Bioinformatics, 36(21), 2020, 5169–5176
doi: 10.1093/bioinformatics/btaa652
Advance Access Publication Date: 27 October 2020
Original Paper

Structural bioinformatics

Single-sequence and profile-based prediction of RNA solvent accessibility using dilated convolutional neural network

Anil Kumar Hanumanthappa[1], Jaswinder Singh [1,*], Kuldip Paliwal[1], Jaspreet Singh[1] and Yaoqi Zhou [2,*]

[1]Signal Processing Laboratory, School of Engineering and Built Environment, Griffith University, Brisbane, QLD 4111, Australia and and [2]Institute for Glycomics and School of Information and Communication Technology, Griffith University, Southport, QLD 4222, Australia

*To whom correspondence should be addressed.

Associate Editor: Yann Ponty
Received on March 14, 2020; revised on June 30, 2020; editorial decision on July 11, 2020; accepted on July 14, 2020

Abstract

Motivation: RNA solvent accessibility, similar to protein solvent accessibility, reflects the structural regions that are accessible to solvents or other functional biomolecules, and plays an important role for structural and functional characterization. Unlike protein solvent accessibility, only a few tools are available for predicting RNA solvent accessibility despite the fact that millions of RNA transcripts have unknown structures and functions. Also, these tools have limited accuracy. Here, we have developed RNAsnap2 that uses a dilated convolutional neural network with a new feature, based on predicted base-pairing probabilities from LinearPartition.

Results: Using the same training set from the recent predictor RNAsol, RNAsnap2 provides an 11% improvement in median Pearson Correlation Coefficient (PCC) and 9% improvement in mean absolute errors for the same test set of 45 RNA chains. A larger improvement (22% in median PCC) is observed for 31 newly deposited RNA chains that are non-redundant and independent from the training and the test sets. A single-sequence version of RNAsnap2 (i.e. without using sequence profiles generated from homology search by Infernal) has achieved comparable performance to the profile-based RNAsol. In addition, RNAsnap2 has achieved comparable performance for protein-bound and protein-free RNAs. Both RNAsnap2 and RNAsnap2 (SingleSeq) are expected to be useful for searching structural signatures and locating functional regions of non-coding RNAs.

Availability and implementation: Standalone-versions of RNAsnap2 and RNAsnap2 (SingleSeq) are available at https://github.com/jaswindersingh2/RNAsnap2. Direct prediction can also be made at https://sparks-lab.org/server/rnasnap2. The datasets used in this research can also be downloaded from the GITHUB and the webserver mentioned above.

Contact: jaswinder.singh3@griffithuni.edu.au or yaoqi.zhou@griffith.edu.au

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

1 Introduction

Solvent accessibility of RNA measures the fraction of the solvent accessible surface area (ASA) of each nucleotide in an RNA chain. It is a 1D structural property important for characterizing RNA interactions with other molecules such as proteins (Mukherjee and Bahadur, 2018), identifying structural signature in RNA thermal adaptation (Jegousse *et al.*, 2017), and analyzing the structural difference between denatured, *in vitro* and *in vivo* RNAs (Mortimer *et al.*, 2014; Rouskin *et al.*, 2014). Precise solvent accessibility can be calculated using RNA 3D structures if available. However, only a few thousand RNA structures have been solved and deposited in protein databank so far (Rose *et al.*, 2017), because the physicochemical properties of RNA structures make them more challenging than proteins to be solved by traditional techniques such as X-ray diffraction (Muñoz-Flores *et al.*, 2014) and nuclear magnetic resonance (Scott and Hennig, 2008). Direct probing of RNA solvent accessibility can also be done by hydroxyl radical footprinting (Hulscher *et al.*, 2016; Kielpinski and Vinther, 2014; Latham and Cech, 1989). However, these experiments remain laborious and costly. It is simply not practical to probe millions of known noncoding RNAs experimentally (RNAcentral, 2016). Thus, it is highly

desirable to develop complementary computational approaches for predicting RNA solvent accessibility.

Unlike RNA solvent accessibility prediction, predicting protein solvent accessibility has a long 30-year history (Zhou and Faraggi, 2010) and evolved from discrete-state prediction (Holbrook *et al.*, 1990; Rost and Sander, 1994) to real-value prediction (Ahmad *et al.*, 2003; Dor and Zhou, 2007). A recent method (Hanson *et al.*, 2020) can achieve >0.8 for the Pearson correlation coefficient (PCC) between predicted and actual solvent accessibility for the test set. This is largely because proteins have large sequence and structural datasets for evolutionary feature extraction and deep contextual learning (Hanson *et al.*, 2019). By comparison, the first method for predicting RNA solvent accessibility (RNAsnap) has only been developed recently by our research group (Yang *et al.*, 2017). The method used support-vector machines (Cortes and Vapnik, 1995) with 89 non-redundant protein-bound RNAs for training and achieved a reasonable performance for protein-bound RNAs but not for protein-free RNAs. RNAsol (Sun *et al.*, 2019) improved over RNAsnap by using a relatively larger training set, improved evolutionary profiles, predicted minimum free energy (MFE) secondary structure from RNAfold (Lorenz *et al.*, 2011) and unidirectional recurrent neural networks with long-short-term-memory (LSTM) cells (Hochreiter and Schmidhuber, 1997). In particular, it can achieve similar performance for protein-bound and protein-free RNAs with an average PCC at about 0.45.

In this work, we established a dilated convolutional neural network for predicting RNA solvent accessibility. Our work was inspired by the fact that a dilated convolutional neural network can learn long-range interactions better than convolutional neural networks and LSTM (Senior *et al.*, 2020) and was demonstrated useful for RNA secondary structure prediction (Singh *et al.*, 2019). In addition, we used predicted base-pair probabilities from LinearPartition-V (Zhang *et al.*, 2020) based on thermodynamic parameters (Mathews *et al.*, 1999; Xia *et al.*, 1998) as a new input feature. We show that the new method (RNAsnap2) can achieve >0.5 for the average PCC value with the same training and test sets used by RNAsol and consistent performance improvement for independent, newly solved RNA structures. Moreover, a single-sequence version of RNAsnap2 can achieve a performance comparable to the profile-based RNAsol.

