What is the best reference state for building statistical potentials in RNA 3D structure evaluation?

YA-LAN TAN,[1] CHEN-JIE FENG,[1] LEI JIN,[1] YA-ZHOU SHI,[2] WENBING ZHANG,[1] and ZHI-JIE TAN[1]

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

[2]Research Center of Nonlinear Science, School of Mathematics and Computer Science, Wuhan Textile University, Wuhan 430073, China

ABSTRACT

Knowledge-based statistical potentials have been shown to be efficient in protein structure evaluation/prediction, and the core difference between various statistical potentials is attributed to the choice of reference states. However, for RNA 3D structure evaluation, a comprehensive examination on reference states is still lacking. In this work, we built six statistical potentials based on six reference states widely used in protein structure evaluation, including averaging, quasi-chemical approximation, atom-shuffled, finite-ideal-gas, spherical-noninteracting, and random-walk-chain reference states, and we examined the six reference states against three RNA test sets including six subsets. Our extensive examinations show that, overall, for identifying native structures and ranking decoy structures, the finite-ideal-gas and random-walk-chain reference states are slightly superior to others, while for identifying near-native structures, there is only a slight difference between these reference states. Our further analyses show that the performance of a statistical potential is apparently dependent on the quality of the training set. Furthermore, we found that the performance of a statistical potential is closely related to the origin of test sets, and for the three realistic test subsets, the six statistical potentials have overall unsatisfactory performance. This work presents a comprehensive examination on the existing reference states and statistical potentials for RNA 3D structure evaluation.

Keywords: RNA 3D structure; knowledge-based potential; reference states

INTRODUCTION

RNA molecules play vital roles in cell life activities such as gene regulations and catalysis (Dethoff et al. 2012; Guttman and Rinn 2012), and their functions are generally relevant to their structures (Watson et al. 2003; Montange and Batey 2008). Therefore, understanding RNA structures, especially RNA three-dimensional (3D) structures, would help to understand their biological functions. RNA 3D structures can be derived through several experimental techniques such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy (Aviv et al. 2006; Baird et al. 2010). However, it is still very expensive and time consuming to experimentally obtain high-resolution RNA 3D structures and consequently, RNA structures deposited in the PDB database (Rose et al. 2017) are still limited up to now.

To complement experimental methods, various computational models have been proposed (Shi et al. 2014a; Miao and Westhof 2017; Schlick and Pyle 2017; Sun et al. 2017), aiming to predict RNA 3D structures in silico, including knowledge-based and physics-based models (Major et al. 1991; Das and Baker 2007; Ding et al. 2008; Parisien and Major 2008; Das et al. 2010; Jossinet et al. 2010; Cao and Chen 2011; Rother et al. 2011; Popenda et al. 2012; Zhang et al. 2012; Zhao et al. 2012; Cragnolini et al. 2013; Xia et al. 2013; Kim et al. 2014; Shi et al. 2014b, 2015, 2018; Xu et al. 2014; Bian et al. 2015; Boniecki et al. 2016; Li et al. 2016; Magnus et al. 2016; Bell et al. 2017; Jain and Schlick 2017; Wang et al. 2017; Jin et al. 2018). Generally, a predictive model can produce an ensemble of folded candidate structures, and consequently, a reliable knowledge-based statistical potential is required to evaluate predicted candidate structures. Furthermore, a reliable statistical potential can be used to guide RNA 3D structure folding (Jonikas et al. 2009; Zhang and Chen 2018).

Knowledge-based statistical potential has been proved to be efficient and effective for evaluating protein tertiary

Corresponding authors: zjtan@whu.edu.cn, wbzhang@whu.edu.cn

Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.069872.118.

structures (Sippl 1990; DeBolt and Skolnick 1996; Thomas and Dill 1996; Samudrala and Moult 1998; Lu and Skolnick 2001; Zhou and Zhou 2002; Shen and Sali 2006; Rykunov and Fiser 2007; Zhang and Zhang 2010; Huang and Zou 2011), protein–protein (Huang and Zou 2008), and protein–ligand docking (Huang and Zou 2006a, 2006b), and there have been six representative statistical potentials developed for protein tertiary structure evaluation, i.e., RAPDF (Samudrala and Moult 1998), KBP (Lu and Skolnick 2001), HA_SRS (Rykunov and Fiser 2007), Dfire (Zhou and Zhou 2002), Dope (Shen and Sali 2006), and RW (Zhang and Zhang 2010). These six statistical potentials for proteins are built based on six different reference states, i.e., averaging (Samudrala and Moult 1998), quasi-chemical approximation (Lu and Skolnick 2001), atom-shuffled (Rykunov and Fiser 2007), finite-ideal-gas (Zhou and Zhou 2002), spherical-noninteracting (Shen and Sali 2006), and random-walk-chain (Zhang and Zhang 2010) reference states, respectively, and the core difference between these statistical potentials mainly originates from the choice of different reference states. For RNA 3D structure evaluation, several statistical potentials have been built based on different reference states. Bernauer et al. (2011) have derived fully differentiable statistical potentials of KB at both all-atom and coarse-grained levels, based on the quasi-chemical approximation reference state. Capriotti et al. (2011) also have built all-atom and coarse-grained statistical potentials of RASP based on the averaging reference state. Recently, for RNA 3D structure evaluation, Wang et al. (2015) obtained a combined statistical potential of 3dRNAscore based on averaging reference state, which is composed of distance-dependent and torsion angle-dependent potentials.

Simultaneously, knowledge-based statistical potentials have also been built to simulate RNA structural folding (Jonikas et al. 2009; Zhang and Chen 2018). Jonikas et al. (2009) have proposed a nucleotide-level coarse-grained potential of bond, angle, dihedral, and nonbond term based on statistical analyses on RNA structure information, and used the potential to model 3D structures of large RNAs based on secondary structure and tertiary contact predictions. Very recently, Zhang and Chen proposed a set of correlated energy functions through an iterative method, and such energy functions can produce the RNA structural parameters very close to those from experimental structures in the PDB database (Zhang and Chen 2018). However, compared to proteins, only the averaging and quasi-chemical approximation reference states were successfully used to construct statistical potentials for RNA 3D structure evaluation, and for RNAs, there is still lacking a comprehensive understanding of the performances of those reference states widely used for proteins. Therefore, we would perform a comprehensive examination on the extensive reference states and try to figure out which reference state is the best one for RNA 3D structure evaluation.

In this work, based on six representative reference states—averaging (Samudrala and Moult 1998), quasi-chemical approximation (Lu and Skolnick 2001), atom-shuffled (Rykunov and Fiser 2007), finite-ideal-gas (Zhou and Zhou 2002), spherical-noninteracting (Shen and Sali 2006), and random-walk-chain (Zhang and Zhang 2010) reference states—we have built six statistical potentials for RNA 3D structure evaluation. Furthermore, we conducted an extensive examination of the six statistical potentials against three RNA test sets, including six subsets, through comparing their ability to identify native structures, identify near-native structures, and rank whole decoy sets. Additionally, we made extensive comparisons with the existing statistical potentials for RNAs and further examined the effect of training sets on the performance of statistical potentials. In order to get a reliable understanding of the reference states, the six statistical potentials were trained by a uniform nonredundant RNA training set and with the same parameters, such as bin width and distance cutoff.

RESULTS

Evaluation metrics

In general, there are two aspects for assessing the performance of a statistical potential: the ability of correctly identifying the native structure from a pool of decoys and ranking the near-native structures reasonably. Thus, in this work, we use the number of native structures with the minimum energy obtained by a statistical potential in the test sets, and we also calculate the ES (enrichment score) (Tsai et al. 2003; Bernauer et al. 2011; Wang et al. 2015) and PCC (Pearson correlation coefficient) (Capriotti et al. 2011) as metrics for near-native structures.

ES is defined as (Bernauer et al. 2011; Wang et al. 2015)

$$\mathrm{ES} = \frac{|E_{\mathrm{top10\%}} \cap R_{\mathrm{top10\%}}|}{0.1 \times 0.1 \times N_{\mathrm{decoys}}}, \tag{1}$$

where $|E_{\mathrm{top10\%}} \cap R_{\mathrm{top10\%}}|$ is the number of structures with energy in the lowest 10% energy range whose rmsd is also in the lowest 10% rmsd range. N_{decoys} is the total number of decoy structures for one native RNA. If the energy is extremely correlated to the rmsd, ES is equal to 10, and if it is completely unrelated to the rmsd, ES is equal to 1.

PCC is given as (DeBolt and Skolnick 1996)

$$\mathrm{PCC} = \frac{\sum_{n=1}^{N_{\mathrm{decoys}}} (E_n - \bar{E})(R_n - \bar{R})}{\sqrt{\sum_{n=1}^{N_{\mathrm{decoys}}} (E_n - \bar{E})^2} \sqrt{\sum_{n=1}^{N_{\mathrm{decoys}}} (R_n - \bar{R})^2}}, \tag{2}$$

where E_n and R_n are the energy and rmsd of the nth structure, respectively. $\bar{E}$ and $\bar{R}$ are the average energy and

average rmsd of decoys, respectively. *N* is the total number of decoy structures for one native RNA. Equation 2 indicates that the closer the value of PCC to 1, the more linear the relationship between the rmsds and energies. If PCC is equal to 1, the relationship between the energies and rmsds is completely linear and the performance of the statistical potential is perfect.

From the above, the number of identified native structures, ES value and PCC value can describe the abilities of a statistical potential in identifying native structures, in identifying near-native structures and in ranking whole decoy structures, respectively. In the following, we evaluate the six statistical potentials using the above-mentioned three evaluation metrics against three different RNA test sets including six test subsets.

Performance on test set I

Test set I, called randstr decoy set (Capriotti et al. 2011), consists of 85 RNAs with decoy structures generated by MODELER (Šali and Blundell 1993) with a set of Gaussian restraints for atom distances and dihedral angles from 85 native structures, which can be downloaded at http ://melolab.org/supmat/RNApot/Sup._Data.html. In test set I, there are 500 decoy structures for each RNA native structure, and the rmsds of decoy structures for test set I are mainly distributed in the range of 0–6 Å; see Figure 1A. Firstly, we examined the six statistical potentials through identifying native structures from decoy structures for test set I, and the numbers of native structures identified by them are summarized in Table 1. As shown in Table 1, all the statistical potentials identify 83 native structures out of the decoys of 85 RNAs. Afterward, we calculated the ES and PCC values for test set I by the six statistical potentials. As shown in Supplemental Table S2 in the Supplemental Material, the average ES values of 85 decoys obtained by Avg-REF, QChA-REF, ASh-REF, FIG-REF, SNI-REF, and RWC-REF are 9.0, 8.9, 9.0, 8.9, 9.0, and 8.9, respectively, and the average PCC values are 0.87, 0.86, 0.87, 0.85, 0.87, and 0.87, respectively. This indicates that for test set I, the correlations between rmsds and energies from the six statistical potentials are all very strong and all reach high performance. Thus, overall, the six statistical potentials all exhibit high performance and are not significantly different in structure evaluation for the test set I. The rmsd-energy scatterplots for all the 85 RNAs in test set I by the six statistical potentials can be found in the Supplemental Material.

Performance on test set II

Test set II is comprised of the decoy structures built by Bernauer et al. (2011) and Das and Baker (2007). The former includes two subsets: decoys for five RNAs generated by replica-exchange molecular dynamics simulations with atom position restrained, called MD subset (Bernauer et al. 2011); and decoys for 15 RNAs generated by normal mode perturbation method, called NM subset (Bernauer et al. 2011). These two subsets can be downloaded from http://csb.stanford.edu/rna/. There are 3500 decoy structures for four RNAs and 2600 decoy structures for one RNA (PDB ID: 1msy) in the MD subset, and the rmsds of the decoy structures are mainly in the range of 0–3 Å; see Figure 1B. Furthermore, there are 500 decoy structures for each RNA in the NM subset, and the majority rmsds of decoy structures are in the range of 1–5 Å; see Figure 1B. The latter is called the FARNA subset (Das and Baker 2007), which includes decoys for 20 RNAs and was generated by the FARNA method (Das and Baker 2007). For each RNA in the FARNA subset, there are about 500 decoy structures, and the rmsds of decoy structures for the FARNA subset are quite dispersed in the range of 4–15 Å. The FARNA subset can be downloaded from https://daslab.stanford .edu/site_data/pub_data/.

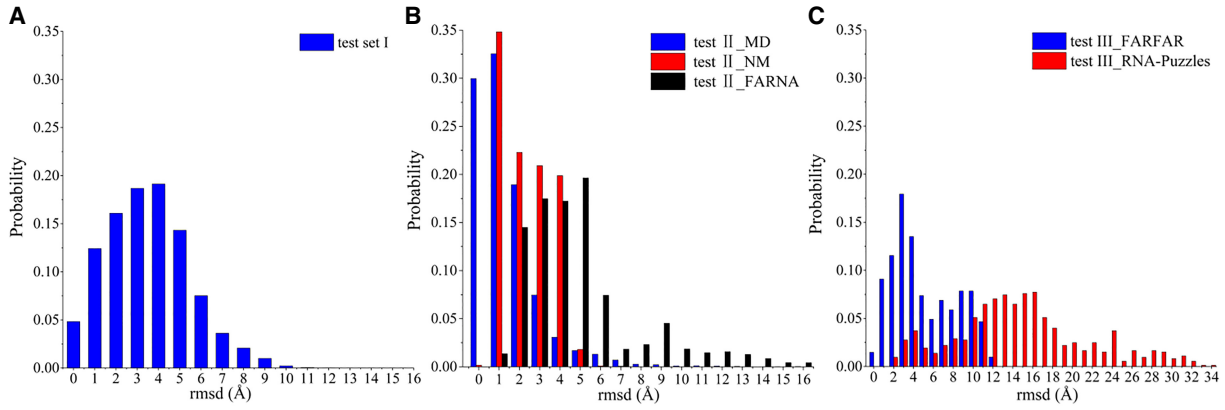

FIGURE 1. (*A*) The rmsd probability distribution of decoys in test set I (Bernauer et al. 2011). (*B*) The rmsd probability distributions of decoys in test set II, which is composed of MD (Bernauer et al. 2011), NM (Bernauer et al. 2011), and FARNA (Das and Baker 2007) subsets within 16 Å. (*C*) The rmsd probability distributions of test set III, which is composed of FARFAR (Das et al. 2010) and RNA-Puzzles (Miao et al. 2017) subsets within 34 Å.

