SCIENCE CHINA
Information Sciences

• LETTER •

January 2019, Vol. 62 019103:1–019103:3
https://doi.org/10.1007/s11432-018-9558-2

Identifying RNA-binding proteins using multi-label deep learning

Xiaoyong PAN[1*], Yong-Xian FAN[2], Jue JIA[3] & Hong-Bin SHEN[4,5*]

[1]*Department of Medical Informatics, Erasmus MC, Rotterdam* 3014ZK, *The Netherlands;*
[2]*Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin* 541004, *China;*
[3]*Affiliated Hospital of Jiangsu University, Zhenjiang* 212001, *China;*
[4]*Department of Automation, Shanghai Jiao Tong University, Shanghai* 200240, *China;*
[5]*Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai* 200240, *China*

Received 10 April 2018/Revised 5 June 2018/Accepted 16 August 2018/Published online 17 December 2018

Citation Pan X Y, Fan Y X, Jia J, et al. Identifying RNA-binding proteins using multi-label deep learning. Sci China Inf Sci, 2019, 62(1): 019103, https://doi.org/10.1007/s11432-018-9558-2

Dear editor,

RNA-binding proteins (RBPs) are involved in both transcriptional and post-transcriptional gene regulation, such as RNA splicing and localization. In addition, their dysregulations are closely associated with many diseases [1]. For example, mutations in the RBPs FUS and TDP-43 that can cause amyotrophic lateral sclerosis [2]. To date, huge volume of experimentally verified RBP binding sites have been collected [3] by high-throughput sequencing. However, they are still time-consuming and high-cost. Many machine learning-based methods have been developed to learn patterns for RNA-protein interactions. For example, RNAcommender trains a recommended system for suggesting RNA targets for RBPs by integrating RNA structures and RBPs domain information [4]. Recently, neural networks and deep learning have been popularly applied in computational biology [5], especially RNA-protein binding sites [6, 7]. For example, DeepBind trains a CNN to identify DNA/RNA binding preferences of proteins [6]. iDeep trains a hybrid deep network with deep belief networks (DBNs) and CNNs using multiple data sources [7]. iDeepE combines a global CNN and a local CNN to predict RNA-protein binding sites and motifs from sequences alone [8]. Considering the structure preference of RBPs, iDeepS takes structures into consideration for RBP binding specificity [9]. It trains two individual CNNs and a long short term memory network (LSTM) for sequences and structures to capture binding sequence and structure motifs of RBPs. All the above methods train RBP-specific models, where each model can only predict RNA targets for one RBP. In addition, the relationship among different RBPs are totally ignored. For example, different RBPs share similar binding domains, which can also be integrated into predict RNA-protein interaction. Another disadvantage of RBP-specific model is that it requires constructing negative samples for each RBP, how to construct negative sites for each RBP has big impact on the trained RBP-specific models. Different strategies to construct negative sets yield different prediction performance. For example, randomly shuffling RNA positive sequences as negative sequences can make trained models yield better performance than randomly shuffling the coordinates of the bound sites within the same gene [8]. As mentioned above, CNNs can extract high-level motif features and LSTMs can learn long-range dependency. Thus, a hybrid CNN/LSTM model that does not require constructing negative sets for individual RBPs is needed. In this study, our goal is to predict which one or multiple RBPs can bind to a given input RNA sequence. To this end, we formulate this prediction problem as a multi-label classifica-

* Corresponding author (email: xypan172436@gmail.com, hbshen@sjtu.edu.cn)

tion problem using deep learning, which does not need construct negative sites for each RBP. Thus, we present a joint computational model based on multi-label deep learning to predict how a RNA sequence is attached by a set of RBPs. We use a CNN to learn shared abstract features across RBPs, which is different from learning shared features across different sources of features in iDeep. Then the learned abstract features are further fed to LSTM under multi-label learning framework. It connects deeply to how to model the dependency and combinations of RBPs. To the best of our knowledge, our work is the first computational method to explore multi-label learning for predicting RNA-binding proteins using deep learning.

