UFold: Fast and Accurate RNA Secondary Structure Prediction with Deep Learning

Yingxin Cao[1,4], Laiyi Fu[1,4], Jie Wu[2], Qing Nie[3], Xiaohui Xie[1*]

[1]Dept. of Computer Science, University of California, Irvine, CA, USA
[2]Dept. of Biological Chemistry, University of California, Irvine, CA, USA
[3]Dept. of Mathematics, University of California, Irvine, CA, USA
[4]These authors contribute equally

Correspondence author:

Prof. Xiaohui Xie
Dept. of Computer Science,
University of California, Irvine
Irvine, CA
USA

Email: xhx@uci.edu
Phone: 949-824-9289

Abstract

For many RNA molecules, the secondary structure is essential for the correction function of the RNA. Predicting RNA secondary structure from nucleotide sequences is a long-standing problem in genomics, but the prediction performance has reached a plateau over time. Traditional RNA secondary structure prediction methods are primarily based on thermodynamic models through free energy minimization. Here we propose a deep learning-based method, called UFold, for RNA secondary structure prediction, trained directly on annotated data without any thermodynamic assumptions. UFold improves substantially upon previous models, achieving approximately 31% improvement over traditional thermodynamic models and 24.5% improvement over other learning-based models. An online web server running UFold is publicly available at https://ufold.ics.uci.edu.

Introduction

The biology of RNA is diverse and complex. Aside from its conventional role as an intermediate between DNA and protein, cellular RNA consists of many other functional classes, including ribosomal RNA (rRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), microRNA, and other noncoding RNAs[1]. Some RNAs possess catalytic functionality, playing a role similar to protein enzymes. The spliceosome, which performs intron splicing, is assembled from several snRNAs. The microRNAs are abundant in many mammalian cell types, targeting approximately 60% of genes[2].

Cellular RNA is typically single stranded. RNA folding is in large part determined by nucleotide base pairing, including canonical base pairing - A-U and C-G, and non-canonical base pairing - primarily G-U pairing[34]. The base-paired structure is often referred to as the secondary structure of RNA. For many RNA molecules, the secondary structure is essential for the correct function of the RNA, in many cases, more so than the primary sequence itself. As an evidence of this, many homologous RNA species demonstrate conserved secondary structures accompanied with diverged sequences characterized by compensatory mutations.

RNA secondary structure can be determined from atomic coordinates obtained from X-ray crystallography, nuclear magnetic resonance (NMR), or cryogenic electron microscopy[5]. However, these methods have low throughput, and only a very small fraction of RNAs have experimentally determined structures. To address this limitation, experimental methods have been proposed to infer base paring by using probes based on enzymes, chemicals, and cross-linking techniques coupled with high-throughput sequencing[6]. Although promising, these methods are still at the early stage of development, unable to provide precise base-pairing at a single nucleotide solution.

Computationally predicting the secondary structure of RNA is a long-standing problem in genomics and bioinformatics. Many methods have been proposed over the past two decades. They can be broadly classified into two categories: 1) single sequence prediction method, and 2) comparative methods. In the first category, the most common method is to search for thermodynamically stable states through free energy minimization. If the secondary structure contains only nested base pairing, the energy minimization can be efficiently solved through dynamic programming, such as those implemented in as Vienna RNAfold[7], MFold[8], RNAstructure[9], and CONTRAfold[10]. The dynamic programming method for nearest neighbor energy functions has a time complexity of $O(n^3)$ with n being the sequence length. Faster algorithms have been proposed, such as Rfold[11], Vienna RNAplfold[12] and LocalFold[13], which can decrease the running time to $O(nL^2)$, but are still limited by sequence length ($L<=150$). Efficient dynamic programming algorithm that samples suboptimal secondary structures from the Boltzmann ensembles of structures has also been proposed (e.g., CentroidFold[14]). However, dynamics programming breaks down when base pairs contain non-nested patterns, called pseudoknots, which contains two stem-loop structures with half of one stem intercalating between the two halves of another stem. Predicting secondary structures with pseudoknots is hard, and has shown to be NP-complete under the energy minimization

framework. Methods in the secondary category utilize covariance methods by aligning related RNA sequences and identifying correlated compensatory mutations. Although the list of proposed methods in each category is long and diverse[15], the performance of these methods has not been significantly improved over time, reaching a performance ceiling of about 80%[16], possibly because they fail to account for base pairing resulting from tertiary interactions[17], unstacked base pairs, pseudoknot, noncanonical base pairing, and others[4].

