Bioinformatics, 37(17), 2021, 2589–2600
doi: 10.1093/bioinformatics/btab165
Advance Access Publication Date: 11 March 2021
Original Paper

Structural bioinformatics

Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning

Jaswinder Singh [1,*], Kuldip Paliwal[1], Tongchuan Zhang [2], Jaspreet Singh [1], Thomas Litfin[2] and Yaoqi Zhou [2,*]

[1]Signal Processing Laboratory, School of Engineering and Built Environment, Griffith University, Brisbane, QLD 4111, Australia 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: Jan Gorodkin

Received on December 1, 2020; revised on February 5, 2021; editorial decision on February 27, 2021; accepted on March 8, 2021

Abstract

Motivation: The recent discovery of numerous non-coding RNAs (long non-coding RNAs, in particular) has transformed our perception about the roles of RNAs in living organisms. Our ability to understand them, however, is hampered by our inability to solve their secondary and tertiary structures in high resolution efficiently by existing experimental techniques. Computational prediction of RNA secondary structure, on the other hand, has received much-needed improvement, recently, through deep learning of a large approximate data, followed by transfer learning with gold-standard base-pairing structures from high-resolution 3-D structures. Here, we expand this single-sequence-based learning to the use of evolutionary profiles and mutational coupling.

Results: The new method allows large improvement not only in canonical base-pairs (RNA secondary structures) but more so in base-pairing associated with tertiary interactions such as pseudoknots, non-canonical and lone base-pairs. In particular, it is highly accurate for those RNAs of more than 1000 homologous sequences by achieving >0.8 F1-score (harmonic mean of sensitivity and precision) for 14/16 RNAs tested. The method can also significantly improve base-pairing prediction by incorporating artificial but functional homologous sequences generated from deep mutational scanning without any modification. The fully automatic method (publicly available as server and standalone software) should provide the scientific community a new powerful tool to capture not only the secondary structure but also tertiary base-pairing information for building three-dimensional models. It also highlights the future of accurately solving the base-pairing structure by using a large number of natural and/or artificial homologous sequences.

Availability and implementation: Standalone-version of SPOT-RNA2 is available at https://github.com/jaswindersingh2/SPOT-RNA2. Direct prediction can also be made at https://sparks-lab.org/server/spot-rna2/. 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

Understanding the functional mechanism of a non-coding RNA requires its three-dimensional (3-D) structure. RNA tertiary structures fold on the preformed secondary structure, which contains a set of canonical base-pairs along with tertiary interactions in non-canonical base-pairs, non-nested base-pairs (pseudoknots), lone-pairs and base-multiplets (Tinoco and Bustamante, 1999). As a result, secondary structure and tertiary base-pairing information have been actively pursued by ever-advancing experimental techniques from one-dimensional to multi-dimensional to high-throughput

probing (Carlson *et al.*, 2018; Kubota *et al.*, 2015; Strobel *et al.*, 2018). However, accurate high-resolution determination of all base-pairs still relies on slow and costly X-ray crystallography, nuclear magnetic resonance (NMR) or cryogenic electron microscopy that are suitable for only a small subset of non-coding RNAs.

To overcome the experimental limitation, many computational methods for RNA secondary structure prediction were developed. They can be classified into single-sequence-based and multiple-sequence alignment-based predictors. Single-sequence based predictors use either the nearest-neighbor model (Schroeder and Turner, 2009) [RNAfold (Lorenz *et al.*, 2011), RNAstructure (Reuter and Mathews, 2010)] or statistically learned parameters [CONTRAfold (Do *et al.*, 2006), CentroidFold (Sato *et al.*, 2009)] to obtain minimum free energy (MFE) or maximum expected accuracy (MEA) secondary structure.

Multiple-sequence alignment-based algorithms predict conserved secondary structure for a set of homologous sequences. These multiple-sequence methods can further be classified into two categories. The first class of the methods include predictors such as RNAalifold (Lorenz *et al.*, 2011), CentroidAlifold (Hamada *et al.*, 2011) and TurboFold II (Tan *et al.*, 2017), which first align the set of homologous sequences and then fold during the secondary structure prediction process. RNAalifold finds the MFE consensus structure formed by a set of input aligned sequences whereas CentroidAlifold computes the centroid structure from the ensemble of structures using an average gamma-centroid estimator approach.

The second class of the methods such as SPARSE (Will *et al.*, 2015) and MXSCARNA (Tabei *et al.*, 2008) simultaneously align and fold input homologous sequences for secondary structure prediction. This approach is based on the Sankoff principle (Sankoff, 1985). According to this principle, structure prediction and alignment of the sequences depend on each other, therefore, should be solved simultaneously. Recently, aliFreeFold (Glouzon and Ouangraoua, 2018) provided a new alignment approach, which predicts secondary structure from each homologous sequence from a set and then splits secondary structure of each sequence into secondary structure motifs such as hairpin loops, stems, bulges and internal loops. It constructs the final secondary structure according to the weighted conserved secondary structure motifs.

Alignment-based predictors are more accurate than the single-sequence-based predictors when more evolutionary information is available. However, the overall performance of these predictors remains low. Moreover, alignment-based predictors simply ignore pseudoknots and can utilize a few hundreds of homologous sequences only even if more homologs are available because of either intensive computational requirement or performance saturation (or decrease) after a few hundreds of homologous sequences.

While the accuracy of these folding-based RNA secondary-structure predictors has been stagnated over the last decade (Hamada, 2015; Zhao et al., 2018), its counterpart in proteins, protein contact-map prediction, has made a significant improvement and led to significant advancement in protein structure prediction (Kryshtafovych *et al.*, 2019). Large improvement in prediction of protein contact maps has resulted from the application of deep learning techniques to sequence profiles and direct mutational coupling derived from homologous sequences (Hanson *et al.*, 2018; Wang *et al.*, 2017b). SPOT-RNA (Singh *et al.*, 2019) was the first to treat RNA secondary structure as a contact-map prediction problem. However, unlike proteins, limited availability of only a few hundreds of non-redundant high-resolution RNA structures makes it risky of overtraining for deep learning. To overcome this limitation, SPOT-RNA makes initial training from a large set of approximate secondary structures collected in bpRNA (Danaee *et al.*, 2018) and performs transfer learning by using a small non-redundant set of RNA crystal structures. This single-sequence-based method yields a substantial improvement over existing single-sequence-based methods for RNA secondary structure prediction in both canonical and non-canonical base-pairs. In fact, it is even more accurate than existing multi-sequence-alignment-based techniques as we shall see late in Section 3.

This work attempts to go beyond SPOT-RNA by using sequence profiles and direct mutational coupling proved successful for protein-contact map prediction. The evolution-derived sequence profile for RNA generated from BLAST-N (Altschul *et al.*, 1997) and INFERNAL (Nawrocki and Eddy, 2013) has previously demonstrated its usefulness for improving the accuracy of RNA solvent accessibility prediction (Hanumanthappa *et al.*, 2021; Sun *et al.*, 2019; Yang *et al.*, 2017). Here, we will further employ RNAcmap (Zhang *et al.*, 2020b) that obtains a set of aligned homologous sequences automatically by BLAST-N (Altschul *et al.*, 1997) and INFERNAL (Nawrocki and Eddy, 2013) and predicted RNA contact maps by direct mutational coupling technique GREMLIN (Kamisetty *et al.*, 2013). The resulting method, called SPOT-RNA2 improves over SPOT-RNA for all types of base-pairs with the largest improvement in tertiary non-canonical, pseudoknot and lone base-pairs. It can even directly employ artificial homologous sequences generated from deep mutational scanning for improving base-pair prediction.

2 Materials and methods

2.1 Datasets

We utilized the same train, validation and test sets as in our previous work SPOT-RNA (Singh *et al.*, 2019) for both initial learning and transfer learning with some minor changes. Specifically, the initial learning data (TR0 for training, VL0 for validation and TS0 for testing) and transfer learning data (TR1, VL1, TS1 and TS2) were from bpRNA (Danaee *et al.*, 2018) (Version 1.0) and protein databank (PDB) (Rose *et al.*, 2017), respectively. The PDB sets TR1, VL1 and TS1 were prepared by downloading all the high-resolution (<3.5 Å) RNA X-ray structures from the PDB on March 2, 2019. Another independent test set TS2 was obtained from NMR structures also from the PDB. In this work, we further prepared a new independent test set (TS3) by downloading (on April 9, 2020) all the high-resolution (<3.5 Å) protein-free and protein-complex RNAs submitted to PDB after March 2, 2019. The numbers of structures for TR1, VL1, TS1, TS2 and TS3 are 120, 30, 67 and 19, respectively, after removing homologous sequences between and within the sets by CD-HIT-EST (Fu *et al.*, 2012) at the lowest allowed sequence identity cut-off of 80% and then by BLAST-N (Altschul *et al.*, 1997) search of test sets against training and validation data (TR0, TR1, VL0 and VL1) with a large E-value cut-off of 10. The numbers of structures for initial learning (TR0, VL0 and TS0) are 10721, 1276 and 1272, respectively. They are slightly smaller than the corresponding numbers of structures used in SPOT-RNA (10814, 1300 and 1305, respectively) due to the removal of potential homologs to TS3, in addition to TR0, VL0 and TS0 by the same criterion above.

To evaluate performance for totally unseen families, we constructed a test set TS-hard (28 RNAs) from high-resolution test sets (TS1 and TS3) by excluding sequences with a covariance model match in the training and validations sets by cmsearch with E-value < 0.1. Covariance model of sequences in test sets TS1 and TS3 was build using BLAST-N (Altschul *et al.*, 1997) and INFERNAL (Nawrocki and Eddy, 2013) tools with consensus secondary structure from SPOT-RNA and NCBI's database as a reference library.

