Bioinformatics, 35(10), 2019, 1686–1691
doi: 10.1093/bioinformatics/bty876
Advance Access Publication Date: 15 October 2018
Original Paper

Structural bioinformatics

Enhanced prediction of RNA solvent accessibility with long short-term memory neural networks and improved sequence profiles

Saisai Sun[1], Qi Wu[1], Zhenling Peng[2,*] and Jianyi Yang [1,*]

[1]School of Mathematical Sciences, Nankai University, Tianjin 300071, China and [2]Center for Applied Mathematics, Tianjin University, Tianjin 300072, China

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on November 29, 2017; revised on September 11, 2018; editorial decision on October 11, 2018; accepted on October 13, 2018

Abstract

Motivation: The *de novo* prediction of RNA tertiary structure remains a grand challenge. Predicted RNA solvent accessibility provides an opportunity to address this challenge. To the best of our knowledge, there is only one method (RNAsnap) available for RNA solvent accessibility prediction. However, its performance is unsatisfactory for protein-free RNAs.

Results: We developed RNAsol, a new algorithm to predict RNA solvent accessibility. RNAsol was built based on improved sequence profiles from the covariance models and trained with the long short-term memory (LSTM) neural networks. Independent tests on the same datasets from RNAsnap show that RNAsol achieves the mean Pearson's correlation coefficient (PCC) of 0.43/0.26 for the protein-bound/protein-free RNA molecules, which is 26.5%/136.4% higher than that of RNAsnap. When the training set is enlarged to include both types of RNAs, the PCCs increase to 0.49 and 0.46 for protein-bound and protein-free RNAs, respectively. The success of RNAsol is attributed to two aspects, including the improved sequence profiles constructed by the sequence-profile alignment and the enhanced training by the LSTM neural networks.

Availability and implementation: http://yanglab.nankai.edu.cn/RNAsol/

Contact: yangjy@nankai.edu.cn or zhenling@tju.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The availability of RNA tertiary structure is in general required to appreciate the molecular mechanisms of RNA's functional roles, such as in gene expression and protein synthesis. However, only a limited number of RNAs have experimentally determined structures. For example, in the Protein Data Bank (PDB) (Rose *et al.*, 2017), there are 1702 entries with RNA structures, which contain 8604 chains (version of July 2018). These chain sequences are highly redundant because they can be clustered into 540 clusters at 80% pairwise sequence identity using the program cd-hit-est (Li and Godzik, 2006). It is thus urgent to develop computational algorithms for RNA structure prediction (Rother *et al.*, 2011; Sharma *et al.*, 2008; Xu *et al.*, 2014; Zhao *et al.*, 2012).

It remains a grand challenge for the *de novo* prediction of RNA structure. In the recent RNA-Puzzle experiments (Miao *et al.*, 2017), it was shown that accurate structure models could be built for targets with homologous templates. However, for targets without homologous templates, the qualities of the predicted structure models were in general poor, especially for those with large size and complex topology. More efforts are in demand for developing new algorithms, especially for the modeling of the non-Watson-Crick interactions in RNA (Miao *et al.*, 2017). Some works were done in

recent years to address this challenge. For example, inspired by the recent advances in residue–residue contact prediction (Wuyun *et al.*, 2018) and its usage in protein structure modeling (Ovchinnikov *et al.*, 2017), the nucleotide-nucleotide contacts predicted from the direct-coupling analysis were used as topological constraints to guide the structure simulations, showing significant improvements in the predicted structure models (De Leonardis *et al.*, 2015; Wang *et al.*, 2017; Weinreb *et al.*, 2016).

Another way to improve the *de novo* prediction of RNA structure is to predict some structural features first, such as the secondary structure (SS) and the solvent accessibility (SA). In fact, almost all methods build the structure models based on the predicted or experimental SS information (Bailor *et al.*, 2011; Ding *et al.*, 2012; Hajdin *et al.*, 2010; Popenda *et al.*, 2012; Rother *et al.*, 2011; Xu *et al.*, 2014; Zhao *et al.*, 2012). Predicted SA should be useful as well because nucleotides with low solvent accessibility tend to have higher probability to be in contact with other nucleotides, as demonstrated in the field of protein structure prediction (Yang *et al.*, 2011, 2015). However, the information of SA has not been explored for use in RNA structure prediction, probably due to the lack of software for SA prediction.