Recently deep learning techniques have started to emerge as an alternative approach to the RNA secondary structure prediction problem[18 19 20 21]. Compared to the thermodynamic model-based approaches, the learning-based methods have the benefits of making few assumptions, allowing pseudoknots, and accounting for tertiary interactions, noncanonical base pairing or other previously unrecognized base pairing constraints. Existing deep learning methods differ in model architectural design and in their choices of model input and output. The majority of these methods treat the input as a sequence, utilizing LSTM or transformer encoder to capture long-range interactions between nucleotides[8 9 10].

Here we propose a simple and yet highly effective deep learning framework for RNA secondary structure prediction. Instead of using the nucleotide sequence itself, the input of our model consists of all possible base pairing maps within the input sequence. Each map, represented by a square matrix of the same dimension as the input sequence length, denotes the occurrences of one of the 16 possible base pairs between the input nucleotides. Under this new representation, the input is treated as a 2D "image" with 16 channels, allowing the model to explicitly consider all long-range interactions and all possible base pairing, including noncanonical ones.

We use an encoder-decoder framework to extract both long- and short-range interaction features of the input sequence, implemented in a U-Net model. For this reason, we will refer to our method as UFold (stands for U-Net based on RNA folding). The output of UFold is the predicted contact score map between the bases of the input sequence. UFold is fully convolutional, and as such it can readily handle input sequences with variable length.

We conduct experiments to compare the performance of UFold against both the traditional energy minimization-based methods and recent learning-based methods. We show that UFold yields substantial performance gain over previous methods, achieving about 75% improvement over energy minimization-based methods, and 17% over the state-of-the-art learning-based methods. In particular, the F1 score reaches 0.96 on the RNAStralign test dataset, highlighting its high potential in solving the RNA secondary structure prediction problem.

UFold is fast with an inference time of about 90ms per sequence for sequences with length up to 1600bp. We have developed an online web server running UFold RNA secondary structure prediction. The sever is freely available at http://ufold.ics.uci.edu, allowing users to enter sequences and see predicted secondary structures.

Methods

The general problem of the RNA secondary structure prediction is to predict base pairing patterns given an input sequence. Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_L)$ with $x_i \in \{'\mathrm{A}', '\mathrm{U}', '\mathrm{C}', '\mathrm{G}'\}$ be an input sequence of length L. The goal is to predict the secondary structure of $\boldsymbol{x}$, represented by a contact matrix $A \in \{0,1\}^{L \times L}$ with $A_{ij} = 1$denoting a base pairing between bases x_i and x_j, and 0 otherwise. UFold utilizes a deep neural network to predict the contact matrix given the input. Next we describe several design choices behind UFold (Figure 1).

Input and output representation

Most existing learning-based methods treat the input as a sequence and use recurrent neural nets (RNNs) to model the interaction between different bases. Gated RNNs, such as LSTMs and GRUs, are often the method of choice for dealing with sequential data because of their ability of modeling long range dependencies. However, RNN models need to be run sequentially, causing issues in both training and inference. Newer RNA structure prediction models based on transformers, which does not require the sequential data to be processed in order, have also been proposed.

Unlike the previous models, UFold converts the input sequence into an "image". This is done by first encoding $\boldsymbol{x}$ with one-hot representation, representing the sequence with an $L \times 4$ binary matrix $X \in \{0,1\}^{L \times 4}$. $\boldsymbol{x}$ is then transformed into a $16 \times L \times L$ tensor through a Kronecker product between $\boldsymbol{x}$ and itself, followed by reshaping dimensions,

$$I = X \otimes X.$$

In this representation, input I can be understood as an image of with size $L \times L$ with 16 color channels. Each channel specifies one of the 16 possible base pairing rules; $I(i, j, k)$ denotes whether bases x_j and x_k are paired according to the *i*-th base pairing rule (e.g., i=2 for A-C pairing).