The secondary structure for all the PDB RNAs (TR1, VL1, TS1, TS2 and TS3) are extracted from their 3-D structure using DSSR (Lu *et al.*, 2015). For NMR-solved structures, model-1 structures were considered as the reference structure. The numbers of different types of base-pairs, the median N_{eff} value, the median sequence length and the maximum sequence length are shown in Supplementary Table S1. Supplementary Table S2 also shows the numbers of nucleotides in different secondary structure motifs. To find the secondary-structure motifs using bpRNA program (Danaee *et al.*, 2018) (available at https://github.com/hendrixlab/bpRNA), we ignored base multiplets. The N_{eff} value for all the dataset was obtained from the GREMLIN tool (Kamisetty *et al.*, 2013) with default parameters.

2.2 Input features

The only input employed in SPOT-RNA (Singh *et al.*, 2019) was one-hot encoding of the RNA sequence, size Lx4, with (1,0,0,0),

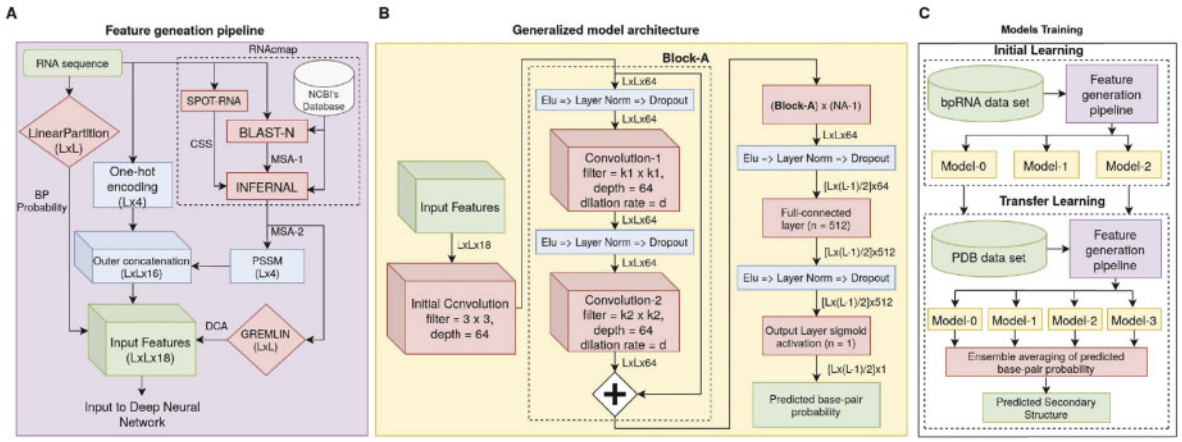

Fig. 1. (**A**) Inputted one dimensional (1-D) and two dimensional (2-D) features employed in SPOT-RNA2 (L is the RNA sequence length; BP is base-pair; CSS is consensus secondary structure). (**B**) An example of the model architecture of SPOT-RNA2. (**C**) The schematic diagram for model pre-training by the bpRNA dataset (TR0) and transfer learning by PDB dataset (TR1)

(0,1,0,0), (0,0,1,0) and (0,0,0,1) for 4 nucleotides (A, U, G and C), where L is the length of the sequence. SPOT-RNA2 adds three more features in addition to one-hot encoding. More specifically, SPOT-RNA2 uses two single-sequence-based and two evolutionary-based features as an input. Single-sequence-based features include one-hot encoding of size Lx4 and predicted base-pair probability from single-sequence-based method LinearPartition (Zhang *et al.*, 2020a) of size LxL. Two evolutionary-based features are Position Specific Score Matrix (PSSM) of size Lx4 and two-dimensional Direct Coupling Analysis (DCA) information of size LxL as shown inFigure 1A.

To avoid any potential bias, we used LinearPartition-V which utilizes the thermodynamic parameters from the Vienna RNAfold package (Lorenz *et al.*, 2011) instead of the default version where the parameters were derived from machine-learning based CONTRAfold (Do *et al.*, 2006). LinearPartition was employed because of its low computational complexity and comparable performance to the existing folding-based secondary-structure prediction algorithms.

The major novel features in SPOT-RNA2 are one-dimensional sequence profiles (PSSM) and two-dimensional coupling information (DCA) inferred from multiple sequence alignment of homologous sequences. These two inputs were produced by RNAcmap (Zhang *et al.*, 2020b) (shown in Fig. 1A), a fully automatic program that performs the initial homology search against the NCBI's database (Coordinators, 2017) using the BLAST-N (Altschul *et al.*, 1997) tool with E-value < 0.001 and a maximal allowed homologous sequences of 50 000 and the second round of sequence-to-profile homologous search using INFERNAL (Nawrocki and Eddy, 2013) tool with the consensus secondary structure (CSS) from the single-sequence-based SPOT-RNA. There is no risk of overtraining because validation and independent test sets were not seen by either SPOT-RNA or SPOT-RNA2. The results from multiple sequence alignment of the second round of homologous sequences were used to obtain PSSM features of size Lx4 and DCA features from GREMLIN (Kamisetty *et al.*, 2013) of size LxL. To save the computational time for feature generation, we do not use CSS during MSA generation of bpRNA dataset.

One-dimensional (1-D) features such as PSSM (Lx4) and one-hot encoding (Lx4) were converted into two-dimensional (2-D) features using an outer-concatenation function as described in RaptorX-Contact (Wang *et al.*, 2017b). All the 1-D features (PSSM, one-hot encoding) after outer-concatenation (LxLx16) and 2-D features (LxL base-pair probability from LinearPartition, LxL DCA from GREMLIN) concatenated together and a feature vector of LxLx18 dimension is used as input to deep neural networks as shown in Figure 1A.

2.3 Deep mutational scanning of CPEB3

For a case study, we obtained additional artificial but functional sequences of CPEB3 (relative activity >0.5) from deep mutational scanning from https://github.com/zh3zh/CODA. These were 18 308 pre-aligned sequences obtained through covariation-induced deviation of activity (CODA) (Zhang *et al.*, 2020c) method.

2.4 Deep neural networks

The deep neural network architecture employed in this work was inspired by the architecture used in our previous work SPOT-RNA (Singh *et al.*, 2019) and RNAsnap2 (Hanumanthappa *et al.*, 2021). We employed an ensemble of deep neural networks as shown in Figure 1B and C for the initial learning and transfer learning for the RNA base-pairing prediction problem. Our previous work, SPOT-RNA used an ensemble of Residual Networks (ResNets) (He *et al.*, 2016), two-dimensional Bidirectional Long Short-Term Memory cells (2-D BLSTMs) (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997) and simple dilation convolutional network (Yu and Koltun, 2015). In this work, we simplified the model architecture by effectively utilizing only the dilated convolutional network. Dilated convolutional networks are better in learning-long range interactions and computationally faster than the LSTMs for both training and inference as shown in our previous work RNAsnap2 (Hanumanthappa *et al.*, 2021). In this work, we explored many models of dilation convolutional networks based on the architecture shown in Figure 1B with different hyper-parameters especially the dilation rate (d) as shown in Supplementary Table S3.

The architecture of each trained model (shown in Fig. 1C and Supplementary Table S3) consists of an initial convolution layer with a filter size of 3x3 and a depth of 64 following by NA number of pre-activated ResNet (He *et al.*, 2016) blocks (Block-A in Fig. 1B), 1 fully connected layer and finally an output layer. A ResNet block consists of two dilated convolutional layers (Yu and Koltun, 2015) with an alternate filter size of $k1$x$k1$ and $k2$x$k2$ and 64 filters. A dilation rate of d was used in each convolutional layer. Input to each layer was activated with ELU (Clevert *et al.*, 2015) function and normalized with layer normalization (Ba *et al.*, 2016) technique. A dropout of 25% was also used to avoid over-fitting on the training data (Srivastava *et al.*, 2014). The output of final ResNet block was also activated with ELU, normalized by layer normalization and dropped out by 25%.

After the NA ResNet blocks, a fully connected (FC) layer with 512 nodes (n) was used as shown in Figure 1B. Again, the output of an FC layer was activated with ELU activation function and normalized using layer normalization technique. A dropout rate of 50% was used to avoid possible over-fitting in FC layer. Finally, an output layer with a single node and sigmoid activation function was used. The single-node output layer with the sigmoid function

converts the feature map from FC layer to upper triangular base-pair probability matrix of size LxL, where L is length of the sequence.

All the deep neural network models were implemented in Google's TensorFlow framework (v1.14) (Abadi *et al.*, 2016) and trained on Nvidia GTX TITAN X graphics processing unit (GPU) to speed up training. For training, ADAM optimization algorithm (Kingma and Ba, 2014) was used with default parameters. For each model, hyper-parameters such as filter sizes ($k1$, $k2$), model depths (d, NA), dilation rates (d) and the number of nodes (n) in the FC layer were optimized for the validation set using the ablation study and is shown in Supplementary Table S3.

2.5 Transfer learning

Here, we used a similar approach as in our previous work SPOT-RNA (Singh *et al.*, 2019) for transfer learning as shown in Figure 1C. Briefly, initial learning was performed using the bpRNA dataset (TR0, VL0 and TS0) based on the deep neural network architecture shown in Figure 1B. Many models were trained using the bpRNA dataset and the final three models were selected based on the best performance on the validation set (VL0). Model hyper-parameters of initially trained models are shown in Supplementary Table S3. Next, transfer learning was performed on initially learned models by further retraining with the TR1 set as shown in Figure 1C. During transfer learning, same hyper-parameters were used except for dilation rate (d) as changing other hyper-parameters did not yield better performance. The hyper-parameters in transfer learning models are shown in Supplementary Table S3. Moreover, all the weights were retrained without freezing any weights because retraining through all the weights performs better than the weights freezing of certain layers. We retrained four models from three initially trained models by varying the dilation rate (d) as shown in Figure 1C and Supplementary Table S3. These four models were optimized for the validation set (VL1) only.