Dataset. We constructed our benchmark dataset from RNAcommender [4], which includes a total of 502178 binding sites derived from CLIP-seq for 67 RBPs, and each RBP has different number of binding sites. Here we treat RBPs as labels. If one RNA binds to multiple RBPs, then we assign multiple labels to this RNA sequence. Besides the 67 RBPs, we also add another label for those RNA sequences that do not bind to any of 67 RBPs. Here we randomly select the number of UTRs (untranslated regions) the same as the RBP with the largest number of binding sequences. We also investigate that how many RNAs have multiple binding proteins, and we find that 74.7% of RNAs have at least two binding proteins. The distribution of number of binding proteins is shown in Figure A1. The data is randomly stratified split, where 80% of the data is used for training and 20% is left for testing.

iDeepM method. We present a deep multi-label learning method called iDeepM for predicting RNA-binding proteins (Figure 1 and Algorithm A1)[1)]. It first encodes sequences to one-hot encoding matrix, which is fed into a CNN, followed by a LSTM layer, where CNNs are used to extract high-level motif features [7] and LSTMs are used to learn dependency among labels. Please refer to Appendix A for more details. The iDeepM is implemented in python using Keras v1.2.0 library. For the CNN, the parameter nb_filter (number of motifs) is 102, which is the total number of verified motifs in CISBP-RNA database. The kernel size of filters is 10, which is used in iDeepM. For the LSTM, we set the hidden size = 68, equaling to the number of classes. The Dropout probability between each layer is 0.5. We learn the model parameters by minimizing the binary cross-entropy loss using Adam optimizer, and batch size is 64. Our experiments are ran on a Ubuntu server with CPUs 2.00 GHz.

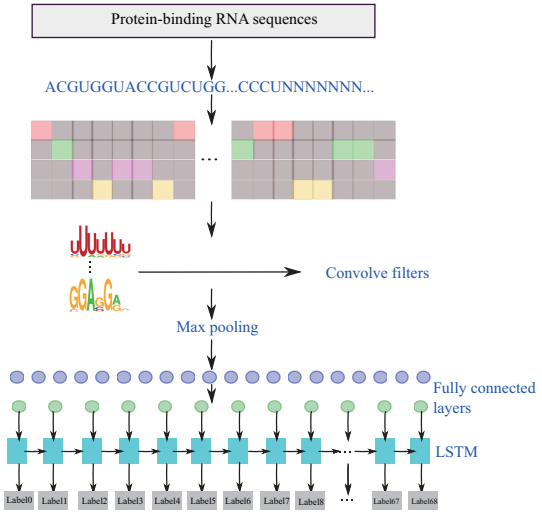

Figure 1 (Color online) The flowchart of iDeepM. iDeepM first converts RNA sequence into one-hot encoded matrix, which is further fed into a CNN, followed by LSTM to learn label dependency under multi-label learning framework.

Results. We evaluate iDeepM on our constructed benchmarked dataset, where it classifies RNA sequences to be bound to 68 RBPs or not, including one artificial one that do not bind to any of 67 RBPs. We first checked the impact of epochs with values (10, 15, 20, 25) on prediction performance. As shown in Table B1, the larger number of epochs yield better performance, but it is more time-consuming. The number of epochs (nb_epoch) has no big impact on the AUC measurement, but it has impacts on the F1 measurement with a larger change. In addition, when nb_epoch is nearby 20, the F1 measurement also approximates to be similar. Considering the trade-off between training time and performance, in this study, we use nb_epoch = 20. In addition, we can also see that the testing time of iDeepM is fast. The iDeepM yields high AUCs nearby 0.90 for Macro-AUC, Micro-AUC and Weighted-AUC, but F1 measurements with a large difference. The results indicate that iDeepM has strong power for classifying RNA-binding proteins. To demonstrate the advantage of deep learning over other conventional machine learning models, we develop another kmer-SVM method to fit a binary SVM classifier for each label against the rest of the labels. The inputs of kmer-SVM is the 4-mer (AAAA, AAAA, ..., UUUU) frequency, and the kernel of the SVM is linear kernel and other parameters are default values in scikit-learn. As shown in Table B1, kmer-SVM yields very bad performance, it

1) https://github.com/xypan1232/iDeepM.

is close to random guessing, and its prediction results are dominated by two majority classes "negatives" and "Ago1". The results indicate that SVM is especially not good at handling the imbalanced data of this study and it needs be combined with other up-sampling or down-sampling techniques. In addition, the testing time of kmer-SVM is much longer than iDeepM.