UFold takes I as input and computes $Y = f(I; \theta)$ with a deep convolutional neural net. The output $Y \in [0,1]^{L \times L}$ is an $L \times L$ matrix, with Y_{ij} denoting the probability score of nucleotides bases x_i and x_j being paired.

The new input representation taken by UFold has several advantages: First, using an image representation allows it to explicitly model all possible long-range interactions. Base pairing between distant sequence segments shows up locally in the image representation. Second, it considers all possible base pairing patterns, making no distinction between canonical and non-canonical base pairs. Third, it allows us to implement a fully convolutional neural model that can handle variable sequence length, eliminating the need of padding the input sequence to a fixed length.

Scoring network architecture

UFold uses an encoder-decoder architecture for computing predicted contact score matrix *Y* (Figure 1). The model consists of a sequence of down-sampling layers (encoder) to derive increasingly complex semantic representations of the input, followed by a sequence of up-sampling layers (decoder), with lateral connections from the encoder to fill in contextual

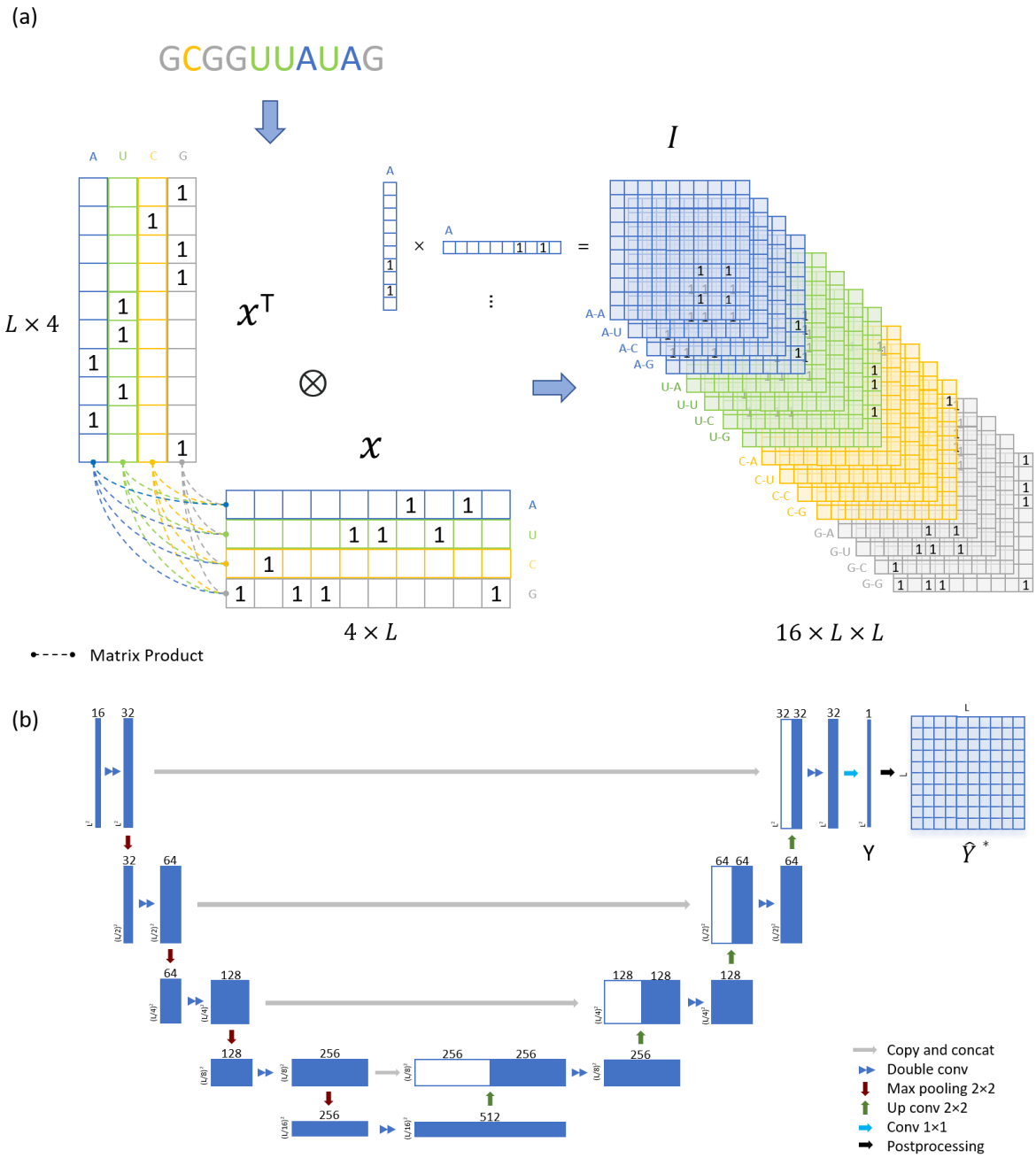


Figure 1. The overall architecture of UFold. (a) A sequence is first converted into one-hot embedding. A novel representation of sequence is then introduced by taking outer product of all combinations of base pair channels, resulting in an image-like representation with 16 channels and the same size as the contact map. (b) Detailed architecture of our framework. The input is a $16 \times L \times L$ tensor representation of original sequence. The U-Net takes the $16 \times L \times L$ input and outputs a $L \times L$ score matrix Y. After postprocessing, matrix $\hat{Y}^*$ is the final prediction for contact map.

information. The overall design follows the U-Net model, widely used in the field of image segmentation.

All operations in UFold are fully convolutional. Thus, the input sequence can be of variable length, with the output matrix changing correspondingly. This feature is especially beneficial for RNA secondary structure as the range of the input sequence length is very large, from

