ncRFP: A Novel end-to-end Method for Non-Coding RNAs Family Prediction Based on Deep Learning

Linyu Wang, Shaoge Zheng, Hao Zhang, Zhiyang Qiu, Xiaodan Zhong, Haiming Liu, and Yuanning Liu

Abstract—Evidence has accumulated enough to prove non-coding RNAs (ncRNAs) play important roles in cellular biological processes and disease pathogenesis. High throughput techniques have produced a large number of ncRNAs whose function remains unknown. Since the accurate identification of ncRNAs family is helpful to the research of their function, it is of necessity and urgency to predict the family of each ncRNAs. Although several traditional excellent methods are applicable to predict the family of ncRNAs, their complex procedures or inaccurate performance remain major problems confronting us. The main idea of those methods is first to predict the secondary structure, and then identify ncRNAs family according to properties of the secondary structure. Unfortunately, the multi-step error superposition, especially the imperfection of RNA secondary structure prediction tools, maybe the cause of low accuracy. In this paper, a novel end-to-end method 'ncRFP' was proposed to complete the prediction task based on Deep Learning. Instead of predicting the secondary structure, ncRFP predicts the ncRNAs family by automatically extracting features from ncRNAs sequences. Compared with other methods, ncRFP not only simplifies the process but also improves accuracy. The source code of ncRFP can be available at https://github.com/linyuwangPHD/ncRFP.

Index Terms—ncRNAs family, ncRNAs sequences, deep learning, end-to-end, ncRFP

1 INTRODUCTION

The expression of protein-coding genes (messenger RNAs: mRNAs) has been the focus of life studies for decades. But in recent years, increasing evidence has shown that ncRNAs play multi vital roles in biological processes [1] and complex diseases [2] by means of replication, transcription or gene expression regulation [3, 4, 5]. With the continuous research, a large number of ncRNAs have been identified based on their sequence length, structure properties or function. The most well-known ncRNAs are transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs). tRNAs, which include the highly conserved secondary structure of the ‘cloverleaf model’, are mainly responsible for the transportation of amino acids in the course of translation [6], and tRNAs mainly take charge of the synthesis of peptide chains [7]. Other famous ncRNAs are microRNAs (miRNAs) and long non-coding RNAs (IncRNAs). miRNAs are small single-strand ncRNAs of approximately 22 nucleotides in length, which are directly bound to the 3'-untranslated region of target genes and negatively regulate them. Hence, miRNAs are essential in many physiological and pathological processes, such as cancer and Alzheimer. Through regulation, they can promote or inhibit cell proliferation, and then act as tumor suppressors or oncogenes [8]. IncRNAs are longer than 200 nt nucleotides [9], which are important signal transduction regulators by various patterns. One of the most acknowledges molecule mechanisms of IncRNAs is to act as “sponges” to modulate the activity of miRNAs [10]. Other important ncRNAs are small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), silencing RNAs (siRNAs), riboswitches, internal ribosome entry sites (IRES) and so on. ncRNAs have a three-hierarchy structure [11]: primary structure, secondary structure, and tertiary structure. The primary structure is the bases sequence. Although it seems that ncRNAs sequences are not as conservative as secondary/tertiary structures, the same family of ncRNAs contains unify seeds. The secondary structure refers to the planar structure formed by a combination of secondary structure elements with a variety of specific shapes through its own folding and base pairing within the sequence. Since the secondary structure with similar functions is conservative, it can be expressed as a graph structure to predict different ncRNAs family. The tertiary structure refers to the stable three-dimensional conformation formed in space by the interaction of secondary structure elements. Although the tertiary structure is conservative, it is not suitable for ncRNAs family prediction because of the difficulty to obtain.