2.6 Output

The output of each model shown in Figure 1C is a 2-dimensional (2-D) LxL upper triangular matrix, where L is the length of the input RNA sequence. This upper triangular matrix represents the likelihood of each nucleotide to be paired with any other nucleotides in a sequence. The outputs from 4 individual models were averaged to obtain the final output. A single threshold value is used to decide whether a nucleotide is making an H-bond with any other nucleotides. The value of threshold was optimized by maximizing the Matthews Correlation Coefficient (MCC) value for the validation set (VL1) only.

2.7 Performance evaluation

F1-score and MCC are the main measurements for performance. F1-score, $F1 = 2(PR * SN)/(PR + SN)$ is a harmonic mean of sensitivity ($SN = TP/(TP + FN)$), and precision ($PR = TP/(TP + FP)$), where TP, FN and FP denote true positives, false negatives and false positives, respectively. MCC is calculated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TN denotes true negatives. Moreover, a precision-recall (sensitivity) curve is used to compare our model with currently available RNA secondary structure predictors. To show the statistical significance of improvement by SPOT-RNA2 over other predictors, a paired t-test was used on F1-score to obtain P-value (Lovric, 2011). The smaller the P-value is, the more significant the difference is between the two predictors.

2.8 Method comparison

SPOT-RNA2 uses single-sequence and alignment-based features. Therefore, we compared SPOT-RNA2 with both single-sequence and alignment-based secondary structure predictors. Single-sequence based predictors includes recent deep learning based predictor SPOT-RNA (Singh *et al.*, 2019) (available at https://github.com/jaswindersingh2/SPOT-RNA), mxfold2 (Sato *et al.*, 2021) version-0.1.0 (available at https://github.com/keio-bioinformatics/mxfold2/releases/), Ufold (Fu *et al.*, 2021) (available at https://ufold.ics.uci.edu/), 2dRNA (Mao *et al.*, 2020) (available at http://biophy.hust.edu.cn/new/2dRNA/) and E2Efold (Chen *et al.*, 2020) (available at https://github.com/ml4bio/e2efold), heuristic approach based LinearPartition (Zhang *et al.*, 2020a) (available at https://github.com/LinearFold/LinearPartition) and LinearFold (Huang *et al.*, 2019) (available at https://github.com/LinearFold/LinearFold), machine learning based CONTRAfold (Do *et al.*, 2006) version 2.02 (available at http://contra.stanford.edu/contrafold/download.html) and mxfold (Akiyama *et al.*, 2018) version-0.0.2 (available at https://github.com/keio-bioinformatics/mxfold/releases/), integer programming based IPknot (Sato *et al.*, 2011) version 0.0.4 (available at https://github.com/satoken/ipknot/releases), maximum expected accuracy (MEA) prediction from partition function based Probknot (Bellaousov and Mathews, 2010) (from RNAstructure package version 6.2, available at http://rna.urmc.rochester.edu/RNAstructure.html), thermodynamic parameter based RNAfold (Lorenz *et al.*, 2011) (from Vienna package version 2.4.14, available at https://www.tbi.univie.ac.at/RNA/) and RNAstructure (Reuter and Mathews, 2010) (from RNAstructure package version 6.2, available at http://rna.urmc.rochester.edu/RNAstructure.html) and γ-centroid estimator based CentroidFold (Sato *et al.*, 2009) version-0.0.16 (available at https://github.com/satoken/centroid-rna-package/releases/). In addition, we also compared SPOT-RNA2 with RNAshapes (available at https://bibiserv.cebitec.uni-bielefeld.de/rnashapes) and pkiss (available at https://bibiserv.cebitec.uni-bielefeld.de/pkiss) from the Shape Studio (Janssen and Giegerich, 2015). Finally, we also added CycleFold (Sloma and Mathews, 2017) predictor which performed relatively better than the other predictors for non-canonical base-pair predictions in literature.

Alignment-based predictors include align-then-fold predictors such as CentroidAlifold (Hamada *et al.*, 2011) version-0.0.16 (available at https://github.com/satoken/centroid-rna-package/releases/), TurboFold II (Tan *et al.*, 2017) (from RNAstructure package version 6.2, available at http://rna.urmc.rochester.edu/RNAstructure.html) and RNAalifold (Lorenz *et al.*, 2011) (from Vienna package version 2.4.14, available at https://www.tbi.univie.ac.at/RNA/) and simultaneous align-and-fold predictors such as SPARSE (Will *et al.*, 2015) (from LocARNA package version 1.9.2.1, available at http://www.bioinf.uni-freiburg.de/Software/SPARSE/), MXSCARNA (Tabei *et al.*, 2008) version 1.9 (available at https://www.ncrna.org/software/mxscarna/dl/). For comparison, we also include alignment free predictor aliFreeFold (Glouzon and Ouangraoua, 2018) (available at https://github.com/UdeS-CoBIUS/alifreefold) and PETfold (Seemann *et al.*, 2011) version 2.1 (available at https://rth.dk/resources/petfold/download.php) which combine energy-based and evolution-based approaches. In addition to alignment-based predictors, we also added direct coupling analysis (DCA) based secondary structure predictors GREMLIN (Kamisetty *et al.*, 2013) (available at https://github.com/sokrypton/GREMLIN_CPP) and CaCoFold (Rivas, 2020) (available at http://eddylab.org/R-scape) for comparison.

To have a fair comparison, all alignment-based predictors provided the same homologous sequences and multiple sequence alignments generated from RNAcmap as in SPOT-RNA2. However, it is almost computationally impossible for most of the existing alignment-based predictors to use all the homologous sequences generated by RNAcmap pipeline shown in Figure 1A. We observed that more aligned sequences do not always improve the secondary structure prediction accuracy for these predictors. For all the predictors (except CentroidAlifold and PETFold), performance starts decreasing after more than 500 homologous sequences as input. Therefore, we restricted the number of aligned input sequences to these predictors to a maximum of 1000 ($1k$) sequences. Furthermore, we made predictions for different numbers of aligned input sequences. A set of 50, 100, 200, 500 and 1000 ($1k$) aligned sequences was used as an input and the final predicted structure was considered from a set

which maximizes F1 on test sets (TS1, TS2 and TS3). The number of input aligned sequences used for the prediction is specified by the superscript in Table 4. We used the same aligned sequences for all the result analysis as shown in Table 4. For TurboFold II, we used a maximum of 200 aligned homologous sequences as an input because it computationally becomes very expensive for 500 aligned input sequences as shown in Supplementary Table S4.

We used the webservers to obtain the predictions for 2dRNA, Ufold, RNAshapes and pkiss predictors while all the remaining predictors were run locally using their standalone versions. For CentroidAlifold, we maximize its accuracy by performing a grid search (see Supplementary Table S5) over three inference engines (McCaskill, CONTRAfold and Alifold) and different values of gamma (γ), which controls precision and sensitivity of predicted structures. As shown in Supplementary Table S5, CONTRAfold inference engine (IE) and a gamma value of 16 yielded the most accurate results on the combined test set. Therefore, these settings were used for CentroidAlifold. Similarly, the CONTRAfold inference engine and a gamma value of 4 were used for CentroidFold after the grid search. For RNAalifold, we made structure prediction based on minimum free energy (MFE) and maximum expected accuracy (MEA). For LinearPartition and LinearFold, we used a machine-learning-based model from CONTRAfold instead of thermodynamic free energy model from Vienna RNAfold as it was more accurate on our test sets. The secondary structure for LinearPartition was extracted from predicted base-pair probabilities with the threshold (0.198) that maximizes the MCC on test set TS1. Predictions for the remaining predictors were made using the default parameters if not explicitly mentioned with the predictor's name in Tables and Figures. The abbreviation MEA and MFE are used along with the predictor's name to show the maximum expected accuracy and minimum free energy structure prediction respectively. If possible, predictions were also made by allowing the non-canonical (NC) base-pair and lone-pairs (LP) for the predictor.

3 Results

3.1 Feature contributions

As shown in Figure 1A the types of features employed include single sequence (one-hot encoding), sequence profiles from BLAST-N (Altschul *et al.*, 1997) and INFERNAL (Nawrocki and Eddy, 2013) [Position specific scoring matrix (PSSM)], and mutational direct coupling analysis (DCA) from GREMLIN (Kamisetty *et al.*, 2013) by RNAcmap, and predicted base-pair probabilities from the single-sequence folding method LinearPartition-V (Zhang *et al.*, 2020a). The contributions of these features were examined by using a baseline model directly trained from a high-resolution non-redundant training set (TR1), and validated by the validation set (VL1), and tested on the test set TS1 from the PDB, developed previously in SPOT-RNA. Potential homologous sequences within and between all datasets were removed by CD-HIT-EST (Fu *et al.*, 2012) at the minimum allowed cut-off of 0.8 and then by BLAST-N (Altschul *et al.*, 1997) search of test sets against training sets at a large E-value cut-off of 10. First, the baseline model (Model-0, based on the architecture shown in Figure 1B and Supplementary Table S3) was trained on the single-sequence (one-hot encoding) feature only. The model achieved reasonable performance with F1-score of 0.577 and 0.557 for VL1 and TS1, respectively, as shown in Table 1. The addition of either sequence profile (PSSM) or direct coupling analysis (DCA) improved the F1-score by more than 20% for VL1 and TS1 by PSSM or more than 16% by DCA. Including both PSSM and DCA features further improves the performance on VL1 and TS1 by an additional 3% and 2%, respectively, compared to adding PSSM or DCA features alone. Moreover, incorporating the single-sequence-based base-pair probability from LinearPartition-V (Zhang *et al.*, 2020a) provides more than 3% additional improvement in F1-score for validation (VL1) and test set (TS1). Finally, the transfer learning by first training on the large bpRNA dataset and retrain on high-resolution PDB data (Fig. 1C) further improved F1-score on both validation (VL1) and test set (TS1) by another 1.5%. Similar trends were observed if Matthews correlation coefficient (MCC) was used to measure the performance as shown in Table 1. Because all single-sequence and evolutionary-information based features made a consistent improvement over validation (VL1) and test set (TS1) on the baseline model (Model-0), all features were employed for subsequent training of additional models for ensembled learning.