ends of nucleotides for small RNAs to thousands of nucleotides for large RNAs. Padding input sequences to the same length as done in other methods would have significantly impacted the efficiency of the algorithm.

UFold is trained by minimizing the cross-entropy between the predicted probability contact matrix and the true contact matrix,

$$Loss(Y, A;\ \theta) = -\sum_{ij}\left[A_{ij}\log\left(Y_{ij}\right) + \left(1 - A_{ij}\right)\log\left(1 - Y_{ij}\right)\right].$$

using stochastic gradient descent.

Postprocessing

After the contact scoring matrix Y is computed by UFold, we use a postprocessing procedure to derive the final secondary structure. The postprocessing procedure takes into account four hard constraints in the secondary structure: 1) the contact matrix should be symmetric; 2) only canonical plus U-G paring rules are allowed (this can be relaxed by including other non-canonical base pairs); 3) no sharp loops are allowed, for which we set $A_{ij} = 0, \forall i, j$ with $|i - j| < 4$; and 4) no overlapping pairs are allowed, that is, $A\mathbf{1} \leq \mathbf{1}$. We follow the steps used in e2efold[19] by encoding constraints 2 & 3 into a matrix M, defined as $M(x)_{\mathrm{ij}} \coloneqq 1$ if nucleotides x_i and x_j can be paired under constraints 2 & 3 and equals to 0 otherwise.

To address the first two constraints, we transform *Y* according to

$$\mathcal{T}(Y) \coloneqq \frac{1}{2}(Y + Y^T) \circ M(x),$$

where $\circ$ denotes element-wise multiplication. It ensures that the transformed *Y* is symmetric and satisfies constraints 1, 2 and 3.

To address the last constraint, we relax it into a linear programming problem,

$$\hat{Y}^* = \underset{\hat{Y} \in R^{L \times L}}{\operatorname{argmax}} \left\langle \hat{Y}, \mathcal{T}(Y) \right\rangle, \qquad \text{subjec to } \hat{Y}\mathbf{1} \leq \mathbf{1}$$

which tries to find an optimal scoring matrix $\hat{Y}$ that is most similar to $\mathcal{T}(Y)$ while at the same time satisfying the nonoverlapping pair constraint. The similarity is measured in terms of the inner product between $\hat{Y}$ and $\mathcal{T}(Y)$.