To the best of our knowledge, RNAsnap is the only available method for RNA SA prediction, which works reasonably well for protein-bound RNAs (Yang *et al.*, 2017). However, its performance on the protein-free RNAs was unsatisfactory with a low level of accuracy. This could be explained by two factors. The first is that protein-free RNAs are more flexible, making them more difficult to predict. The second is that the research on the prediction of RNA SA is still in its infancy and more efforts will be needed.

In this work, we propose RNAsol, an algorithm to predict the SA in RNA. RNAsol was developed based on improved sequence profiles and trained on data of protein-bound RNAs and protein-free RNAs using the long short-term memory neural networks. Benchmark tests suggest that RNAsol consistently outperforms RNAsnap, which is especially significant for the protein-free RNAs.

2 Materials and methods

2.1 Benchmark datasets

We constructed our benchmark datasets from PDB with a similar procedure in the work of RNAsnap. First, we downloaded 1112 RNA-containing structures from PDB (March 2017) with resolution < 4 Å and chain length > 32. These structures contain 853 protein-bound RNAs (2871 chains) and 259 protein-free RNAs (334 chains). Second, we used the cd-hit-est program (Li and Godzik, 2006) to remove highly similar RNA sequences at 80% pairwise sequence identity (the lowest cutoff available), followed by another filtering with BLASTclust (Altschul *et al.*, 1990) at 30% pairwise sequence identity. Here cd-hit-est was used together with BLASTclust because the latter alone failed to remove some highly similar short sequences in our experiments. A total of 234 non-redundant sequences remained after these processes. Third, we removed the sequences that share more than 30% sequence identity with any of the sequences in TS44 and CN48 (the test sets from RNAsnap). After this, we got 165 sequences, which consist of 135 protein-bound RNAs and 30 protein-free RNAs. Each chain structure (rather than the complex structure) was submitted to the POPS package (Cavallo *et al.*, 2003) with the probe diameter of 1.5 Å to calculate all nucleotide-specific accessible surface areas (ASAs). The ASAs were converted to relative accessible surface areas (RSAs) after normalization by the maximum ASA of the corresponding nucleotide (i.e. A, G = 400 Å^2, U, C = 350 Å^2). As done in the prediction of protein SA, we will predict the RSA rather than the ASA. A nucleotide is regarded as being buried inside the molecule rather than binding with others, if it is predicted with a low value of RSA. Finally, these chains were randomly divided into two subsets: one for training (120 sequences, TR120) and the other for test (45 sequences, TS45).

2.2 Feature extraction from improved sequence profiles

It was shown that profile-based model significantly outperformed sequence-based model in the work of RNAsnap. The profile was represented in the form of a multiple sequence alignment (MSA). In RNAsnap, the MSA was obtained based on the sequence-sequence alignment tool BLASTN (Altschul *et al.*, 1990), by searching the query sequence through the NCBI's non-redundant nucleotide sequence database (nt) to identify homologous sequences. In the field of protein structure prediction, it was well accepted that sequence-profile and profile-profile alignments are much more sensitive than sequence-sequence alignment (Eddy, 2009; Edgar and Sjolander, 2004; Karplus *et al.*, 1998; Remmert *et al.*, 2012; Yan *et al.*, 2013). We thus conclude that the RSA prediction in RNA would be enhanced if the sequence-profile or profile-profile alignments are used for building the sequence profiles.

Figure 1 shows the overall flowchart of the proposed algorithm. We propose to construct a new sequence profile from Infernal, a tool for sequence-profile alignment (Nawrocki and Eddy, 2013) using the covariance models. In Infernal, the query sequence is first searched through the nt database by BLASTN to construct a seed MSA. A covariance model is calculated from the seed MSA and used to search the nt database again to find more homologous sequences, which form a new MSA.

