

Received October 10, 2019, accepted October 28, 2019, date of publication November 11, 2019, date of current version December 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2952621

# iRBP-Motif-PSSM: Identification of RNA-Binding Proteins Based on Collaborative Learning

XIN GAO<sup>1,\*</sup>, DONGHUA WANG<sup>2,\*</sup>, JUN ZHANG<sup>1</sup>, QING LIAO<sup>1</sup>, AND BIN LIU<sup>3,4</sup> 

<sup>1</sup>School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518057, China

<sup>2</sup>Department of General Surgery, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin 150088, China

<sup>3</sup>School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100811, China

<sup>4</sup>Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing 100811, China

Corresponding authors: Qing Liao (liaoqing@hit.edu.cn) and Bin Liu (bliu@biliblab.net)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672184 and Grant 61822306, in part by the Fok Ying-Tung Education Foundation for Young Teachers in the Higher Education Institutions of China under Grant 161063, and in part by the Scientific Research Foundation in Shenzhen under Grant JCYJ20180306172207178, Grant JCYJ20180306172156841, and Grant JCYJ20180507183608379.

\*Xin Gao and Donghua Wang are co-first authors.

**ABSTRACT** RNAs and RNA-binding proteins (RBPs) in cells can bind with each other to form a nuclear ribonucleoprotein (RNP) complex, playing important roles in life processes, and gene regulation. How to accurately predict the RNA-binding proteins is a big challenge and hot research task. Here, we proposed a new computational predictor called iRBP-Motif-PSSM for identifying RNA-binding proteins by combining the motif information and the evolutionary information extracted from the Position Specific Scoring Matrixes. Collaborative Learning was employed to address the instability problem of the predictor. The experimental results showed that iRBP-Motif-PSSM showed better performance than other existing methods for identifying RNA-binding proteins, indicating that iRBP-Motif-PSSM is a useful tool for biological analysis.

**INDEX TERMS** RNA-binding proteins, Motif-PSSM, collaborative learning.


## I. INTRODUCTION

RNA is an important molecular playing many important functions [1]. RNA molecules can interact with RNA-binding proteins (RBPs). In general, the majority of RNAs and RNA-binding proteins (RBPs) in cells can bind with each other to form a nuclear ribonucleoprotein (RNP) complex, playing an important role in life processes, and gene regulation. Because a RBP may have many corresponding target RNAs, investigation of the interaction between RNA and protein is the key to explore RNA functions. If the regulation is abnormal, it will lead to various diseases, such as cancer [2], [3], myeloid leukaemia [4], [5], etc. How to accurately predict the functions of these RNA-binding proteins is a big challenge and hot research task in the field of genomic function annotation. Because of the limitations of the biological experimental methods (time consuming and expensive), it is desired to develop predictors to detect these RNA-binding proteins only based on the sequence information. In the past few years, several experimental predictors

[6]–[11] were proposed to identify RNA-binding proteins. They are mainly based on nucleic acid and amino acid physicochemical properties, for example SPOT-Seq-RNA [6] is a template-based technique using Z-score and energy to distinguish RBPs. RNAPred [11] uses the evolutionary information extracted from PSSMs, and outperforms other sequence-based methods. The RBPPred predictor [10] incorporates the representative physicochemical properties.

All these methods have obviously facilitated the development of this important field. However, with the fast growth of the number of protein sequences, there are still some problems should be addressed in this field: (1) In general, the available features failed to accurately represent the protein sequences for RBP prediction, preventing the performance improvement of the existing predictors. (2) How to efficiently combine the different features and classifiers to construct a more accurate predictor is still a challenging problem.

As short conserved patterns in protein sequences, motifs are critical for the structural and functional activities of proteins [12], [13]. Furthermore, the profiles contain the evolutionary information [14], such as PSSMs. Both motifs and PSSMs are important for RBP identification. Can we

The associate editor coordinating the review of this manuscript and approving it for publication was Dariusz Mrozek .

combine these two important features to introduce a more discriminative feature? To answer this question, in this study, we introduced a new feature called Motif-PSSM sharing the advantages of both motifs and PSSMs. The Motif-PSSM features and some other sequence-based features were combined via the framework of Collaborative Learning based on Support Vector Machines (SVMs). Finally a predictor called iRBP-Motif-PSSM was proposed to predict the RBPs. Experimental results showed that iRBP-Motif-PSSM outperformed other competing methods. Furthermore, iRBP-Motif-PSSM is also useful for identifying new RBPs in human proteome.