2 Materials and methods

2.1 Datasets

We directly used the RNAsol benchmark datasets for training, validating and testing of our neural network (Sun *et al.*, 2019). The RNAsol training set consisted of 120 (119 effectively as 1 RNA appeared twice) high-resolution RNAs, which were randomly separated into 95 training (TR95) and 24 validation (VL24) RNAs. The RNAsol test set (TS45) contains 45 RNAs. All RNAs in these three datasets (TR95, VL24 and TS45) have >32 sequence length (L) and <4 Å X-ray resolution. They are non-redundant from each other according to CD-HIT-EST (Fu *et al.*, 2012) with the identity cut-off of 0.8 followed by BLASTclust (Altschul *et al.*, 1990) with 30% identity cut-off. Most RNA sequences (about 80–90%) in each dataset are protein-complexes, as shown in Supplementary Table S1. This simply reflects the fact that there are more non-redundant protein–RNA complexes in the PDB (Protein Data Bank) (Rose *et al.*, 2017) as compared to protein-free RNAs. Combining all three datasets, 98 out of 164 RNAs can be completely annotated to its secondary structure [using DSSR (Lu *et al.*, 2015)] from PDB 3D structure while the remaining RNAs can be partially annotated to its secondary structure as the 3D structure for some nucleotides are missing. Supplementary Table S1 also shows the maximum, minimum and average sequence lengths in TR95, VL24 and TS45 dataset. The distribution of the number of Adenine (A), Uracil (U), Guanine (G) and Cytosine (C) nucleotides varies between 24% and 27%, 18% and 23%, 29% and 33% and 21% and 25%, respectively, for these three datasets (see Supplementary Table S1).

In addition to TS45, we prepared an additional test set by downloading (on January 29, 2020) 366 RNA sequences (672 chains) which were submitted to the PDB after March 2017, the previous date for obtaining TR95, VL24 and TS45. These 672 chains were filtered using CD-HIT-EST and BLASTclust with 0.8 and 30% identity cut-off, respectively, so that the new set is non-redundant from the train (TR95), validation (VL24) and test (TS45) sets and between each other. The final high-resolution set (<4 Å X-ray resolution) has 31 RNA sequences, denoted as TS31. Each sequence in TS31 is a protein-free RNA and 22 can be completely annotated to its secondary structure, as shown in Supplementary Table S1. Also, TS31 consists of relatively shorter RNA sequences but the distribution of nucleotides is similar to that of TS45 (see Supplementary Table S1).

To obtain the true ASA labels for TS31, we used the same approach as RNAsol. First, the 3D structure of the individual chain was extracted using Biopython (Cock *et al.*, 2009) from the 3D structure of multiple chains and protein complexed RNAs. Then, the POPS package (Cavallo, 2003) with a probe diameter of 1.4 Å was used to obtain the true labels of ASAs for every RNA chain in TS31. The ASA values were further normalized by the highest ASA value of the corresponding nucleotide (i.e. A, G = 400 Å^2 and U, C = 350 Å^2) and converted to relative accessible surface areas (RSAs). The ASAs values for TR95, VL24 and TS45 are directly obtained from the RNAsol webserver.

The true ASA labels for all the protein-complex RNAs were obtained from individual 3D chain structures instead of protein–RNA complex structures. Therefore, these ASA labels do not account for the interactions with proteins in protein–RNA complexes. To observe how the performance of all the predictors will be affected if the ASA is calculated in the presence of proteins, we also obtained the new protein-present ASA values from the protein–RNA complex structures using the POPS package. However, our methods were trained on ASA values in the absence of proteins.

2.2 Feature extraction

Both RNAsnap (Yang *et al.*, 2017) and RNAsol (Sun *et al.*, 2019) have shown that using evolution-derived sequence profiles as input results in better accuracy in predicting solvent accessibility when compared with using the single sequence as input to the model. RNAsol further demonstrated that the sequence profile obtained from the sequence-profile alignment performs better than that from the sequence–sequence alignment in predicting RSA. Here, we used the same sequence profile as RNAsol. More specifically, a homology search for a query RNA sequence was first made using BLASTN (Altschul *et al.*, 1990) with *E*-value <0.001 and a maximum of 50 000 homologous sequences from the NCBI's nucleotide database. The resulting homologous sequences were aligned for building a covariance model (CM) by cmbuild from Infernal (Nawrocki and Eddy, 2013). The CM was then utilized as input to the Infernal tool for the second round of homology search and yield a new sequence profile based on multiple sequence alignment of new homologous sequences.

In addition, we utilized an additional feature of the RNA secondary structure base-pair probability from LinearPartition (Zhang *et al.*, 2020). LinearPartition is a heuristic algorithm that evaluates RNA base-pair probabilities in linear time. In addition to its computational efficiency, it is more accurate on longer sequences (base-pairs 500+ nts apart) when compared to existing algorithms like RNAfold (Lorenz *et al.*, 2011) and CONTRAfold (Do *et al.*, 2006). Supplementary Table S2 directly compares the secondary structure performance of LinearPartition and RNAfold [both MFE and maximum expected accuracy (MEA)] on test sets TS45, TS45* (a subset of TS45, see Supplementary Table S1) and TS31 for canonical base-pairs. LinearPartition achieves better accuracy on TS45 and TS45* and comparable performance on the TS31 which consists of relatively short sequences (see Supplementary Table S1). The preference of the LinearPartition over RNAfold (MFE) is further validated by precision–recall curves obtained from the base-pair probabilities of both predictors. As shown in Supplementary Figure S1, the precision–recall curve for LinearPartition is defined for almost all

threshold values while the RNAfold (MFE) precision–recall curve becomes undefined at higher threshold values. This indicates that the highest possible precision for RNAfold (MFE) is much lower than LinearPartition. From the above analysis (Supplementary Table S2 and Fig. S1), we preferred using LinearPartition over RNAfold (MFE and MEA) for the ASA prediction problem on our datasets.