The features extracted from the new MSA are similar to those in RNAsnap but with the consideration of the background frequencies of nucleotides. First, for the ith column of the new MSA, the frequency f_{ij} for the jth nucleotide (A, C, G or U) is calculated according to Eq. (1).

$$f_{ij} = (n_{ij} + s_{ij}) / \sum_{k=1}^{4} (n_{ik} + s_{ik}) \quad (1)$$

where n_{ij} is the total number of the jth nucleotide at the ith column; s_{ij} is a pseudo count, which is 9 if the nucleotide is of the same type with the query nucleotide, and 0.3 otherwise. Second, the log odds ratio p_{ij} is calculated according to Eq. (2) and used as a feature.

$$p_{ij} = \ln \frac{f_{ij}}{b_j} \quad (2)$$

where b_j is the background frequency for the jth nucleotide, estimated from the nt database ($b_A = 0.27$, $b_C = 0.23$, $b_G = 0.23$ and $b_U = 0.27$). A score greater/less than zero means that the nucleotide occurs more/less frequently than expectation. These scores are linearly scaled to the range of $[-1, 1]$ to make them comparable between different RNAs. In addition, the RNA sequence is submitted to RNAfold (Zuker and Stiegler, 1981) to predict its SS, which indicates if a nucleotide forms base pairing or not. The SS of a nucleotide is used as another feature.

As the nucleotides in a sequence are not independent with each other, to encode the ith nucleotide, its neighbors inside a window of size w on each side, are also employed. Thus, the total number of features used to encode a nucleotide is $(2 \times w + 1) \times 5$. For terminal nucleotides, the values of the features for null neighbors are set to -1.

2.3 Long short-term memory neural networks

As shown in Figure 1, the machine learning algorithm of our method is a deep neural network, which consists of four hidden layers,

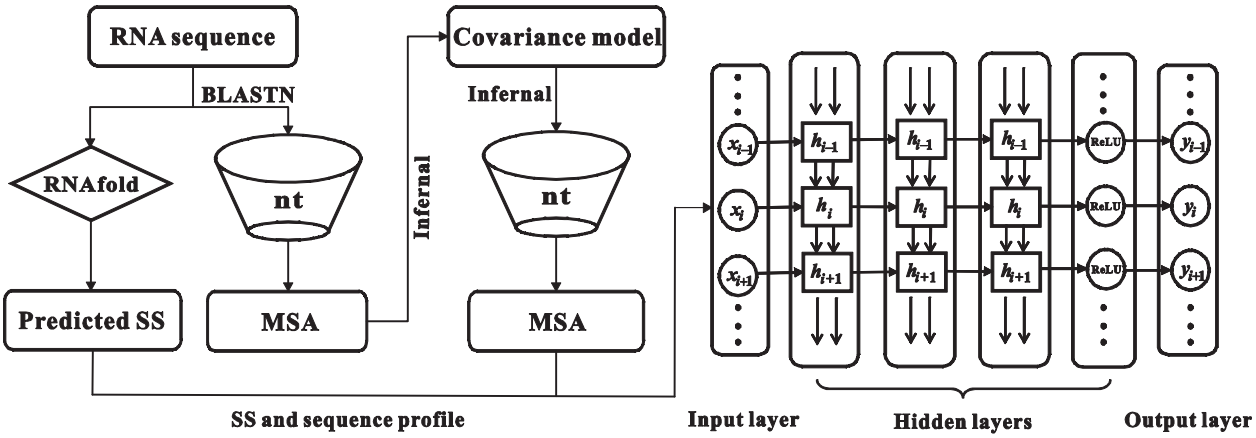

Fig. 1. The overall architecture of the RNAsol. x_i represents the feature vector for the ith nucleotide; y_i is the predicted RSA for the ith nucleotide; and h_i is a vector used for the information transfer

including three long short-term memory (LSTM) layers and one fully connected layer. We built our network architecture using the deep learning library Keras (https://keras.io/). Each layer used the ReLU (Rectified Linear Units) as the activation function (Nair and Hinton, 2010). The weights were initialized with a uniform distribution. The mean squared error was used as the loss function with the Adamax optimizer (Kingma and Ba, 2014). To reduce overfitting, a dropout rate was added for each layer.

2.4 Performance evaluation

The accuracy of the predicted RSA is measured with three metrics. The first two are the Pearson's correlation coefficient (PCC) and the Spearman's correlation coefficient (SCC) between the predicted and the real RSAs. The third is defined based on ASA to measure the area-based error. To this end, the ASA values for the predictions were recovered by multiplying the predicted RSA values by the maximum ASA of the corresponding nucleotide. Then the mean absolute error (MAE) between the recovered ASA values and the native ASA values was calculated.