TABLE 1. The numbers of native structures identified by the different statistical potentials

RNA data sets[a]	Statistical potentials[b]					
	Avg	QChA	ASh	FIG	SNI	RWC
Test set I	**83/85**	**83/85**	**83/85**	**83/85**	**83/85**	**83/85**
Test set II_MD	4/5	4/5	**5/5**	**5/5**	4/5	**5/5**
Test set II_NM	**13/15**	**13/15**	**13/15**	**13/15**	**13/15**	**13/15**
Test set II_FARNA	15/20	15/20	15/20	16/20	15/20	**17/20**
Test set III_FARFAR	19/32	19/32	19/32	**22/32**	21/32	**22/32**
Test set III_RNA-Puzzles	4/18	5/18	5/18	**6/18**	5/18	4/18

[a]The statistical potentials were built based on six different reference states: Avg-REF, the averaging reference state (Samudrala and Moult 1998); QChA-REF, the quasi-chemical approximation reference state (Lu and Skolnick 2001); ASh-REF, the atom-shuffled reference state (Rykunov and Fiser 2007); FIG-REF, the finite-ideal-gas reference state (Zhou and Zhou 2002); SNI-REF, the spherical-noninteracting reference state (Shen and Sali 2006); RWC-REF, the random-walk-chain reference state (Zhang and Zhang 2010). Here, the suffix of REF is not shown, due to limited space. The numbers of identified native structures are bolded for the best values among the six statistical potentials.
[b]Test set I is the randstr decoy set (Capriotti et al. 2011); MD subset in test set II was generated by the replica-exchange molecular dynamics simulation (Bernauer et al. 2011); NM subset in test set II was generated by normal mode perturbation method (Bernauer et al. 2011); FARNA subset in test set II was predicted by FARNA method (Das and Baker 2007); FARFAR subset in test set III was obtained by FARFAR method (Das et al. 2010); and RNA-Puzzles subset in test set III is from RNA-Puzzles (Miao et al. 2017).

For the MD subset in test set II, as shown in Table 1, ASh-REF, FIG-REF, and RWC-REF can identify five native structures out of the decoys of five RNAs, while Avg-REF, QChA-REF, and SNI-REF identified four native structures. As shown in Table 2, the ES values from the statistical potentials derived by six reference states are all around 8.0 and do not differ much: ASh-REF and SNI-REF are very slightly higher than others and FIG-REF is slightly lower than the other five. However, for the PCC value, FIG-REF has the best performance with 0.85 and RWC-REF is very slightly lower than FIG-REF. The performances of Avg-REF, QChA-REF, ASh-REF, and SNI-REF are similar, and are slightly lower than FIG-REF and RWC-REF. Therefore, overall, FIG-REF and RWC-REF slightly outperform others for the MD subset, and the higher performance on PCC values of FIG-REF as well as RWC-REF may mainly come from their high performance on the decoy structures with relatively large rmsds (see Figure 2 and Supplemental Figure S2 in the Supplemental Material for the rmsd-energy scatter-plots from the six potentials for RNAs of PDB IDs 1f27 and 434d).

For the NM subset, as shown in Table 1, all the statistical potentials can identify 13 native structures out of the decoys for 15 RNAs, and the two unidentified native structures are the RNAs of PDB IDs of 1esy and 1kka, whose native structures were solved by NMR spectroscopy at low salt (Amarasinghe et al. 2000; Cabello-Villegas et al. 2002). The experiment condition may be the main reason that these two native structures cannot be identified, since the deformation of RNA structure can be strongly influenced by cation concentration of solution and its structure will become less compact at lower salt due to the polyanionic nature (Jin et al. 2018; Shi et al. 2018). As shown in Table 2, the average ES value of RWC-REF is slightly higher than those of other potentials, and the average ES of FIG-REF is slightly lower than those of others. However, for average PCC values shown in Table 2, FIG-REF and RWC-REF have the best performances, and the performances of Avg-REF, QChA-REF, ASh-REF, and SNI-REF are similar and very slightly lower than those of FIG-REF and RWC-REF. Therefore, based on the overall assessment of both ES and PCC values, RWC-REF has the very slightly better performance for the NM subset. For the two largest RNAs (PDB IDs of 1x9k and 1i9v), on which RWC-REF has the best performance of the PCC value, the rmsd-energy scatter-plots by these six potentials are shown in Figure 3, and in Supplemental Figure S3 in the Supplemental Material, respectively.

For the FARNA subset, as shown in Table 1, the numbers of identified native structures of Avg-REF, QChA-REF, ASh-REF, FIG-REF, SNI-REF, and RWC-REF are 15, 15, 15, 16, 15, and 17 out of the decoys for 20 RNAs, respectively, and RWC-REF can identify the most native structures for the FARNA subset. As shown in Table 2 for ES and PCC values, the six statistical potentials all have unsatisfactory performances with mean ES values <3 and PCC values <0.4, respectively. QChA-REF has the slightly best performance with mean ES value of 2.56 and PCC value of 0.38, and FIG-REF has the worst performance with mean ES value of 1.83 and PCC value of 0.20. Overall, QChA-REF slightly outperforms other potentials on evaluation metrics for near-native structures in the FARNA subset. Besides, it is shown that the rmsds of decoys are generally large (e.g., between ~8 Å and ~15 Å for 1j6s) in this subset, and this may be the major reason that the statistical potentials all have unsatisfactory performance for the FARNA subset (see Fig. 4 and Supplemental Fig. S4 in the Supplemental Material for the rmsd-energy scatter-

TABLE 2. The ES and PCC values calculated by the different statistical potentials for test set II[a]

			Enrichment score (ES)						Pearson correlation coefficient (PCC)					
Decoy sets[b]	RNA	Length	Avg	QChA	ASh	FIG	SNI	RWC	Avg	QChA	ASh	FIG	SNI	RWC
Test set II_MD	1duq	26	8.3	8.3	8.3	7.9	8.4	8.1	0.66	0.68	0.67	0.86	0.66	0.82
	1f27	30	8.7	8.6	8.7	8.7	8.7	8.7	0.57	0.59	0.58	0.82	0.56	0.77
	1msy	27	7.3	7.3	7.3	7.2	7.2	7.3	0.89	0.89	0.89	0.84	0.90	0.86
	1nuj	24	7.6	7.6	7.6	7.7	7.7	7.7	0.64	0.62	0.66	0.88	0.64	0.84
	434d	14	8.2	8.2	8.3	8.2	8.2	8.2	0.74	0.73	0.74	0.84	0.74	0.82
Average value			8.02	7.99	**8.03**	7.92	**8.03**	8.01	0.70	0.70	0.71	**0.85**	0.70	0.82
Test set II_NM	1duq	26	7.7	7.7	7.7	7.5	7.7	7.5	0.84	0.86	0.85	0.89	0.84	0.89
	1esy	19	4.1	4.1	4.1	4.9	4.5	4.7	0.83	0.81	0.84	0.93	0.85	0.92
	1f27	30	5.9	5.9	5.9	5.3	5.3	5.7	0.81	0.80	0.82	0.92	0.81	0.90
	1i9v	76	5.7	6.1	5.7	3.7	5.5	5.1	0.87	0.88	0.88	0.88	0.87	0.90
	1kka	17	5.7	5.7	5.7	5.9	5.9	5.9	0.85	0.87	0.86	0.90	0.86	0.89
	1msy	27	4.9	4.9	4.7	5.1	4.7	5.7	0.90	0.90	0.91	0.92	0.91	0.93
	1nuj	24	7.5	7.3	7.7	7.5	7.7	7.9	0.86	0.85	0.87	0.93	0.86	0.92
	1qwa	21	3.7	3.7	3.9	3.3	3.3	3.5	0.88	0.88	0.89	0.90	0.89	0.91
	1x9k	62	5.4	5.4	5.8	2.9	5.2	4.8	0.84	0.85	0.85	0.88	0.82	0.91
	1xjr	46	8.5	8.5	8.5	8.5	8.1	8.3	0.92	0.93	0.93	0.94	0.92	0.94
	1ykq	19	4.4	4.6	4.6	3.9	4.6	4.1	0.81	0.84	0.82	0.90	0.82	0.90
	1zih	12	6.9	7.1	6.9	7.3	6.9	7.3	0.88	0.88	0.89	0.93	0.89	0.92
	28sp	28	5.3	4.9	5.1	5.9	5.5	6.3	0.83	0.83	0.85	0.90	0.86	0.91
	2f88	34	6.6	6.8	6.4	5.5	6.6	6.6	0.90	0.90	0.91	0.93	0.90	0.93
	434d	14	7.9	7.9	7.9	7.9	7.9	7.9	0.89	0.89	0.90	0.91	0.90	0.91
Average value			6.02	6.04	6.04	5.69	5.96	**6.10**	0.86	0.86	0.87	**0.91**	0.87	**0.91**
Test set II_FARNA	157d	24	3.2	3.4	3.2	3.0	2.6	3.2	0.57	0.57	0.56	0.53	0.53	0.56
	1a4d	41	2.4	4.2	2.2	1.8	2.2	1.6	0.25	0.59	0.25	0.14	0.19	0.13
	1csl	28	2.2	2.0	1.8	1.0	1.8	1.2	0.50	0.49	0.47	0.17	0.44	0.30
	1dqf	19	4.2	4.2	4.0	2.8	4.0	3.8	0.68	0.68	0.65	0.35	0.64	0.48
	1esy	19	4.2	4.2	4.4	2.6	4.2	3.0	0.60	0.56	0.61	0.33	0.61	0.43
	1i9x	26	3.2	3.4	3.0	2.4	2.8	3.2	0.50	0.52	0.47	0.39	0.42	0.45
	1j6s	24	0.6	0.6	0.6	2.8	0.4	1.8	−0.14	−0.10	−0.12	0.45	−0.12	0.36
	1kd5	22	2.4	2.2	2.4	0.8	2.6	0.8	0.30	0.27	0.30	0.03	0.30	0.13
	1kka	17	1.0	1.0	1.0	0.2	1.2	0.4	0.05	0.00	0.02	−0.37	0.10	−0.21
	1l2x	27	0.4	0.4	0.4	3.2	0.4	2.6	−0.34	−0.30	−0.31	0.61	−0.30	0.43
	1mhk	32	2.2	2.0	2.2	1.0	1.6	1.0	0.31	0.30	0.29	0.07	0.28	0.17
	1q9a	27	2.2	2.2	2.0	0.4	2.8	1.0	0.52	0.51	0.50	0.02	0.53	0.20
	1qwa	21	1.8	1.8	1.8	0.4	2.2	1.0	0.46	0.42	0.43	−0.20	0.46	0.04
	1xjr	46	2.8	2.8	2.6	2.8	2.8	3.0	0.48	0.44	0.48	0.32	0.45	0.43
	1zih	12	5.2	5.2	5.2	5.0	5.2	5.0	0.58	0.57	0.57	0.42	0.59	0.49
	255d	24	2.2	2.2	1.6	0.6	2.0	0.8	0.38	0.37	0.33	−0.24	0.36	−0.05
	283d	24	1.0	1.0	1.0	1.4	1.2	1.4	0.18	0.16	0.21	0.22	0.21	0.21
	28sp	28	3.0	3.0	3.0	2.0	3.0	2.8	0.65	0.64	0.64	0.11	0.63	0.31
	2a43	26	1.6	1.8	1.8	1.4	2.0	1.4	0.24	0.28	0.26	0.44	0.29	0.43
	2f88	34	3.6	3.8	3.2	1.0	3.4	1.8	0.54	0.55	0.52	0.12	0.52	0.27
Average value			2.47	**2.56**	2.37	1.83	2.42	2.04	0.37	**0.38**	0.36	0.20	0.36	0.28

[a]The statistical potentials were built based on six different reference states: Avg-REF, the averaging reference state (Samudrala and Moult 1998); QChA-REF, the quasi-chemical approximation reference state (Lu and Skolnick 2001); ASh-REF, the atom-shuffled reference state (Rykunov and Fiser 2007); FIG-REF, the finite-ideal-gas reference state (Zhou and Zhou 2002); SNI-REF, the spherical-noninteracting reference state (Shen and Sali 2006); RWC-REF, the random-walk-chain reference state (Zhang and Zhang 2010). Here, the suffix of REF was not indicated due to limited space. The average ES and PCC values are bolded for the best values among the six statistical potentials.

[b]Test set II is composed of three test subsets: MD subset was generated by the replica-exchange molecular dynamics simulation (Bernauer et al. 2011); NM subset was generated by normal mode perturbation method (Bernauer et al. 2011); and FARNA subset was predicted by FARNA method (Das and Baker 2007).

plots for RNAs of PDB IDs 1j6s and 1a4d). The rmsd-energy scatterplots for all 40 RNAs in test set II by the six statistical potentials can be found in the Supplemental Material.