In this era of high-throughput technology [12], a great number of unknown ncRNAs have been found and the function research on these ncRNAs has imposed great challenges to researchers. For the reason that different families of ncRNAs have distinct functions, accurately predict ncRNAs family is conducive to the research on ncRNAs function. Identifying the ncRNAs family via biological experimental methods is time-consuming, laborious and expensive, which cannot meet the needs of high throughput technology. Therefore, it is inevitable to use computational methods to predict the family of ncRNAs. There are several traditional excellent methods (GraPPLE [13], RNAcon [14], nRC [15]) can complete the prediction task. GraPPLE uses machine learning methods to predict ncRNAs family based on the secondary structure features. RNAcon considers 20 graph features obtained from the predicted ncRNAs secondary structure and adopts an RF classifier. nRC currently represents the state-of-art method, where the secondary structure typical features are first extracted by Moss [16] and processed into one-hot code, and then the convolutional neural network is employed to identify ncRNAs. Those traditional methods are all required to get the secondary structure of ncRNAs at the beginning and then identify ncRNAs family with the secondary structure features. Because the real secondary structure of ncRNAs is difficult to obtain, those traditional methods need to use the computational tools (iFknot [17], RNAfold [18]) to get the secondary structure of ncRNAs. There are errors in the process of predicting the ncRNAs secondary structure, which will have disadvantage impact on other steps. Therefore, the low prediction accuracy of traditional methods may be caused by multi-step error superposition, especially the imperfection of secondary structure prediction tools.

In those traditional methods, ncRNAs secondary structure is predicted based on ncRNAs sequences, the characteristics of secondary structure are also hidden in the ncRNAs sequences. Therefore, so long as the method of extracting features is appropriate, perfect family features can be extracted from ncRNAs sequences to identify their family. In accordance with the requirement of ncRNAs family prediction and the shortcomings of traditional methods, a novel end-to-end method 'ncRFP' was proposed to predict the family of ncRNAs. The main novelty of ncRFP is that it is different from those traditional methods to predict the family process, which directly predicts the family based on ncRNAs sequences. With ncRNAs sequences as input, ncRFP extracts features automatically and learn them with the assistance of a deep learning...
Fig. 1. Prediction accuracy of multiple models in different length of input. The violet, green and red curve represents the accuracy of RNN, CNN, and DNN respectively.

model. In this paper, three models were created based on RNN, CNN, and DNN respectively. After compared those models, the prediction performance of RNN is greater than the other two models. Fig. 1 shows the performance comparison among those models. Hence, we chose the RNN model, which contains Bi-LSTM [19], Attention Mechanism [19], and fully connected neural networks [20], as the final model. Compared with other methods, ncRFP reduces method steps and error sources. Therefore, it not only pre-digests the process but also has the potential to improve the prediction accuracy.

2 MATERIALS AND METHODS

2.1 Data Collection and progressing

The data used in this paper comes from the recent literature [15], which was collected from the Rfam database [21]. It includes: microRNAs, 55_rRNA, 5.8S_rRNA, ribozymes, CD-box, HACA-box, scaRNA, tRNA, Intron gpl, Intron gplII, IRES, leader and riboswitch 13 different types of ncRNAs. There are 6320 non-redundant ncRNAs sequences, which IRES contains 320 ncRNAs sequences, and each of the other family contains 500 ncRNAs sequences. All sequences of each family were randomly divided into 10 parts. One part of each family is randomly selected to form a test set and the rest to form a train set, so that all ncRNAs sequences can form 10-fold cross-validation train sets and test sets. In this paper, two encoding methods were selected to convert ncRNAs sequences into matrixes. The first one is to convert each of base into a vector of 1’8 [22], another one is to convert each of base into a vector of 1’4 (one-hot encoding). Tables 1 and 2 show the conversion rules of those two encoding methods.

Because of the different lengths of various ncRNAs, it is convenient for training and testing to process different lengths of ncRNAs into the same length. In this paper, a new intercepting/padding method (IPM) was put forward to process various ncRNAs into a fixed length. ncRNAs with length greater than the fixed length are truncated from the beginning to the fixed length, and with length less than the fixed length are filled with ‘N’ at the tail to the fixed length. In order to choose an excellent encoding method and appropriate fixed length, we randomly chose a fold train set and test set to training and testing ncRFP. Fig. 2 shows the prediction accuracy in different lengths and different codes. It can be seen that the accuracy of 1’8 is greater than 1’4 in all lengths, and when the length is 500 achieved maximum accuracy. Therefore, the length of 500 and the encoding method of 1’8 were selected. Hence, each of the ncRNAs sequences was processed into 500 bases by IPM and converted into a matrix of 8’500.

2.2 Method

ncRFP is a Deep Learning model composed of Bi-LSTM, attention mechanism (AM), and fully connected network. Bi-LSTM and AM are mainly responsible to encode different ncRNAs into fixed format data, and the fully connected network is to decode the output of Bi-LSTM and AM. Fig. 3 displays the architecture of ncRFP. It can be found that after the original ncRNAs sequences are converted into matrixes, they are used as the input of Bi-LSTM and AM. Bi-LSTM encodes each of base at different locations into fixed size data based on the context. AM primarily facilitates the model to focus on the import locations of the different ncRNAs and to encode the output of Bi-LSTM into the same format size. The fully connected network, which majorly undertakes the task of decoding the output of Bi-LSTM and AM into the corresponding family of ncRNAs.