The final predicted binary contact map is taken to be $\hat{Y}^*$ after thresholding it with an offset, which is chosen through a grid search.

Results

We compare the performance UFold with traditional thermodynamic models and recent learning-based models. Our experiments show that UFold outperforms the previous methods by a large margin in both prediction accuracy and speed.

Datasets

Two benchmark datasets are used in this study: a) RNAStralign[22], which contains 30,451 unique sequences from 8 RNA families and is one of the most comprehensive RNA structure datasets available; and b) ArchiveII[23], which contains 3,975 sequences from 10 RNA families and is the most widely used dataset for benchmarking RNA structure prediction performance. The RNAStralign dataset is randomly split into training, validation and test sets, with 24,895, 2,702, and 2,854 samples, respectively. UFold is trained on the RNAStralign training set with hyperparameters tuned on the RNAStralign validation set. ArchiveII is only used for performance evaluation. During the training, stratified sampling[20] is applied to the training set to balance the number of training samples from each RNA family.

Evaluation metrics

For evaluation, we report three metrics: F1 score, precision and recall. Precision is defined as $Prec = \frac{TP}{TP+FP}$, evaluated on all predicted base pairs. Recall is defined as $Recall = \frac{TP}{TP+FN}$. And F1 score is the harmonic mean of precision and recall, defined as $F1 = 2 \cdot \frac{Prec \cdot Reacll}{Prec+Recall}$. In the literature, precision and recall are sometimes calculated after allowing for a small local mismatch. For this reason, we also report Recall(S), Prec(S) and F1(S), in which one position shift is tolerated.

Experimental results

Table 1 summarizes the performance of UFold on the RNAStralign test set, together with the performances of a collection of both traditional energy-based and recent learning-based methods (the performance numbers are quoted from the E2Efold[20] paper). The traditional methods achieve a F1 score in the range of 0.42 to 0.61. Recent learning-based methods improve the F1 score to 0.82 (E2Efold). UFold is able to further improve the performance, achieving a F1 score of 0.96. Compared with the state-of-the-art method E2Efold, UFold achieves a 13.9% increase in F1 score, a 17% increase in recall, and a 9.7% increase in precision.

Table 2 summarizes the evaluation results on the ArchiveII dataset, which is invisible during UFold training. The performances of the energy-based methods are similar to those in the RNAStralign dataset. The performance of E2EFold is significantly reduced, achieving a F1 score of 0.686, although still being better than the other previous methods. By contrast, UFold achieves a F1 score of 0.931 on this dataset, corresponding to a 24.5% improvement over the state-of-the-art and 31% improvements over traditional methods. Improvements on recall and precision are also substantial.

Predicting secondary structures with pseudoknots is especially challenging for thermodynamic models. Next we validate the performance of UFold on predicting base pairing in the presence of pseudoknots. For this purpose, we pull out all RNA structures with pseudoknots in the RNAStralign test set, on which we then benchmark UFold against two

other methods that can predict pseudoknots, including E2Efold and RNAstructure. As shown in Table 3, both the performances of E2Efold and RNAstructure significantly drop in the presence of pseudoknots. By contrast, UFold still achieves a high F1 score of 0.944, highlighting the robustness of UFold in the presence of pseudoknots.

Table 1: Evaluation results on the RNAStralign test set.

Method	Prec	Rec	F1	Prec(S)*	Rec(S)	F1(S)
UFold	**0.963**	**0.958**	**0.960**	**0.969**	**0.964**	**0.967**
E2Efold	0.866	0.788	0.821	0.880	0.798	0.833
CDPfold	0.633	0.597	0.614	0.720	0.677	0.697
LinearFold	0.620	0.606	0.609	0.635	0.622	0.624
Mfold	0.450	0.398	0.420	0.463	0.409	0.433
RNAstructure	0.537	0.568	0.550	0.559	0.592	0.573
RNAfold	0.516	0.568	0.540	0.533	0.587	0.558
CONTRAfold	0.608	0.663	0.633	0.624	0.681	0.650

*"S" indicates results when one-position shift is allowed.

Table 2: Evaluation results on the ArchiveII dataset.