Performance on test set III

Test set III is composed of the FARFAR subset (Das et al. 2010) and RNA-Puzzles subset (Cruz et al. 2012; Miao

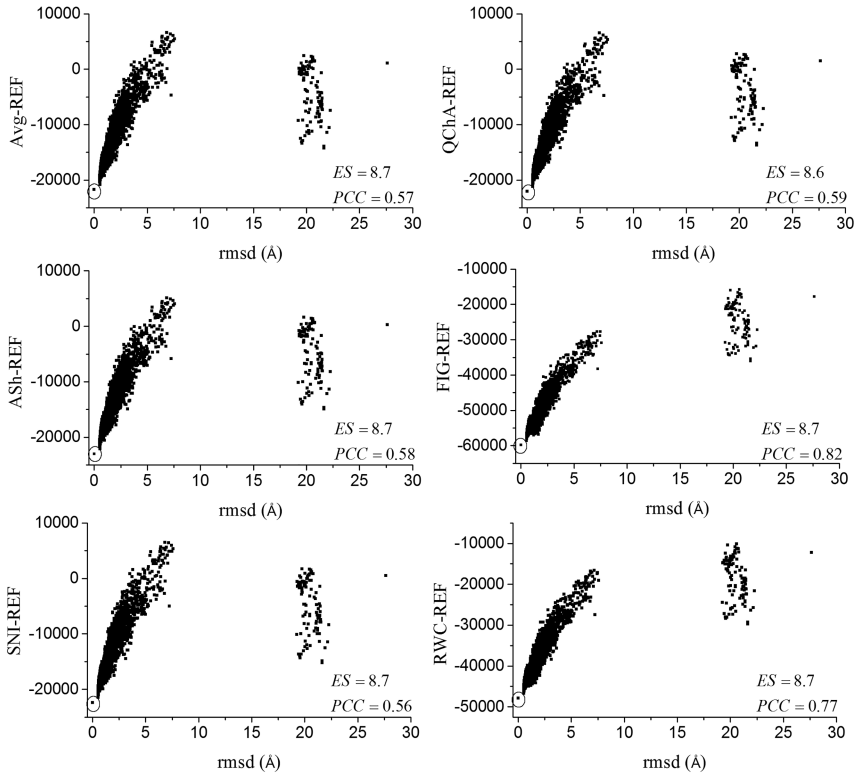

FIGURE 2. The rmsd-energy scatter-plot of 1f27 in RNA test set II_MD. Here, the energy was calculated by the statistical potentials based on six reference states: Avg-REF, the averaging reference state (Samudrala and Moult 1998); QChA-REF, the quasi-chemical approximation reference state (Lu and Skolnick 2001); ASh-REF, the atom-shuffled reference state (Rykunov and Fiser 2007); FIG-REF, the finite-ideal-gas reference state (Zhou and Zhou 2002); SNI-REF, the spherical-noninteracting reference state (Shen and Sali 2006); RWC-REF, the random-walk-chain reference state (Zhang and Zhang 2010). The native structure is highlighted by an empty circle, and ES and PCC values are shown in the respective panels.

et al. 2017). The former subset was obtained by RNA modeling with the FARFAR method (Das et al. 2010), and it contains the five lowest energy clusters of structure models for each of the 32 RNA motifs containing noncanonical base pairs (Das et al. 2010). Thus, FARFAR decoys can be used to assess the ability of statistical potentials to evaluate RNAs with noncanonical base pairs. As shown in Figure 1C, the rmsds for decoy structures in the FARFAR subset are mainly in the range of 1–11 Å. The latter subset was obtained from RNA-Puzzles (Miao et al. 2017), which is a CASP-like evaluation of blind 3D RNA structure predictions (Miao et al. 2017). Thus, the RNA-Puzzles subset contains various predicted decoy structures from the different RNA prediction models and can be a realistic test set for demonstrating the performance of a statistical potential in evaluating RNA 3D structures. There are dozens of predicted decoy structures for 18 different RNAs in the RNA-Puzzles subset, and the rmsds are distributed in the wide range of ~2–34 Å (see Fig. 1C). FARFAR and RNA-Puzzles subsets can be downloaded from https://daslab.stanford.edu/site_data/pub_data/ and https://github.com/RNA-Puzzles/RNA-Puzzles-Normalized-submissions, respectively. Since there are only a dozen or several dozens of predicted structures for each RNA in FARFAR and RNA-Puzzles subsets, we only calculated the rmsd of the predicted structure that had the lowest energy for each RNA instead of ES value, as well as PCC values.

For the FARFAR subset, as shown in Table 1, the numbers of identified native structures by Avg-REF, QChA-REF, ASh-REF, FIG-REF, SNI-REF, and RWC-REF are 19, 19, 19, 22, 21, and 22 for 32 RNAs, respectively. FIG-REF and RWC-REF slightly outperform SNI-REF, while Avg-REF, QChA-REF, and ASh-REF have a slightly worse performances than the others. As shown in Table 3, for the average rmsd of predicted structures with the lowest energy, the six different statistical potentials are similar. Moreover, different statistical potentials also have very similar PCC values, while ASh-REF, SNI-REF, and RWC-REF are very slightly better than the others. Overall, for the FARNA subset, the six statistical potentials have similar and unsatisfactory performances.

For the RNA-Puzzles subset, as shown in Table 1, FIG-REF still performs the best in identifying native structures, and it can identify six native structures out of the decoys of 18 RNAs. Next, QChA-REF, ASh-REF, and SNI-REF can identify five native structures for 18 RNAs, and Avg-REF and RWC-REF can only identify four native structures. Furthermore, as shown in Table 4, FIG-REF has the best performance with the lowest average rmsd of the predicted structures with the lowest energy and the maximum average PCC value of 0.47, and RWC-REF has a slightly lower performance with an average PCC value of 0.43. However, for the RNA-Puzzles subset, all six statistical potentials do not have a satisfactory performance in identifying and ranking near-native structures. As described above, 10 RNA structures in the RNA-Puzzles subset are also in the training set; thus we benchmarked the effect of the 10 RNA structures on the performance of six reference states on the decoys of these 10 RNA structures by using the leave-one-out or jackknife method (Capriotti et al. 2011). In other words, for each one of the 10 RNAs, we rebuilt a statistical potential based on the training set with the remaining 107 native RNA structures by removing the specific RNA structure to assess

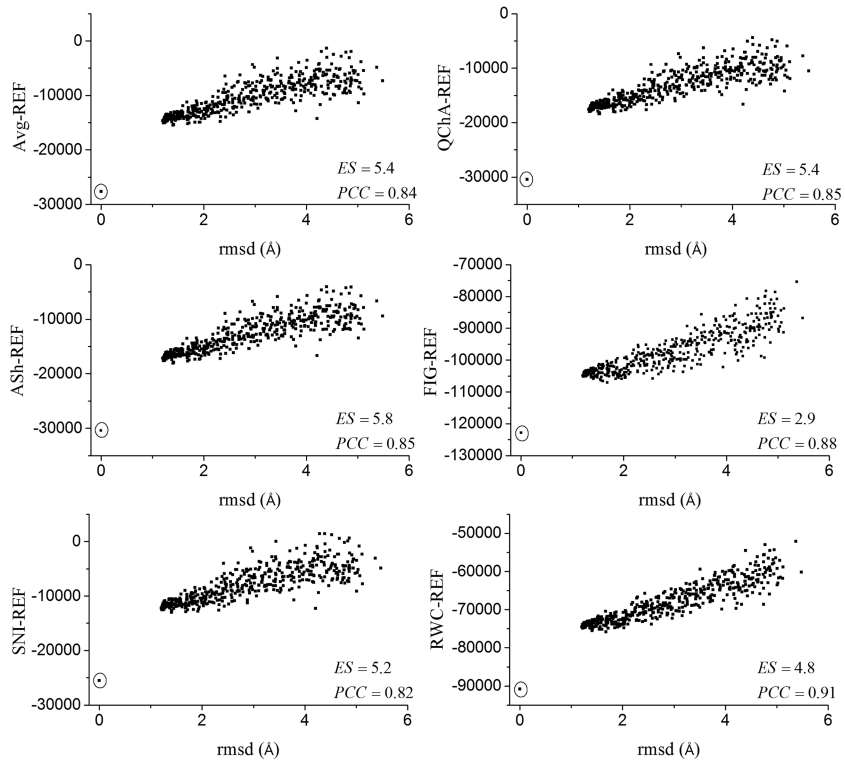

FIGURE 3. The rmsd-energy scatter-plot of 1x9k in RNA test set II_NM. Here, the energy was calculated by the statistical potentials based on six reference states: Avg-REF, the averaging reference state (Samudrala and Moult 1998); QChA-REF, the quasi-chemical approximation reference state (Lu and Skolnick 2001); ASh-REF, the atom-shuffled reference state (Rykunov and Fiser 2007); FIG-REF, the finite-ideal-gas reference state (Zhou and Zhou 2002); SNI-REF, the spherical-noninteracting reference state (Shen and Sali 2006); RWC-REF, the random-walk-chain reference state (Zhang and Zhang 2010). The native structure is highlighted by an empty circle, and ES and PCC values are shown in the respective panels.

decoy structures of this RNA. The results obtained by the above-described leave-one-out method are shown in Supplemental Tables S3 and S4. Compared with those from the training set of 108 RNA native structures, the rmsds of predicted structures with minimum energy and PCC values are almost exactly the same except for the subtle change in the average PCC value of ASh-REF. Such subtle effect due to the leave-one-out method is not surprising since the percentage of each one of these 10 RNAs in our training set is <2.6% in nucleotide number.

Overall performance on test sets

Identifying native structures

As shown in Table 1, FIG-REF and RWC-REF can identify the most native structures for five subsets. Next, ASh-REF identifies the most native structures for three subsets. After that, Avg-REF, QChA-REF, and SNI-REF can identify the most native structures for two subsets. On the overall level, Avg-REF, QChA-REF, ASh-REF, FIG-REF, SNI-REF, and RWC-REF can recognize 138, 139, 140, 145, 141, 144 native structures for 175 RNAs. Therefore, the performances of the different statistical potentials based on six reference states in identifying the native structures follow the order: FIG-REF $\gtrsim$ RWC-REF > SNI-REF $\gtrsim$ ASh-REF $\gtrsim$ QChA-REF $\gtrsim$ Avg-REF.

It should be noted that the ability of the six statistical potentials in identifying RNA native structures is still weak, e.g., for the RNA-Puzzles subset, even FIG-REF with the best performance can only identify six native structures out of the decoys of 18 RNAs. Therefore, a good statistical potential is still highly desired for identifying native structures out of the predicted candidates from computational models for RNA 3D structures.

Identifying near-native structures

Equation 1 indicates that ES value reflects the ability of a statistical potential in identifying 10% of near-native structures from whole decoy structures. For only a dozen or several dozens of decoys corresponding to each native RNA in FARFAR and RNA-Puzzles subsets, we cannot calculate ES value for these two subsets, and we use the rmsd of predicted structure with the lowest energy instead of ES value. As shown in Table 2, QChA-REF, ASh-REF, SNI-REF, and RWC-REF have very slightly highest mean ES values for one subset (FARNA subset for QChA-REF, MD subset for ASh-REF and SNI-REF**,** and NM subset for RWC-REF) in test set II, which contains three subsets in total. Nevertheless, it is noted that the overall difference between various potentials in ES value is rather slight. For example, the maximum mean difference in ES value between different statistical potentials is 0.11 (between 7.92 for FIG-REF and 8.03 for ASh-REF and SNI-REF), 0.41 (between 5.69 for FIG-REF and 6.10 for RWC-REF), and 0.73 (between 1.83 for FIG-REF and 2.56 for QChA-REF) for MD, NM, and FARNA subsets, respectively. In addition, except for the MD subset, the statistical potentials all have relatively low ES values for NM and FARNA subsets, especially for the FARNA subset, i.e., mean ES value is as low as <2.6 for the FARNA subset. For the FARFAR subset, as shown in Table 3, there is still no distinctive difference between different statistical potentials for the average rmsd of predicted structure with the lowest energy. Besides, for identifying the nearest-native structure, which has minimum rmsd and energy simultaneously

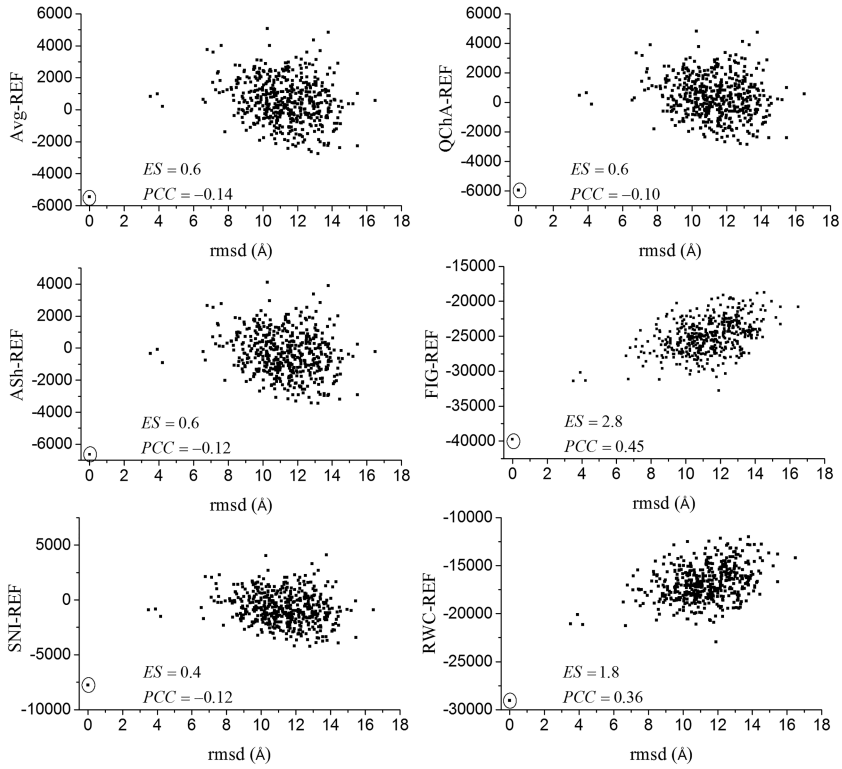

FIGURE 4. The rmsd-energy scatterplot of 1j6s in RNA test set II_FARNA. Here, the energy was calculated by the statistical potentials based on six reference states: Avg-REF, the averaging reference state (Samudrala and Moult 1998); QChA-REF, the quasi-chemical approximation reference state (Lu and Skolnick 2001); ASh-REF, the atom-shuffled reference state (Rykunov and Fiser 2007); FIG-REF, the finite-ideal-gas reference state (Zhou and Zhou 2002); SNI-REF, the spherical-noninteracting reference state (Shen and Sali 2006); RWC-REF, the random-walk-chain reference state (Zhang and Zhang 2010). The native structure is highlighted by an empty circle, and ES and PCC values are shown in the respective panels.

excluding its native structure, except that RWC-REF can identify eight nearest-native structures from the decoys of 32 RNAs, other statistical potentials can identify nine nearest-native structures. For the RNA-Puzzles subset, as shown in Table 4, all six statistical potentials have very unsatisfactory performance, and FIG-REF has the relatively better performance compared to the others. FIG-REF can identify three nearest-native structures from the decoys of 18 RNAs. This also suggests that a statistical potential of high performance is still highly desired for identifying near-native structures for predicted RNA 3D structures.