For each of the above metrics, there are two ways of evaluations and let us take the PCC as an example. The first is to put all nucleotides from all RNAs together and then calculate a single PCC. This has been used in the work of RNAsnap (Yang *et al.*, 2017). The second is to calculate the PCC of each RNA and then evaluate the mean PCC over all RNAs. The former may be inappropriate as the nucleotides from different RNAs will affect each other. Supplementary Figure S1 in the supplementary materials illustrates that even if the PCCs for two RNAs are low (around 0, Supplementary Fig. S1A and B), a high single PCC (>0.8) can be obtained by putting the two RNAs together (Supplementary Fig. S1C). It is apparent that such high single PCC does not reflect the real performance of a method. On the contrary, the mean PCC for the two RNAs is around 0.

To make the above analysis statistically meaningful, the following experiment was conducted. First, we ran the RNAsnap program to predict the RSA for the RNAs in its test set TS44. Then we randomly removed four RNAs (about 10%) from TS44 and calculated the corresponding PCCs for the remaining 40 RNAs. This was repeated by 100 times to draw the PCC distribution, which is presented in Figure 2. It shows that the single PCC ranges between 0.54 and 0.66. However, the mean PCC forms a normal distribution with a smaller variation between 0.3 and 0.37, which is expected from random sampling-based experiments. Based on such observation, we decided to use the second way of evaluation in the subsequent experiments.

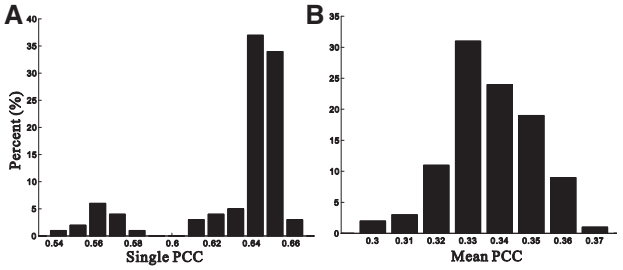

Fig. 2. The distribution of single PCCs and mean PCCs in 100 experiments with random removal of 10% RNAs from the set TS44

3 Results and discussion

3.1 Parameter optimization

All parameters including the window size (w), the batch size (b), the learning rate (l), the number of cells in each layer (n) and the dropout rate (d) were tuned to maximize the PCC on the training set TR120 based on the 5-fold cross validation. To speed up the optimization, the window size was first fixed to 40, taken from the work of RNAsnap (Yang *et al.*, 2017). Other parameters were tuned using the strategy of grid search and the ranges of the parameters are listed in Supplementary Table S1.

After optimization, the batch size, dropout rate and learning rate were 100, 0.3 and 0.002, respectively. There were 512, 128, 128 and 64 cells in the four hidden layers, respectively. The window size w was further optimized after the values of these parameters were fixed. Figure 3 shows the influence of the window size to the performance of RNAsol. We can see that the PCC is the lowest (0.36), when the window size is zero, i.e. no neighboring nucleotides are used. The PCC improves significantly to more than 0.5 when the window size is enlarged to 10. The PCC drops when the window size is greater than 10. Therefore, the optimal window size was set to 10. We performed a statistical test for the results at this optimal window size and the initial window size (i.e. 40). The test indicates that the difference is significant with a P-value of 10^{-3}.

3.2 Comparison with two baseline predictors

The prediction model was built on the training set TR120, with the optimal parameters listed above. The model was assessed on the independent test set TS45 and the results are listed in Table 1. It shows that the PCC, SCC and MAE of the predictions are 0.49, 0.48 and

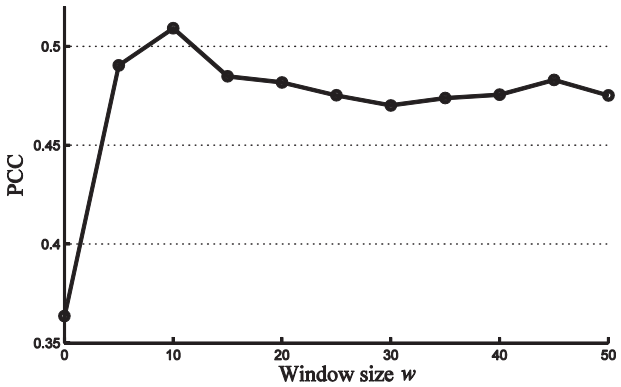

Fig. 3. Optimization of the window size based on the 5-fold cross validation on the training set TR120

Table 1. The comparison of RNAsol with two baseline predictors on the test set TS45

Method	PCC	SCC	MAE (Å^2)
RNAsol	0.49	0.48	34.3
Random	0 (1.7×10^{-11})	0 (7.9×10^{-11})	40.3 (7.1×10^{-8})
Template	0.06 (1.5×10^{-11})	0.06 (1.2×10^{-10})	39 (4.2×10^{-7})

Note: The values inside the parentheses are the *P*-values from the statistical tests of the differences between RNAsol and other methods.