Furthermore, to avoid possible overfitting, we used the version of LinearPartition-V based on thermodynamic parameters as in Vienna RNAfold (Lorenz *et al.*, 2011), rather than the default version based on the parameters from machine learning. These 2-D features were converted into 1-D features by simple summation of all the probabilities for a given position. Besides the features above, the traditional one-hot encoding of the RNA sequence (a 4D vector with 1 for the nucleotide type at the sequence position and 0 for other dimensions) was also used. The input features were standardized to have zero mean and unity variance (according to training set TR95) before being input to the model. Unlike RNAsol and RNAsnap, we did not use windowing features as our model can itself learn contextual relationships with neighboring nucleotides. This reduced the number of input features per nucleotide for RNAsnap2 by a factor of 10 relative to that for RNAsol.

2.3 Dilated convolutional neural network

The neural network architecture of RNAsnap2 is shown in Figure 1. It consists of an initial 1D convolutional layer with a kernel size (k) of 3 and 64 filters. The initial convolutional layer (Conv1D) is followed by three residual blocks. Each residual block (marked by dashed red lines in Fig. 1) consists of two 1D dilated convolutional layers (Yu and Koltun, 2015) having alternating kernel size (k) of 5 and 7, respectively. Each layer in the residual block has 64 filters. The dilation factor/rate (DF) for each layer is determined by 2^i, where i is the position of the convolutional layer. The input to each layer in the residual block is preactivated using the exponential linear unit (Elu) activation function (Clevert *et al.*, 2015). Preactivation results in improved accuracy, as depicted by He *et al.* (2016). The output of each convolutional layer is normalized with batch instance normalization (BIN) (Nam and Kim, 2018). The final residual block is followed by the single-node output layer with a sigmoidal activation function. To avoid overfitting, a dropout rate (d) of 40% was used before each convolutional layer (except for the initial convolutional). The order of operations before each layer (except for the initial convolution) was normalization (BIN), activation (Elu) and finally dropout (d).

The neural network was implemented in Google's TensorFlow framework (v1.14) (Abadi *et al.*, 2016) and trained using the RMSProp (Tieleman and Hinton, 2012) optimization algorithm with a learning rate of 0.001 and a mini-batch of the size of 8. The mean square error between the predicted RSA and the actual RSA was used as a loss function. The model hyperparameters such as the kernel size (k), the number of filters, the number of residual units, the activation function, the normalization technique, the dropout rate (d), the choice of the optimizer and the learning rate were optimized based on the model's performance on the validation set (VL24).

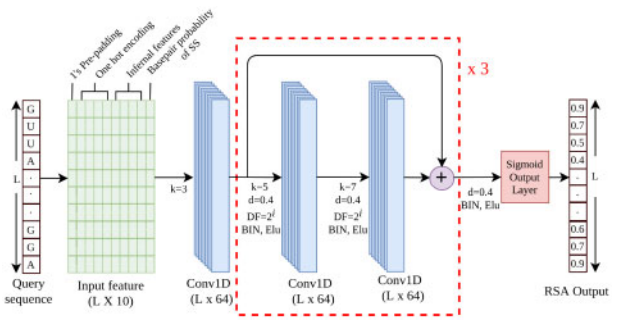

Fig. 1. The network architecture of RNAsnap2. The residual block is shown within dashed red line. k, d, DF and BIN are the size of filter, dropout rate, dilation factor and batch instance normalization, respectively, and L is the length of the input RNA. Scalar 10 and 64 represent the number of features per nucleotide and the number filters in each convolutional layer, respectively. (Color version of this figure is available at *Bioinformatics* online.)

2.4 Performance evaluation

The performance of our method was evaluated using the same measures used previously (Sun *et al.*, 2019; Yang *et al.*, 2017). These include the PCC between predicted and actual RSA values and mean absolute error (MAE) between predicted and actual ASA. These values are evaluated for each RNA chain and then the average value over all chains is reported. In addition, we used one-tailed paired t-test to obtain the P-value (Lovric, 2011) to verify the statistical significance of improvement made by RNAsnap2 over other predictors. The smaller the P value is, the more significant the difference is between the two predictors. The P-values were calculated by our own code implemented in Python which is publicly available at https://github.com/jaswindersingh2/RNAsnap2.

2.5 Method comparison

We compared RNAsnap2 with the only two available RNA solvent accessibility predictors. We downloaded standalone-version of RNAsol (available at https://yanglab.nankai.edu.cn/RNAsol/) and RNAsnap (available at https://sparks-lab.org/downloads/) to obtain the results for TS45 and TS31. We used default parameters for both predictors to get results on test sets.

3 Results

Table 1 compares the performance of RNAsnap2 using different feature combinations for VL24, TS45 and TS31 datasets. The single sequence alone (one-hot encoding) can yield a reasonable performance as average PCC values between 0.45 and 0.49 for the three datasets. The addition of a single-sequence-based prediction of secondary structure (LinearPartition) further improves the average PCC value between 0.48 and 0.51. Incorporating sequence profiles generated from Infernal provides additional improvement by increasing the average PCC values between 0.51 and 0.55. As a comparison, if replacing LinearPartition base-pair probabilities by RNAfold (MFE) secondary structure as in RNAsol, we found a poorer performance across all three datasets (PCC values between 0.47 and 0.51), confirming the importance of base-pair probability estimates in the overall performance of RNAsnap2. We noted that the consistent performance across validation and two test sets indicate the robustness of the method performance for those unseen RNA chains. For convenience, we will denote the profile-based model as RNAsnap2, whereas RNAsnap2 (SingleSeq) denotes the model with one-hot encoding and LinearPartition.