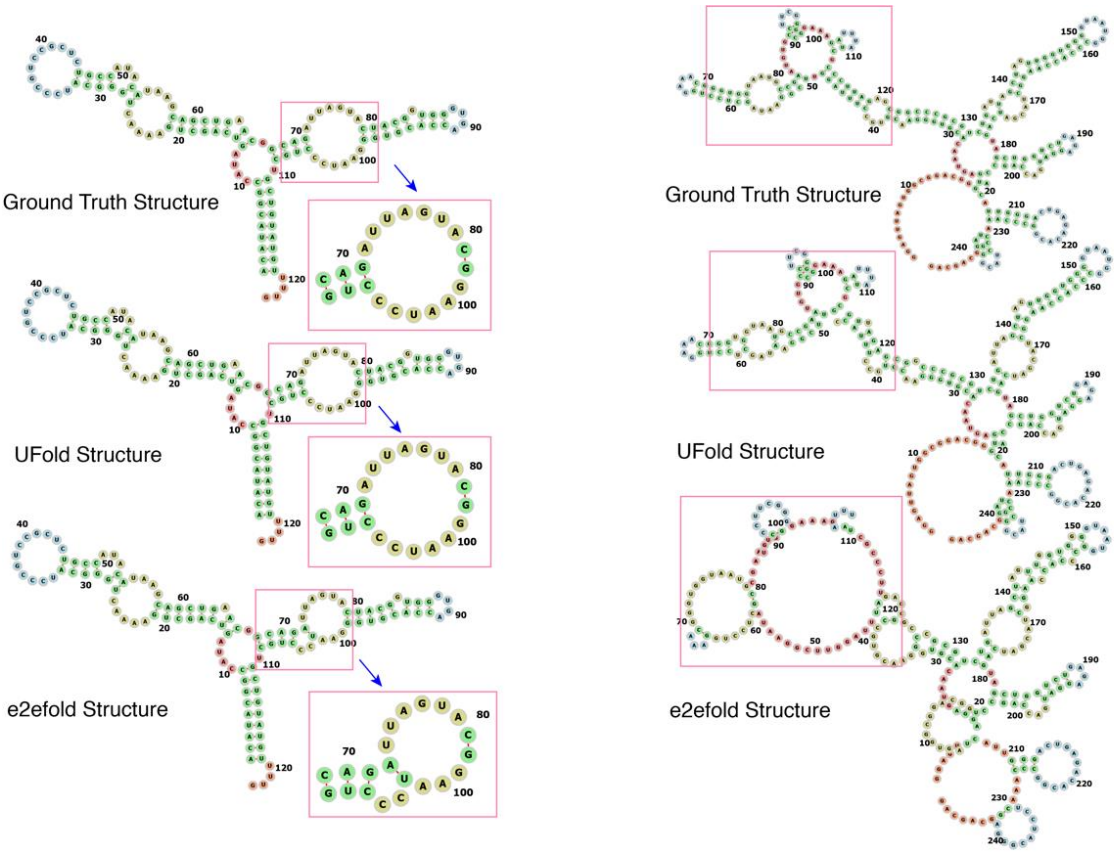
Method	Prec	Rec	F1	Prec(S)	Rec(S)	F1(S)
UFold	**0.931**	**0.935**	**0.931**	**0.942**	**0.947**	**0.943**
E2Efold	0.734	0.660	0.686	0.758	0.676	0.704
CDPfold	0.557	0.535	0.545	0.612	0.585	0.597
LinearFold	0.641	0.617	0.621	0.668	0.644	0.647
Mfold	0.428	0.383	0.401	0.450	0.403	0.421
RNAstructure	0.563	0.615	0.585	0.590	0.645	0.613
RNAfold	0.565	0.627	0.592	0.586	0.652	0.615
CONTRAfold	0.607	0.679	0.638	0.629	0.705	0.662

Table 3: Pseudoknot prediction results on the RNAStralign test set.