## II. MATERIALS AND METHOD

### A. BENCHMARK DATASET

In this study, we used the benchmark dataset  $\mathbb{S}$  constructed by Zhang and Liu [10] containing  $\mathbb{S}^+$  and  $\mathbb{S}^-$ .

$$\mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^- \quad (1)$$

where ‘ $\cup$ ’ represents the ‘union’;  $\mathbb{S}^+$  represents RNA-binding proteins with 2780 samples;  $\mathbb{S}^-$  indicates non-RNA-binding proteins with 7093 samples.

### B. INDEPENDENT DATASET

An independent dataset including 68 RBPs and 100 non-RBPs reported in [11] was used to further test the generalization capability of our predictor.

### C. FEATURE EXTRACTION STRATEGY

We introduced a novel feature extraction method named Motif-PSSM. However, a single feature usually fails to accurately represent the proteins in all situations [12], [1], and it will be also affected by the data fluctuation. To cope with such limitations, some state-of-the-art sequence-based features were combined with the proposed Motif-PSSM feature to accurately capture the characteristics of proteins, including ACC-PSSM [15], Kmer [16], [17], Top-n-gram [15], and CKSAAP [18], [19].

#### 1) MOTIF-PSSM

Given a protein  $\mathbf{P}$  whose length is  $L$ .  $\mathbf{P}$  can be represented as:

$$\mathbf{P} = A_1 A_2 A_3 A_4 A_5 A_6 A_7 \cdots A_L \quad (2)$$

where  $A_1$  is the first amino acid,  $A_2$  is the second amino acid, etc.

For the given protein sequence as shown in Eq. 2, its corresponding PSSM matrix was calculated by PSI-BLAST [20] based on the NRDB 90 database [21] with default parameters.

Previous studies showed that motifs have impact on the structural and functional activities of proteins [22], [23]. The motifs are short protein subsequences, which are the structural components of specific functions with specific spatial conformations [24]. In the past few decades, various computational methods, such as MEME SUITE [25] have been proposed for identifying, characterizing, and searching the motifs. According to the characteristics, the motifs

are mainly divided into three categories including sequence motifs, structural motifs, and short linear motifs. Sequence motifs are sequence patterns of residues in protein sequences. Structural motifs represent structural patterns in protein structures. Short linear motif is subsequence mediating the process of protein–protein interaction. In this study, we focused on the short linear motifs and structural motifs. From the MegaMotifBase [26], 301 structural motifs related to RNA-binding proteins were extracted and converted to Meme Motif Format [25] by using Multiple Sequence Alignment. The 164 short linear motifs with fixed length were extracted from ELM database [27], [28]. The detailed information for these motifs is given in **Supplementary Information S1**.

To construct a powerful predictor for RBP prediction, one of the keys is to represent protein sequences with an effective mathematical expression reflecting their intrinsic correlation with the characteristics to be identified [12], [29]. In this regard, we proposed the Motif-PSSM feature (see Fig. 1). The process of generating the Motif-PSSM feature will be introduced in the following section.

(1) All the motifs were converted into frequency matrices. For the 164 short linear motifs, their corresponding frequency matrices were downloaded from ELM database [27], [28]. For the 301 structural motifs, their corresponding Multiple Sequence Alignments (MSAs) were converted into frequency matrices.

(2) For each motif, its corresponding frequency matrix ( $\mathbf{M}$ ) was searched against the PSSM segments of the target protein  $\mathbf{P}(\mathbf{S})$  with the same size as the motif by using a sliding window approach with step size as 1. The  $\mathbf{MS}$  was calculated by:

$$\mathbf{M}_i \mathbf{S}_j = \mathbf{M}_i \times \mathbf{S}_j \quad (3)$$

The total value ( $val$ ) of the elements in  $\mathbf{M}_i \mathbf{S}_j$  was calculated by:

$$val_j = \sum \mathbf{M}_i \mathbf{S}_j \quad (4)$$

where  $\mathbf{M}_i$  is the  $i$ -th motif frequency matrix,  $j$  is the  $j$ -th search for  $\mathbf{S}$ .