3.2 Effect of ensemble learning and transfer learning

Table 2 compares the results of final 4 models after transfer learning and their ensemble on VL1 and TS1 according to MCC and F1-

Table 1. Performance of the baseline model according to Matthews Correlation Coefficient (MCC), F1-score, precision and sensitivity on validation and test sets (VL1 and TS1) of PDB structures by using different combinations of features with direct training on the PDB dataset (TR1)

Baseline model	Feature	VL1				TS1			
(Model-0 only)	Type	MCC	F1	Precision	Sensitivity	MCC	F1	Precision	Sensitivity
Direct Training (DT)	Single Sequence (SS)	0.576	0.577	0.649	0.519	0.562	0.557	0.677	0.473
DT	SS + Sequence Profile (SP)	0.700	0.699	0.777	0.635	0.684	0.685	0.749	0.630
DT	SS + DCA (GREMLIN)	0.679	0.675	0.782	0.594	0.668	0.664	0.766	0.587
DT	SS + SP + DCA (GREMLIN)	0.728	0.722	0.853	0.625	0.708	0.699	0.849	0.594
DT	SS + SP + DCA + LinearPartition (LP)	0.745	0.742	0.838	0.665	0.731	0.729	0.814	0.661
Transfer Learning (TL)	SS + SP + DCA + LP	0.754	0.753	0.822	0.695	0.738	0.739	0.788	0.696

Table 2. Performance of the ensemble and its individual models according to Matthews Correlation Coefficient (MCC), F1-score, precision and sensitivity on VL1 and TS1

Predictor	VL1				TS1			
	MCC	F1	Precision	Sensitivity	MCC	F1	Precision	Sensitivity
Model-0	0.754	0.753	0.822	0.695	0.738	0.739	0.788	0.696
Model-1	0.761	0.762	0.820	0.711	0.735	0.736	0.782	0.695
Model-2	0.753	0.753	0.816	0.699	0.741	0.743	0.784	0.705
Model-3	0.763	0.761	0.845	0.693	0.742	0.742	0.803	0.690
Ensemble	0.765	0.764	0.840	0.700	0.756	0.756	0.823	0.699

Table 3. F1-scores given by single-sequence-based SPOT-RNA and sequence-profile and mutation-coupling based SPOT-RNA2 for different categories for three test sets (TS1, TS2 and TS3)

	TS1		TS2		TS3	
	SPOT-RNA	SPOT-RNA2	SPOT-RNA	SPOT-RNA2	SPOT-RNA	SPOT-RNA2
All base-pairs	0.701	0.751	0.790	0.775	0.701	0.774
Canonical	0.782	0.826	0.864	0.852	0.784	0.859
Non-canonical	0.282	0.416	0.292	0.325	0.216	0.377
Nested	0.728	0.763	0.819	0.739	0.735	0.795
Pseudoknots	0.217	0.504	0.597	0.316	0.421	0.326
Stem	0.775	0.824	0.849	0.827	0.765	0.808
Hairpin loop	0.720	0.765	0.816	0.680	0.675	0.612
Bulge	0.386	0.512	0.424	0.531	0.296	0.370
Internal loop	0.265	0.355	0.276	0.254	0.145	0.136
Multiloop	0.463	0.659	0.228	0.255	0.361	0.589
Exterior loop	0.651	0.809	0.859	0.512	0.783	0.789

score. The ensemble model shows a small improvement in the validation set but a larger improvement (2% in F1-score and MCC) in the test set. This might be because model hyper-parameters were optimized for the VL1, therefore, slightly more accurate prediction of individual models for VL1 makes it more difficult to improve by the ensemble. Moreover, similar performance on the validation and test set after ensemble average indicates a better generalization capability.

To examine the usefulness of transfer learning, we also trained all 4 models directly on a small PDB dataset. Supplementary Table S6 shows the performance of each model on validation (VL1) and test (TS1) set after direct training on TR1. All 4 models achieve 1-3% inferior F1-score than the transfer learning models (Table 2) on VL1 and TS1. Supplementary Table S7 shows the performance of initially trained models prior to transfer learning (Fig. 1C). They were trained on TR0, validated on VL0 and tested on TS0, all these datasets from bpRNA. These models achieved significant and similar performance on VL0 and TS0 with F1-score between 0.726 and 0.738 but comparatively poor performance on crystal structure test set TS1 with F1-score of between 0.62 to 0.66 for all 3 models. This result confirms less than perfect annotations in the bpRNA dataset and the necessity of transfer learning due to the limited number of high-resolution structures.

3.3 Comparison to SPOT-RNA

We first compare to the single-sequence-based method SPOT-RNA (Singh *et al.*, 2019) as SPOT-RNA2 and SPOT-RNA employed essentially the same training data. The test sets TS1 and TS2 built for testing SPOT-RNA have 67 high-resolution (<3.5 Å) X-ray structures and 39 NMR structures, respectively. These test sets are non-redundant from bpRNA and PDB data (training and validation) according to CD-HIT-EST (Fu *et al.*, 2012) at the lowest allowed cut-off of 0.8 followed by BLAST-N (Altschul *et al.*, 1997) filtering of TS1 and TS2 at large E-value cut-off of 10 against training (TR0, TR1) and validation (VL0, VL1) datasets. We further prepared a new test set TS3, which consists of 19 newly submitted (after 2 March 2019 up to 9 April 2020) PDB structures that are non-redundant from the existing training, validation and test sets according to the same criteria as for TS1 and TS2.

Table 3 compares the performance of these two methods for different types of base-pairs and secondary-structural motifs on the three different test sets. SPOT-RNA2 improves F1-score by more than 7% and 10% for all base-pairs over SPOT-RNA on the test set TS1 and TS3, respectively. SPOT-RNA2 slightly underperforms as compare to SPOT-RNA on TS2. We found that this underperformance is due to the lack of homologous information in TS2, as shown in Supplementary Table S1. The median number of effective homologous sequences (N_{eff}) value for TS2 was only 6 as compared to 61 and 15 for TS1 and TS3, respectively. TS2 is also made of RNAs with relatively short (easier to predict) sequences (Supplementary

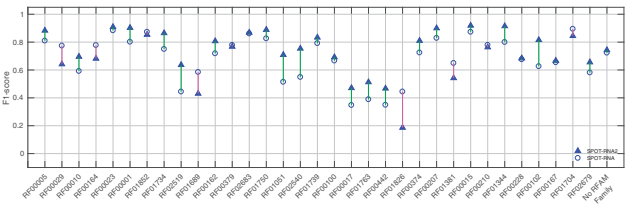

Fig. 2. Performance comparison between SPOT-RNA2 and SPOT-RNA on 125 RNAs from TS1 + TS2 + TS3 by mapping to RFAM families. The color green indicates the improvement over SPOT-RNA by SPOT-RNA2 whereas the color magenta indicates a lack of improvement over to SPOT-RNA

Table S1). Moreover, TS2 contains the structures solved by NMR whereas the method was trained for X-ray structures. The performance of SPOT-RNA on TS2, similar to other folding-based predictors (Singh *et al.*, 2019), is much higher than TS1 and TS3, whereas SPOT-RNA2 shows more consistent performance across three test sets, confirming the robustness of SPOT-RNA2 for RNAs of different lengths, different number of homologous sequences and even different experimental techniques.

Base-pairs are made of canonical and non-canonical ones. SPOT-RNA2 improved F1-score for canonical base-pairs by more than 5% and 9% over SPOT-RNA on TS1 and TS3, respectively, with comparable performance on TS2 (Table 3). A much larger improvement is on non-canonical pairs. F1-score from SPOT-RNA2 is improved over SPOT-RNA by 47%, 11% and 74% for TS1, TS2 and TS3, respectively. SPOT-RNA2 made a large improvement in predicting pseudoknot base-pairs on TS1 but underperforms for TS2 and TS3 in comparison to SPOT-RNA. Such a large fluctuation of improvement is largely due to the small number of pseudoknot base-pairs in each set (160 in TS1, but only 41 in TS2 and 52 pairs in TS3, see Supplementary Table S1).

Further, we compare the performance of SPOT-RNA2 and SPOT-RNA for different RNA secondary structure motifs. SPOT-RNA2 achieves a better F1-score on the majority of structural motifs on test sets TS1 and TS3 on comparison to SPOT-RNA. Again, SPOT-RNA performs better on TS2 because of low N_{eff} and short sequences which suit the single-sequence based predictors.