Figure 2 compares the performance of RNAsnap, RNAsol and RNAsnap2 as well as the single-sequence-based RNAsnap2 (SingleSeq) according to median PCC values, 25th and 75th percentiles. In addition to the results for TS45 and TS31 sets, we also make a TS45* set after removing 19 sequences from TS45 which were in RNAsnap training set. Statistically, RNAsnap2 significantly improves over RNAsol with P-values at 7.1×10^{-3} for TS45 and 3.7×10^{-4} for TS31, respectively. This reflected the fact that RNAsnap2 not only has a higher median PCC value but also has a narrower distribution of PCC values than RNAsol. The improvement of RNAsnap2 over RNAsol is 11% for TS45 and 22% for TS31, respectively, in median values and 15 and 54%, respectively in 25th percentiles. In particular, RNAsnap2 (SingleSeq) has a comparable performance to RNAsol for TS45 and a better performance for TS31 despite that RNAsnap2 (SingleSeq) does not employ evolutionary information. This highlights the power of the new network architecture for extracting non-local structural information. In Figure 2, we noticed that the performance of RNAsnap2 (SingleSeq) increases significantly from TS45* to TS31 test set. A similar improvement was also observed in RNAsol. We found that this

Table 1. Performance of RNAsnap2 on VL24, TS45 and TS31 using different combinations of features

Model name	Feature Type	VL24 PCC	VL24 MAE (Å^2)	TS45 PCC	TS45 MAE (Å^2)	TS31 PCC	TS31 MAE (Å^2)
–	Single Sequence (SS)	0.493 (5.2×10^{-03})	34.74 (6.4×10^{-01})	0.466 (5.4×10^{-08})	34.04 (5.0×10^{-03})	0.447 (1.3×10^{-04})	33.32 (5.4×10^{-02})
RNAsnap2 (SingleSeq)	SS + LinearPartition (LP)	0.511 (1.2×10^{-03})	36.29 (1.8×10^{-04})	0.500 (4.0×10^{-07})	33.91 (4.6×10^{-07})	0.483 (1.1×10^{-04})	33.53 (1.0×10^{-05})
–	SS + Sequence Profile (SP)	0.540 (3.9×10^{-01})	34.99 (5.0×10^{-01})	0.518 (4.2×10^{-02})	32.85 (4.0×10^{-01})	0.459 (1.3×10^{-04})	34.32 (3.8×10^{-03})
–	SS + SP + RNAfold (MFE)	0.504 (1.7×10^{-02})	35.84 (1.2×10^{-01})	0.510 (4.2×10^{-02})	33.16 (8.9×10^{-02})	0.470 (7.2×10^{-03})	34.35 (1.1×10^{-02})
RNAsnap2	SS + SP + LP	**0.548**	**34.93**	**0.539**	**32.80**	**0.509**	**32.79**

Note: The values inside the parentheses are the *P*-value obtained through paired *t*-test to show the statistical significance of improvement made by RNAsnap2 features over other different feature types by taking RNAsnap2 as a reference predictor. Bold indicates the method with the best performance.

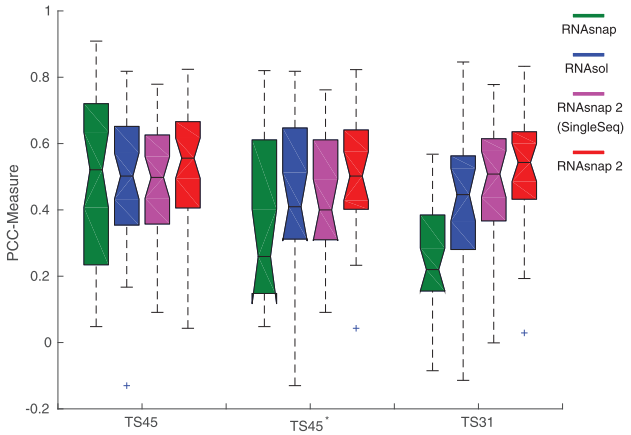

Fig. 2. Distribution of PCC score for individual RNA chains on test sets TS45, TS45* and TS31. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The outliers are plotted individually using the '+' symbol

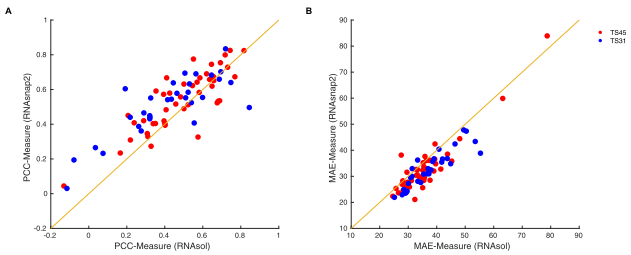

Fig. 3. (**A**) Performance comparison between RNAsnap2 and RNAsol for the PCC values of individual RNA chains on TS45 (in red) and TS31 (in blue). (**B**) Performance comparison between RNAsnap2 and RNAsol for the MAE values of individual RNA chains on TS45 (in red) and TS31 (in blue). (Color version of this figure is available at *Bioinformatics* online.)

improvement is mainly caused by the improvement in single-sequence-based secondary structure used in RNAsol [RNAfold (MFE)] and RNAsnap2 (LinearPartition). A more accurate prediction of secondary structure for TS31 is because 29 out of 31 RNAs have a sequence length shorter than 150 and single-sequence-based secondary structure predictors are more accurate on small RNAs (see Supplementary Table S2).