Therefore, for each motif, the corresponding feature vector can be generated by combining all the  $val$  for each search.

$$\mathbf{MOTIF}_i = [val_1, val_2, val_3, \dots, val_{L-l+1}] \quad (5)$$

where  $L$  is the length of  $\mathbf{S}$ , and  $l$  is the length of  $\mathbf{M}_i$ .

The average and max values of  $\mathbf{MOTIF}_i$  can be calculated by Eq6 and Eq7, respectively.

$$ave_i = \sum \mathbf{MOTIF}_i / N_i \quad (6)$$

$$max_i = \text{Max}(\mathbf{MOTIF}_i) \quad (7)$$

Finally, the resulting feature vector based on 465 motifs was generated by:

$$\mathbf{P} = [max_1, ave_1, max_2, ave_2 \dots, max_{465}, ave_{465}] \quad (8)$$

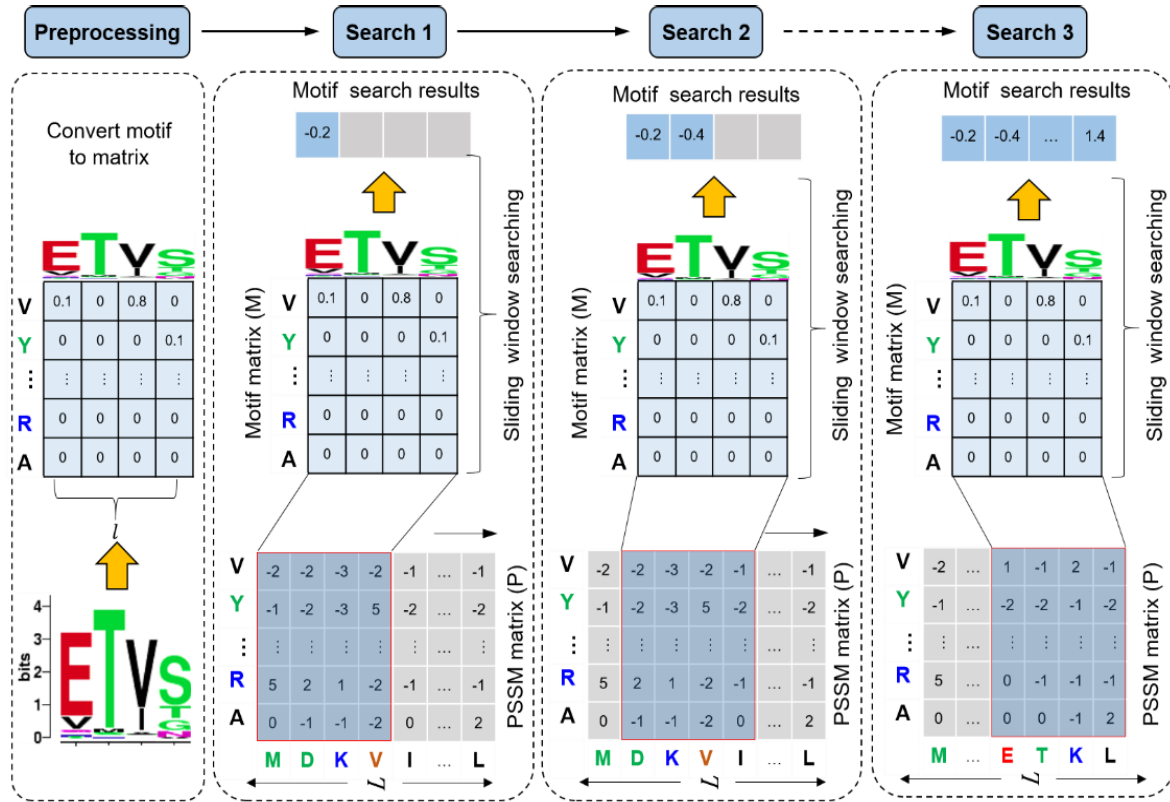


FIGURE 1. The process of generating Motif-PSSM feature based on PSSM matrix.

