* **375** 3851
[3] Gong S, Wang Y and Zhang W 2015 *J. Chem. Phys.* **143** 045103
[4] Aviv T, Amborski A N, Zhao X S, Kwan J J, Johnson P E and Sicheri F 2006 *J. Mol. Biol.* **356** 274
[5] Cruz JA, Blanchet MF, Boniecki M, Bujnicki JM, Chen S J and Cao S 2012 *RNA* **18** 610
[6] Laing C and Schlick T 2011 *Curr. Opin. Struc. Biol.* **21** 306
[7] Parisien M and Major F 2008 *Nature* **452** 51
[8] Das R and Baker D 2007 *Proc. Natl. Acad. Sci. U.S.A.* **104** 14664
[9] Sharma S, Ding F and Dokholyan NV 2008 *Bioinf.* **24** 1951
[10] Zhou H and Skolnick J 2011 *Biophys. J.* **101** 2043
[11] Cao S and Chen S J 2011 *J. Phys. Chem. B* **115** 4216
[12] Zhao Y, Huang Y, Gong Z, Wang Y, Man J and Xiao Y 2012 *Sci. Rep.* **2** 734
[13] Shi Y Z, Wu Y Y, Wang F H and Tan Z J 2014 *Chin. Phys. B* **23** 078701
[14] Shi Y Z, Wang F H, Wu Y Y and Tan Z J 2014 *J. Chem. Phys.* **141** 105102
[15] Tan Z J, Zhang W, Shi Y Z, and Wang F H 2015 *Adv. Exp. Med. Biol.* **827** 143
[16] Kerpedjiev P, Höner Z S C and Hofacker I L 2015 *RNA* **21** 1110
[17] Shi Y Z, Jin L, Wang F H, Zhu X L and Tan Z J 2015 *Biophys. J.* **109** 2654
[18] Li J, Zhang J, Wang J, Li W and Wang W 2016 *PLoS Comput. Biol.* **12** e1005032
[19] Boniecki M J, Lach G, Dawson W K, Tomala K, Lukasz P and Soltysinski T 2016 *Nucleic Acids Res.* **44** e63
[20] Lorenz R, Wolfinger M T, Tanzer A and Hofacker I L 2016 *Nat. Methods* **103** 86
[21] Wang J, Mao K, Zhao Y, Zeng C, Xiang J and Zhang Y 2017 *Nucleic Acids Res.* **45** 11
[22] Shapiro B A, Yingling Y G, Kasprzak W and Bindewald E 2007 *Curr. Opin. Struct. Biol.* **17** 157
[23] Kim D E, Chivian D and Baker D 2004 *Nucleic Acids Res.* **32** W526.
[24] Shen M and Sali A 2006 *Protein Sci.* **15** 2507
[25] Singh H, Singh S and Raghava G P 2014 *PLOS ONE* **9** 8
[26] Setny P, Bahadur R P and Zacharias M 2012 *Bmc Bioinf.* **13** 228
[27] Pan X and Shen H B 2017 *Bmc Bioinf.* **18** 136
[28] Jonikas M A, Radmer R J, Laederach A, Das R, Pearlman S, Herschlag D and Altman R B 2009 *RNA* **15** 189
[29] Das R, Karanicolas J and Baker D 2010 *Nat. Methods* **7** 291
[30] Capriotti E, Norambuena T, Marti-Renom M A and Melo F 2011 *Bioinf.* **27** 1086
[31] Norambuena T, Cares J F, Capriotti E and Melo F 2013 *Bioinf.* **29** 2649
[32] Bernauer J, Huang X, Sim A Y and Levitt M 2011 *RNA* **17** 1066
[33] Wang J, Zhao Y, Zhu C and Xiao Y 2015 *Nucleic Acids Res.* **43** e63
[34] Zheng S, Robertson T A and Varani G 2007 *FEBS J.* **274** 6378
[35] Holbrook SR 2008 *Annu. Rev. Biophys.* **37** 445