2.2.1 Bi-LSTM

ncRNAs are context sensitive text data [22], so it is necessary to record the context of each base when predicting their family. Because the bidirectional RNN can effectively record the past and future characteristics of each base, we chose it as the first part of the model to process each base into the same format data combined with its context. The memory and storage capability of ordinary RNN is limited, which will lose the ability to learn information and fall into gradient vanishing with the increase of sequences length. As a special RNN, LSTM solves the problem of gradient vanishing that arises in the ordinary RNN by introducing the memory cell and gate mechanism, which makes it perform better in representing the past information, future information and extracting long-distance dependencies of elements in sequence data. The LSTM memory cell could be implemented as follows:

\[
 i_t = \sigma(w_i x_t + w_h h_{t-1} + w_c C_{t-1} + b_i) \\
 f_t = \sigma(1 + w_f x_t + w_f h_{t-1} + w_c C_{t-1} + b_f) \\
 o_t = f_tC_{t-1} + i_t \tanh(w_o x_t + w_o h_{t-1} + b_o) \\
 h_t = o_t \tanh(C_t),
\]

where \(\sigma\) is the logistic sigmoid function, \(i, f, o, c\) are the input gate, forget gate, output gate, and cell vector, respectively, all of which are at the same dimension as the hidden vector \(h\). Meanwhile, \(w\) denotes the weight matrices and \(b\) indicates the bias vectors. Hence, the Bi-LSTM was selected in ncRFP, which
principle is that the same bases connect two LSTMs with opposite directions. The forward LSTM can record the past information and the backward LSTM can record the future data information. The hidden state $H_t$ of Bi-LSTM at time $t$ includes forward $h_t^f$ and backward $h_t^b$.

2.2.2 Attention Mechanism

In recent years, AM has gradually turned to be a research hotspot in the field of deep learning. AM originates from the simulation of attention characteristics of the human brain. The core idea of AM is to allocate more attention to imported information and less to other information, so as to ingeniously and reasonably change the attention to the information from the outside world, ignoring the irrelevant information and amplifying the desirable information. Thus, the receiving sensitivity and processing speed of information in the focused attention area are greatly improved. Ref. [23] first applied AM to investigate representations and proposed that global attention and local attention models should be applied to machine translation. Besides, Ref. [24] applied the AM to image description and obtained satisfying results. It was also mentioned in Ref. [25] that, in the field of automatic document summarization AM was capable of creating the sentence and document embeddings, which in turn enhanced the document summarization task. The same family of ncRNAs have consistent seed sequences, which can be used to distinguish them. Therefore, AM was selected as the second part of ncRFP, which is responsible for focusing more attention of ncRFP on the consistent seed sequences. So as to improve prediction accuracy.

2.2.3 Fully Connected Network

After Bi-LSTM and AM, it is no doubt that it is necessary to decode the output of Bi-LSTM and AM into the corresponding family of each ncRNAs. In this paper, a four-layer fully connected neural network was proposed to accomplish the task. Fig. 2 contains the architecture which consists of one input layer, two hidden layers, and one output layer, in which ReLU [26] was regarded as the activation function. The fully connected neural network could be implemented as follows:

$$y = ReLU(wx + b),$$

where $w$ is the weight matrices, $b$ is the bias vectors, and $x$ as well as $y$ is the input and output between any two layers.

2.3 Parameters of ncRFP

In this paper, the Adaptive moment estimation (Adam) optimizer [27] was used to train the method, which dynamically adjusts the learning rate of each parameter. In each iteration, the learning rate has a clear range, so as to effectively prevent the gradient from

---

**Fig. 2.** The prediction accuracy of ncRFP in different length of input data. The violet and green curve represents the accuracy in the coding of 1’8 and 1’4 respectively.