2) ACC-PSSM

ACC-PSSM [30] measures the correlations between two amino acids based on PSSMs [31]. ACC contains two parts, including AC and CC. A protein sequence can be represented by ACC by combining AC and CC. The AC-PSSM is calculated by [30]

$$AC(i, d) = \sum_{j=1}^{L-d} \frac{(s_{i,j} - \bar{s}_i)(s_{i,j+d} - \bar{s}_i)}{(L-d)} \quad (9)$$

where  $i$  is one amino acid residue,  $d$  represents the distance between two amino acids along the protein,  $L$  indicates the protein length,  $s_{i,j}$  represents the score of residue  $i$  to appear at position  $j$  in the corresponding PSSM matrix,  $\bar{s}_i$  represents the average score of residue  $i$ .

The AC-PSSM is calculated by [30]

$$CC(i_1, i_2, d) = \sum_{j=1}^{L-d} \frac{(s_{i_1,j} - \bar{s}_{i_1})(s_{i_2,j+d} - \bar{s}_{i_2})}{(L-d)} \quad (10)$$

where  $i_1, i_2$  represent two amino acids,  $d$  represents the distance between two amino acids in the protein,  $s_{i_1,j}$  represents the score of residue  $i_1$  to appear at position  $j$  in PSSM matrix,  $s_{i_2,j+d}$  represents the score of residue  $i_2$  to appear at position  $j+d$  in PSSM matrix and  $\bar{s}_{i_1}$  and  $\bar{s}_{i_2}$  represent average scores of residues  $i_1$  and  $i_2$  in the protein, respectively.

Since ACC-PSSM combines AC-PSSM and CC-PSSM, the number of AC-PSSM variables is  $20 \times D$ , the number of CC-PSSM variables is  $380 \times D$ . Therefore, the the number

of ACC-PSSM variables is  $400 \times D$ . In the study, we set the  $D$  as 7.

3) KMER

Kmer [16], [32], [33] is a widely used feature. A kmer is a subsequence in a protein with  $k$  amino acids. The parameter  $k$  in this study was set as 2. The frequencies of kmers can be calculated by:

$$f(r) = \frac{N(r)}{L}, r \in \{AC, AD, AE, \dots, YV, YW, YY\} \quad (11)$$

where  $N(r)$  represents the total number of kmer type  $r$ ,  $\in$  means “member of” and  $L$  is the protein length.

4) TOP-N-GRAM

Top- $n$ -gram [34] incorporates the evolutionary information, and has been applied to solve many tasks in bioinformatics. In this study, the Top- $n$ -gram feature was calculated by BioSeq-Analysis2.0 with default parameters except that the parameter in Top- $n$ -gram was set as 2.

5) CKSAAP

The CKSAAP [18] calculates the frequencies of amino acid pairs with distance  $k$  in the protein. In this study, the  $k$  was set as 5. The detailed information has been introduced in [35]–[38].

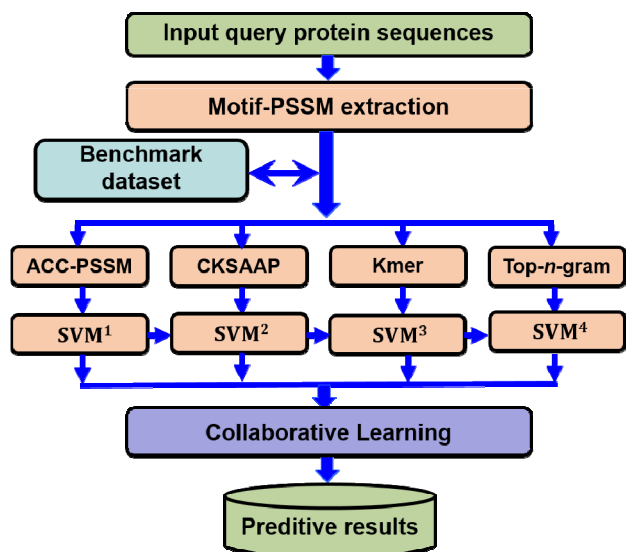


FIGURE 2. A flowchart to show how the SVM-based collaborative learning works.

#### D. CLASSIFIER CONSTRUCTION

##### 1) SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) was used as the classifier to construct the predictor in this study. SVM was trained by using the encoded features to represent the model for prediction, which has been widely used in bioinformatics [24], [39]–[44]. The publicly available Scikit-learn package was employed as the implementation of SVM algorithm [45] with RFB kernel with two parameters: one is penalty parameter  $C$  for the regularization and another is  $\gamma$  for the kernel width. The grid search was used to optimize these parameters.