Ranking decoy structures in test sets

An important aim of a statistical potential is to be used to guide RNA folding or structure prediction, and thus there should be a positive and strong correlation between rmsds and the corresponding energies evaluated by a statistical potential of high performance. As shown in Tables 2–4, for PCC values, FIG-REF performs the best for three subsets (MD, NM, and RNA puzzles subsets). After that, RWC-REF has the best performance for two subsets (NM and FARFAR subsets) and QChA-REF, ASh-REF and SNI-REF have the best performance for one subset (FARNA subset for QChA-REF, FARFAR subset for ASh-REF and SNI-REF). Furthermore, for MD and RNA-Puzzles subsets, PCC values from RWC-REF are only slightly smaller than those from FIG-REF, and the other four statistical potentials have visibly lower performance than FIG-REF and RWC-REF. It is noted that FIG-REF has the best performance for test set III_RNA-Puzzles, which is from blind RNA 3D structure predictions, using extensive computational models (Miao et al. 2017). Therefore, the performances of the statistical potentials in ranking near-native structures follow the order: FIG-REF $\gtrsim$ RWC-REF > QChA-REF ~ ASh-REF ~ SNI-REF ~ Avg-REF. It is also noted that the six statistical potentials globally have unsatisfactory performances for FARNA, FARFAR, and RNA-Puzzles subsets, with PCC values <0.5.

From the above shown performances on identifying native structures and ranking structure decoys, we can roughly rank that FIG-REF and RWC-REF are slightly superior to other statistical potentials, although for three subsets (FARNA, FARFAR, and RNA-Puzzles), the performances of FIG-REF and RWC-REF do not reach a satisfactory level.

Ability of capturing base-base interactions

Base-pairing and base-stacking interactions are critical to the stability of RNA 3D structure (Wang et al. 2016, 2018). The ability of capturing the base-pairing and base-stacking interactions is also an important criterion for assessing the quality of a statistical potential. Figure 5 shows the potentials between the N2 atom of guanine and the O2 atom of cytosine derived based on six reference states. There are two apparent wells for all of the six potentials: The first well at the distance of ~3.0 Å is corresponding to the base-pairing interaction, and the second one at ~8.0 Å is corresponding to the indirect base-stacking interaction between next-nearest residues. However, only FIG-REF, SNI-REF, and RWC-REF capture the significant base-stacking interaction between nearest bases at ~3.6 Å, and the wells of FIG-REF and SNI-REF at 3.6 Å are slightly more distinctive. Similar phenomena

TABLE 3. The rmsds of predicted structures with minimum energy and PCCs between energy and rmsd of decoy structures calculated by the different statistical potentials for test set III_FARFAR[a]

		The rmsd of predicted structure with minimum energy (Å)						Pearson correlation coefficient (PCC)					
Motif name	Length	Avg	QChA	ASh	FIG	SNI	RWC	Avg	QChA	ASh	FIG	SNI	RWC
G-A base pair	6	**1.190**	**1.190**	**1.190**	**1.190**	**1.190**	**1.190**	0.99	0.99	0.99	0.99	0.99	0.99
Fragment with G/G and G/A pairs, SRP helix VI	8	**3.270**	**3.270**	**3.270**	4.695	**3.270**	4.695	0.76	0.75	0.75	0.65	0.74	0.69
Helix with A/C base pairs	12	3.306	3.306	3.306	3.306	3.306	3.306	0.57	0.59	0.55	0.40	0.50	0.46
Four-way junction, HCV IRES	13	11.201	11.201	11.201	11.658	11.658	11.658	0.56	0.57	0.54	−0.27	0.45	0.03
Loop 8, A-type Ribonuclease P	7	3.440	3.440	3.440	3.440	3.440	3.440	−0.62	−0.62	−0.62	−0.58	−0.59	−0.59
Helix with U/C base pairs	8	2.095	**2.095**	**2.095**	**2.095**	**2.095**	**2.095**	0.96	0.96	0.96	0.96	0.96	0.96
Curved helix with G/A and A/A base pairs	12	3.146	3.146	3.146	3.146	3.146	3.146	−0.04	−0.05	−0.02	0.17	0.04	0.10
Precatalytic conformation, hammerhead ribozyme	19	**7.659**	**7.659**	**7.659**	12.966	**7.659**	**7.659**	0.39	0.37	0.41	−0.03	0.40	0.16
Loop E motif, 5S RNA	18	2.269	2.269	2.269	2.269	2.269	2.269	0.40	0.41	0.41	0.50	0.45	0.47
UUCG tetraloop	6	**1.122**	**1.122**	**1.122**	**1.122**	**1.122**	**1.122**	0.99	0.99	0.98	0.99	0.99	0.99
Rev response element high affinity site	9	4.078	4.078	4.078	4.078	4.078	4.078	0.84	0.84	0.84	0.82	0.84	0.83
Fragment with A/C pairs, SRP helix VI	12	**1.832**	**1.832**	**1.832**	**1.832**	**1.832**	**1.832**	0.93	0.93	0.93	0.93	0.93	0.93
Signal recognition particle Domain IV	12	2.346	2.346	2.346	2.346	2.346	2.346	0.76	0.77	0.76	0.79	0.79	0.79
Bulged G motif, sarcin/ricin loop	13	5.109	5.109	5.109	5.160	5.109	5.160	−0.57	−0.57	−0.58	−0.66	−0.59	−0.62
Tertiary interaction, hammerhead ribozyme	16	9.863	9.863	9.863	9.863	9.863	9.863	0.37	0.35	0.40	0.54	0.39	0.48
GAGA tetraloop from sarcin/ricin loop	6	**0.819**	**0.819**	**0.819**	**0.819**	**0.819**	**0.819**	1.00	1.00	1.00	1.00	1.00	1.00
Pentaloop from conserved region of SARS genome	7	1.098	1.098	1.098	1.098	1.098	1.098	0.78	0.78	0.78	0.80	0.77	0.79
L2/L3 tertiary interaction, purine riboswitch	18	9.461	9.590	9.461	9.590	9.461	9.461	0.14	0.16	0.16	0.20	0.20	0.18
L3, thiamine pyrophosphate riboswitch	7	**1.995**	**1.995**	**1.995**	**1.995**	**1.995**	**1.995**	0.79	0.79	0.79	0.83	0.83	0.82
Kink-turn motif from SAM-I riboswitch	13	8.075	8.075	8.075	8.782	8.075	8.782	−0.23	−0.27	−0.22	−0.30	−0.24	−0.29
Active site, hammerhead ribozyme	17	11.187	11.187	11.187	11.187	11.187	11.187	−0.12	−0.11	−0.10	0.29	−0.06	0.14
P1/L3, SAM-II riboswitch	23	12.289	12.289	12.289	10.685	12.289	10.685	0.14	0.16	0.17	0.58	0.12	0.49
J4/5 from P4–P6 domain, Tetrahymena ribozyme	9	2.352	2.352	2.352	2.352	2.352	2.352	0.71	0.71	0.72	0.73	0.72	0.72
Stem C internal loop, L1 ligase	12	3.353	3.353	3.353	**2.240**	2.416	2.416	0.66	0.67	0.66	0.68	0.69	0.69

Continued

TABLE 3. *Continued*

Motif name	Length	The rmsd of predicted structure with minimum energy (Å)						Pearson correlation coefficient (PCC)					
		Avg	QChA	ASh	FIG	SNI	RWC	Avg	QChA	ASh	FIG	SNI	RWC
J5/5a hinge, P4–P6 domain, Tetr. ribozyme	17	10.973	10.973	10.973	10.273	10.973	10.273	–0.08	–0.12	–0.05	0.13	–0.01	0.07
Three-way junction, purine riboswitch	13	7.126	7.126	7.126	6.214	7.126	6.214	0.29	0.29	0.34	0.59	0.37	0.55
J4a/4b region, metal-sensing riboswitch	14	4.507	4.507	4.507	4.507	4.507	4.507	0.62	0.62	0.62	0.47	0.61	0.56
Kink-turn motif	15	9.085	9.085	9.085	9.085	9.085	9.085	0.11	0.08	0.16	0.33	0.18	0.30
Tetraloop/helix interaction, L1 ligase crystal	10	**0.857**	**0.857**	**0.857**	**0.857**	**0.857**	**0.857**	0.98	0.98	0.98	0.94	0.97	0.96
Hook-turn motif	11	7.616	7.616	4.228	**1.718**	4.228	4.228	–0.48	–0.48	–0.47	–0.46	–0.47	–0.47
Tetraloop/receptor, P4–P6 domain, Tetr. ribozyme	15	3.312	3.312	3.312	6.770	3.312	3.312	0.77	0.76	0.77	0.64	0.79	0.72
Pseudoknot, domain III, CPV IRES	18	3.547	3.547	3.547	3.808	3.547	3.808	0.70	0.69	0.70	0.48	0.69	0.54
Average value		4.987 **(9/32)**	4.991 **(9/32)**	4.881 **(9/32)**	5.036 **(9/32)**	4.866 **(9/32)**	**4.842** (8/32)	0.44	0.44	**0.45**	0.44	**0.45**	**0.45**

[a]The statistical potentials were built based on six different reference states: Avg-REF, the averaging reference state (Samudrala and Moult 1998); QChA-REF, the quasi-chemical approximation reference state (Lu and Skolnick 2001); ASh-REF, the atom-shuffled reference state (Rykunov and Fiser 2007); FIG-REF, the finite-ideal-gas reference state (Zhou and Zhou 2002); SNI-REF, the spherical-noninteracting reference state (Shen and Sali 2006); RWC-REF, the random-walk-chain reference state (Zhang and Zhang 2010). Test set III_FARFAR was obtained by the FARFAR method (Das et al. 2010). The total counts of identified nearest-native structures (decoys with both lowest rmsd and energy) are presented in parentheses in the last line of the table. The rmsd of predicted structure with minimum energy in each row is bolded when it is the nearest-native structure, and the average values of the rmsds of predicted structures with minimum energy and the PCC values are bolded for the best values among the six statistical potentials. Here, the suffix of REF is not indicated due to limited space.

TABLE 4. The rmsds of predicted structures with lowest energy and PCCs between energy and rmsd of decoy structures calculated by the different statistical potentials for test set III_RNA-Puzzles[a]

		The rmsd of predicted structure with minimum energy (Å)						Pearson correlation coefficient (PCC)					
Puzzle-X[b]	Length	Avg	QChA	ASh	FIG	SNI	RWC	Avg	QChA	ASh	FIG	SNI	RWC
1 (3MEI)	46	5.710	5.710	5.710	5.710	5.710	5.710	0.24	0.23	0.25	0.51	0.22	0.41
2 (3P59)	100	3.660	3.660	3.660	3.660	3.660	3.660	0.33	0.33	0.32	0.32	0.33	0.34
3 (—)	84	14.250	14.250	14.250	14.250	14.250	14.250	0.37	0.39	0.36	0.45	0.35	0.42
4 (3V7E)	126	3.372	3.372	3.374	4.124	3.372	4.124	0.43	0.43	0.43	0.42	0.43	0.43
5 (4P8Z)	188	26.970	20.270	24.720	**9.030**	24.720	20.270	0.52	0.52	0.54	0.63	0.44	0.73
6(4GXY)	168	14.110	14.110	14.110	14.110	14.110	14.110	0.24	0.22	0.24	0.62	0.21	0.52
7 (4R4V)	185	26.140	26.140	26.140	24.890	26.140	24.890	0.08	0.07	0.08	0.46	0.05	0.36
8 (4L81)	96	12.270	12.270	12.270	11.780	12.170	11.780	0.40	0.40	0.42	0.73	0.37	0.67
10 (4LCK)	171	10.270	10.270	10.270	10.270	10.270	10.270	0.31	0.30	0.33	0.70	0.25	0.62
12 (4QLN)	125	13.308	16.771	13.308	15.874	16.771	15.874	0.41	0.40	0.41	0.49	0.39	0.54
13 (4XW7)	71	14.839	12.502	14.839	**5.410**	14.839	5.553	0.50	0.49	0.51	0.84	0.46	0.78
14-Bound (5DDP)	61	**5.977**	**5.977**	**5.977**	11.566	**5.977**	11.566	0.50	0.50	0.49	0.29	0.51	0.38
14-Free (5DDO)	61	11.442	11.442	11.442	**6.862**	15.940	**6.862**	0.10	0.12	0.11	0.46	0.07	0.37
15 (5DI4)	68	19.777	19.777	19.777	15.459	19.777	14.768	0.40	0.40	0.40	0.56	0.38	0.52
17 (5K7C)	62	8.642	8.642	8.642	8.779	8.642	8.779	0.42	0.42	0.42	0.52	0.41	0.49
18 (5TPY)	71	16.346	16.346	13.362	9.820	16.346	9.820	0.07	0.12	0.07	0.49	0.04	0.37
19 (5T5A)	62	16.489	16.489	16.489	16.602	16.489	16.489	−0.12	−0.13	−0.11	0.13	−0.13	0.05
21 (5NZ6)	41	18.030	18.030	18.030	16.050	18.030	18.030	−0.53	−0.49	−0.52	−0.19	−0.52	−0.32
Average value		13.422 (1/18)	13.113 (1/18)	13.132 (1/18)	**11.347 (3/18)**	13.734 (1/18)	12.045 (1/18)	0.26	0.26	0.27	**0.47**	0.24	0.43

