Front.Comput.Sci.
DOI

Appendix of DeepCRBP

Appendix A Details about the datasets

To evaluate the performance of our proposed DeepCRBP and the comparison methods, we use the same benchmark datasets from [1]. In the datasets, The circRNA sequences are extracted from the circRNA Interactome database (https://circinteractome.nia.nih.gov/) which contains 120,000 human circRNA sequences. To filter the redundant circRNA sequence, CD-HIT [2] package with the recognition threshold of 0.8 is used for preventing classification bias from high sequence similarity. Subsequently, reading peak from each binding site corresponding to the CLIP-seq, 50 nucleotides (nts) are expanded to the upstream and downstream around the center of RBP binding site, then sequence fragments with a length of 101 nts are extracted [3]. As for the selection of negative samples, they are the segments that do not interact with RBPs and are selected with the same length as that of positive samples. It is noteworthy that CLIP-seq is a terrific scientific method to experimentally resolve transcriptome-wide binding sites of RNA-RBPs [4] and it is often applied in the research process [5–11]. As a result, there are 335,976 positive and 335,976 negative samples assiciated with 37 RBPs in total. The largest dataset includes 20,000 samples and the smallest dataset includes 446 samples, with an average of 9,080 circRNAs corresponding to each protein.

Appendix B Experimental Setting

We utilize five-fold cross-validation and the same random number seed to evaluate the performance of the predictive methods. When training each dataset, we select 20% of the samples as an independent test set. In the remaining samples, we divide 80% of the samples as the train set, and the rest of 20% as the validation set for optimizing model parameter.

Appendix C Comparison with Baseline Methods

To demonstrate the identification performance of our proposed DeepCRBP, three existing efficient methods are assessed with the evaluation metrics (AUC and ACC).

CRIP [1] encodes one-dimension circRNA sequences by a pseudo-amino acid stacking coding scheme and then employs CNN and BiLSTM to learn circRNA representations.

Debcan [12] is featured by hybrid double embeddings for representing one-dimension circRNA sequences and a cross-branch attention neural network for classification.

DHN [13] proposes CircRNA2Vec and k-tuple nucleotide frequency pattern to extract different feature subsets and uses a deep multi-scale residual network to identify binding sites of circRNA-RBP by combining the feature subsets.

DMSK [14] integrates sequence feature and pseudo-secondary structure feature of circRNAs and utilizes a multi-view classification method consisted of multi-view deep learning, subspace learning and multi-view classifier for predicting the binding sites of circRNA-RBP.

Appendix D Case Study

To validate the generalization ability of DeepCRBP, we implement experiments to predict the binding sites of circRNA-RBPs. Specifically, we use the DeepCRBP to train and collect 82 test samples which are circRNA sequences of length 101 nts from eight datasets to evaluate. Then, the test samples are fed into the trained model and the corresponding prediction scores are obtained. After that, with the threshold of 0.5 which is the dividing line between positive and negative samples, we derive the results that DeepCRBP predicted 11 out of 13, 7 out of 8, 11 out of 12, 7 out of 8, 11 out of 12 and 6 out of 8 existing binding sites of hsa_circ_0001339 with IGF2BP2, hsa_circ_0001711 with EWSR1, hsa_circ_0001146 with ZC3H7B, hsa_circ_0002203 with U2AF65, hsa_circ_0000039 with EIF4A3 and hsa_circ_0001711 with MOV10. In particular, the predicted results of hsa_circ_0000005 with AGO2 and hsa_circ_0001585 with CAPRIN1 are all correct with 10 samples. The results are shown in Figure 1, where stars are the truly predicted binding sites of circRNA-RBPs and diamonds are the results of wrong predictions. It should be noted that the horizontal coordinate indicates each circRNA sequence fragment involved in the prediction. In order to measure the quality of the model and its classification accuracy for positive and negative examples, we calculated the results of AUC and ACC, which are shown in Table 1. The above experiments further confirm the outstanding