##### 2) SVM-BASED COLLABORATIVE LEARNING

An ensemble predictor combining various individual predictors will achieve better performance than a single predictor [43], [46]–[52]. There are two major problems in constructing ensemble predictors. i) How to select the individual classifiers with low correlation; ii) how to construct an ensemble classifier by assembling the selected classifiers [43]. In the study, we used different features and parameters to generate several predictors so as to ensure the low correlation among the predictors, and we employed the SVM-based Collaborative Learning, which passes messages among meta-predictors in training process, and combines the results of each meta-predictor for the final training. In the study, we constructed four meta-predictors for SVM-based Collaborative Learning. The first meta-predictor is based on Motif-PSSM and ACC-PSSM. The second meta-predictor is based on Motif-PSSM, CKSAAP and the message passed from the first predictor, the third meta-predictor is based on Motif-PSSM, Kmer and the message passed from the second meta-predictor, the fourth meta-predictor is based on Motif-PSSM, Top- $n$ -gram, and the message passed from the third predictor. This process was shown in Fig. 2.

TABLE 1. The performance of different meta-predictors combination orders on the benchmark dataset with 10-fold cross-validation<sup>a</sup>.

Combination order	Sn (%)	Sp (%)	Acc (%)	MCC
F <sub>1</sub> F <sub>2</sub> F <sub>3</sub> F <sub>4</sub>	87.93	97.27	94.63	0.87
F <sub>1</sub> F <sub>4</sub> F <sub>3</sub> F <sub>2</sub>	87.68	97.34	94.61	0.87
F <sub>4</sub> F <sub>3</sub> F <sub>2</sub> F <sub>1</sub>	87.31	97.30	94.47	0.86
F <sub>2</sub> F <sub>4</sub> F <sub>1</sub> F <sub>3</sub>	87.31	97.32	94.48	0.86

<sup>a</sup> F<sub>1</sub> represents the meta-predictor based on Motif-PSSM and ACC-PSSM, F<sub>2</sub> represents the meta-predictor based on Motif-PSSM and CKSAAP, F<sub>3</sub> represents the meta-predictor based on Motif-PSSM and Kmer, F<sub>4</sub> represents the meta-predictor based on Motif-PSSM and Top- $n$ -gram.

#### E. PERFORMANCE MEASURES AND CROSS VALIDATION

We employed the 10-fold cross-validation to evaluate the performance of the proposed method [53].

We used four metrics to evaluate a predictor's quality [46], [48], [53]–[61] as shown in Eq. 12.

$$\left\{ \begin{array}{l} Sn = \frac{TP}{TP + FN} \\ Sp = \frac{TN}{TN + FP} \\ Acc = \frac{TP + TN}{TP + FN + TN + FP} \\ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TN + FN)(TP + FP)(TN + FP)}} \end{array} \right. \quad (12)$$

where TP is the number of true positive samples (correctly predicted RBP samples); TN is the number of true negative samples (correctly predicted non-RBP samples); FP represents the number of false positive samples (non-RBP samples wrongly predicted as RBP samples); FN represents the number of false negative samples (RBP samples wrongly predicted as non-RBP samples).

### III. RESULT AND DISCUSSION

We introduced a new predictor called iRBP-Motif-PSSM based on SVM-based Collaborative Learning method by combining the Motif-PSSM and various sequence-based features. The results showed that iRBP-Motif-PSSM was highly comparable, and even outperformed the other existing methods for identification of RNA-binding proteins, indicating that iRBP-Motif-PSSM will be helpful for biological analysis.

#### A. THE PERFORMANCE OF iRBP-MOTIF-PSSM BASED ON DIFFERENT COMBINATION ORDERS OF META-PREDICTORS

We investigated the impact of the combination order of the meta-predictors on the predictive performance. Four different combination orders were tested, and their predictive results were shown in Table 1, from which we can see that the proposed iRBP-Motif-PSSM predictor achieved stable performance with different combination orders of meta-predictors. The combination order F<sub>1</sub>F<sub>2</sub>F<sub>3</sub>F<sub>4</sub> was selected,



**TABLE 2.** The performance of various methods on the benchmark dataset (cf. Eq. 1) with 10-fold cross validation test.

Methods	Sn (%)	Sp (%)	Acc (%)	MCC
RBPPred <sup>a</sup>	82.77	96.50	92.64	0.81
iRBP-Motif-PSSM <sup>b</sup>	87.89	97.29	94.63	0.87

<sup>a</sup>The predictor reported in [10].