34.3 Å^2, respectively. As RNAsnap's training set has some overlap with TS45, it may be unfair to compare with RNAsnap on this test set. Instead, we compared RNAsol with two baseline predictors: random prediction and template-based prediction. For random prediction, the RSA of each nucleotide was predicted as the mean RSA of the nucleotides in the training set TR120. For template-based prediction, the RSAs were copied from the returned templates in the training set TR120, based on the query-template alignments. When no templates were detected, the predictions were the same as random predictions.

From Table 1, we can see that for random prediction, the PCC and SCC are both zero and the MAE is 40.3 Å^2, much worse than those of RNAsol. For template-based prediction, homologous templates are available for five targets, making the prediction slightly better than random prediction but still significantly worse than RNAsol. For these targets, head-to-head comparisons were presented in Supplementary Table S2. It shows that the accuracy of template-based prediction on these targets is much higher than that listed in Table 1, but still lower than or comparable to RNAsol's predictions. This is probably because the coverage of the alignment is not high as local rather than global alignments were generated by BLASTN. Another reason is the sequence identity between the test set and the training set is less than 30%, making template-based prediction difficult. Statistical tests were conducted to compare the differences between the data in Table 1 and the corresponding *P*-values are listed in the same table. It shows that for all measures, the improvements of RNAsol over both random and template-based predictors are significant (with *P*-values $< 10^{-6}$).

Though the improvement of RNAsol over random prediction was shown statistically significant above, we note that the MAE value for RNAsol is still not very satisfactory (i.e. 34.3 Å^2). A deep analysis was made to appreciate the high MAE value in two aspects as follows.

The first is done by dividing the RNAs into two groups: easy and hard. A target is defined as an easy target if more than 70% of its nucleotides are exposed; and hard otherwise. At this definition, two targets are easy and the remaining 43 targets are hard. Here, the 'easy' targets are trivial to predict for human experts, which can be done by setting all nucleotide as exposed. Such targets are of less interest and they are shown to be more difficult to predict with RNAsol. The performance comparison between RNAsol and random prediction on these targets is presented in Supplementary Table S3. It shows that for both types of targets, RNAsol consistently outperforms random prediction. In addition, we can see that the easy targets are in fact more difficult to predict than the hard targets, as reflected by the higher MAE values of RNAsol and random prediction (i.e. 73.6 and 96.4 Å^2). Predictions for the easy targets should be trivial by simply setting high RSAs for all nucleotides. However, our method does not take this factor into consideration and thus does not perform well on the easy targets. On the contrary, the MAE values on the hard targets are less than half of that on the easy targets. Unfortunately, the MAE value for RNAsol is still as high as 32.4 Å^2, though lower than that of random prediction (37.6 Å^2), which will be analyzed below.

The second is by dividing the nucleotides of the hard targets into two groups: good and bad. A nucleotide is defined as bad/good if the absolute difference between the predicted and the real ASAs is greater/less than the MAE value in the dataset. Under this definition, for random prediction, 38.8%/61.2% nucleotides are assigned into the bad/good group and the MAE value for these nucleotides is 75.5/23.6 Å^2. In comparison, for RNAsol, 31.7%/68.3% nucleotides are assigned into the bad/good group and the MAE value for these nucleotides is 66.2/13.1 Å^2. These data suggest that RNAsol in fact predicts reasonably well with 13.1 Å^2 MAE for most nucleotides (68.3%), which is compared with 23.6 Å^2 MAE for 61.2% nucleotides by random prediction. We conclude that the high MAE value in Table 1 (i.e. 34.3 Å^2) was mainly caused by the minority group of nucleotides with poor predictions. It remains a challenge to improve the prediction for these nucleotides.