One interesting question is how these two methods perform on different Rfam families. Combining three test sets (125 RNAs) allows us to map into 33 different Rfam families using Rfam (Kalvari *et al.*, 2018) webserver (https://rfam.xfam.org/). As shown in Figure 2, SPOT-RNA2 improved F1-score over SPOT-RNA on 24 Rfam families (shown by green lines in Fig. 2) while underperforms for only nine Rfam families (shown by magenta lines in Fig. 2). For these nine families, there are only five with large difference (RF00029, RF00164, RF01689, RF01826 and RF01381). RF01689 does not have any homologous sequences with $N_{eff} = 1$. Among the remaining four Rfam families, RF01826 and RF00164

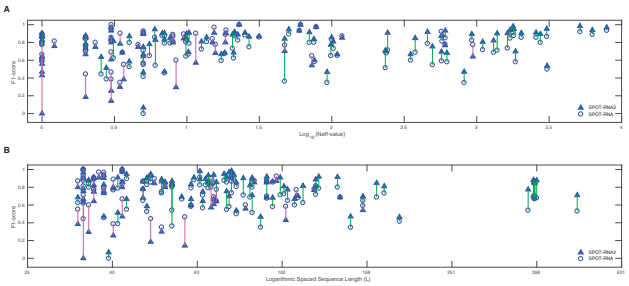

Fig. 3. Performance comparison between SPOT-RNA2 and SPOT-RNA as a function of (**A**) the number of effective homologous sequences (N_{eff}) (**B**) RNA sequence length (L). The color green indicates the improvement over SPOT-RNA by SPOT-RNA2 whereas the color magenta indicates a lack of improvement over to SPOT-RNA

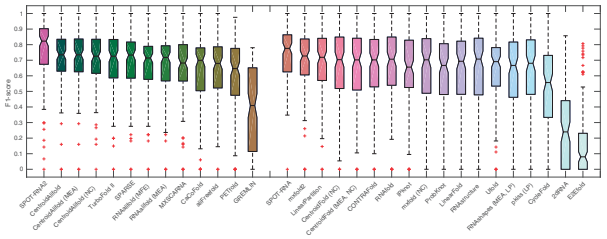

Fig. 4. Distribution of F1-scores for individual RNAs on the combined test sets TS1, TS2 and TS3 given by various methods as labeled. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The outliers are plotted individually by using the '+' symbol

have very low N_{eff} value of 2 and 15.73 respectively. The only exception is that SPOT-RNA2 did not perform better with a moderate N_{eff} value of 76.15 for RF01381 and high N_{eff} of 757.46 for RF00029. This could be due to the possibility that homologous sequences may bring in noises rather than useful information.

To examine the impact of the number of homologous sequences in the NCBI's sequence library on method performance, Figure 3A shows the performance as a function of N_{eff} value by combining three test sets. As expected, SPOT-RNA2 achieved comparatively low F1-score for low N_{eff} value RNAs and significantly improved over SPOT-RNA for medium and higher N_{eff} values. When $N_{eff} > 1000$, F1-scores given by SPOT-RNA2 are greater than 0.8 for 14 out of 16 RNAs. One with a low F1-score of 0.54 is tRNA (4v8n, chain-AW with N_{eff} =3062). For this case, SPOT-RNA2 only manages to improve over SPOT-RNA slightly, indicating that some evolutionary information may contain noises, perhaps due to inaccurate, automatic sequence alignment of tRNA.

To examine the impact of sequence length on method performance, we added nine long RNAs ($300 < L < 500$) to existing three test sets from PDB by extending the X-ray resolution to 4 Å as there were no other long high-resolution, non-redundant RNA in PDB. Figure 3B shows F1-score for 134 individual RNA (125 from 3 test sets + 9 long RNAs) as a function of sequence length. Except for short sequences ($32 < L < 50$), SPOT-RNA2 performed better than SPOT-RNA in most cases.

3.4 Comparison with existing techniques

We compared our SPOT-RNA2 with 9 existing alignment-based predictors and 17 single-sequence-based predictors on three independent test sets (TS1, TS2 and TS3). To compare with other predictors, TS1 reduced from 67 to 65 RNAs and TS2 reduced from 39 to 36 RNAs as few predictors were unable to predict for sequences containing invalid or missing nucleotides. Table 4 shows that SPOT-RNA2 improved F1-score by 10%, 2% and 9% over second-best alignment-based predictors on test sets TS1, TS2 and TS3, respectively. In comparison to alignment-based predictors, single-sequence-based predictors achieve comparatively low F1-score on the test set TS1 and TS3 because of more evolutionary information (median N_{eff} >14) available for these test sets as shown in Supplementary Table S1. As expected, single-sequence-based predictors perform better than the alignment-based predictors on TS2 because of limited evolutionary information (median N_{eff} <7) in this test set. Also, TS2 consists of smaller number of non-canonical and pseudoknot base-pairs (as shown in Supplementary Table S1) in comparison to TS1 and TS3 which makes TS2 easier for prediction for the majority of the predictors. Importantly, SPOT-RNA2 shows consistent performance across three test sets irrespective of different distributions. The performance improvement observed for three test sets is statistically significant as shown by the *P*-value obtained through paired t-test in Supplementary Table S8.

Figure 4 shows the distribution of F1-score among individual RNAs in terms of median, 25^{th}, and 75^{th} percentile for all the predictors on 120 RNAs from three test sets. SPOT-RNA2 achieved the highest median F1-score with the least spread around the median value as compared to other predictors. This shows the stable performance of SPOT-RNA2 in comparison to other predictors. Furthermore, Figure 5 shows the performance comparison by Precision-Recall (PR) curve. Predictors with base-pair probability are shown by the curves while the predictors with discrete outputs are shown by a single point in the PR-curve. Furthermore, alignment-based predictors are shown by filled symbols whereas single-sequence-based predictors are shown by open symbols. SPOT-RNA2 outperforms other predictors for most thresholds values. This shows the robustness of the SPOT-RNA2 predictor. For example, at high precision of 80%, the sensitivity of SPOT-RNA2 is 72%, compared to 66%, by SPOT-RNA.

We further compared methods for different Rfam families (Supplementary Figs S1 and S2) and for different types of base-pairs with the combined test sets (Supplementary Table S9). SPOT-RNA2 outperforms these predictors for majority of the Rfam families mapped RNAs. It also performs better than existing predictors on canonical, non-canonical and pseudoknot base-pairs. SPOT-RNA2 improved significantly for lone-pairs and base-triples which are either less explored or completely ignored by most existing RNA secondary structure predictors.

To evaluate performance for totally unseen families, we compared all the predictors on test set TS-hard (as shown in Table 5). The maximum number of homologous sequences for all the alignment-based predictors was restricted to 1000, including SPOT-RNA2. As expected, the performance of almost all the predictors including homology-modelling baseline significantly reduced on the TS-hard set , indicating that these cases are completely independent of those used for model training. Homology modelling baseline was obtained by assigning labels to sequences based on the top-scoring match (E-value < 10) with the training and validation sets. The search was conducted by cmsearch from the covariance model build using the predicted secondary structure by SPOT-RNA. Under this setting, SPOT-RNA2 remains the top-performing model, and shows a particularly strong advantage over those other methods that also attempt to predict non-canonical base-pairs.

Two examples are illustrated in Figures 6 (from test set TS1) and 7 (from test set TS-hard) to compare the performance by SPOT-RNA2 with the second-best single-sequence (SPOT-RNA) and the second-best alignment-based predictor (CentroidAlifold). In each figure, correctly predicted canonical, non-canonical and pseudoknot base-pairs are shown in blue, orange and green, respectively and incorrectly predicted base-pairs are shown in magenta. Figure 6 shows the prediction of high N_{eff} value (N_{eff} =1803) 70S ribosome (released on 18 April 2018, chain 1Y in PDB ID 6cae) (Pantel *et al.*, 2018) from TS1 by SPOT-RNA2, SPOT-RNA and CentroidAlifold in comparison to its native structure. SPOT-RNA2 nearly predict all native base-pairs correctly for this RNA with an F1-score of 0.95. SPOT-RNA2 successfully predicted non-canonical (in orange), pseudoknots (in green), lone-pair (U51-A55) and base-triples (G10-C23 and G10-G42; C13-G20 and G20-G43; A9-A21 and U12-A21) shown in Figure 6C. In comparison, SPOT-RNA and CentroidAlifold predicted structure with F1-scores of 0.88 and 0.81 respectively. Figure 7 demonstrates prediction of a synthetic

Table 4. Performance comparison between SPOT-RNA2 and other single-sequence and alignment-based predictors on three test sets (TS1, TS2 and TS3)

	TS1			TS2			TS3		
	F1[a]	Precision	Sensitivity	F1[a]	Precision	Sensitivity	F1[a]	Precision	Sensitivity
Multi-sequence-based									
SPOT-RNA2	**0.756**	0.823	**0.699**	**0.774**	0.869	**0.698**	**0.774**	0.828	**0.727**
CentroidAlifold^{1k}	0.688	0.845	0.580	0.733	0.856	0.641	0.667	0.785	0.579
CentroidAlifold^{1k} (MEA)	0.683	0.824	0.584	0.738	0.856	0.649	0.668	0.776	0.587
CentroidAlifold^{1k} (NC)	0.675	0.816	0.576	0.740	0.852	0.655	0.687	0.807	0.598
Turbofold II^{200} (MEA)	0.637	0.831	0.517	0.762	**0.927**	0.647	0.712	**0.881**	0.598
SPARSE^{200}	0.663	0.849	0.544	0.701	0.891	0.578	0.655	0.822	0.545
RNAalifold^{100} (MFE)	0.659	0.885	0.525	0.666	0.877	0.537	0.648	0.855	0.522
RNAalifold^{100} (MEA)	0.658	**0.920**	0.512	0.667	0.894	0.532	0.623	0.867	0.486
MXSCARNA^{500}	0.629	0.801	0.518	0.662	0.831	0.550	0.645	0.815	0.534
CaCoFold	0.641	0.847	0.515	0.629	0.864	0.495	0.635	0.805	0.525
aliFreefold^{200}	0.624	0.812	0.506	0.656	0.860	0.531	0.603	0.777	0.492
PETfold^{1k}	0.574	0.821	0.441	0.649	0.862	0.520	0.592	0.825	0.461
GREMLIN	0.463	0.651	0.359	0.258	0.370	0.198	0.425	0.588	0.332
Single-sequence-based									
SPOT-RNA	**0.702**	**0.855**	**0.596**	**0.789**	0.881	**0.714**	**0.701**	0.805	**0.621**
mxfold2	0.651	0.835	0.533	0.771	**0.946**	0.651	0.684	0.852	0.571
LinearPartition	0.626	0.750	0.537	0.757	0.864	0.675	0.696	0.818	0.606
CentroidFold (NC)	0.596	0.785	0.480	0.778	0.933	0.667	0.681	0.846	0.570
CentroidFold (MEA, NC)	0.593	0.756	0.488	0.778	0.925	0.671	0.682	0.836	0.576
CONTRAFold	0.591	0.729	0.497	0.776	0.898	0.683	0.677	0.808	0.582
RNAfold	0.576	0.713	0.484	0.777	0.923	0.671	0.688	0.828	0.589
IPknot	0.581	0.777	0.463	0.751	0.929	0.630	0.682	**0.857**	0.567
mxfold (NC)	0.578	0.738	0.474	0.764	0.924	0.651	0.677	0.836	0.568
ProbKnot	0.570	0.696	0.482	0.754	0.890	0.653	0.668	0.786	0.581
LinearFold	0.579	0.778	0.462	0.738	0.902	0.625	0.635	0.828	0.516
RNAstructure	0.556	0.693	0.464	0.767	0.914	0.661	0.654	0.796	0.556
Ufold	0.601	0.790	0.485	0.710	0.871	0.599	0.573	0.729	0.472
RNAshapes (MEA, LP)	0.552	0.633	0.490	0.754	0.851	0.677	0.631	0.699	0.575
pkiss (LP)	0.534	0.653	0.452	0.765	0.893	0.670	0.620	0.733	0.537
2dRNA	0.364	0.848	0.232	0.232	0.780	0.136	0.156	0.655	0.089
E2Efold	0.243	0.438	0.168	0.094	0.184	0.063	0.103	0.192	0.070