A more direct comparison between RNAsnap2 and RNAsol is made in Figure 3A for each RNA chain. RNAsnap2 has 56 RNA chains with higher PCC values but only 18 RNA chains with lower PCC values than RNAsol. For some RNAs, RNAsol even yields negative PCC values but not RNAsnap2. Interestingly, there is a high correlation between predicted PCC values by RNAsnap2 and those by RNAsol. The correlation coefficient is 0.82. This indicates the level of difficulty of solvent accessibility prediction for an RNA chain is similar for RNAsnap2 and RNAsol. An even higher correlation coefficient of 0.91 was observed between RNAsnap2 and RNAsnap2 (SingleSeq). Similar trends are observed for performance comparison between RNAsnap2 and RNAsol using MAE as the performance measure (Fig. 3B). RNAsnap2 yields lower MAE values for 68 RNAs but higher MAE values only for 8 RNA chains, when compared to RNAsol. A high correlation coefficient of 0.90 between the RNAsnap2 and RNAsol is observed.

To understand the difficulty in the prediction of each chain, Figure 4A shows PCC values as a function of the number of effective homologous sequences per nucleotide (N_{eff}/L) in a logarithmic plot for both test sets (TS45 and TS31). For those sequences with $N_{\text{eff}}/L > 0.3$, PCC > 0.35. However, for those sequences with $N_{\text{eff}}/L < 0.3$, RNAsnap2 performs poorly for 7 sequences. This

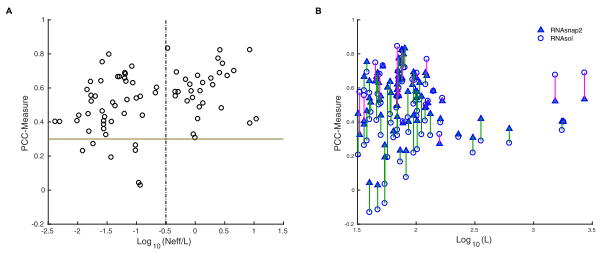

Fig. 4. (**A**) The PCC values for individual RNA chains in the test sets (TS45 and TS31) by RNAsnap2 as a function of the number of effective homologous sequences per nucleotides (N_{eff}/L). (**B**) The PCC values for individual RNA chain by RNAsnap2 and RNAsol as a function of sequence length L. The color green indicates the improvement over RNAsol by RNAsnap2, whereas the color magenta indicates the lack of improvement over RNAsol. (Color version of this figure is available at *Bioinformatics* online.)

suggests that lacking sufficient evolution information is only responsible for inaccurate prediction for a few cases. The majority has a reasonable predicted PCC value independent of N_{eff}/L.

Another possibility is that longer RNAs are more difficult to predict. Figure 4B shows the PCC as a function of the length of the sequence for RNAsnap2 and RNAsol. The performance of both the predictors slightly decreases with the increase in the length of the sequence. Still, the PCC value is >0.30 for the longer sequences for both predictors. Moreover, the small number of long RNA chains makes it difficult to draw any conclusions. It is likely, this performance drop for long RNA sequences is due to the lack of training data on the longer sequences. There are only 9 sequences out of 95 with length >300 nucleotides in training data TR95. We also examined the performance of RNAsnap2 according to the fraction of buried nucleotides (Supplementary Fig. S2) and no significant correlation was observed.

Furthermore, the performance of the predictors was analyzed at the secondary-structure-motif level. Supplementary Table S3 shows PCC values of nucleotides in the stem, hairpin-loop, bulge, internal-loop, multi-loop and exterior loop regions on 57 RNAs in TS45* and TS31. TS45 was deliberately excluded from this comparison as the RNAsnap training set has some overlapping with TS45. Secondary structure motifs are obtained from known RNA structure, using bpRNA (Danaee *et al.*, 2018). RNAsnap2 achieved better PCC scores among all the predictors for all secondary structure motifs. The predicted RSA is the least accurate in multiloop regions. This is mainly because predicted secondary structures are the least accurate for the same region (Singh *et al.*, 2019). We also analyzed the performance of all predictors on nucleotides involved in the tertiary interactions like pseudoknot base-pairs and base multiplets (see Supplementary Table S3). As expected, the low performance was observed for all the predictors for RSA prediction of these nucleotides. However, we did not find any significant correlation between the performance of RNAsnap2 and the number or fraction of the bases in multiloops, pseudoknots and multiplets because the fractions of bases are small in those regions. We also did not observe a significant correlation between the accuracy of predicted secondary structures and the RNAsnap2 performance.

One interesting question is that if there is a performance difference between protein-bound RNAs and protein-free RNAs. This performance difference was observed in the previous work RNAsnap but not as much in RNAsol. The large performance difference in RNAsnap is likely due to a lack of protein-free RNAs in the training set of RNAsnap. Here, TR95 contains 18 protein-free RNAs and 77 protein-bound RNAs. Table 2 compares the performance of four methods by splitting TS45 into protein complex structures and protein-free structures. All RNAs in TS31 are protein-free structures. Table 2 shows that the performance of RNAsnap2 is more consistent than the RNAsol and RNAsnap regardless if RNAs are complexed with proteins or not. PCC values given by RNAsnap2 are greater than 0.48 for all sets. Nevertheless, the

Table 2. Performance comparison of RNAsnap2 with other predictors for protein-bound and protein-free RNAs in TS45 and TS31

	TS45				TS31	
	Protein complex (37 RNAs)		Protein free (8 RNAs)		Protein free (31 RNAs)	
	PCC	MAE (Å^2)	PCC	MAE (Å^2)	PCC	MAE (Å^2)
RNAsnap	0.548 (5.2×10^{-01})	**31.66** (8.6×10^{-01})	0.215 (4.6×10^{-04})	35.46 (5.1×10^{-04})	0.250 (1.1×10^{-10})	39.35 (2.3×10^{-07})
RNAsol	0.514 (5.5×10^{-02})	35.66 (5.8×10^{-05})	0.387 (1.5×10^{-03})	36.88 (1.3×10^{-03})	0.416 (3.7×10^{-04})	37.54 (8.2×10^{-10})
RNAsnap2 (SingleSeq)	0.510 (9.9×10^{-06})	34.27 (2.5×10^{-06})	0.434 (7.6×10^{-03})	32.28 (3.9×10^{-02})	0.483 (1.1×10^{-04})	33.53 (1.0×10^{-05})
RNAsnap2	**0.550**	32.98	**0.488**	**31.95**	**0.509**	**32.79**

Note: The values inside the parentheses are the *P*-value obtained through paired *t*-test to show the statistical significance of improvement made by RNAsnap2 over other predictors by taking RNAsnap2 as a reference predictor. Bold indicates the method with the best performance.