Method	F1	TP	FP	TN	FN
UFold	**0.944**	**1247**	**49**	**1491**	**3**
E2Efold	0.710	1312	242	1271	0
RNAstructure	0.472	1248	307	983	286

Visualization

Next, we visualize the RNA secondary structures predicted by UFold. For this purpose, the predicted contact maps were first converted to a dot-bracket format according to base pair positions. Raw sequences with the corresponding predicted dot-bracket sequences were fed into the ViennaRNA web server Forna[24] to obtain the visualization result. As a comparison, we also show the predicted structures from E2Efold as well as the ground-truth structures. Two examples are shown in Figure 2. In both cases, UFold generates RNA secondary structures more similar to the ground-truth.



(a) 5S_rRNA_database/Eukaryota/E01165.ct (b) 16S_rRNA_database/Alphaproteobacteria/L41814.ct

Figure 2. Visualization of two example UFold RNA secondary structure predictions. From top to bottom: ground truth, UFold prediction, and E2Efold prediction. Two RNA sequences are (a) 5S_rRNA/Eukaryota/E01165.ct and (b) 16S_rRNA/Alphaproteobacteria/L41814.ct. Major prediction differences are highlighted by red boxes. In both cases, UFold produces predictions more aligned with the ground-truth.

Inference time

The speed of the prediction algorithm is an important factor in RNA secondary structure prediction. Traditional energy minimization-based methods tend to be slow because of the time complexity of the minimization algorithm. UFold inference, on the other hand, runs on feedforward neural nets, with operations highly parallelizable.

The average inference time per sequence of UFold on the RNAStralign test set is reported in Table 4, together with the average running times of other methods. UFold is much faster than both learning-based and energy-based methods. UFold is nearly five-times faster than E2Ffold, and orders-of-magnitude faster than RNAstruture, a traditional method that can handle pseudoknots.

Table 4: Inference time on the RNAStralign test set.

Method	Time per seq
UFold (Pytorch)	**0.09s (GPU)**
E2Efold (Pytorch)	0.40s (GPU)
CDPfold (Pytorch)	300.107s
LinearFold (C)	0.43s

Mfold (C)	7.65s
RNAstructure (C)	142.02s
RNAfold (C)	0.55s
CONTRAfold (C)	30.58s

Web server

For the users' convenience, we also developed a web server running UFold on the backend and made it freely available for anyone who would like to try out UFold for RNA secondary structure predictions. Users can type in or upload their own RNA sequences in FASTA format. Our server predicts RNA secondary structures using the pre-trained UFold model and stores predicted structures in a dot-bracket file for end-users to download. The server also provides an interface to connect to the ViennaRNA web for visualizing predicted structures. The UFold website is available at https://ufold.ics.uci.edu.

Discussion

In this study, we present UFold, a new deep learning-based model for RNA secondary structure prediction. We show that UFold outperforms previous methods by a large margin, achieving 31% performance improvement over traditional thermodynamic methods and 24.5% improvement over the state-of-the-art learning-based method. UFold achieves a F1 score of 0.960 on the RNAStralign dataset and 0.931 on the ArchiveII dataset, bringing in substantial gains in RNA secondary prediction accuracy. In addition, UFold is fast, being able to generate predictions at roughly 90ms per sequence.

A key difference between UFold and previous learning-based methods is its architectural design. Instead of using raw sequences as input, UFold converts sequences into "images", explicitly modeling all possible base pairing between the nucleotides of the input sequence. This choice of input representation has several important implications: First, base pairing patterns between distant sequence segments show up locally in the image representation, making the detection and learning of these distant base pairing patterns easier. Second, all base pairing patterns are explicitly represented in the input, allowing the model to pick up all potential base pairing rules that might contribute to the formation of the secondary structure. Lastly, but perhaps most importantly, the image representation allows us to implement a fully convolutional model to pick up base pairing features across multiple scales through an encoder-decoder architecture. This implement is not only efficient, with operations highly parallelable and allowing for variable input sequence length, but also highly effective in combining both local and global features for the final prediction.