[a]Harmonic mean of precision and sensitivity. MEA is maximum expected accuracy structure prediction. MFE is minimum free energy structure prediction. NC is structure prediction by allowing non-canonical base-pairs. LP is structure prediction by allowing lone-pairs. Superscript with the name of alignment-based predictor shows the number of aligned sequences used for the prediction. Refer 'Methods comparison' section for detail. Bold indicates the predictor with the best performance.

construct RNA (released on 26 June 2019, chain H in PDB ID 6dvk) (Yesselman *et al.*, 2019) with very low N_{eff} value (N_{eff} =1) from the TS-hard test set. SPOT-RNA2 predicted a structure close to native structure with an F1-score of 0.86, including three non-canonical base-pairs (in orange) but missed long-distance pseudoknots. SPOT-RNA and CentroidAlifold predicted structure with F1-score of 0.85 and 0.81, respectively.

Supplementary Figures S3 and S4 show two challenging RNA examples from the test set TS3 and TS-hard respectively. Supplementary Figure S3 shows the prediction of a pistol ribozyme (released on 18 December 2019, chain A, B in PDB ID 6ufj) (Teplova *et al.*, 2020) by SPOT-RNA2, SPOT-RNA and CentroidAlifold. SPOT-RNA2 predicts the structure with 1 non-canonical base-pair (G42-G61 in orange) and 1 pseudoknot stem (A1-U16 to G5-C12 shown in green) with an overall F1-score of 0.67. In comparison, SPOT-RNA and CentroidAlifold predict structure with F1-scores of 0.60 and 0.47, respectively. Supplementary Figure S4 shows the prediction of a Mango-III aptamer (released on 17 April 2019, chain A in PDB ID 6e8s) (Trachman *et al.*, 2019) by these 3 predictors. SPOT-RNA2, SPOT-RNA and CentroidAlifold predict structures with a poor F1-score of 0.47, 0.46 and 0.44 respectively, for this synthetic RNA. However, the most missed predictions are non-canonical base-pairs. If these non-canonical base-pairs are ignored, F1-scores for pistol ribozyme would be 0.60, 0.76 and 0.79, for CentroidAlifold, SPOT-RNA and SPOT-RNA2, respectively. For mango-III aptamer, F1-scores would be 0.62, 0.70 and

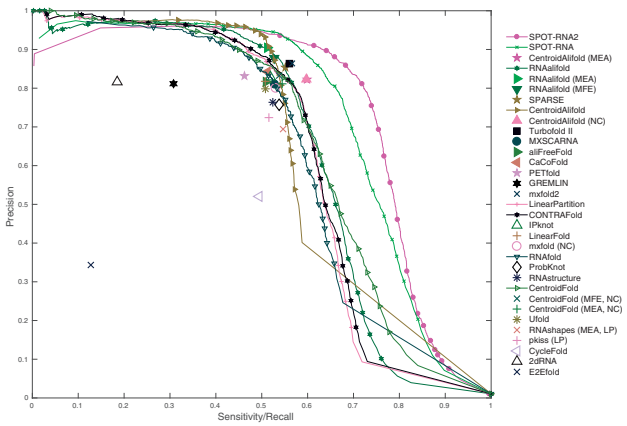

Fig. 5. Precision-recall curves on the combined test sets TS1, TS2 and TS3 by SPOT-RNA2 (shown in magenta) along with 10 alignment-based predictors and 17 single-sequence-based predictors. Precision and sensitivity results from alignment-based predictors are shown by filled symbols and results from single-sequence-based predictors are shown with open symbols

Table 5. Performance comparison between all the predictors on the TS-hard for all, canonical and non-canonical base-pairs

	All base-pairs			Canonical base-pairs			Non-canonical base-pairs		
	F1[a]	Precision	Sensitivity	F1[a]	Precision	Sensitivity	F1[a]	Precision	Sensitivity
Multi-sequence-based									
SPOT-RNA2	**0.678**	0.731	**0.632**	0.760	0.768	**0.752**	0.278	0.446	0.202
SPOT-RNA2^{1k}	0.675	0.728	0.629	0.756	0.765	0.747	**0.283**	0.451	**0.206**
CentroidAlifold^{1k}	0.653	0.765	0.570	0.752	0.796	0.713	0.102	0.294	0.061
CentroidAlifold^{1k} (MEA)	0.641	0.740	0.566	0.742	0.780	0.707	0.098	0.238	0.061
CentroidAlifold^{1k} (NC)	0.657	0.759	0.579	**0.762**	0.821	0.711	0.158	0.277	0.110
Turbofold II^{200} (MEA)	0.657	0.833	0.543	0.758	0.833	0.695	–	–	–
SPARSE^{200}	0.639	0.807	0.528	0.737	0.815	0.673	0.024	0.286	0.012
RNAalifold^{100} (MFE)	0.616	0.824	0.492	0.715	0.826	0.630	0.000	0.000	0.000
RNAalifold^{100} (MEA)	0.604	**0.890**	0.457	0.706	**0.890**	0.585	0.000	0.000	0.000
MXSCARNA^{500}	0.624	0.761	0.530	0.723	0.789	0.668	0.063	0.231	0.037
CaCoFold	0.549	0.763	0.429	0.650	0.812	0.542	0.041	0.133	0.025
aliFreefold^{200}	0.601	0.788	0.485	0.697	0.800	0.618	0.023	0.222	0.012
PETfold^{1k}	0.545	0.794	0.415	0.638	0.798	0.532	0.000	0.000	0.000
GREMLIN	0.208	0.276	0.167	0.285	0.490	0.201	0.038	0.033	0.043
Template-based Modelling	0.012	0.021	0.008	0.013	0.069	0.007	0.014	0.018	0.012
Single-sequence-based									
SPOT-RNA	0.649	0.786	0.552	0.734	0.795	0.682	0.160	0.600	0.092
mxfold2	0.656	0.807	0.552	0.754	0.807	0.707	–	–	–
LinearPartition	0.655	0.807	0.551	0.753	0.807	0.706	–	–	–
CentroidFold (NC)	0.662	0.854	0.540	0.756	0.860	0.675	0.112	**0.667**	0.061
CentroidFold (MEA, NC)	0.658	0.826	0.547	0.751	0.834	0.683	0.111	0.588	0.061
CONTRAFold	0.647	0.781	0.552	0.738	0.795	0.688	0.118	0.478	0.067
RNAfold	0.626	0.759	0.532	0.718	0.759	0.682	–	–	–
IPknot	0.631	0.823	0.512	0.730	0.823	0.656	–	–	–
mxfold (NC)	0.640	0.809	0.530	0.738	0.809	0.678	0.000	0.000	0.000
ProbKnot	0.616	0.726	0.535	0.705	0.726	0.685	–	–	–
LinearFold	0.631	0.852	0.501	0.732	0.852	0.642	–	–	–
RNAstructure	0.606	0.739	0.513	0.696	0.739	0.657	–	–	–
Ufold	0.538	0.683	0.444	0.651	0.762	0.568	0.000	0.000	0.000
RNAshapes (MEA, LP)	0.592	0.646	0.546	0.672	0.646	0.699	0.000	0.000	0.000
pkiss (LP)	0.617	0.732	0.534	0.707	0.732	0.683	–	–	–
CycleFold	0.473	0.464	0.483	0.551	0.533	0.570	0.178	0.184	0.172
2dRNA	0.133	0.554	0.075	0.164	0.554	0.096	–	–	–
E2Efold	0.058	0.114	0.039	0.069	0.114	0.050	–	–	–

[a]Harmonic mean of precision and sensitivity. MEA is maximum expected accuracy structure prediction. MFE is minimum free energy structure prediction. NC is structure prediction by allowing non-canonical base-pairs. LP is structure prediction by allowing lone-pairs. Superscript with the name of alignment-based predictor shows the number of aligned sequences used for the prediction. Refer 'Methods comparison' section for detail. Bold indicates the predictor with the best performance.

0.73, for CentroidAlifold, SPOT-RNA and SPOT-RNA2, respectively. This indicates that SPOT-RNA2 can produce correct secondary structure even when tertiary non-canonical base-pairs are difficult to capture.