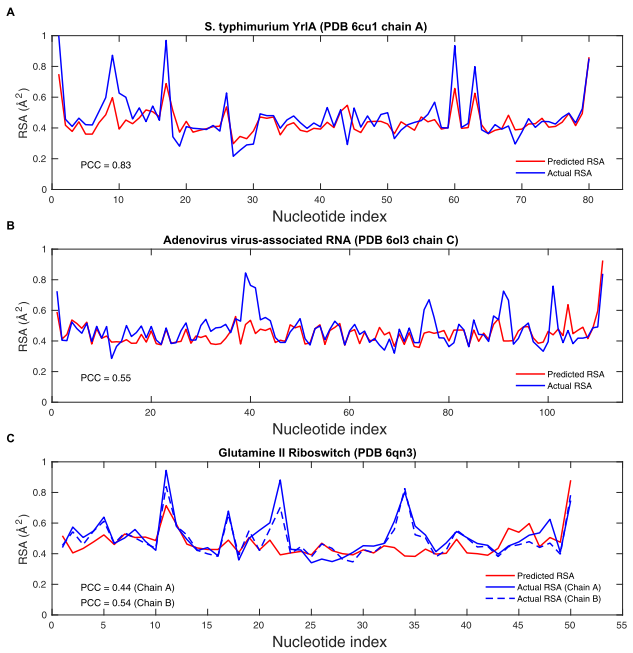

Fig. 5. Predicted (red) versus actual (blue) relative solvent accessible surface area (RSA) of the three RNAs in (**A**), (**B**) and (**C**) as labeled. In (C), Chain A (Solid blue) and Chain B (Dashed blue) of Glutamine II riboswitch form a homodimeric domain-swapped structure with a slightly different conformation for each chain. (Color version of this figure is available at *Bioinformatics* online.)

performance on protein-bound RNAs remains slightly higher than the protein-free RNAs. A more accurate prediction of RSA for protein–RNA complexes is likely due to the following reasons. First, our training set contains more RNAs that are complexed with proteins (77/95) as compare to protein-free RNAs (18/95). This simply reflects the fact that there are more non-redundant RNAs complexed with proteins in the protein databank. The second reason is that RNAs complexed with proteins have more homologous sequences than protein-free RNAs. The average N_{eff}/L for 37 protein-complexed RNAs is 4.08, much higher than 0.44 for the 39 protein-free RNAs in TS45+TS31.

We further analyzed the protein-complex RNAs from TS45 using the ASA labels obtained in the presence of protein. Supplementary Table S4 shows the performance comparison among all the predictors for the ASA labels in the presence and absence of proteins. There is a large drop in performance when the protein-present ASA labels were used. This is somewhat expected because all the predictors are trained on RNA chains in the absence of proteins.

Figure 5 illustrates three examples of predicted versus actual RSA values in TS31. These three RNAs are recently solved structures (after 2018) with high X-ray resolution ($\leq$3.0 Å) and low N_{eff}/L (<0.4). Figure 5A shows an excellent prediction by RNAsnap2 for *Salmonella typhimurium YrlA* RNA (chain A in PDB ID 6cu1, released on October 31, 2018) (Wang *et al.*, 2018) with a high PCC value of 0.83. By comparison, RNAsol and RNAsnap predicted RSA with PCC values of 0.72 and 0.38, respectively. This is a noncoding Y RNA. Figure 5B shows a case of median performance for an adenovirus virus-associated RNA (chain C in PDB ID 6ol3, released on March 7, 2019) (Hood *et al.*, 2019). It is important to know that this RNA was part of the RNA-puzzle dataset used for blind prediction of RNA 3D structures (Miao *et al.*, 2017). RNAsnap2 predicted RSA for this RNA with PCC value of 0.55 while RNAsol and RNAsnap were only able to achieve PCC values of 0.33 and 0.20, respectively. Interestingly for this RNA, RNAsnap2 (SingleSeq) predicted RSA with the highest PCC value of 0.63. This shows that sometimes the evolutionary profile adds noises instead of information to the input feature. Figure 5C shows a case of poor prediction for the Glutamine II Riboswitch (chain A in PDB ID 6qn3, released on June 12, 2019) (Huang *et al.*, 2019). RNAsnap2 predicted RSA with PCC values of 0.44, whereas PCC values are 0.32 for RNAsol and 0.38 for RNAsnap, respectively. During the process of crystallization of Glutamine II Riboswitch, this RNA goes through the process of dimerization. This results in the exchange of strands at the 5′ end of P2 and linking strand at the C19: G41 [refer to Fig. 1C in the original paper (Huang *et al.*, 2019) for notations]. If we use true RSA labels obtained by considering this strand swapping, we would have a significant increase in PCC value to 0.54 by RNAsnap2, compared to 0.41 by both RNAsol and RNAsnap. We also note that RNAsnap2 does not perform poorly for riboswitches in general. For instance, SAM-III riboswitch (PDB ID 6C27, chain A) and glyQ T-box riboswitch (PDB ID 6PMO, chain B) have PCC values of 0.74 and 0.75, respectively.