Although UFold demonstrates a great potential in solving the RNA secondary structure prediction problem, as a learning-based method, its performance is inevitably closely attached to the quality of training data. Unfortunately, the number of experimentally resolved RNA secondary structures through X-ray crystallography or NMR remains small. Many secondary structures in the RNAStralign dataset are computationally generated by

aligning homologous sequences. Because the secondary structures from different RNA families are often highly dissimilar, evaluating UFold on a previously unseen RNA family would provide a much more stringent criterion for testing its generalization ability.

We should note that the method presented here can potentially be applied for protein structure prediction as well. The number of amino acids is much higher than the number of bases. It is worth to explore whether all amino acid pairs, which is 400, or a subset of them should be considered in the input representation.

In summary, we show the promising potential of deep learning in solving the long-standing RNA secondary structure problem. The new framework presented here brings in significant performance improvement. If the results are further validated on other independent datasets, it would represent a milestone in boosting the accuracy of RNA secondary structure prediction.

References

1. Geisler, S. & Coller, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.* **14**, 699–712 (2013).
2. Gebert, L. F. & MacRae, I. J. Regulation of microRNA function in animals. *Nat. Rev. Mol. Cell Biol.* **20**, 21–37 (2019).
3. Fallmann, J. *et al.* Recent advances in RNA folding. *J. Biotechnol.* **261**, 97–104 (2017).
4. Westhof, E. & Fritsch, V. RNA folding: beyond Watson–Crick pairs. *Structure* **8**, R55–R65 (2000).
5. Strobel, E. J., Yu, A. M. & Lucks, J. B. High-throughput determination of RNA structures. *Nat. Rev. Genet.* **19**, 615–634 (2018).
6. Bevilacqua, P. C., Ritchey, L. E., Su, Z. & Assmann, S. M. Genome-wide analysis of RNA secondary structure. *Annu. Rev. Genet.* **50**, 235–266 (2016).
7. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
8. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
9. Mathews, D. H. & Turner, D. H. Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.* **16**, 270–278 (2006).
10. Do, C. B., Woods, D. A. & Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**, e90–e98 (2006).
11. Kiryu, H., Kin, T. & Asai, K. Rfold: an exact algorithm for computing local base pairing probabilities. *Bioinformatics* **24**, 367–373 (2008).
12. Bernhart, S. H., Hofacker, I. L. & Stadler, P. F. Local RNA base pairing probabilities in large sequences. *Bioinformatics* **22**, 614–615 (2006).

13. Lange, S. J. *et al.* Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res.* **40**, 5215–5226 (2012).

14. Sato, K., Hamada, M., Asai, K. & Mituyama, T. CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Res.* **37**, W277–W280 (2009).

15. Kings Oluoch, I., Akalin, A., Vural, Y. & Canbay, Y. A Review on RNA Secondary Structure Prediction Algorithms. in *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)* 18–23 (IEEE, 2018). doi:10.1109/IBIGDELFT.2018.8625347.

16. Seetin, M. G. & Mathews, D. H. RNA structure prediction: an overview of methods. in *Bacterial Regulatory RNA* 99–122 (Springer, 2012).

17. Nowakowski, J. & Tinoco Jr, I. RNA structure and stability. in *Seminars in virology* vol. 8 153–165 (Elsevier, 1997).

18. Zhang, H. *et al.* A New Method of RNA Secondary Structure Prediction Based on Convolutional Neural Network and Dynamic Programming. *Front. Genet.* **10**, 467 (2019).

19. Wang, L. *et al.* DMFold: A novel method to predict RNA secondary structure with pseudoknots based on deep learning and improved base pair maximization principle. *Front. Genet.* **10**, 143 (2019).

20. Chen, X., Li, Y., Umarov, R., Gao, X. & Song, L. RNA secondary structure prediction by learning unrolled algorithms. in *International Conference on Learning Representations* (2019).

21. Singh, J., Hanson, J., Paliwal, K. & Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.* **10**, 1–13 (2019).

22. Tan, Z., Fu, Y., Sharma, G. & Mathews, D. H. TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Res.* **45**, 11570–11581 (2017).

23. Sloma, M. F. & Mathews, D. H. Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. *RNA* **22**, 1808–1818 (2016).

24. Kerpedjiev, P., Hammer, S. & Hofacker, I. L. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics* **31**, 3377–3379 (2015).