[a]The statistical potentials were built based on six different reference states: Avg-REF, the averaging reference state (Samudrala and Moult 1998); QChA-REF, the quasi-chemical approximation reference state (Lu and Skolnick 2001); ASh-REF, the atom-shuffled reference state (Rykunov and Fiser 2007); FIG-REF, the finite-ideal-gas reference state (Zhou and Zhou 2002); SNI-REF, the spherical-noninteracting reference state (Shen and Sali 2006); RWC-REF, the random-walk-chain reference state (Zhang and Zhang 2010). Test set III_RNA-Puzzles is from RNA-Puzzles (Miao et al. 2017). The total counts of identified nearest-native structures (decoys with both lowest rmsd and energy) are presented in parentheses in the last line of the table. The rmsd of predicted structure with minimum energy in each row is bolded when it is the nearest-native structure, and the average values of the rmsds of predicted structures with minimum energy and the PCC values are bolded for the best values among the six statistical potentials. Here, the suffix of REF was not indicated due to limited space.

[b]PDB ID is presented for each puzzle except for Puzzle-3 whose PDB ID is not available from the PDB database.

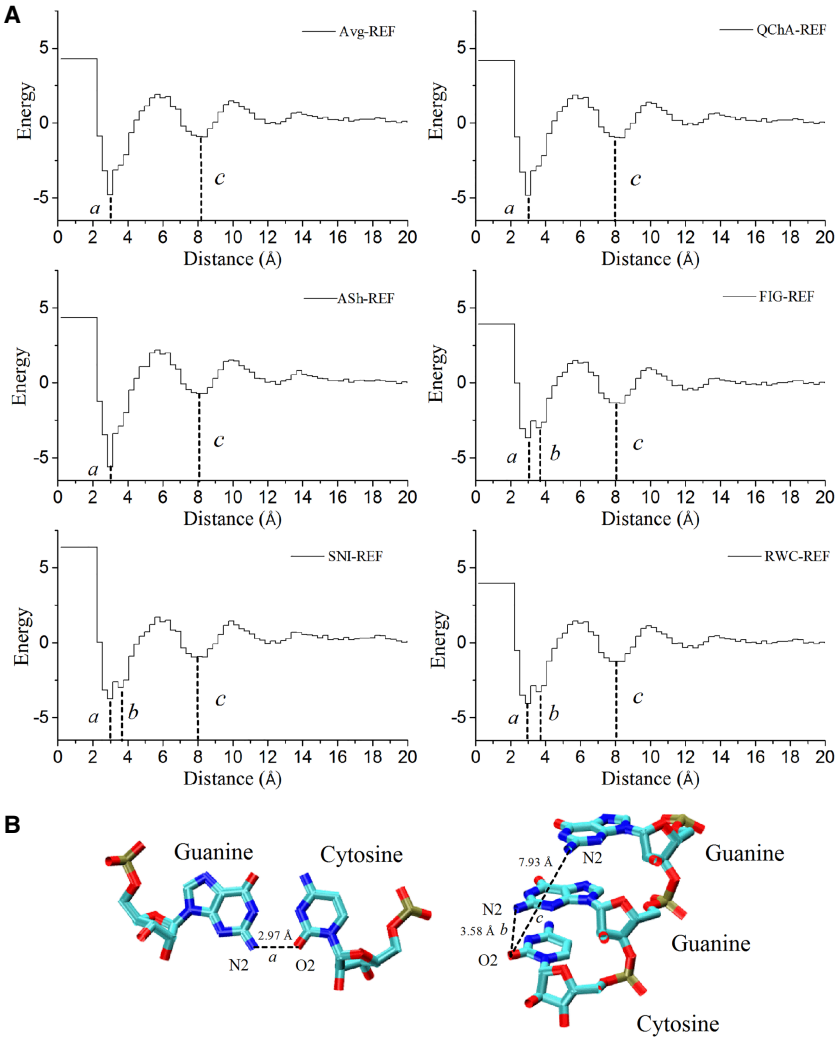

FIGURE 5. (*A*) The statistical potentials between N2 atom of guanine and O2 atom of cytosine based on different reference states: Avg-REF, the averaging reference state (Samudrala and Moult 1998); QChA-REF, the quasi-chemical approximation reference state (Lu and Skolnick 2001); ASh-REF, the atom-shuffled reference state (Rykunov and Fiser 2007); FIG-REF, the finite-ideal-gas reference state (Zhou and Zhou 2002); SNI-REF, the spherical-noninteracting reference state (Shen and Sali 2006); RWC-REF, the random-walk-chain reference state (Zhang and Zhang 2010). For the situation in which atom pairs were not observed within a certain bin width, the statistical potentials in these distance bins were set as the most unfavorable value over the whole range of the corresponding statistical potential. (*B*) *a*, *b*, and *c* illustrate the three representative distances for the base-pairing, the nearest base-stacking, and the next-nearest base-stacking interactions between N2 atom of guanine and O2 atom of cytosine, respectively.

were found in the potentials between the N1 atom of adenine and the N3 atom of uracil derived based on six reference states, which were also shown in Supplemental Figure S5. Therefore, FIG-REF, RWC-REF, and SNI-REF can capture all of the base-pairing, the nearest-neighbor and the next nearest-neighbor base stacking interactions rather than Avg-REF, QChA-REF, and ASh-REF, especially FIG-REF and RWC-REF. This may be the reason why FIG-REF, RWC-REF, and SNI-REF are the top three statistical potentials in identifying native structures.

Comparison with existing statistical potentials

Here, we make the comparison with three existing well-known statistical potentials which have relatively good performances for RNA 3D structure evaluation: KB potential of all-atom version based on quasi-chemical approximation reference state (Bernauer et al. 2011), RASP of all-atom version based on averaging reference state (Capriotti et al. 2011), and 3dRNAscore with involving torsion angle potential and based on averaging reference state (Wang et al. 2015). The data of 3dRNAscore, RASP, and KB potentials for the comparisons on test set I and test set II are directly taken from Bernauer et al. (2011) and Wang et al. (2015). Since the existing statistical potentials have not been examined against test set III completely and there is no complete data available for comparisons on test set III, we only compare their performances with QChA-REF, FIG-REF, SNI-REF, and RWC-REF for test set I and test set II including four subsets.

As shown in Figure 6A, the numbers of identified native structures for test set I by various statistical potentials follow the order: 3dRNAscore $\gtrsim$ QChA-REF = FIG-REF = SNI-REF = RWC-REF > KB > RASP. Overall, these statistical potentials all have excellent performance except for KB and RASP, which can identify 80 and 79 native structures for 85 RNAs in test set I. For test set II, the numbers of identified native structures by the statistical potentials follow the order: RWC-REF3 $\gtrsim$ dRNAscore = KB = RASP = FIG-REF > QChA-REF = SNI-REF. Furthermore, we examined the statistical potentials through calculating ES values. As shown in Figure 6B, the mean ES value of 3dRNAscore is very similar to that of QChA-REF, FIG-REF, SNI-REF, and RWC-REF for the MD subset, and RASP and KB potential have apparently lower ES values. For the NM subset, 3dRNAscore has a very slightly higher ES value than QChA-REF, SNI-REF, and RWC-REF, while KB and RASP both have visibly lower ES values. For the FARNA subset, QChA-REF and SNI-REF have very slightly higher ES values than 3dRNAscore, and the ES values of KB, RASP, and FIG-REF are slightly lower

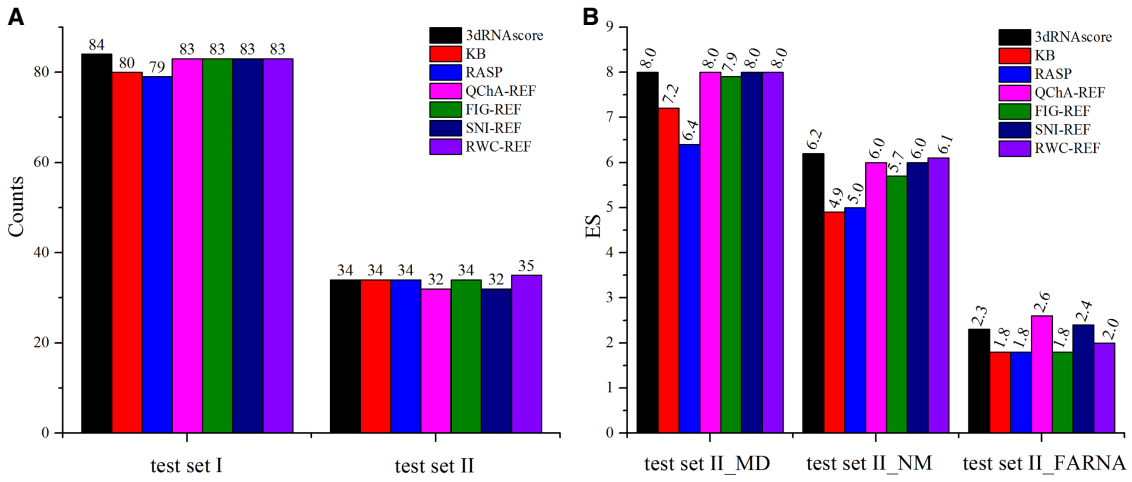

FIGURE 6. (*A*) The numbers of identified native structures for test set I and test set II and (*B*) the average ES values for test set I and test set II calculated by three existing statistical potentials: 3dRNAscore (Wang et al. 2015), KB potential (Bernauer et al. 2011), and RASP (Capriotti et al. 2011), and four current statistical potentials: QChA-REF, the quasi-chemical approximation reference state (Lu and Skolnick 2001); FIG-REF, the finite-ideal-gas reference state (Zhou and Zhou 2002); SNI-REF, the spherical-noninteracting reference state (Shen and Sali 2006); RWC-REF, the random-walk-chain reference state (Zhang and Zhang 2010). The data of 3dRNAscore, KB potential, and RASP are taken from Bernauer et al. (2011) and Wang et al. (2015). Test set I consists of decoy structures of 85 RNAs (Capriotti et al. 2011) and test set II is composed of decoy structures of 40 RNAs (Das et al. 2010; Bernauer et al. 2011).

than others. The detailed data in Figure 6 are also shown in Supplemental Tables S5 and S6.

From the above, the 3dRNAscore is similar to the present statistical potentials of FIG-REF and RWC-REF in identifying native structures for test sets I and II, and the other existing statistical potentials of KB and RASP outperform slightly worse than others. For identifying near-native structures, the 3dRNAscore is similar to QChA-REF and SNI-REF, which have a very slightly better performance than other potentials. It is noted that RASP and KB potential have visibly lower ES values. Therefore, the 3dRNAscore has a similar performance to the top statistical potentials derived from the above described six reference states, while RASP and KB have visibly lower performance than others. In the following, we try to understand the difference in performance between the existing statistical potentials (KB, RASP, and 3dRNAscore) and those built in the current work. First, it was mentioned above that RASP, 3dRNAscore, and Avg-REF are all based on the averaging reference state. However, RASP of the all-atom version is only for 23 clustered atom types, while 3dRNAscore and Avg-REF both are for 85 heavy atom types. The significantly lower resolution of atom types might be the main reason that the performance of RASP is visibly lower than others. In contrast, the (relatively good) performance of 3dRNAscore (Wang et al. 2015) might be attributed to the explicit emphasis on the local torsional structure feature of RNA backbone, which is important for RNA 3D structures. Second, KB and QChA-REF are both based on the quasi-chemical approximation reference state and for 85 heavy atom types, while their training sets consist of 77 and 108 RNA native structures, respectively. Consequently, the lower performance of KB than QChA-REF possibly arises from the different training sets, and the effect of the training set will be discussed in the following subsection.

Effect of training set

The training set involved in this work is a nonredundant set of 108 RNAs excluding RNA structures in RNA–protein and RNA–DNA complexes. We examine the effect of the training set on the performance of a statistical potential, by involving the 3dRNAscore training set, which is composed of 317 RNAs and can be downloaded from http://biophy.hust.edu.cn/3dRNAscore.html (Wang et al. 2015).