4 Discussion

We present a new method called RNAsnap2 for predicting RNA solvent accessibility. We have demonstrated that RNAsnap2 significantly improves over existing methods in the accuracy of solvent-accessibility prediction based on the correlation to and the mean absolute difference from measured solvent accessibility. RNAsnap2 differs from the second-best RNAsol in using predicted base-pair probability from LinearPartition, rather than predicted secondary structure from RNAfold (MFE) and dilated convolutional neural network, instead of unidirectional LSTM Recurrent Neural Network. Unlike RNAsol, RNAsnap2 can predict solvent accessibility with or without sequence profiles. The single-sequence-based RNAsnap2 version is comparable to or more accurate than RNAsol (Table 2).

RNAsnap2 with the same features [one-hot encoding, RNAfold (MFE) and sequence profile] as RNAsol improves over RNAsol. The PCC values are 0.51 (P-value 1.2×10^{-01}) for TS45 and 0.47 (P-value 7.2×10^{-04}) for TS31 (Table 1), compared to 0.49 and 0.42 (Table 2), respectively. This indicates the usefulness of the new neural network architecture. RNAsol used a unidirectional RNN with LSTM cells. Unidirectional RNNs only consider the feature maps of current and previous nucleotides when evaluating the RSA of the current nucleotide. The solvent accessibility of a nucleotide can be affected by neighboring nucleotides from both sides. Therefore, to consider the effect of neighboring nucleotides from both sides, RNAsol included the features of the five preceding and five succeeding nucleotides along with the features from current nucleotide as input. In contrast, RNAsnap2 more efficiently considers the effect of neighboring nucleotides using a dilated convolutional architecture with a wide receptive field. The receptive field is the total number of surrounding nucleotide feature maps under consideration for calculating the RSA of a given nucleotide. The receptive field for the RNAsnap2 neural network (Fig. 1) is 339, which means that when predicting the RSA of a given nucleotide, RNAsnap2 considers a feature map of 169 nucleotides on both sides. In addition to considering a feature map from both sides, a dilated convolutional architecture is better at learning long-range dependencies than an LSTM network (Bai *et al.*, 2018).

Another reason for significant improvement in performance by RNAsnap2 can be attributed to the use of LinearPartition with a 6–8% improvement over the use of RNAfold (MFE) (Table 1). LinearPartition provides more accurate base-pair probability (especially on longer sequences) as compared to existing predictors such as RNAfold (MFE) (Lorenz *et al.*, 2011), CONTRAfold (Do *et al.*, 2006) and CentroidFold (Sato *et al.*, 2009). Another reason to choose LinearPartition is because of the linear scaling of its computational time as a function of the length of the sequence. In contrast, the computational time of other predictors discussed above grows exponentially as a function of sequence length. Therefore, LinearPartition reduces the computational complexity in addition to improvement in performance accuracy.

The above improvement is not due to overtraining. This reflects from the fact that a more consistent performance given by

RNAsnap2 across validation and two test sets (Table 1) and between protein-bound and protein-free structures (Table 2), compared to RNAsol. Moreover, RNAsnap2 uses only 1/13 of the trainable parameters used by RNAsol. RNAsnap2 has 151 233 parameters, compared to 1 905 793 parameters used by RNAsol for training on the same training data. Using fewer trainable parameters reduces the risk of model over-fitting on small training data (Ying, 2019).

It is interesting to know why some RNAs are more difficult to predict than others. This level of difficulty seems to be independent of the number of homologous sequences (Fig. 4A) and sequence length (Fig. 4B). The performance analysis of nucleotides existing in different structural motifs shows that all the predictors are least accurate in the multiloop regions (see Supplementary Table S3). It is also difficult to predict accurate RSA for the nucleotides involved in the tertiary interactions like pseudoknots and multiplets (see Supplementary Table S3). However, we did not find a correlation between the performance of RNAsnap2 and the number of nucleotides involved in multi-loops, pseudoknots and multiplets. We also did not find a correlation to the accuracy of predicted RNA secondary structure. Thus, more studies are needed to identify the reason why some RNAs are more difficult to predict than others.

RNAsnap2 used LinearPartition that relies on single-sequence only. The next possible improvement is to replace LinearPartition by secondary structure predictors that use homologous sequence information, such as TurboFold-II (Tan *et al.*, 2017), LocARNA (Will *et al.*, 2007), RNAalifold (Lorenz *et al.*, 2011) and CentroidAlifold (Hamada *et al.*, 2011). However, using these predictors will come with a significantly higher computational cost. The work in this area is still in progress.

Predicting RNA solvent accessibility may require genome-scale studies (Yang *et al.*, 2017). Thus, a computationally efficient program will be important. Excluding computing times for feature generations, RNAsnap2 is about 17% faster than RNAsol for an RNA chain of 1000 nucleotides [8.9 versus 10.5 s on a single thread of Intel Xeon(R) CPU E5-2630 with a clock frequency of 2.3 GHz]. Both are linearly scaled with sequence length with the rate of increase two times smaller by RNAsnap2 than by RNAsol. However, the most time-consuming part of both RNAsnap2 and RNAsol is the generation of sequence profiles, which depends on the number of homologous sequences found and the length of the query RNA sequence. For time-sensitive calculations, RNAsnap2 (SingleSeq) can provide a fast calculation with reasonable performance (PCC = 0.5 for TS45 and 0.48 for TS31). Both RNAsnap2 and RNAsnap2 (SingleSeq) are available as a downloadable package at https://github.com/jaswindersingh2/RNAsnap2 and as a server at https://sparks-lab.org/server/rnasnap2.

Acknowledgements

The authors gratefully acknowledge the use of the High Performance Computing Cluster Gowonda to complete this research, and the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF). They also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

Funding

This work was supported by Australia Research Council [DP180102060 to Y.Z. and K.P.].

Conflict of Interest: none declared.