<sup>b</sup>The proposed predictor combining four meter-predictors in the order of F<sub>1</sub>F<sub>2</sub>F<sub>3</sub>F<sub>4</sub> via Collaborative Learning (see the footnote of Table 1).

**TABLE 3.** Performance of various methods on the independent test set.

Methods	SN (%)	SP (%)	Acc (%)	MCC
RNAPred <sup>a</sup>	69.57	76.00	73.37	0.45
RNAPred <sup>b</sup>	73.91	76.00	75.15	0.49
RNAPred <sup>c</sup>	72.46	79.00	76.33	0.51
iRBP-Motif-PSSM <sup>d</sup>	60.29	92.00	79.17	0.57

<sup>a</sup>The predictor based on PSSM-400 feature reported in [11].

<sup>b</sup>The predictor based on Hybrid1 feature reported in [11].

<sup>c</sup>The predictor based on Hybrid2 feature reported in [11].

<sup>d</sup>The proposed predictor combining four meter-predictors in the order of F<sub>1</sub>F<sub>2</sub>F<sub>3</sub>F<sub>4</sub> via Collaborative Learning (see the footnote of Table 1).

because the corresponding predictor can achieve the highest Acc.

### B. COMBINING MOTIF-PSSM AND SEQUENCE-BASED FEATURES CAN IMPROVE THE PREDICTIVE PERFORMANCE

In order to solve the aforementioned two problems in the field of RNA-binding protein identification, we combined the proposed Motif-PSSM feature and various sequence-based features via the SVM-based Collaborative Learning, and a predictor called iRBP-Motif-PSSM was proposed. Its performance was directly compared with another state-of-the-art predictor RBPPred [10] on the benchmark dataset, and the results were listed in **Table 2**, from which we can obviously observe that iRBP-Motif-PSSM outperformed RBPPred in terms of all the four performance measures listed in Eq. 12. Therefore, we concluded that the Motif-PSSM feature and the SVM-based Collaborative Learning contributed to the performance improvement of iRBP-Motif-PSSM.

### C. INDEPENDENT TEST

Independent test is a way to validate the generalization ability of the predictor [54]. In this study, we employed a widely used independent test set reported in [11] containing 68 RNA-binding protein and 100 non-RNA-binding proteins to further evaluate the performance of the proposed predictor.

In order to avoid overestimating the performance of our method, we removed the sequences sharing more than 25% sequence similarity with the sequences in the independent dataset from the benchmark dataset, and retrained the model based on the removed benchmark dataset. The trained model was used to predict the proteins in the independent dataset to give the final predictive results, and the corresponding results were shown in **Table 3**. These results further confirmed that the iRBP-Motif-PSSM is better than RNAPred.

**TABLE 4.** Performance of various methods on Gerstberger dataset.

Methods	Dataset	Sn (%)
RBPPred <sup>a</sup>	Gerstberger-1284	68.00
iRBP-Motif-PSSM <sup>b</sup>	Gerstberger-1396	86.60

<sup>a</sup>The predictor reported in [10].

<sup>b</sup>The proposed predictor combining four meter-predictors in the order of F<sub>1</sub>F<sub>2</sub>F<sub>3</sub>F<sub>4</sub> via Collaborative Learning (see the footnote of Table 1).

### D. APPLICATION OF iRBP-MOTIF-PSSM TO IDENTIFY RBPs IN HUMAN PROTEOME

The iRBP-Motif-PSSM was applied to predict the RBPs in the human proteome on the Gerstberger dataset [62], which is a census of 1396 RBPs in human proteome [62] extracted from Pfam database [63]. After removing the overlapping proteins between the benchmark dataset and the Gerstberger dataset from the benchmark dataset, the iRBP-Motif-PSSM was retrained with the benchmark dataset to predict the proteins in the Gerstberger dataset, and the results were shown in **Table 4**, from which we can see that iRBP-Motif-PSSM obviously outperformed RBPPred.