As shown in Figure 7A, the statistical potentials trained by 3dRNAscore and the present training sets can identify similar numbers of native structures for test set I, while for test set II, the statistical potentials from the present training set can identify slightly more native structures. Furthermore, we examined the effect of training sets by calculating ES values for test set II. As shown in Figure 7B, the statistical potentials based on the current training set generally have slightly higher ES values than those based on the 3dRNAscore training set, except for SNI-REF for the FARNA subset. For the MD subset, the mean ES value can decrease by ~0.45 if the current training set is replaced by the 3dRNAscore training set. Such a decrease in mean ES value becomes ~0.3 for the NM subset and ~0.08 for the FARNA subset, respectively. It is not strange that the statistical potentials from the current

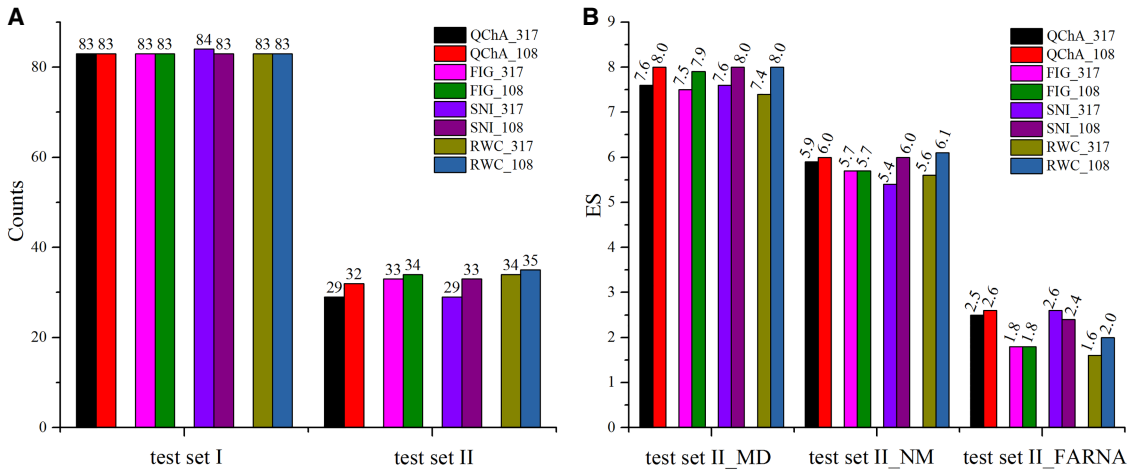

FIGURE 7. (*A*) The numbers of identified native structures for test set I and test set II and (*B*) the average ES values for test set I and test set II calculated by the statistical potentials from the 3dRNAscore training set with 317 RNAs (Wang et al. 2015) and the present training set with 108 RNAs, respectively: QChA-REF, the quasi-chemical approximation reference state (Lu and Skolnick 2001); FIG-REF, the finite-ideal-gas reference state (Zhou and Zhou 2002); SNI-REF, the spherical-noninteracting reference state (Shen and Sali 2006); RWC-REF, the random-walk-chain reference state (Zhang and Zhang 2010). Here, for simplicity, we use 317 and 108 to denote 3dRNAscore and the current training sets, respectively. In addition, for FIG_317, the parameter α was taken as 1.0 to build a relatively uniform atom distribution for 317 RNA spheres (Zhou and Zhou 2002), and for RWC_317, the Kuhn length l was also set to 310. Test set I consists of decoy structures of 85 RNAs (Capriotti et al. 2011), and test set II is composed of decoy structures of 40 RNAs (Das et al. 2010; Bernauer et al. 2011).

training set with 108 RNAs are slightly better than those from the 3dRNAscore training set of 317 RNAs. This may be because our training set excludes those RNAs complexed with protein or DNA, although the number of RNAs in our training set is much smaller than that in the 3dRNAscore training set. Note that after removing those RNA structures complexed with DNA or protein in the 3dRNAscore training set, 42 RNA structures remained. However, it needs to be noted that the current nonredundant training set of 108 RNAs may still be inadequate for training a satisfactory statistical potential. Additionally, with the increasing number of RNA structures in the PDB database (Rose et al. 2017), the statistical potentials can be further improved in the future. Finally, a high-quality training set is required for generating a high-performance statistical potential, and it is also necessary to examine the minimal number of RNA structures for a satisfactory training set when there are plenty of available RNA 3D structures in the PDB database. The detailed data in Figure 7 were also presented in Supplemental Tables S7 and S8.

DISCUSSION

Significance of reference states

The ideal reference state for statistical potentials refers to the state in which there are no interactions between atoms, and such a state should contain all possible RNA chain conformations in phase space, including extended and compact ones (Rykunov and Fiser 2007). The purpose of a statistical potential involving a reference state is for extracting the structural features to distinguish the native structure from a set of nonnative conformations (Samudrala and Moult 1998), or for guiding RNA folding/structure predictions. The six reference states examined in this work can be classified into two types: based on measured RNA 3D structures in the PDB database and based on physical modeling. The former includes Avg-REF (averaging reference state) (Samudrala and Moult 1998), QChA-REF (quasi-chemical approximation reference state) (Lu and Skolnick 2001), and ASh-REF (atom-shuffled reference state) (Rykunov and Fiser 2007), which use the structures in the PDB database to simulate reference states through summation on atom types, involving molar fraction of atom types and shuffling atom/residue types, respectively. In contrast, the latter includes FIG-REF (finite-ideal-gas reference state) (Zhou and Zhou 2002), SNI-REF (spherical-noninteracting reference state) (Shen and Sali 2006), and RWC-REF (random-walk-chain reference state) (Zhang and Zhang 2010), which discard the 3D structures in the PDB database and use an ideal gas model, spherical cluster model, and random-walk-chain to simulate reference states. The performances of the six statistical potentials should be tightly related to the corresponding reference states.

The principles of the 3D-structure-based reference states are similar to each other by using various methods of mixing atom types for 3D structures deposited in the PDB database, and consequently the corresponding performances are also similar. In the physical-model-based reference states, various physical models such as random-walk-chain and finite-ideal-gas are approximated as

reference states, which can better capture the conformation space than the 3D-structure-based reference states despite the simplification of local structure details. Furthermore, the physical-model-based reference states involve undetermined parameters in simulating reference states, e.g., dimension parameter α in FIG-REF and Kuhn length l in RWC-REF, which may also contribute to the higher performance of FIG-REF and RWC-REF than other statistical potentials. However, the above described 3D-structure-based reference states are based only on the ensemble of native structures, which naturally excludes a huge number of nonnative conformations in phase space. Thus, the 3D-structure-based reference states would differ significantly from the ideal reference states in conformation space, and consequently the corresponding statistical potentials do not work very well. It is also noted that the performance of FIG-REF and RWC-REF does not reach a satisfactory level (e.g., for the latter three test subsets), which may be attributed to the oversimplified structure models, e.g., ideal gas or random-walk-chain.

Effect of the origin of test sets

As shown in the Results section, a statistical potential can have different performances for different test sets, and as shown in Supplemental Table S9, even for the decoys from different test sets with the same native structures, the same statistical potential can have distinctly different performances. For example, for some RNAs (PDB IDs: 1kka, 1qwa, 1xjr, 28sp, and 2f88), the statistical potentials have overall much better performance for the NM subset than for the FARNA subset. This may come from the different origins for generating test sets (e.g., perturbation method for the NM subset [Bernauer et al. 2011] and fragment assembly for the FARNA subset [Das and Baker 2007]). For test set I, MD and NM subsets in test set II, the six statistical potentials overall have relatively good performance in identifying native structures and ranking near-native structures. These three subsets were produced by MODELLER with Gaussian constraints (Capriotti et al. 2011), replica-exchange MD (Bernauer et al. 2011), and the normal mode perturbation method (Bernauer et al. 2011), respectively, and thus the produced decoy structures are generally very close to the native structures. For example, as shown in Figure 1A,B, the rmsds of decoy structures in test set I, test set II_MD, and test set II_NM are mainly in the range of 0–6, 0–3, and 1–5 Å, respectively. In contrast, the rmsds of decoy structures in test set II_FARNA produced by the fragment assembly with FARNA (Das and Baker 2007) are quite dispersed in the range of 4–15 Å. Thus, all six statistical potentials overall have unsatisfactory performance for test set II_FARNA.

Test set III_FARFAR from the fragment assembly with FARFAR (Das et al. 2010) is composed of small RNA segments of 6–23 nt, and the rmsds of the decoy structures are rather dispersed relative to their length (Fig. 1C). Test set III_RNA-Puzzles is a special test set with a small number of decoys of large RNAs that were from the blind CASP-like RNA 3D structure predictions of various computational models (Miao et al. 2017), and the rmsds of the decoy structures are very dispersed, e.g., the rmsd distribution of this subset extends to ~34 Å. Thus, test set III is very challenging and the six statistical potentials overall have unsatisfactory performances on test set III_FARFAR and test set III_RNA-Puzzles. Since the purpose of a statistical potential is for realistic application, the test set from realistic predictions such as the RNA-Puzzles subset can be a more realistic examination for a statistical potential, rather than other test sets by near-native perturbation methods.

Limitation of current statistical potentials and perspective

First, the number of RNA molecules in a nonredundant RNA training set is still limited, which leads to insufficient data information for training a statistical potential. This is an inevitable problem for building all kinds of statistical potentials, and the problem can be gradually overcome in the future with the increase of the number of RNA structures deposited in the PDB database. Second, as mentioned above, the six reference states are either based on RNA native structures deposited in PDB or based on ideal physical models. Such oversimplified approximations may cause the produced reference states to significantly deviate from the ideal reference state, which may be the main reason for the unsatisfactory performance on the decoys with dispersed rmsd distributions. A more advanced approximation for a reference state is still highly required. For example, for circumventing the reference states, Huang and Zou developed an iterative method to build a scoring function for protein–ligand docking, and such an iterative method may be useful for building a statistical potential for RNA structure evaluation (Huang and Zou 2006a,b, 2008, 2011). Third, knowledge-based statistical potentials have been widely used and proven to be effective in protein structure evaluation. However, RNA structure is distinctively different from protein, and the involvement of RNA structure characteristics in a statistical potential may improve its performance. For example, RNAs are strongly charged polyanionic chains, whose structure can be affected by intrachain Columbic repulsion. Thus, RNA 3D structures can be highly sensitive to ion conditions (Tan and Chen 2010, 2011; Wu et al. 2015; Xi et al. 2018), and the involvement of ion-electrostatic interaction might be helpful for the performance of a statistical potential. Finally, the statistical potentials are generally pairwise for different atom pair types, while in principle the effect of other atoms is already involved. Consequently, the summation on such pairwise potentials for calculating total energy would bring double-counting. The development of many-body

statistical potentials (e.g., through developing many-body contact potentials to supplement the pairwise statistical potentials; Li and Liang 2010) or removal of the effect of other atoms in building a pairwise statistical potential might bring the improvement of the statistical potential on RNA 3D structure evaluation.

Conclusions

In this work, we used six representative reference states widely used for proteins to construct statistical potentials for RNA 3D structure evaluation, and we have made extensive comparisons between them against three test sets including six subsets. We found that, overall, on identifying native structures and ranking decoy structures, the performances of FIG-REF and RWC-REF are slightly better than other ones and on identifying near-native structures, very slight difference exists between the six reference states. In addition, compared with three existing RNA statistical potentials, the top statistical potentials derived from six reference states have a similar performance to 3dRNAscore (Wang et al. 2015), while RASP (Capriotti et al. 2011) and KB (Bernauer et al. 2011) have a visibly lower performance than others. Furthermore, the performance of a statistical potential could be apparently dependent on the training set. Finally, we found that the performance of a statistical potential is closely related to the origin of the test sets.

However, the overall performance of the six statistical potentials is still not at a high level for realistic test sets from structure prediction models, and thus an applicable statistical potential with high performance still remains to be improved, through proposing more realistic reference states, circumventing the problem of reference states, or combining a physical potential. Besides, previous studies for proteins show that the combination of structural clustering may improve the performance of statistical potentials (Zhang and Skolnick 2004; Zhang 2009; Xu et al. 2011; Deng et al. 2012). Furthermore, involving the unique characteristics of RNA, such as local structure feature (Wang et al. 2015) or ion electrostatic interactions in a statistical potential, can possibly improve its performance. Moreover, a multibody statistical potential (Singh et al. 1996; Feng et al. 2007; Masso 2018) can possibly capture more structural features than conventional pairwise ones, while generally involving a higher computational cost. Finally, machine-learning methods can be applied in building the statistical potential to dig critical information not easily detected for RNA structures (Li et al. 2018).

Nevertheless, this work presents a comprehensive and critical survey on the performances of the existing reference states and statistical potentials for RNA 3D structure evaluation. Therefore, the present study can be considered as a benchmark work and can serve as a basis for further development on advanced knowledge-based statistical potentials of high performance for RNA 3D structure evaluation.

MATERIALS AND METHODS

Knowledge-based statistical potential

A knowledge-based statistical potential is generally derived based on Boltzmann or Bayesian formulations, and any kind of structure features that are able to distinguish the native conformation from a set of structural decoys can be used to derive a statistical potential (Samudrala and Moult 1998). Here, we still focus on a conventional all-heavy atom distance-dependent statistical potential, which can be given by (Deng et al. 2012)

$$u_{ij}(r) = -k_B T \ln \frac{f_{ij}^{obs}(r)}{f_{ij}^{ref}(r)}, \quad (3)$$

where k_B is the Boltzmann constant, T is the Kelvin temperature, $f_{ij}^{obs}(r)$ is the observed probability for the pair of atom types i and j residing within the distance bin of $[r,r+dr]$:

$$f_{ij}^{obs}(r) = \frac{N_{ij}^{obs}(r)}{N_{ij}^{obs}}. \quad (4)$$

$f_{ij}^{ref}(r)$ is the probability for the pair of atom types i and j within the distance bin of $[r,r+dr]$ from a conformation ensemble of the so-called reference state (Deng et al. 2012), and in principle, an ideal reference state can be obtained from a nonredundant and complete reference decoy conformation ensemble where interactions between atoms are assumed to vanish:

$$f_{ij}^{ref}(r) = \frac{N_{ij}^{ref}(r)}{N_{ij}^{ref}}. \quad (5)$$

Unfortunately, an ideal reference decoy database might not exist (Deng et al. 2012). Hence, people generally use various approximations based on experimental structure database or statistical physics models to build the reference states (Samudrala and Moult 1998; Lu and Skolnick 2001; Zhou and Zhou 2002; Shen and Sali 2006; Rykunov and Fiser 2007; Zhang and Zhang 2010; Deng et al. 2012). For building statistical potentials for proteins, there have been six reference states, which are introduced briefly as follows: averaging (Samudrala and Moult 1998), quasi-chemical approximation (Lu and Skolnick 2001), atom-shuffled (Rykunov and Fiser 2007), finite-ideal-gas (Zhou and Zhou 2002), spherical-noninteracting (Shen and Sali 2006), and random-walk-chain (Zhang and Zhang 2010) reference states.