3.3 Comparison with the method RNAsnap

We compare our method with RNAsnap, the only available method for RNA SA prediction. To make the comparison as fair as possible, we re-built the prediction model with the same training set of RNAsnap (i.e. the protein-bound RNA set TR89), using the same set of parameters in Section 3.1. The model is then assessed on the independent test sets of RNAsnap (i.e. the protein-bound RNA set TS44 and the protein-free RNA set CN48). For RNAsnap, its standalone software was downloaded and ran locally to make predictions on the test sets, so that we can calculate the values of the metrics defined in this work. The single PCCs from the local running were very similar to the ones reported in the original publication of RNAsnap. Thus, such way of comparison with RNAsnap should be fair.

The data in Table 2 show that on the test sets TS44 and CN48, RNAsol achieves 0.43 and 0.26 PCCs, which are 26.5 and 136.4% higher than RNAsnap, respectively. The PCC distributions of RNAsol and RNAsnap on these test sets are presented in Supplementary Figure S2, which indicates that RNAsol has a greater number of RNAs with PCC higher than 0.3. When measured by SCC, the improvements of RNAsol over RNAsnap increase to 31.2 and 177.8% on TS44 and CN48, respectively. On TS44/CN48, the MAE value for RNAsnap is 36.4/34.9 Å^2, which reduces to 34/32.8 Å^2 in RNAsol. Statistical tests were performed on the differences between RNAsnap and RNAsol and the *P*-values are listed in Table 2. It shows that for all metrics, RNAsol's improvements over RNAsnap are significant at the significance level 0.001 on both test sets. The

Table 2. Comparison with RNAsnap on the test sets TS44 and CN48

Method	PCC	SCC	MAE (Å^2)
RNAsol[a]	0.43	0.42	34
RNAsnap[a]	0.34 (1.6×10^{-5})	0.32 (3.8×10^{-5})	36.4 (1.1×10^{-3})
RNAsol[b]	0.26	0.25	32.8
RNAsnap[b]	0.11 (2.1×10^{-8})	0.09 (3.2×10^{-7})	34.9 (4.8×10^{-6})

Note: The values for the metrics of RNAsnap were obtained by evaluating the predictions from its standalone software. The values inside the parentheses are the *P*-values from the statistical tests of the differences between RNAsnap and RNAsol.

[a]Results on TS44.

[b]Results on CN48.

improvements can be attributed to the improved sequence profiles by Infernal and the enhanced models trained by the LSTM neural networks, which will be illustrated further in the next section.

We note that the accuracy for the set CN48 is lower than TS44 for both RNAsnap and RNAsol. One of the possible reasons is that the training set TR89 only contains protein-bound RNAs. We validated this hypothesis using the new model built with the training set TR120, which consists of 100 protein-bound RNAs and 20 protein-free RNAs. With this new model, the PCCs for TS44 and CN48 increased to 0.49 and 0.46, respectively. The PCC distributions of the new predictions are presented in the last row of Supplementary Figure S2, showing that all RNAs are predicted with non-negative PCCs. Because the pairwise sequence identity between the RNAs in TR120 and the RNAs in TS44 and CN48 is less than 30%, the improvements can be accounted by the enlarged size of the training set and the increased coverage of more data types (i.e. both protein-bound RNAs and protein-free RNAs).

3.4 Factors affecting the performance of RNAsol

We analyze the following factors that may affect the performance of RNAsol: BLASTN-based versus Infernal-based profiles and predicted versus native SSs. The test set TS44 from RNAsnap was used here for this discussion so that the results of RNAsnap in Table 2 could be used as a reference. The results are summarized in Table 3. We can see that when the profile from BLASTN is used, the accuracy of the predictions reduces to a similar level of RNAsnap's accuracy. For example, the PCC decreases from 0.43 to 0.36. A statistical test shows that the improvement of PCC from Infernal-based profile over BLASTN-based profile is significant with a *P*-value of 1.7×10^{-7}. In addition, as expected, when the native SS is used, the predictions are significantly improved, as indicated by the increased values of all metrics and the corresponding *P*-values. A close comparison between the predicted and the native SSs shows that the improvement mainly comes from the unpaired nucleotides, 34% of which were predicted as paired by RNAfold. In comparison, only 12% of the paired nucleotides were predicted as unpaired. This suggests that the RSA predictions can be further enhanced by improving the SS prediction, especially for the unpaired nucleotides.