### IV. CONCLUSION

In this study, we introduced a new computational predictor for identifying RBPs. Compared with other existing predictors, it has the following advantages: 1) It incorporated a new feature called Motif-PSSM, considering both the evolutionary information from the PSSM and the function and structure information from the structural motifs and linear motifs; 2) It combined various meta-predictors via the SVM-based Collaborative Learning. It can be anticipated that the proposed Collaborative Learning framework would be applied to solve many important problems in bioinformatics, such as protein disordered protein prediction [64], DNA replication origin prediction, protein post-translational modification sites [65], etc.

### ACKNOWLEDGMENT

(Xin Gao and Donghua Wang are co-first authors.)

### REFERENCES

- [1] Q. Zou, P. Xing, L. Wei, and B. Liu, "Gene2vec: Gene subsequence embedding for prediction of mammalian N<sup>6</sup>-methyladenosine sites from mRNA," *RNA*, vol. 25, no. 2, pp. 205–218, 2019.
- [2] J. Chen, M. Guo, X. Wang, and B. Liu, "A comprehensive review and comparison of different computational methods for protein remote homology detection," *Briefings Bioinf.*, vol. 9, no. 2, pp. 231–244, 2016.
- [3] Z. Liao, Q. Zou, D. Li, L. Li, and X. Wang, "Cancer diagnosis through IsomiR expression with machine learning method," *Current Bioinf.*, vol. 13, no. 1, pp. 57–63, 2018.
- [4] J. Fukunaga, Y. Nomura, Y. Tanaka, R. Amano, T. Tanaka, Y. Nakamura, G. Kawai, T. Sakamoto, and T. Kozu, "The runt domain of AML1 (RUNX1) binds a sequence-conserved RNA motif that mimics a DNA element," *RNA*, vol. 19, no. 7, pp. 927–936, 2013.
- [5] W. H. Hudson and E. A. Ortlund, "The structure, function and evolution of proteins that bind DNA and RNA," *Nature Rev. Mol. Cell Biol.*, vol. 15, no. 11, pp. 749–760, 2014.
- [6] Y. Yang, H. Zhao, J. Wang, and Y. Zhou, "SPOT-Seq-RNA: Predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction," *Methods Mol. Biol.*, vol. 1137, no. 2, pp. 119–130, 2014.