**Fig. 3.** The structure diagram of ncRFP. It consists of data processing, Bi-LSTM Layer, Attention mechanism, and fully connected neural network. Those four parts cooperate to predict ncRNAs family.
disappearing and exploding, and to make the parameters change smoothly. The parameters of ncRFP are given below:

1) Bi-LSTM hidden layer had 128 nodes and the output layer had 256 nodes. The number of nodes in fully connected layers were 128, 64 and 13.
2) All of the weighting matrices were initialized by the Gauss distribution (mean value is 0, the standard deviation is 0.05).
3) The dropout [28] layer was added to prevent over-fitting by reducing the number of points involved in the calculation. The dropout parameter was 0.3 in the Bi-LSTM and 0.4 in the fully connected neural network.

3 RESULT

In this section, we presented the training results and prediction results of ncRFP. In order to verify the stability of ncRFP, we used the 10-fold cross-validation to train ncRFP and obtain the training process of each epoch. After training ncRFP, we used it to get the prediction results of 10-fold cross-validation test sets and compared it with nRC and RNAcon in two aspects. The first aspect was compared to the average results of all test sets, another aspect was compared to the average results of each family. Fig. 4 shows the average accuracy and loss of each epoch in the 10-fold cross-validation experiments. It can be seen that in the last few epochs, the accuracy and loss of the test sets tend to be stable, which indicates that ncRFP can successfully complete the prediction of ncRNAs family. Tables S1-S3 (supplementary document, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2020.2982873) is the detail average prediction results of ncRFP, nRC, and RNAcon in the 10-fold cross-validation test sets. In Table S3, available online, because the prediction results of RNAcon do not contain scaRNA, we deleted scaRNA in the test sets when RNAcon was used to predict ncRNAs families. The green parts of each table are the correct prediction number of each ncRNAs family.

In order to estimate the performance of each method, the indexes of Accuracy, Sensitivity, Precision, F-score and MCC are used to compare ncRFP, nRC, and RNAcon. Accuracy is the proportion of predicted correct ncRNAs in the total ncRNAs, Sensitivity (Equation (1)), Precision (Equation (2)), F-score (Equation (3)) and MCC (Equation (4)), which are well known and have been applied.
In the process of multiple methods comparison, our method can achieve the optimal in the whole ncRNAs sequences and 11 single families: 5S_rRNA, 5.8S_rRNA, tRNA, CD-box, miRNA, Intron_gpII, HACA-box, IRES, riboswitch, leader, and scaRNA. In the whole ncRNAs sequences, Accuracy, Sensitivity, Precision, F-score and MCC increased by 13.88, 14.36, 14.92, 14.61 and 16.40 percent respectively. Table 4 shows the detail improvement of indexes in each family. There are two main reasons for the best performance of our method in the whole ncRNAs and most families. On the one hand, our method is a deep learning model, which can comprehensively extract the features of ncRNAs for prediction and the model including Bi-LSTM and AM can effectively focus on the important sentences of ncRNAs sequences. On the other hand, our method is an end-to-end method, which reduces the intermediate process. In Tables S1 and S2, available online, ribozyme and Intron_gpI have the most mispredictions between each other, which means that Intron_gpI and ribozyme are similar in sequence and secondary structure. There are two main reasons why the performance of ncRFP is not optimal in ribozyme and Intron_gpI. On the one hand, the high similarity between ribozyme and Intron_gpI makes the extracted features unable to predict their families. On the other hand, our method has some loss when intercepting ncRNAs sequences, which makes our method not optimal.

In future work, we will create a model, which combines RNN and CNN to simultaneously learn the primary and secondary structural features of ncRNAs to improve the prediction accuracy. The new model will be predicting more types of ncRNAs to provide more calculation support for biological experiments. Like nRC, we will also create a publicly available web service for the family prediction of unlabeled non-coding RNA sequences.

**ACKNOWLEDGMENTS**

The authors would like to acknowledge the support of the National Natural Science Foundation of China under Grant No. 61471181, the Natural Science Foundation of Jilin Province under Grant No. 2019C053-2, 2019C053-6.

**REFERENCES**

[1] B. C. Meyers, M. Matzke, and V. Sundaresan, “The RNA world is alive and well,” *Trends Plant Sci.*, vol. 13, no. 7, pp. 311–313, 2008.

[2] W.-T. Wang, C. Han, Y.-M. Sun, T.-Q. Chen, and Y.-Q. Chen, “Noncoding RNAs in cancer therapy resistance and targeted drug development,” *J. Hematol. Oncol.*, vol. 12, no. 1, 2019, Art. no. 55.

[3] M. Mason, A. Schuller, and E. Skordalakes, “Telomerase structure function,” *Curr. Opinion Structural Biol.*, vol. 21, no. 1, pp. 92–100, 2011.