Reference states

Averaging reference state

The averaging reference state was developed by Samudrala and Moult (1998). They used the average distribution of all kinds of atom pair types in experiment structures to approximately simulate the distribution of different atom pair types in the reference state. Thus, the probability $f_{ij}^{ref}(r)$ can be approximated as

(Samudrala and Moult 1998)

$$f_{ij}^{\text{ref}}(r) = \frac{N_{ij}^{\text{ref}}(r)}{N_{ij}^{\text{ref}}} = \frac{\sum_{ij} N_{ij}^{\text{obs}}(r)}{\sum_{ij}\sum_{r} N_{ij}^{\text{obs}}(r)} = \frac{N^{\text{obs}}(r)}{N^{\text{obs}}}, \quad (6)$$

where $N^{\text{obs}}(r)$ is the observed number of atom pairs within the distance bin of $[r,r+dr]$ regardless of atom types. N^{obs} is the total number of atom pairs over all distance bins. Given an RNA structure database, the averaging reference state is easy to use to derive a statistical potential.

Quasi-chemical approximation reference state

Considering that the counts of a certain atom pair type of the ideal reference state should be proportional to the mole fraction of the corresponding one in the experiment structures, Lu and Skolnick (2001) proposed the quasi-chemical approximation reference state, and $f_{ij}^{\text{ref}}(r)$ can be calculated by (Lu and Skolnick 2001)

$$f_{ij}^{\text{ref}}(r) = \frac{N_{ij}^{\text{ref}}(r)}{N_{ij}^{\text{ref}}} = \frac{x_i x_j N^{\text{obs}}(r)}{N_{ij}^{\text{obs}}}. \quad (7)$$

Here x_i is the mole fraction of atom type i and can be obtained from a whole experimental structure database. N_{ij}^{obs} is the number of the pair of atom types i and j over all the distance bins, and N^{obs} is the total number of atom pairs over all distance bins.

Atom-shuffled reference state

Shuffled reference states were proposed by Rykunov and Fiser (2007) to simulate the reference decoys. There are three shuffling modes, including residue-shuffled, sequence-shuffled, and atom-shuffled (Rykunov and Fiser 2007). Here, we used the atom-shuffled reference state in which all the atoms are fixed in coordinates while randomly exchanged in the identity of them, and we shuffled every experimental structure more than 3 million times, randomly. Then, $f_{ij}^{\text{ref}}(r)$ can be given by Rykunov and Fiser (2007):

$$f_{ij}^{\text{ref}}(r) = \frac{N_{ij}^{\text{shuffled}}(r)}{N_{ij}^{\text{shuffled}}}, \quad (8)$$

where $N_{ij}^{\text{shuffled}}(r)$ and N_{ij}^{shuffled} are acquired from all the structures after being shuffled. This reference state provides an extremely stochastic reference conformation space and eliminates the effect of the connection of chemical bond.

Finite-ideal-gas reference state

Zhou and Zhou (2002) proposed the finite-ideal-gas reference state by applying the pair distribution function to the protein macromolecule system. The pair distribution function is written as (Friedman 1985)

$$g_{ij}(r) = \frac{N_{ij}^{\text{obs}}(r)/4\pi r^2\Delta r}{N_i N_j/V}, \quad (9)$$

where $N_{ij}^{\text{obs}}(r)$ is the observed number of pairs of atoms types i and j within the spherical shell of the radius bin of $[r,r+dr]$. N_i and N_j are the total number of atom types i and j over all the distance bins, respectively. V is the volume of a protein system. The atomic pairwise potential $u_{ij}(r)$ is equal to $-k_B T \ln g_{ij}(r)$ (Friedman 1985), and $u_{ij}(r)$ can be expressed as follows:

$$u_{ij}(r) = -k_B T \ln \frac{N_{ij}^{\text{obs}}(r)/N_{ij}^{\text{obs}}}{N_i N_j 4\pi r^2 \Delta r/V N_{ij}^{\text{obs}}}. \quad (10)$$

In general, for the distance between two considered atoms longer than the cutoff distance r_c, the interaction between them would decrease to zero. That is, $u_{ij}(r)\approx 0$ for $r \geq r_c$. Thus, $f_{ij}^{\text{ref}}(r)$ can be given by (Zhou and Zhou 2002)

$$f_{ij}^{\text{ref}}(r) = \frac{N_{ij}^{\text{obs}}(r_c)}{N_{ij}^{\text{obs}}}\left(\frac{r}{r_c}\right)^{\alpha} = f_{ij}^{\text{obs}}(r_c)\left(\frac{r}{r_c}\right)^{\alpha}. \quad (11)$$

Here, α is a dimension parameter since the systems of macromolecules are not ideal gases even at high temperature. In our RNA system, α was taken as 1.39 to build a relatively uniform atom distribution for all RNA spheres (Zhou and Zhou 2002), and the relative fluctuation of the atom distribution function has been shown in Supplemental Figure S1.

Spherical-noninteracting reference state

Shen and Sali (2006) developed the spherical-noninteracting reference state for proteins, in which a native structure is represented by a sample sphere with the same radius of gyration as the native structure. Thus, this reference state takes the native structures of different sizes into account. $f_{ij}^{\text{ref,p}}(r)$ can be expressed by (Shen and Sali 2006)

$$f_{ij}^{\text{ref,p}}(r) = f^{\text{ref,p}}(r,a) = \begin{cases} \dfrac{3r^2(r-2a)^2(r+4a)}{16a^6} & r_c > 2a; \\ \dfrac{6r^2(r-2a)^2(r+4a)}{r_c^3(r_c^3 - 18a^2 r_c + 32a^3)} & r_c \leq 2a, \end{cases} \quad (12)$$

where a is defined as the radius of an effective sphere, which has the same radius of gyration (R_g) as the experimental protein structure. Here, r_c also means the cutoff distance. Based on this reference state, one needs to make the statistics from the experimental structures one by one, and p represents a sampled protein structure. Thus, the final statistical potential can be calculated by (Shen and Sali 2006)

$$u_{ij}(r) = -k_B T \ln\left[\sum_{p} w_p \frac{N_{ij}^{\text{obs,p}}(r)}{f^{\text{ref,p}}(r,a) N_{ij}^{\text{obs,p}}}\right], \quad (13)$$

where the weight w_p of the sampled experiment structure is given by the ratio between the number of all atom pairs in this sampled structure and the number of atom pairs in all samples, regardless of the pair types.

Random-walk-chain reference state

The random-walk-chain reference state was proposed by Zhang and Zhang (2010) to simulate the inherent connectivity of protein chains. According to the polymer theory of freely joined chain models, $f_{ij}^{\text{ref,p}}(r)$ can be written as (Zhang and Zhang 2010)

$$f_{ij}^{\text{ref,p}}(r) = f^{\text{ref,p}}(r) = \int f^{\text{ref,p}}(r, n) dn = \sum_{n=1}^{N} 4\pi r^2 \left(\frac{3}{2\pi n l^2}\right)^{3/2} \exp\left(-\frac{3r^2}{2nl^2}\right)\Delta r, \quad (14)$$

where N is the number of residues in an experimental structure, l is Kuhn length, and l is considered as an adjustable parameter to match the scale of free-joint chain to that of a realistic protein chain. In addition, p also represents a sampled protein structure, and $f_{ij}^{\mathrm{ref}}(r)$ is equal to the sum of $f_{ij}^{\mathrm{ref},p}(r)$ over all the experimental protein structures. Similarly, $u_{ij}(r) \approx 0$ at $r = r_c$, and $f_{ij}^{\mathrm{ref}}(r)$ can be given by (Zhang and Zhang 2010)

$$f_{ij}^{\mathrm{ref}}(r) = \sum_{p} f_{ij}^{\mathrm{obs},p}(r_c)\left(\frac{r}{r_c}\right)^2 \frac{\sum_{n=1}^{N} \exp(-3r^2/2nl^2)/n^{3/2}}{\sum_{n=1}^{N} \exp(-3r_c^2/2nl^2)/n^{3/2}}. \quad (15)$$

In this work for RNAs, the value of l^2 is set to 310, in which case the potential based on random-walk-chain has the best performance for RNAs.

Training set and parameters

In this work, we established our nonredundant training set based on the RNA 3D Hub nonredundant set (Release 2.121, 2017-03-31), which can ensure that the sequence identity between any two chains in the set is <95% (Leontis and Zirbel 2012). First, we collected 1245 representative RNA chains of all the different clusters with X-ray resolution <3.5 Å from RNA 3D Hub list (Release 2.121, 2017-03-31), which can be downloaded from http://rna.bgsu.edu/rna3dhub/nrlist. This list shows that all of the RNA redundant chains have been divided into many clusters, and each cluster has a representative RNA chain. Next, what we needed to do was discard the representative chain whose structure complexes with protein or DNA and replace it with another member in this cluster whose structure is without protein and DNA, to avoid the possibly significant influence of complexed proteins or DNAs on RNA structures. Afterward, we removed the RNA structures with sequence identity >80% and coverage >80% using the BLASTN program (Altschul et al. 1990). However, since sequence identity is not equal to structure identity, and at sequence matching regions, we still kept those RNAs that have different 2D structures at sequence matching regions, even though the value of their sequence identity reaches the criterion. The 2D structure of different RNAs can be viewed from http://rnafrabase.cs.put.poznan.pl/ (Popenda et al. 2010). Finally, through the prior operation steps, our training set contained 108 RNA structures and we downloaded them from the Protein Data Bank (PDB) (Rose et al. 2017) in the form of biological assembly, which is believed to be the functional form of the macromolecule (Leontis and Zirbel 2012). It should be pointed out that our training set does not contain RNAs in test set I, test set II and the FARFAR subset in test set III, while there are 10 complicated native structures for the RNA-Puzzles subset in test set III. These 10 RNAs are riboswitches (3OX0, 4GXY, 4L81, 4QLM, 4XWF), ribozymes (4R4V, 5EAQ), exonuclease resistant RNA (5TPY), RNA Nanosquare (3P59), and regulatory motifs from mRNA (3MEI). The PDB IDs of these 108 RNAs are presented in Supplemental Table S1, and the PDB IDs of the 10 RNAs in the RNA-Puzzles subset are bolded. In building the six statistical potentials, 85 heavy atom types were considered. The distance cutoff was set to 20 Å and the distance bin width was taken as 0.3 Å, according to a previous study (Wang et al. 2015). For the situation that some atom pairs were not observed within a certain bin width, the corresponding potentials were set to the value of highest potential in the whole potential, and k_BT was taken as the unit of potential energy in this work. Also, for convenience, we used the following abbreviations to represent six different reference states: Avg-REF (averaging reference state) (Samudrala and Moult 1998), QChA-REF (quasi-chemical approximation reference state) (Lu and Skolnick 2001), AShREF (atom-shuffled reference state) (Rykunov and Fiser 2007), FIG-REF (finite-ideal-gas reference state) (Zhou and Zhou 2002), SNI-REF (spherical-noninteracting reference state) (Shen and Sali 2006), and RWC-REF (random-walk-chain reference state) (Zhang and Zhang 2010).

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We are grateful to Professors Shi-Jie Chen (University of Missouri) and Jian Zhang (Nanjing University) for valuable discussions. This work was supported by grants from the National Science Foundation of China (11774272, 11575128, and 11605125). Parts of the numerical calculations in this work were performed on the supercomputing system in the Supercomputing Center of Wuhan University.