In addition, we made a statistical test on the differences between using and not using background frequencies in Eq. (2). It shows that improvement was observed but not significant at 5% significance level. This is probably because our estimation for the background frequencies from the nt database may not be accurate enough as the database contains many kinds of nucleotide sequences, not just RNAs. More efforts are required for improving the estimation in future work.

Table 3. The performance of RNAsol on the test set TS44 trained by replacing the default features by others

Feature	PCC	SCC	MAE (Å^2)
Default	0.43	0.42	34
BLASTN	0.36 (1.7×10^{-7})	0.3 (2.2×10^{-8})	36 (6.6×10^{-6})
Native SS	0.5 (3.3×10^{-8})	0.45 (8.1×10^{-4})	30.3 (2.3×10^{-6})

Note: BLASTN means using the profiles generated by BLASTN. The values inside the parentheses are the *P*-values from the statistical tests of the differences between the default features and others.

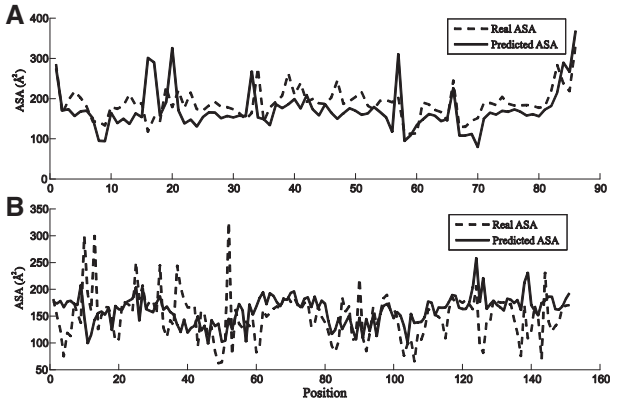

Fig. 4. Two examples demonstrating the relationship between the real ASAs and the predicted ASAs by RNAsol. The predicted ASA values were obtained by multiplying the predicted RSA values by the maximum ASA of the corresponding nucleotide

3.5 Case studies

It is encouraging from Supplementary Figure S2 that many RNAs are predicted with relatively high accuracy. Figure 4 gives two examples from the test sets TS44 and CN48 to illustrate the correlation between the predicted and the real values. These examples were selected based on the average PCCs on the test sets TS44 and CN48 (i.e. 0.43 and 0.26, respectively). The first example (Fig. 4A) is a '*Bacterial Ribonuclease P Holoenzyme Complex*' (PDB ID: 3Q1Q_C), which is a mature tRNA with 87 nucleotides from the protein-bound RNA from the set TS44. For this RNA, there are ~14 000 homologous sequences identified by Infernal, which is much greater than that by BLASTN (i.e. 32). This helps RNAsol to predict the RSAs with 0.47 PCC. The second example (Fig. 4B) is a protein-free RNA from the set CN48, which is a '*Specificity domain of Ribonuclease P of the A-type*' with 161 nucleotides (PDB ID: 1U9S_A). Though many homologous sequences were identified by Infernal (~5000), the predicted RSAs for this RNA have a relatively low PCC (i.e. 0.21). When the training set TR89 was expanded to include both protein-bound and protein-free RNAs in TR120, the PCC for this target increases to 0.46.

As can be seen from Supplementary Figure S2, we did note some targets with low or even negative PCCs for both RNAsol and RNAsnap. We investigated these targets and found that there are three possible reasons for this. The first has been discussed in the Section 3.2, i.e. existence of special targets with virtually all exposed nucleotides (e.g. 2GTT_X, 4TVX_L). Some of these targets have special conformations (e.g. a ring-like topology in 2GTT_X) to maintain their biological functions (e.g. binding with proteins). Special attentions will be needed to improve the prediction for such targets. The second is that the numbers of homologous sequences

are still not enough with Infernal for constructing accurate sequence profiles for some targets. For example, no homologous sequences were detected for the targets 4XJN_N and 4FRG_B, which results in poor predictions with PCCs of −0.007 and −0.235, respectively. To solve this issue, more sensitive alignment tool is required to detect more homologous sequences for building more accurate profile. The last one is that the secondary structure was not predicted well, which has been mentioned in the Section 3.4. For example, the accuracy of the predicted secondary structure for the target 2CZJ_B is 0.6, much lower than the average accuracy (∼0.8), making its PCC very low (i.e. −0.224). The PCC increases to 0.31, after replacing the predicted secondary structure by its native secondary structure. We think there is still much more room for improvement in future, to make the RSA prediction to be useful for *de novo* modeling of RNA tertiary structure.