References

Abadi,M. *et al.* (2016) TensorFlow: a system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283. USENIX Association, Savannah, GA.

Ahmad,S. *et al.* (2003) Real value prediction of solvent accessibility from amino acid sequence. *Proteins Struct. Funct. Bioinf.*, **50**, 629–635.

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

Bai,S. *et al.* (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv Preprint arXiv:1803.01271v2*.

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

Clevert,D.-A. *et al.* (2015) Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv Preprint arXiv:1511.07289*.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

Danaee,P. *et al.* (2018) bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Res.*, **46**, 5381–5394.

Do,C.B. *et al.* (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.

Dor,O. and Zhou,Y. (2007) Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. *Proteins Struct. Funct. Bioinf.*, **68**, 76–81.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Hamada,M. *et al.* (2011) Improving the accuracy of predicting secondary structure for aligned RNA sequences. *Nucleic Acids Res.*, **39**, 393–402.

Hanson,J. *et al.* (2019) Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics*, **35**, 2403–2410.

Hanson,J. *et al.* (2020) Getting to know your neighbor: protein structure prediction comes of age with contextual machine learning. *J. Comput. Biol.*, **27**, 796–814.

He,K. *et al.* (2016) Identity mappings in deep residual networks. In: Leibe,B. *et al.* (eds.) *Computer Vision "EUR" ECCV 2016*. Springer International Publishing, Cham., pp. 630–645.

Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput.*, **9**, 1735–1780.

Holbrook,S.R. *et al.* (1990) Predicting surface exposure of amino acids from protein sequence. *Protein Eng. Des. Select.*, **3**, 659–665.

Hood,I.V. *et al.* (2019) Crystal structure of an adenovirus virus-associated RNA. *Nat. Commun.*, **10**, 2871.

Huang,L. *et al.* (2019) Structure and ligand binding of the glutamine-II riboswitch. *Nucleic Acids Res.*, **47**, 7666–7675.

Hulscher,R.M. *et al.* (2016) Probing the structure of ribosome assembly intermediates in vivo using DMS and hydroxyl radical footprinting. *Methods*, **103**, 49–56.

Jegousse,C. *et al.* (2017) Structural signatures of thermal adaptation of bacterial ribosomal RNA, transfer RNA, and messenger RNA. *PLoS One*, **12**, e0184722.

Kielpinski,L.J. and Vinther,J. (2014) Massive parallel-sequencing-based hydroxyl radical probing of RNA accessibility. *Nucleic Acids Res.*, **42**, e70.

Latham,J. and Cech,T. (1989) Defining the inside and outside of a catalytic RNA molecule. *Science*, **245**, 276–282.

Lorenz,R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

Lovric, M. (ed.) (2011) *International Encyclopedia of Statistical Science*. Springer, Berlin Heidelberg.

Lu,X.-J. *et al.* (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.*, **43**, e142.

Mathews,D.H. *et al.* (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

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

Mortimer,S.A. *et al.* (2014) Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.*, **15**, 469–479.

Mukherjee,S. and Bahadur,R.P. (2018) An account of solvent accessibility in protein–RNA recognition. *Sci. Rep.*, **8**, 10546.

Muñoz-Flores,B.M. *et al.* (2014) Synthesis, X-ray diffraction analysis and nonlinear optical properties of hexacoordinated organotin compounds derived from Schiff bases. *J. Organomet. Chem.*, **769**, 64–71.

Nam,H. and Kim,H.-E. (2018) Batch-instance normalization for adaptively style-invariant neural networks. *arXiv Preprint arXiv:1805.07925*.

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

RNAcentral. (2016) RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Res.*, **45**, D128–D134.

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.

Rost,B. and Sander,C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins Struct. Funct. Bioinf.*, **20**, 216–226.

Rouskin,S. *et al.* (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505**, 701–705.

Sato,K. *et al.* (2009) CentroidFold: a web server for RNA secondary structure prediction. *Nucleic Acids Res.*, **37**, W277–W280.

Scott,L.G. and Hennig,M. (2008) *RNA Structure Determination by NMR*. Humana Press, Totowa, NJ, pp. 29–61.

Senior,A.W. *et al.* (2020) Improved protein structure prediction using potentials from deep learning. *Nature*, **577**, 706–710.

Singh,J. *et al.* (2019) RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.*, **10**, 5407.

Sun,S. *et al.* (2019) Enhanced prediction of RNA solvent accessibility with long short-term memory neural networks and improved sequence profiles. *Bioinformatics*, **35**, 1686–1691.

Tan,Z. *et al.* (2017) TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Res.*, **45**, 11570–11581.

Tieleman,T. and Hinton,G. (2012) *Lecture 6.5—RMSProp: Divide the Gradient by a Running Average of Its Recent Magnitude. COURSERA: Neural Networks for Machine Learning*.

Wang,W. *et al.* (2018) Structural basis for tRNA mimicry by a bacterial Y RNA. *Structure*, **26**, 1635–1644.e3.

Will,S. *et al.* (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLOS Comput. Biol.*, **3**, e65.

Xia,T. *et al.* (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry*, **37**, 14719–14735.

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.

Ying,X. (2019) An overview of overfitting and its solutions. *J. Phys. Conf. Ser.*, **1168**, 022022.

Yu,F. and Koltun,V. (2015) Multi-scale context aggregation by dilated convolutions. *arXiv Preprint arXiv:1511.07122*.

Zhang,H. *et al.* (2020) LinearPartition: linear-time approximation of RNA folding partition function and base pairing probabilities. *Bioinformatics*, **36**, i258–i267.

Zhou,Y. and Faraggi,E. (2010) Prediction of One-Dimensional Structural Properties of Proteins by Integrated Neural Networks, In: Rangwala,H. *et al.* (eds.) *Protein Structure Prediction: Method and Algorithms*, Hoboken, NJ, Wiley. Chapter 4, pp. 45–74.