E-mail:

Table 1 The results of AUC and ACC on eight datasets

Case Study	ACC	AUC
hsa_circ_0001339 with IGF2BP2	0.768	0.860
hsa_circ_0001711 with EWSR1	0.919	0.977
hsa_circ_0000005 with AGO2	0.856	0.976
hsa_circ_0001146 with ZC3H7B	0.720	0.885
hsa_circ_0001585 with CAPRIN1	0.853	0.950
hsa_circ_0002203 with U2AF65	0.882	0.939
hsa_circ_0000039 with EIF4A3	0.786	0.917
hsa_circ_0001711 with MOV10	0.925	0.921

predictive capability and generalization performance of DeepCRBP.

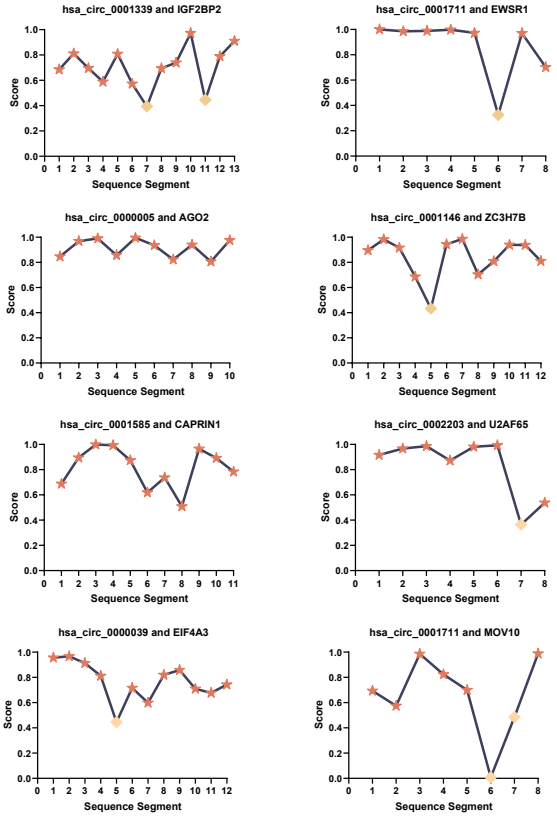

Fig. 1 The results of DeepCRBP for predicting binding sites of circRNA-RBP on eight datasets.

Appendix E Motif analysis

In order to deepen the understanding of the sequence patterns characterizing the binding sites of circRNA-RBPs, we conduct a motif experiment on the positive sequence fragments utilizing the MEME suite [15]. Specifically, we extract motifs for each RBP dataset, and subsequently visualize the most significant motif in Figure 2. Notably, the motifs are assigned a width of 10bp. Our investigation reveals that certain RBPs exhibit a shared binding pattern on their target sites, specifically the sequence "GAAGAAG". This pattern is observe across multiple RBPs, including AGO1, C22ORF28, IGF2BP3, SFRS1, TNRC6, WTAP and so on. Specially, this pattern has been previously associated with RNA modification [16]. It is evident that certain regions of the RNAs hold crucial significance, and they display resemblance to established motifs of RBPs present in the CISBP-RNA database [17]. Through motif searching via the MEME suite, we are able to identify some conserved patterns that align with the motifs found in the CISBP-RNA database, namely "UUUUU" associated with U2AF65 binding, "GAAUG" linked to TDP43 binding, and "GAAGAA" correlated with SFRS1 binding [18].

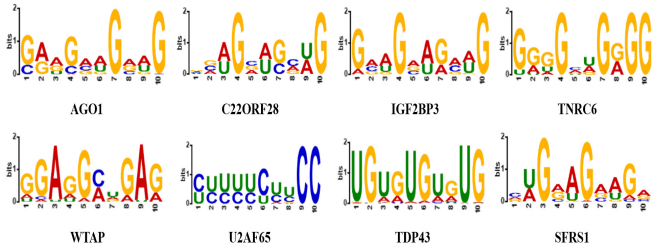

Fig. 2 The motifs on AGO1, C22ORF28, IGF2BP3, SFRS1, TNRC6, WTAP, U2AF65 and TDP43 circRNAs.