[36] Reiter N J, Chan C W and Mondrag'on A 2011 *Curr. Opin. Struct. Biol.* **21** 319
[37] Zhang C, Liu S, Zhou H and Zhou Y 2004 *Biophys. J.* **86** 3349
[38] Gultyaev A P, van Batenburg F H and Pleij C W 1995 *J. Mol. Biol.* **250** 37
[39] Schroeder R, Barta A and Semrad K 2004 *Nat. Rev. Mol. Cell Bio.* **5** 908
[40] Fagegaltier D, Lescure A, Walczak R, Carbon P and Krol A 2000 *Nucleic Acids Res.* **28** 2679
[41] Kao S C and Bobst A M 1985 *Biochem.* **24** 5465
[42] Anfinsen CB 1973 *Science* **181** 223
[43] Sippl MJ 1990 *J. Mol. Biol.* **213** 859
[44] Altschul S F, Gish W, Miller W, Myers E W and Lipman D J 1990 *J. Mol. Biol.* **215** 403
[45] Sali A and Blundell TL 1993 *J. Mol. Biol.* **234** 779
[46] Huang E S, Subbiah S, Tsai J and Levitt M 1996 *J. Mol. Biol.* **257** 716
[47] Hansmann U H and Okamoto Y 1999 *Curr. Opin. Struct. Biol.* **9** 177
[48] Summa C M and Levitt M 2007 *Proc. Natl. Acad. Sci. U.S.A.* **104** 3177

[49] Tsai J, Bonneau R, Morozov A V, Kuhlman B, Rohl C A and Baker D 2003 *Proteins: Struct. Funct. Genet.* **53** 76
[50] Leonarski F, Trovato F, Tozzini V, Les A and Trylska J 2013 *J. Chem. Theory Comput.* **9** 4874
[51] Amarasinghe G K, Guzman R N D, Turner R B and Summers M F 2000 *J. Mol. Biol.* **299** 145
[52] Cabello V J, Winkler M E and Nikonowicz E P 2002 *J. Mol. Biol.* **319** 1015
[53] Finger LD, Trantirek L, Johansson C and Feigon J 2003 *Nucleic Acids Res.* **31** 6461
[54] Wang Y, Gong S, Wang Z and Zhang W 2016 *J. Chem. Phys.* **144** 115101.
[55] Frappat L, Sorba P and Sciarrino A 1998 *Phys. Lett. A* **250** 214
[56] Morgado C A, Svozil D, Turner D H and Šponer J 2012 *Phys. Chem. Chem. Phys.* **14** 12580
[57] Yang H, Jossinet F, Leontis N, Chen L, Westbrook J and Berman H 2003 *Nucleic Acids Res.* **31** 3450
[58] Deng H, Jia Y, Wei Y and Zhang Y 2012 *Proteins: Struct. Funct. Bioinf.* **80** 2311
[59] Huang S Y and Zou X 2006 *J. Comput. Chem.* **27** 1866
[60] Thomas P D and Dill K A 1996 *J. Mol. Biol.* **257** 457
[61] Wang J and Xiao Y 2016 *Phys. Rev. E* **94** 040401
[62] Zheng W M 2015 *Chin. Phys. B* **24** 22
[63] Chang X X, Xu L F and Shi H L 2015 *Chin. Phys. B.* **24** 128703
[64] Tan Z J and Chen S J 2011 *Biophys. J.* **101** 176
[65] Tan Z J and Chen S J 2010 *Biophys. J.* **99** 1565
[66] Bao L, Zhang X, Jin L and Tan Z J 2016 *Chin. Phys. B* **25** 313
[67] Bao L, Zhang X, Shi Y Z, Wu Y Y and Tan Z J 2017 *Biophys. J.* **112** 1094
[68] Zhang Z L, Wu Y Y, Xi K, Sang J P and Tan Z J 2017 *Biophys. J.* **113** 517