[4] Z. Yang, Q. Zhu, K. Luo, and Q. Zhou, “The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription,” *Nature*, vol. 414, no. 6861, pp. 317–322, 2001.

[5] M. Lageo-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschi, “Identification of novel genes coding for small expressed RNAs,” *Science*, vol. 294, no. 5543, pp. 853–858, 2001.

[6] F. Tuorto and R. Parlato, “tRNA and tRNA bridges to neuronal homeostasis in health and disease,” *J. Mol. Biol.*, vol. 431, pp. 1763–1779, 2019.

[7] N. Sonenberg and A. G. Hinnebusch, “Regulation of translation initiation in eukaryotes: mechanisms and biological targets,” *Cell*, vol. 136, no. 4, pp. 731–745, 2009.

[8] J. Lu et al., “MicroRNA expression profiles classify human cancers,” *Nature*, vol. 435, no. 7043, pp. 834–838, 2005.

[9] R.-W. Yao, Y. Wang, and L.-C. Chen, “Cellular functions of noncoding RNAs,” *Nature Cell Biol.*, vol. 21, no. 5, pp. 542–551, 2019.
[10] H. Liu et al., “Integrative analysis of dysregulated lncRNA-associated ceRNA network reveals functional lncRNAs in gastric cancer,” Genes, vol. 9, no. 6, 2018, Art. no. 305.

[11] H. F. Noller, “RNA structure: Reading the ribosome,” Science, vol. 309, no. 5740, pp. 1508–1514, 2005.

[12] J. A. Reuter, D. V. Spacek, and M. P. Snyder, “High-throughput sequencing technologies,” Mol. Cell, vol. 58, no. 4, pp. 586–597, 2015.

[13] L. Childs, Z. Nikoloski, P. May, and D. Walther, “Identification and classification of ncRNA molecules using graph properties,” Nucleic Acids Res., vol. 37, no. 9, pp. e66–e66, 2009.

[14] B. Panwar, A. Arora, and G. P. Raghava, “Prediction and classification of ncRNAs using structural information,” BMC Genomics, vol. 15, no. 1, 2014, Art. no. 127.

[15] A. Fiannaca, M. La Rosa, L. La Paglia, R. Rizzo, and A. Urso, “nCRC: Non-coding RNA Classifier based on structural features,” BioData Mining, vol. 10, no. 1, 2017, Art. no. 27.

[16] C. Borgelt, T. Meinl, and M. Berthold, “MoSS: A program for molecular substructure mining,” in Proc. 1st Int. Workshop Open Source Data Mining/Bart Goethals, 2005, pp. 6–15.

[17] K. Sato, Y. Kato, M. Hamada, T. Akutsu, and K. J. B. Asai, “IPknot: Fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming,” Bioinformatics, vol. 27, no. 13, pp. i85–i93, 2011.

[18] M. Zuker and P. Stiegler, “Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information,” Nucleic Acids Res., vol. 9, no. 1, pp. 133–148, 1981.

[19] S. Wang, X. Wang, S. Wang, and D. Wang, “Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting,” Int. J. Elect. Power Energy Syst., vol. 109, pp. 470–479, 2019.

[20] I. A. Basheer and M. Hajiemeer, “Artificial neural networks: Fundamentals, computing, design, and application,” J. Microbiol. Methods, vol. 43, no. 1, pp. 3–31, 2000.

[21] E. P. Nawrocki et al., “Rfam 12.0: Updates to the RNA families database,” Nucleic Acids Res., vol. 43, no. D1, pp. D130–D137, 2014.

[22] L. Wang 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., vol. 10, 2019, Art. no. 145.

[23] M.-T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” 2015, arXiv:1508.04025.

[24] L. Li, S. Tang, Y. Zhang, L. Deng, and Q. Tian, “GLA: Global-local attention for image description,” IEEE Transactions on Multimedia, vol. 20, no. 3, pp. 726–737, Mar. 2018.

[25] K. Al-Sabahi, Z. Zuping, and M. Nadher, “A hierarchical structured self-attentive model for extractive document summarization (HSSAS),” IEEE Access, vol. 6, pp. 24205–24212, 2018.

[26] D. Yarotsky, “Error bounds for approximations with deep ReLU networks,” Neural Netw., vol. 94, pp. 103–114, 2017.

[27] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.