We also further investigate the performance of different RBPs. The developed iDeepM can yield very high AUC for all RBPs. However, for F1-score, the performance across RBPs varies a lot. Thus, we list those RBPs with top 10 F1-score (Table B2). We can see that for some RBPs from the same protein family, iDeepM yields similar performance, like TIAL1 and TIA1. Especially, iDeepM can classify RNAs that do not bind to any of 67 RBPs with high AUC 0.87. Furthermore, we investigate the relationship between binding domains and the performance. Thus we scan each RBP sequence against the HMM models of the Pfam-A v. 28.0 database, and select all domains with e-value $\leqslant$ 1.0 for this RBP. Of the top 10 RBPs, TIAL1, TIA1 and RBP47 have domain PF00076 (RNA recognition motif), LIN28A and LIN28B have shared domain PF00098 (Zinc finger). We also collect those domains that are shared by at least three RBPs of the 67 RBPs, then we plot the boxplot of AUC and AUPRC with these domains (Figure B1). For domain PF12235 (Fragile X-related 1 protein core C terminal) shared by five RBPs, iDeepM yield almost the same high AUCs and AUPRCs on these 5 RBPs. However, for some domains, iDeepM can yield different performance for different RBPs sharing the same domain, e.g., PF00035 (double-stranded RNA binding motif). The potential reason is that our training and testing RNAs are all single-stranded, thus iDeepM cannot learn binding specificities for domain PF00035 from them.

Conclusion and future work. We propose a novel approach for RNA-binding proteins prediction that formulates the task as a multi-label classification problem, which is different from previous RBP-specific models. To the best of our knowledge, it is the first study to predict RNA-binding proteins under multi-label learning framework. In addition, we also combine the power of deep learning to better learn high-level features with strong discriminate power for RNA-binding proteins prediction. Please refer to Appendix C for future work and discussion about iDeepM. It is expected that this study provides a new avenue for predicting RNA-binding proteins.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61725302, 61671288, 61603161, 61462018, 61762026, 81500351), Science and Technology Commission of Shanghai Municipality (Grant Nos. 16JC1404300, 17JC1403500), Jiangsu Province's Young Medical Talents Project (Grant No. QNRC2016842), and "5123 Talents Project" of Affiliated Hospital of Jiangsu University (Grant No. 51232017305).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Pan X Y, Shen H B. OUGENE: a disease associated over-expressed and under-expressed gene database. Sci Bull, 2016, 61: 752–754
2 Colombrita C, Onesto E, Megiorni F, et al. TDP-43 and FUS RNA-binding proteins bind distinct sets of cytoplasmic messenger RNAs and differently regulate their post-transcriptional fate in motoneuron-like cells. J Biol Chem, 2012, 287: 15635–15647
3 Li J H, Liu S, Zhou H, et al. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucl Acids Res, 2014, 42: 92–97
4 Corrado G, Tebaldi T, Costa F, et al. RNAcommender: genome-wide recommendation of RNA-protein interactions. Bioinformatics, 2016, 42: 3627–3634
5 Li Y Y, Zou X F. Identifying disease modules and components of viral infections based on multi-layer networks. Sci China Inf Sci, 2016, 59: 070102
6 Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol, 2015, 33: 831–838
7 Pan X Y, Shen H B. RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. BMC BioInf, 2017, 18: 136
8 Pan X Y, Shen H B. Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks. Bioinformatics, 2018, 34: 3427–3436
9 Pan X Y, Rijnbeek P, Yan J C, et al. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC Genomics, 2018, 19: 511