For those RNAs with low N_{eff}, artificial but functional homologous sequences from deep mutational scanning may be useful for improving base-pairing prediction (Salehi-Ashtiani *et al.*, 2006; Zhang *et al.*, 2020c). To demonstrate the case, we downloaded the sequences of the CPEB3 ribozyme with relative activity of greater than 0.5 generated from deep mutational scanning (Zhang *et al.*, 2020c). N_{eff} for this ribozyme is small ($N_{eff} = 94$). Figure 8 compares predicted secondary structure by SPOT-RNA, SPOT-RNA2 (with default MSA from RNAcmap), and SPOT-RNA2 with MSA from RNAcmap and Deep Mutational Scanning (DMS) (Zhang *et al.*, 2020c). As SPOT-RNA2 is not trained for handling highly homologous sequences from DMS, secondary structures were obtained from predicted base-pair probability by optimizing MCC on this specific RNA to have a fair comparison between all the methods. SPOT-RNA2 incorporated with deep mutational scanning MSA (Fig. 8C) can detect all four stems in CPEB3 ribozyme including a pseudoknot stem (in green) with an F1-score of 0.80 although, it missed two non-canonical base-pairs and a lone-pair. This is a large improvement over the default SPOT-RNA2 (without the DMS data) and SPOT-RNA which have F1-scores of 0.72, and 0.67, respectively.

4 Discussion

Inspired by the success of protein contact map prediction, this work employed evolution-derived sequence profiles and mutational coupling for RNA secondary structure and tertiary base-pairing prediction. The method, called SPOT-RNA2, improves over the single-sequence-based method SPOT-RNA using the same transfer-learning approach for those sequences with homologous sequences. The improvement is most significant for RNAs with more complex base-pairing patterns containing tertiary contacts such as non-canonical base-pairs, pseudoknots, lone-pairs and base triplets (Supplementary Table S9). More importantly, SPOT-RNA2 makes a highly accurate prediction (F1-score >0.8) for the majority of those sequences with N_{eff} >1000 (14/16 RNAs, 87.5%). Thus, evolution-derived sequence profiles and mutational coupling are important for high accuracy RNA base-pairing prediction.

RNA molecules within the same family have a highly conserved consensus secondary structure (albeit with some sequence-specific

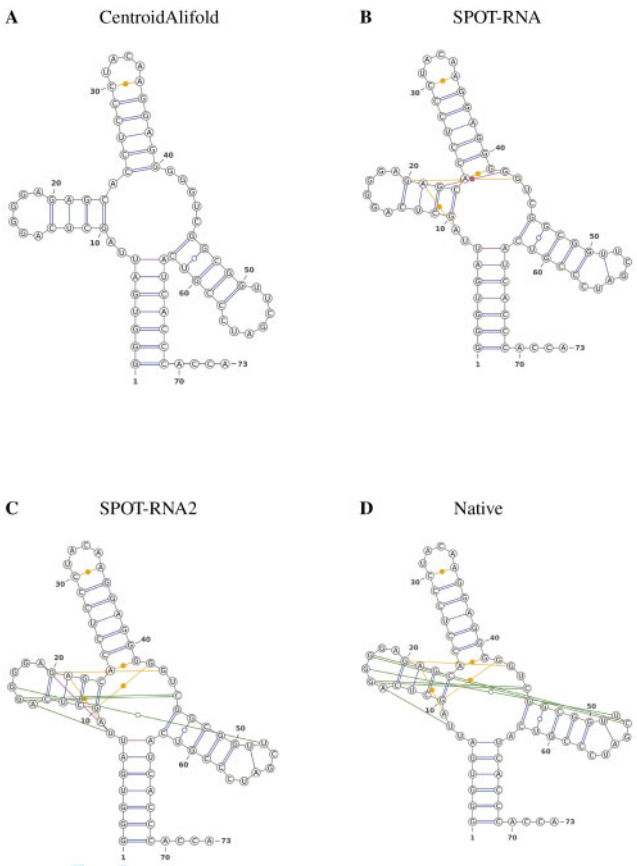

Fig. 6. Comparison of SPOT-RNA2, SPOT-RNA and CentroidAlifold prediction with the native structure of a 70S ribosome (chain 1Y in PDB ID 6cae, $N_{eff} = 1803$). (**A**) predicted structure by CentroidAlifold, with 96% precision and 70% sensitivity. (**B**) predicted structure by SPOT-RNA, with 93% precision and 83% sensitivity. (**C**) predicted structure by SPOT-RNA2, with 94% precision and 97% sensitivity. (**D**) Native structure. [VARNA (Darty *et al.*, 2009) was used for plotting.]

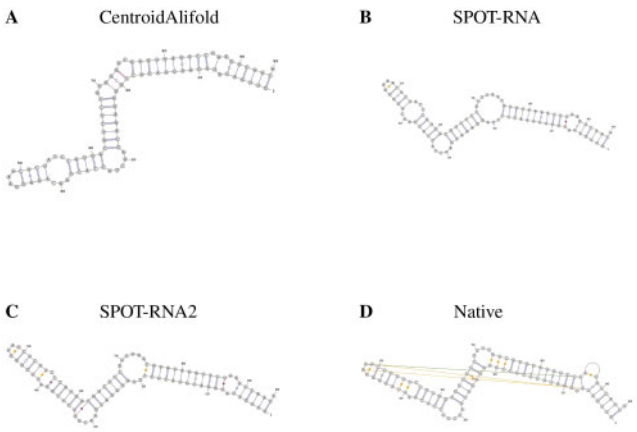

Fig. 7. Comparison of SPOT-RNA2, SPOT-RNA and CentroidAlifold prediction with the native structure of a synthetic construct RNA (chain H in PDB ID 6dvk, $N_{eff} = 1$). (**A**) Predicted structure by CentroidAlifold, with 89% precision and 74% sensitivity. (**B**) Predicted structure by SPOT-RNA, with 97% precision and 77% sensitivity. (**C**) Predicted structure by SPOT-RNA2, with 92% precision and 81% sensitivity. (**D**) Native structure. [VARNA (Darty *et al.*, 2009) was used for plotting.]

variation). This means that knowledge of a labeled RNA sequence from the same family can be an advantage at inference time. The simplest realization of this advantage is using template-based modelling (TBM) by assigning labels to the query based on the most likely evolutionary alignment. In this work, we show that, contrary to classical thermodynamic models, SPOT-RNA2 can benefit from evolutionary similarity to the training set when it is available, without sacrificing performance for totally unseen families as shown in Table 5. Furthermore, this performance is not limited to a simple reproduction of training labels. We compared SPOT-RNA2 to a homology modelling baseline by assigning labels to sequences based on the top-scoring match (E-value < 10) with the training and validation sets. The search was conducted by cmsearch from the covariance model build using the predicted secondary structure by SPOT-RNA. In Supplementary Figure S5, the SPOT-RNA2 model is shown to capture sequence-specific secondary structure preferences beyond the homology modelling, even when high confidence matches can be found in the training set.

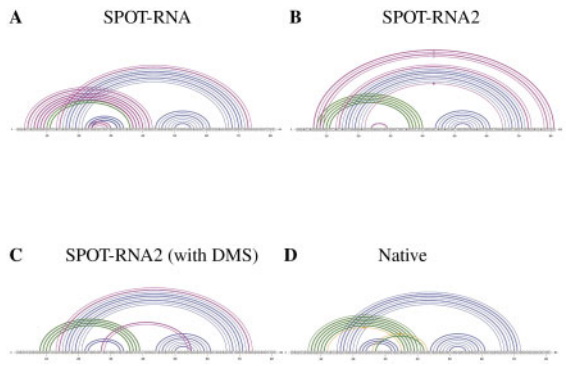

Fig. 8. Performance comparison between SPOT-RNA2 with additional Deep Mutational Scanning (DMS) sequencing data, SPOT-RNA2 and SPOT-RNA with the native structure of a CPEB3 ribozyme RNA. (**A**) predicted structure by SPOT-RNA, with 61% precision and 73% sensitivity. (**B**) predicted structure by SPOT-RNA2, with 70% precision and 73% sensitivity. (**C**) predicted structure by SPOT-RNA2 with additional deep mutational sequencing data, with 83% precision and 77% sensitivity. (**D**) Native structure. [VARNA (Darty *et al.*, 2009) was used for plotting.]

One limitation due to the use of evolutional information in SPOT-RNA2 is the challenge to derive the information for sequences longer than 1000. SPOT-RNA2 relies on RNAcmap, the first tool that automatically takes in an RNA sequence and then performs BLAST-N and INFERNAL for the first round of sequence-based homology search and the second round of sequence-profile and secondary-structure-based search. This pipeline becomes computationally prohibitive for the sequences longer than 1000. As a result, SPOT-RNA2 currently limits to RNA of <1000 nucleotides long. SPOT-RNA does not have this limitation, although its performance is not as accurate as those RNAs with <500 nucleotides long because it is trained on the sequences with <500 nucleotides. As shown in Figure 3B, there is no obvious size dependence from 60 to 500 for SPOT-RNA2 or SPOT-RNA.

Another limitation, also due to the use of evolutionary information, is computing time requirement. Locating homologous sequences and performing multiple sequence alignments are time-consuming, in particular as the RNA sequence library expands exponentially (Coordinators, 2017). It takes about 5–6 h for an RNA of 500 nucleotides with a median number of homologous sequences (N_{eff} =1000) in reference database when 40 thread of Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20 GHz are used. After that, the time for complete prediction by SPOT-RNA2 is relatively short. As one illustrative example, Supplementary Table S4 compares the time requirement of several alignment-based techniques for a 500 nucleotides long RNA and varying number of homologous sequences after homologous sequences were found. For 1000 homologous sequences, SPOT-RNA2 takes 1513 s on a single thread of Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20 GHz. This is much longer than 569 s by CentroidAlifold (MEA), but much shorter than CentroidAlifold (NC), MXSCARNA and TurboFold II (MEA). Unlike other alignment-based predictors (except TurboFold II), SPOT-RNA2 prediction can be easily parallelized by using multiple threads of CPU. For instance, if all the predictor allowed to run on 40 CPU threads, SPOT-RNA2 is the third quickest method after RNAalifold (MFE) and RNAalifold (MEA) as shown in Supplementary Table S4. Thus, SPOT-RNA2 is relatively fast for getting results after the homolog search is done.