4 Conclusions

Accurate prediction of RNA solvent accessibility provides an opportunity to address the challenge in the *de novo* prediction of RNA structure. We have developed RNAsol, a new algorithm to predict RNA solvent accessibility. Experiments show that RNAsol consistently outperforms RNAsnap, the only available method for RNA solvent accessibility prediction. A couple of factors contribute to the success of RNAsol, including the improved sequence profiles constructed by sequence-profile alignment and the enhanced training by the long short-term memory neural networks. A web server for RNA solvent accessibility prediction is available at: http://yanglab.nankai.edu.cn/RNAsol/.

Acknowledgment

We thank Dr. Lukasz Kurgan for the insightful discussion on the comparison of RNAsol with baseline predictors.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC 11501306, 11501407, 11871290 and 61873185), the Fok Ying-Tong Education Foundation (161003), the Fundamental Research Funds for the Central Universities, the China Scholarship Council and the Thousand Youth Talents Plan of China.

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bailor,M.H. *et al.* (2011) Topological constraints: using RNA secondary structure to model 3D conformation, folding pathways, and dynamic adaptation. *Curr. Opin. Struct. Biol.*, **21**, 296–305.

Cavallo,L. *et al.* (2003) POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res.*, **31**, 3364–3366.

De Leonardis,E. *et al.* (2015) Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res.*, **43**, 10444–10455.

Ding,F. *et al.* (2012) Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nat. Methods*, **9**, 603–608.

Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inf. Int. Conf. Genome Inf.*, **23**, 205–211.

Edgar,R.C. and Sjolander,K. (2004) COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics*, **20**, 1309–1318.

Hajdin,C.E. *et al.* (2010) On the significance of an RNA tertiary structure prediction. *RNA*, **16**, 1340–1349.

Karplus,K. *et al.* (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Kingma,D. and Ba,J. (2014) Adam: a method for stochastic optimization. *arXiv Preprint arXiv*, 1412.6980.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Miao,Z. *et al.* (2017) RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme, *RNA*. **23**, 655–672.

Nair,V. and Hinton,G.E. (2010) Rectified linear units improve restricted boltzmann machines. In: *International Conference on International Conference on Machine Learning*, pp. 807–814.

Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Ovchinnikov,S. *et al.* (2017) Protein structure determination using metagenome sequence data. *Science*, **355**, 294–298.

Popenda,M. *et al.* (2012) Automated 3D structure composition for large RNAs. *Nucleic Acids Res.*, **40**, e112.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Rose,P.W. *et al.* (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.*, **45**, D271–D281.

Rother,M. *et al.* (2011) ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res.*, **39**, 4007–4022.

Sharma,S. *et al.* (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, **24**, 1951–1952.

Wang,J. *et al.* (2017) Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide-nucleotide interactions from direct coupling analysis. *Nucleic Acids Res.*, **45**, 6299–6309.

Weinreb,C. *et al.* (2016) 3D RNA and functional interactions from evolutionary couplings. *Cell*, **165**, 963–975.

Wuyun,Q. *et al.* (2018) A large-scale comparative assessment of methods for residue–residue contact prediction. *Brief. Bioinf.*, **19**, 219–230.

Xu,X. *et al.* (2014) Vfold: a web server for RNA structure and folding thermodynamics prediction. *PLoS One*, **9**, e107504.

Yan,R. *et al.* (2013) A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction, *Sci. Rep.*, **3**, 2619.

Yang,J. *et al.* (2015) The I-TASSER Suite: protein structure and function prediction. *Nat. Methods*, **12**, 7–8.

Yang,Y. *et al.* (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics*, **27**, 2076–2082.

Yang,Y. *et al.* (2017) Genome-scale characterization of RNA tertiary structures and their functional impact by RNA solvent accessibility prediction. *RNA*, **23**, 14–22.

Zhao,Y. *et al.* (2012) Automated and fast building of three-dimensional RNA structures. *Sci. Rep.*, **2**, 734.

Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.