Received December 8, 2018; accepted April 6, 2019.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410. doi:10.1016/S0022-2836(05)80360-2

Amarasinghe GK, De Guzman RN, Turner RB, Summers MF. 2000. NMR structure of stem-loop SL2 of the HIV-1 Ψ RNA packaging signal reveals a novel A-U-A base-triple platform. *J Mol Biol* **299:** 145–156. doi:10.1006/jmbi.2000.3710

Aviv T, Amborski AN, Zhao XS, Kwan JJ, Johnson PE, Sicheri F, Donaldson LW. 2006. The NMR and X-ray structures of the *Saccharomyces cerevisiae* Vts1 SAM domain define a surface for the recognition of RNA hairpins. *J Mol Biol* **356:** 274–279. doi:10.1016/j.jmb.2005.11.066

Baird NJ, Ludtke SJ, Khant H, Chiu W, Pan T, Sosnick TR. 2010. Discrete structure of an RNA folding intermediate revealed by cryo-electron microscopy. *J Am Chem Soc* **132:** 16352–16353. doi:10.1021/ja107492b

Bell DR, Cheng SY, Salazar H, Ren P. 2017. Capturing RNA folding free energy with coarse-grained molecular dynamics simulations. *Sci Rep* **7:** 45812. doi:10.1038/srep45812

Bernauer J, Huang X, Sim AYL, Levitt M. 2011. Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA* **17:** 1066–1075. doi:10.1261/rna.2543711

Bian Y, Zhang J, Wang J, Wang J, Wang W. 2015. Free energy landscape and multiple folding pathways of an H-type RNA pseudoknot. *PLos One* **10:** e0129089. doi:10.1371/journal.pone.0129089

Boniecki M J, Lach G, Dawson W K, Tomala K, Lukasz P, Soltysinski T, Rother K M, Bujnicki JM. 2016. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res* **44:** e63. doi:10.1093/nar/gkv1479

Cabello-Villegas J, Winkler ME, Nikonowicz EP. 2002. Solution conformations of unmodified and $A_{37}N^6$-dimethylallyl modified

anticodon stem-loops of Escherichia coli $tRNA^{Phe}$. *J Mol Biol* **319:** 1015–1034. doi:10.1016/S0022-2836(02)00382-0

Cao S, Chen SJ. 2011. Physics-based de novo prediction of RNA 3D structures. *J Phys Chem B* **115:** 4216–4226. doi:10.1021/jp112059y

Capriotti E, Norambuena T, Martirenom MA, Melo F. 2011. All-atom knowledge-based potential for RNA structure prediction and assessment. *Bioinformatics* **27:** 1086–1093. doi:10.1093/bioinformatics/btr093

Cragnolini T, Derreumaux P, Pasquali S. 2013. Coarse-grained simulations of RNA and DNA duplexes. *J Phys Chem B* **117:** 8047–8060. doi:10.1021/jp400786b

Cruz JA, Blanchet MF, Boniecki M, Bujnicki JM, Chen SJ, Cao S, Das R, Ding F, Dokholyan NV, Flores SC, et al. 2012. RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA* **18:** 610–625. doi:10.1261/rna.031054.111

Das R, Baker D. 2007. Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci* **104:** 14664–14669. doi:10.1073/pnas.0703836104

Das R, Karanicolas J, Baker D. 2010. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat Methods* **7:** 291–294. doi:10.1038/nmeth.1433

DeBolt SE, Skolnick J. 1996. Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. *Protein Eng* **9:** 637–655. doi:10.1093/protein/9.8.637

Deng H, Jia Y, Wei Y, Zhang Y. 2012. What is the best reference state for designing statistical atomic potentials in protein structure prediction? *Proteins* **80:** 2311–2322. doi:10.1002/prot.24121

Dethoff EA, Chugh J, Mustoe AM, Alhashimi HM. 2012. Functional complexity and regulation through RNA dynamics. *Nature* **482:** 322–330. doi:10.1038/nature10885

Ding F, Sharma S, Chalasani P, Demidov VV, Broude NE, Dokholyan NV. 2008. Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA* **14:** 1164–1173. doi:10.1261/rna.894608

Feng Y, Kloczkowski A, Jernigan RL. 2007. Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. *Proteins* **68:** 57–66. doi:10.1002/prot.21362

Friedman HL. 1985. *A course in statistical mechanics*. Prentice-Hall, Englewood Cliffs, NJ.

Guttman M, Rinn JL. 2012. Modular regulatory principles of large noncoding RNAs. *Nature* **482:** 339–346. doi:10.1038/nature10887

Huang SY, Zou XQ. 2006a. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J Comput Chem* **27:** 1866–1875. doi:10.1002/jcc.20504

Huang SY, Zou XQ. 2006b. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J Comput Chem* **27:** 1876–1882. doi:10.1002/jcc.20505

Huang SY, Zou XQ. 2008. An iterative knowledge-based scoring function for protein-protein recognition. *Proteins* **72:** 557–579. doi:10.1002/prot.21949

Huang SY, Zou XQ. 2011. Statistical mechanics-based method to extract atomic distance-dependent potentials from protein structures. *Proteins* **79:** 2648–2661. doi:10.1002/prot.23086

Jain S, Schlick T. 2017. F-RAG: generating atomic coordinates from RNA graphs by fragment assembly. *J Mol Biol* **429:** 3587–3605. doi:10.1016/j.jmb.2017.09.017

Jin L, Shi YZ, Feng CJ, Tan YL, Tan ZJ. 2018. Modeling structure, stability and flexibility of double-stranded RNAs in salt solutions. *Biophys J* **115:** 1403–1416. doi:10.1016/j.bpj.2018.08.030

Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D, Altman RB. 2009. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA* **15:** 189–199. doi:10.1261/rna.1270809

Jossinet F, Ludwig TE, Westhof E. 2010. Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics* **26:** 2057–2059. doi:10.1093/bioinformatics/btq321

Kim N, Laing C, Elmetwaly S, Jung S, Curuksu J, Schlick T. 2014. Graph-based sampling for approximating global helical topologies of RNA. *Proc Natl Acad Sci* **111:** 4079–4084. doi:10.1073/pnas.1318893111

Leontis N, Zirbel C. 2012. Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In *RNA 3D structure analysis and prediction* (ed. Leontis N, Westhof E), Vol. 27, pp. 281–298. Springer, Berlin, Heidelberg.

Li X, Liang J. 2010. Geometric cooperativity and anticooperativity of three-body interactions in native proteins. *Proteins* **60:** 46–65. doi:10.1002/prot.20438

Li J, Zhang J, Wang J, Li W, Wang W. 2016. Structure prediction of RNA loops with a probabilistic approach. *PLoS Comput Biol* **12:** e1005032. doi:10.1371/journal.pcbi.1005032

Li J, Zhu W, Wang J LW, Gong S, Zhang J, Wang W. 2018. RNA3DCNN: local and global quality assessments of RNA 3D structures using 3D deep convolutional neural network. *PLoS Comput Biol* **14:** e1006514. doi:10.1371/journal.pcbi.1006514

Lu H, Skolnick J. 2001. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins* **44:** 223–232. doi:10.1002/prot.1087

Magnus M, Boniecki MJ, Dawson W, Bujnicki JM. 2016. SimRNAweb: a web server for RNA 3D structure modeling with optional restraints. *Nucleic Acids Res* **44:** W315–W319. doi:10.1093/nar/gkw279

Major F, Turcotte M, Gautheret D. 1991. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science* **253:** 1255–1260. doi:10.1126/science.1716375

Masso M. 2018. All-atom four-body knowledge-based statistical potential to distinguish native tertiary RNA structures from nonnative folds. *J Theor Biol* **453:** 58–67. doi:10.1016/j.jtbi.2018.05.022

Miao Z, Westhof E. 2017. RNA structure: advances and assessment of 3D structure prediction. *Annu Rev Biophys* **46:** 483–503. doi:10.1146/annurev-biophys-070816-034125

Miao Z, Adamiak RW, Antczak M, Batey RT, Becka AJ, Biesiada M, et al. 2017. RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA* **23:** 655–672. doi:10.1261/rna.060368.116

Montange RK, Batey RT. 2008. Riboswitches: emerging themes in RNA structure and function. *Annu Rev Biophys* **37:** 117–133. doi:10.1146/annurev.biophys.37.032807.130000

Parisien M, Major F. 2008. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452:** 51. doi:10.1038/nature06684

Popenda M, Szachniuk M, Blazewicz M, Wasik S, Burke EK, Blazewicz J, Adamiak RW. 2010. RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics* **11:** 231. doi:10.1186/1471-2105-11-231

Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, Blazewicz J, Adamiak RW. 2012. Automated 3D structure composition for large RNAs. *Nucleic Acids Res* **40:** e112. doi:10.1093/nar/gks339

Rose PW, Prlic A, Altunkaya A, Bi C, Bradley AR, Christie CH, Di Costanzo L, Duarte JM, Dutta S, Feng Z, et al. 2017. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res* **45:** D271–D281. doi:10.1093/nar/gkw1042

Rother M, Rother K, Puton T, Bujnicki JM. 2011. ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res* **39:** 4007–4022. doi:10.1093/nar/gkq1320

Rykunov D, Fiser A. 2007. Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. *Proteins* **67:** 559–568. doi:10.1002/prot.21279

Šali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234:** 779–815. doi:10.1006/jmbi.1993.1626

Samudrala R, Moult J. 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol* **275:** 895–916. doi:10.1006/jmbi.1997.1479

Schlick T, Pyle AM. 2017. Opportunities and challenges in RNA structural modeling and design. *Biophys J* **113:** 225–234. doi:10.1016/j.bpj.2016.12.037

Shen MY, Sali A. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15:** 2507–2524. doi:10.1110/ps.062416606

Shi YZ, Wu YY, Wang FH, Tan ZJ. 2014a. RNA structure prediction: progress and perspective. *Chin Phys B* **23:** 078701. doi:10.1088/1674-1056/23/7/078701

Shi YZ, Wang FH, Wu YY, Tan ZJ. 2014b. A coarse-grained model with implicit salt for RNAs: predicting 3D structure, stability and salt effect. *J Chem Phys* **141:** 105102. doi:10.1063/1.4894752

Shi YZ, Jin L, Wang FH, Zhu XL, Tan ZJ. 2015. Predicting 3D structure, flexibility, and stability of RNA hairpins in monovalent and divalent ion solutions. *Biophys J* **109:** 2654–2665. doi:10.1016/j.bpj.2015.11.006

Shi YZ, Jin L, Feng CJ, Tan YL, Tan ZJ. 2018. Predicting 3D structure and stability of RNA pseudoknots in monovalent and divalent ion solutions. *PLos Comput Biol* **14:** e1006222. doi:10.1371/journal.pcbi.1006222

Singh RK, Tropsha A, Vaisman II. 1996. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. *J Comput Biol* **3:** 213–221. doi:10.1089/cmb.1996.3.213

Sippl MJ. 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. doi:10.1016/S0022-2836(05)80269-4

Sun LZ, Zhang D, Chen SJ. 2017. Theory and modeling of RNA structure and interactions with metal ions and small molecules. *Annu Rev Biophys* **46:** 227–246. doi:10.1146/annurev-biophys-070816-033920

Tan ZJ, Chen SJ. 2010. Predicting ion binding properties for RNA tertiary structures. *Biophys J* **99:** 1565–1576. doi:10.1016/j.bpj.2010.06.029

Tan ZJ, Chen SJ. 2011. Salt contribution to RNA tertiary structure folding stability. *Biophys J* **101:** 176–187. doi:10.1016/j.bpj.2011.05.050

Thomas PD, Dill KA. 1996. Statistical potentials extracted from protein structures: How accurate are they? *J Mol Biol* **257:** 457–469. doi:10.1006/jmbi.1996.0175

Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. 2003. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins* **53:** 76–87. doi:10.1002/prot.10454

Wang J, Zhao Y, Zhu C, Xiao Y. 2015. 3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures. *Nucleic Acids Res* **43:** e63. doi:10.1093/nar/gkv141

Wang Y, Gong S, Wang Z, Zhang W. 2016. The thermodynamics and kinetics of a nucleotide base pair. *J Chem Phys* **144:** 115101. doi:10.1063/1.4944067

Wang J, Mao K, Zhao Y, Zeng C, Xiang J, Zhang Y, Xiao Y. 2017. Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide-nucleotide interactions from direct coupling analysis. *Nucleic Acids Res* **45:** 6299–6309. doi:10.1093/nar/gkx386

Wang Y, Wang Z, Liu T, Gong S, Zhang W. 2018. Effects of flanking regions on HDV cotranscriptional folding kinetics. *RNA* **24:** 1229–1240. doi:10.1261/rna.065961.118

Watson JD, Baker T, Bell S, Gann A, Levine M, Losick R. 2003. *Molecular biology of the gene*. Pearson/Benjamin Cummings, San Francisco.

Wu YY, Zhang ZL, Zhang JS, Zhu XL, Tan ZJ. 2015. Multivalent ion-mediated nucleic acid helix-helix interactions: RNA versus DNA. *Nucleic Acids Res* **43:** 6156–6165. doi:10.1093/nar/gkv570

Xi K, Wang FH, Xiong G, Zhang ZL, Tan ZJ. 2018. Competitive binding of Mg^{2+} and Na^{+} ions to nucleic acids: from helices to tertiary structures. *Biophys J* **114:** 1776–1790. doi:10.1016/j.bpj.2018.03.001

Xia Z, Bell DR, Shi Y, Ren PY. 2013. RNA 3D structure prediction by using a coarse-grained model and experimental data. *J Phys Chem B* **117:** 3135–3144. doi:10.1021/jp400751w

Xu D, Zhang J, Roy A, Zhang Y. 2011. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based *ab initio* folding and FG-MD-based structure refinement. *Proteins* **79:** 147–160. doi:10.1002/prot.23111

Xu X, Zhao P, Chen SJ. 2014. Vfold: a web server for RNA structure and folding thermodynamics prediction. *PLos One* **9:** e107504. doi:10.1371/journal.pone.0107504

Zhang Y. 2009. I-TASSER: fully automated protein structure prediction in CASP8. *Proteins* **77:** 100–113. doi:10.1002/prot.22588

Zhang D, Chen SJ. 2018. IsRNA: an iterative simulated reference state approach to modeling correlated interactions in RNA folding. *J Chem Theory Comput* **14:** 2230–2239. doi:10.1021/acs.jctc.7b01228

Zhang Y, Skolnick J. 2004. SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem* **25:** 865–871. doi:10.1002/jcc.20011

Zhang J, Zhang Y. 2010. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLos One* **5:** e15386. doi:10.1371/journal.pone.0015386

Zhang J, Bian Y, Lin H, Wang W. 2012. RNA fragment modeling with a nucleobase discrete-state model. *Phys Rev E* **85:** 021909. doi:10.1103/PhysRevE.85.021909

Zhao Y, Huang Y, Gong Z, Wang Y, Man J, Xiao Y. 2012. Automated and fast building of three-dimensional RNA structures. *Sci Rep* **2:** 734104. doi:10.1038/srep00734

Zhou HY, Zhou YQ. 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* **11:** 2714–2726. doi:10.1110/ps.0217002