One way to reduce computing time is to set a smaller number for the maximum number of homologous sequences. It is currently set at 50000. To understand the impact of the prediction accuracy by presetting a smaller number for allowed homologs, we plotted the average F1-scores of 20 RNAs with at least 10000 homologs as a function of the number of homologs used in generating sequence profiles in Supplementary Figure S6. As the figure shows, there is a slow but steady increase in the average performance when more homologous sequences were employed. Thus, more homologous sequences are better for secondary structure prediction at the expense of computing time.

One more limitation of SPOT-RNA2 is that many RNAs do not have many sequence homologs. In this case, SPOT-RNA is more reliable for those RNAs with N_{eff} <10, in particular. Thus, we strongly recommend comparing SPOT-RNA and SPOT-RNA2 results when SPOT-RNA2 shows N_{eff} <10 for a given sequence and its sequence length is shorter than 50 at the same time (Fig. 3). One recent advancement shows, however, that it is possible to perform deep mutational scanning to generate artificial homologous sequences. These artificial functional and non-functional sequences can produce accurate base-pairing information (Rollins *et al.*, 2019; Zhang *et al.*, 2020c). Indeed, without any modification, SPOT-RNA2 can combine existing homologous sequences with artificial but functional sequences from deep mutational scanning to generate sequence profiles and mutational coupling. The resulting new input leads to improved base-pairing prediction (Fig. 8). This result highlights the ability of SPOT-RNA2 to extract useful information from both natural and artificial homologous sequences.

A more accurate prediction of RNA base-pairs by SPOT-RNA2 offers the potential for improving RNA structure prediction. Using predicted base-pairs or secondary structures as restraints is a common practice for RNA structure prediction. Recent work has shown that using contacts generated from direct coupling analysis of pre-aligned Rfam sequences yield large improvement in predicted three-dimensional structures (De Leonardis *et al.*, 2015; Wang *et al.*, 2017a; Weinreb *et al.*, 2016). This work has an added significance because SPOT-RNA2 makes a large improvement over the direct coupling method such as GREMLIN (Supplementary Table S9) by 84% for canonical base-pairs, 521% for non-canonical base-pairs, 582% for pseudoknot base-pairs and 1521% for lone-pairs, all in F1-scores. More significantly, our method is not limited to the RNAs listed in Rfam families, which currently have 3024 families only (<4% known RNAs). Thus, SPOT-RNA2 will help expand the RNAs whose three-dimensional structures can be accurately predicted with secondary structure restraints.

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 and DP210101875 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)*. USENIX Association, Savannah, GA, pp. 265–283.

Akiyama,M. *et al.* (2018) A max-margin training of RNA secondary structure prediction integrated with the thermodynamic model. *J. Bioinf. Comput. Biol.*, **16**, 1840025.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ba,J.L. *et al.* (2016) Layer normalization. *Preprint arXiv: 1607.06450*.

Bellaousov,S. and Mathews,D.H. (2010) ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA*, **16**, 1870–1880.

Carlson,P.D. *et al.* (2018) SnapShot: RNA structure probing technologies. *Cell*, **175**, 600–600.e1.

Chen,X. *et al.* (2020) RNA secondary structure prediction by learning unrolled algorithms. *Preprint arXiv: 2002.05810*.

Clevert,D.-A. *et al.* (2015) Fast and accurate deep network learning by exponential linear units (ELUs). *Preprint arXiv: 1511.07289*.
Coordinators,N.R. (2017) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **46**, D8–D13.
Danaee,P. *et al.* (2018) bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Res*., **46**, 5381–5394.
Darty,K. *et al.* (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.
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.
Do,C.B. *et al.* (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.
Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
Fu,L. *et al.* (2021) Ufold: fast and accurate RNA secondary structure prediction with deep learning. *Preprint arXiv*, https://www.biorxiv.org/content/10.1101/2020.08.17.254896v3.
Glouzon,J.-P.S. and Ouangraoua,A. (2018) aliFreeFold: an alignment-free approach to predict secondary structure from homologous RNA sequences. *Bioinformatics*, **34**, i70–i78.
Hamada,M. (2015) *RNA Secondary Structure Prediction from Multi-Aligned Sequences*. Springer, New York, NY, pp. 17–38.
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.* (2018) Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*, **34**, 4039–4045.
Hanumanthappa,A.K. *et al.* (2021) Single-sequence and profile-based prediction of RNA solvent accessibility using dilated convolutional neural network. *Bioinformatics*, **36**, 5169–5176.
He,K. *et al.* (2016) Identity mappings in deep residual networks. In: Leibe,B. *et al.* (eds.) *Computer Vision – ECCV 2016*. Springer International Publishing, Cham, pp. 630–645.
Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput*., **9**, 1735–1780.
Huang,L. *et al.* (2019) LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. *Bioinformatics*, **35**, i295–i304.
Janssen,S. and Giegerich,R. (2015) The RNA shapes studio. *Bioinformatics*, **31**, 423–425.
Kalvari,I. *et al.* (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res*., **46**, D335–D342.
Kamisetty,H. *et al.* (2013) Assessing the utility of coevolution-based residue–-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. USA*, **110**, 15674–15679.
Kingma,D.P. and Ba,J. (2014) Adam: a method for stochastic optimization. *Preprint arXiv: 1511.07122*.
Kryshtafovych,A. *et al.* (2019) Critical assessment of methods of protein structure prediction (casp)–Round XIII. *Proteins Struct. Funct. Bioinf*., **87**, 1011–1020.
Kubota,M. *et al.* (2015) Progress and challenges for chemical probing of RNA structure inside living cells. *Nat. Chem. Biol*., **11**, 933–941.
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–e142.
Mao,K. *et al.* (2020) Prediction of RNA secondary structure with pseudoknots using coupled deep neural networks. *Biophys. Rep*., **6**, 146–154.
Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
Pantel,L. *et al.* (2018) Odilorhabdins, antibacterial agents that cause miscoding by binding at a new ribosomal site. *Mol. Cell*, **70**, 83–94.e7.
Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.
Rivas,E. (2020) RNA structure prediction using positive and negative evolutionary information. *PLOS Comput. Biol*., **16**, e1008387–25.
Rollins,N.J. *et al.* (2019) Inferring protein 3D structure from deep mutation scans. *Nat. Genet*., **51**, 1170–1176.
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.
Salehi-Ashtiani,K. *et al.* (2006) A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. *Science*, **313**, 1788–1792.
Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math*., **45**, 810–825.
Sato,K. *et al.* (2009) CentroidFold: a web server for RNA secondary structure prediction. *Nucleic Acids Res*., **37**, W277–W280.
Sato,K. *et al.* (2011) IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, **27**, i85–i93.
Sato,K. *et al.* (2021) RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat. Commun*., **12**, 941.
Schroeder,S.J. and Turner,D.H. (2009) Chapter 17 – optical melting measurements of nucleic acid thermodynamics. In: Herschlag,D. (ed.) *Biophysical, Chemical, and Functional Probes of RNA Structure, Interactions and Folding: Part A, Volume 468 of Methods in Enzymology*. Academic Press, pp. 371–387.
Schuster,M. and Paliwal,K.K. (1997) Bidirectional recurrent neural networks. *IEEE Trans. Signal Process*., **45**, 2673–2681.
Seemann,S.E. *et al.* (2011) The PETfold and PETcofold web servers for intra- and intermolecular structures of multiple RNA sequences. *Nucleic Acids Res*., **39**, W107–W111.
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.
Sloma,M.F. and Mathews,D.H. (2017) Base pair probability estimates improve the prediction accuracy of RNA non-canonical base pairs. *PLoS Comput. Biol*., **13**, e1005827.
Srivastava,N. *et al.* (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res*., **15**, 1929–1958.
Strobel,E.J. *et al.* (2018) High-throughput determination of RNA structures. *Nat. Rev. Genet*., **19**, 615–634.
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.
Tabei,Y. *et al.* (2008) A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, **9**, 33.
Tan,Z. *et al.* (2017) TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Res*., **45**, 11570–11581.
Teplova,M. *et al.* (2020) Crucial roles of two hydrated Mg2+ ions in reaction catalysis of the pistol ribozyme. *Angew. Chem. Int. Ed*., **59**, 2837–2843.
Tinoco,I. and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol*., **293**, 271–281.
Trachman,R.J. *et al.* (2019) Structure and functional reselection of the Mango-III fluorogenic RNA aptamer. *Nat. Chem. Biol*., **15**, 472–479.
Wang,J. *et al.* (2017a) Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide-nucleotide interactions from direct coupling analysis. *Nucleic Acids Res*., **45**, 6299–6309.
Wang,S. *et al.* (2017b) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol*., **13**, e1005324–34.
Weinreb,C. *et al.* (2016) 3D RNA and functional interactions from evolutionary couplings. *Cell*, **165**, 963–975.
Will,S. *et al.* (2015) SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. *Bioinformatics*, **31**, 2489–2496.
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.
Yesselman,J.D. *et al.* (2019) Computational design of three-dimensional RNA structure and function. *Nat. Nanotechnol*., **14**, 866–873.
Yu,F. and Koltun,V. (2015) Multi-scale context aggregation by dilated convolutions. *Preprint arXiv: 1511.07122*.
Zhang,H. *et al.* (2020a) LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. *Bioinformatics*, **36**, i258–i267.
Zhang,T. *et al.* (2020b) RNAcmap: a fully automatic method for predicting contact maps of RNAs by evolutionary coupling analysis. *Preprint arXiv: 10.1101/2020.08.08.242636*.
Zhang,Z. *et al.* (2020c) Accurate inference of the full base-pairing structure of RNA by deep mutational scanning and covariation-induced deviation of activity. *Nucleic Acids Res*., **48**, 1451–1465.
Zhao,Y. *et al.* (2018) Evaluation of RNA secondary structure prediction for both base-pairing and topology. *Biophys. Rep*., **4**, 123–132.