References

1. Zhang K, Pan X, Yang Y, Shen H B. CRIP: predicting circRNA–RBP-binding sites using a codon-based encoding and hybrid deep neural networks. Rna, 2019. 25(12):1604–1615.
2. Li W, Cd-hit G A. A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics [Internet]. 2006; 22 (13): 1658–9.
3. Dudekula D B, Panda A C, Grammatikakis I, De S, Abdelmohsen K, Gorospe M. CircInteractome: a web tool for exploring circular RNAs and their interacting proteins and microRNAs. RNA biology, 2016. 13(1):34–42.
4. Li Y E, Xiao M, Shi B, Yang Y C T, Wang D, Wang F, et al. Identification of high-confidence RNA regulatory elements by combinatorial classification of RNA–protein binding sites. Genome biology, 2017. 18(1):1–16.
5. Zhang M, Wang T, Xiao G, Xie Y. Large-scale profiling of RBP-circRNA interactions from public CLIP-seq datasets. Genes, 2020. 11(1):54.
6. Li J H, Liu S, Zhou H, Qu L H, Yang J H. starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. Nucleic acids research, 2014. 42(D1):D92–D97.
7. Li B, Zhang X Q, Liu S R, Liu S, Sun W J, Lin Q, et al. Discovering the Interactions between Circular RNAs and RNA-binding Proteins from CLIP-seq Data using circScan. bioRxiv, 2017. 115980.

8. Zhang X Q, Yang J H. Discovering circRNA-microRNA interactions from CLIP-Seq data. Circular RNAs: Methods and Protocols, 2018. 193–207.
9. Yang Y, Hou Z, Wang Y, Ma H, Sun P, Ma Z, et al. HCRNet: high-throughput circRNA-binding event identification from CLIP-seq data using deep temporal convolutional network. Briefings in Bioinformatics, 2022. 23(2).
10. Ghosal S, Das S, Sen R, Basak P, Chakrabarti J. Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. Frontiers in genetics, 2013. 4:283.
11. Teng X, Chen X, Xue H, Tang Y, Zhang P, Kang Q, et al. NPInter v4. 0: an integrated database of ncRNA interactions. Nucleic acids research, 2020. 48(D1):D160–D165.
12. Yuan L, Yang Y. DeCban: prediction of circRNA-RBP interaction sites by using double embeddings and cross-branch attention networks. Frontiers in Genetics, 2021. 1826.
13. Yang Y, Hou Z, Ma Z, Li X, Wong K C. iCircRBP-DHN: identification of circRNA-RBP interaction sites using deep hierarchical network. Briefings in Bioinformatics, 2021. 22(4):bbaa274.
14. Li H, Deng Z, Yang H, Pan X, Wei Z, Shen H B, et al. circRNA-binding protein site prediction based on multi-view deep learning, subspace learning and multi-view classifier. Briefings in Bioinformatics, 2022. 23(1):bbab394.
15. Bailey T L, Boden M, Buske F A, Frith M, Grant C E, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic acids research, 2009. 37(suppl_2):W202–W208.
16. Dominissini D, Nachtergaele S, Moshitch-Moshkovitz S, Peer E, Kol N, Ben-Haim M S, et al. The dynamic N 1-methyladenosine methylome in eukaryotic messenger RNA. Nature, 2016. 530(7591):441–446.
17. Ray D, Kazan H, Cook K B, Weirauch M T, Najafabadi H S, Li X, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature, 2013. 499(7457):172–177.
18. Cook K B, Kazan H, Zuberi K, Morris Q, Hughes T R. RBPDB: a database of RNA-binding specificities. Nucleic acids research, 2010. 39(suppl_1):D301–D308.