- [7] T. Glisovic, J. L. Bachorik, J. Yong, and G. Dreyfuss, "RNA-binding proteins and post-transcriptional gene regulation," *Febs Lett.*, vol. 582, no. 14, pp. 1977–1986, Jun. 2008.
- [8] C. Z. Cai, L. Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen, "SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3692–3697, 2003.
- [9] X. Shao, Y. Tian, L. Wu, Y. Wang, L. Jing, and N. Deng, "Predicting DNA- and RNA-binding proteins from sequences with kernel methods," *J. Theor. Biol.*, vol. 258, no. 2, pp. 289–293, May 2009.
- [10] X. Zhang and S. Liu, "RBPPred: Predicting RNA-binding proteins from sequence using SVM," *Bioinformatics*, vol. 33, no. 6, pp. 854–862, Mar. 2017.
- [11] M. Kumar, M. M. Gromiha, and G. P. S. Raghava, "SVM based prediction of RNA-binding proteins using binding residues and evolutionary information," *J. Mol. Recognit.*, vol. 24, no. 2, pp. 303–313, 2011.
- [12] B. Liu, X. Gao, and H. Zhang, "BioSeq-Analysis2.0: An updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches," *Nucleic Acids Res.*, vol. 47, no. 20, p. e127, Nov. 2019.
- [13] B. Liu, C.-C. Li, and K. Yan, "DeepSVM-fold: Protein fold recognition by combining Support Vector Machines and pairwise sequence similarity scores generated by deep learning networks," *Briefings Bioinf.*, to be published, doi: [10.1093/bib/bbz098](https://doi.org/10.1093/bib/bbz098).
- [14] B. Liu and Y. Zhu, "ProtDec-LTR3.0: Protein remote homology detection by incorporating profile-based features into learning to rank," *IEEE Access*, vol. 7, pp. 102499–102507, 2019.
- [15] B. Liu, "BioSeq-Analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches," *Briefings Bioinf.*, vol. 20, no. 4, pp. 1280–1294, Dec. 2017, doi: [10.1093/bib/bbx165](https://doi.org/10.1093/bib/bbx165).
- [16] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou, "Pse-in-one: A Web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W65–W71, Jul. 2015.
- [17] X.-J. Zhu, C.-Q. Feng, H.-Y. Lai, W. Chen, and L. Hao, "Predicting protein structural classes for low-similarity sequences by evaluating different features," *Knowl.-Based Syst.*, vol. 163, pp. 787–793, Jan. 2019.
- [18] Z. Chen, P. Zhao, F. Li, T. T. Marquez-Lago, A. Leier, J. Revote, Y. Zhu, D. R. Powell, T. Akutsu, G. I. Webb, K.-C. Chou, A. I. Smith, R. J. Daly, J. Li, and J. Song, "iLearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data," *Briefings Bioinf.*, to be published, doi: [10.1093/bib/bbz041](https://doi.org/10.1093/bib/bbz041).
- [19] Z. Chen, P. Zhao, F. Li, A. Leier, T. T. Marquez-Lago, Y. Wang, G. I. Webb, A. I. Smith, R. J. Daly, K.-C. Chou, and J. Song, "iFeature: A Python package and Web server for features extraction and selection from protein and peptide sequences," *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, Jul. 2018.
- [20] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.
- [21] L. Holm and C. Sander, "Removing near-neighbour redundancy from large protein sequence collections," *Bioinformatics*, vol. 14, no. 5, pp. 423–429, Jun. 1998.
- [22] C. Bousquet-Antonelli and J.-M. Deragon, "A comprehensive analysis of the La-motif protein superfamily," *RNA*, vol. 15, no. 5, pp. 750–764, May 2009.
- [23] C. Maris, C. Dominguez, and F. H.-T. Allain, "The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression," *FEBS J.*, vol. 272, no. 9, pp. 2118–2131, May 2005.
- [24] C.-C. Li and B. Liu, "MotifCNN-fold: Protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks," *Briefings Bioinf.*, to be published, doi: [10.1093/bib/bbz133](https://doi.org/10.1093/bib/bbz133).
- [25] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble, "The MEME suite," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W39–W49, 2015.
- [26] G. Pugalenti, P. N. Sanganthan, R. Sowdhamini, and S. Chakrabarti, "MegaMotifBase: A database of structural motifs in protein families and superfamilies," *Nucleic Acids Res.*, vol. 36, pp. D218–D221, Jan. 2008.
- [27] M. Gouw et al., "The eukaryotic linear motif resource—2018 update," *Nucleic Acids Res.*, vol. 46, no. 1, pp. D428–D434, 2018.
- [28] P. Puntervoll et al., "EDLM server: A new resource for investigating short functional sites in modular eukaryotic proteins," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3625–3630, 2003.
- [29] J. Chen, X. Wang, and B. Liu, "iMiRNA-SSF: Improving the identification of MicroRNA precursors by combining negative sets with different distributions," *Sci. Rep.*, vol. 6, Jan. 2016, Art. no. 19062.
- [30] Q. Dong, S. Zhou, and J. Guan, "A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation," *Bioinformatics*, vol. 25, no. 20, pp. 2655–2662, Oct. 2009.
- [31] X. Ru, L. Li, and Q. Zou, "Incorporating distance-based top-n-gram and random forest to identify electron transport proteins," *J. Proteome Res.*, vol. 18, no. 7, pp. 2931–2939, May 2019.
- [32] J.-X. Tan, S.-H. Li, Z.-M. Zhang, C.-X. Chen, W. Chen, H. Tang, and H. Lin, "Identification of hormone binding proteins based on machine learning methods," *Math. Biosci. Eng.*, vol. 16, no. 4, pp. 2466–2480, 2019.
- [33] X.-X. Chen, H. Tang, W.-C. Li, H. Wu, W. Chen, H. Ding, and H. Lin, "Identification of bacterial cell wall lyases via pseudo amino acid composition," *Biomed. Res. Int.*, vol. 2016, May 2016, Art. no. 1654623.
- [34] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, "A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis," *BMC Bioinf.*, vol. 9, Dec. 2008, Art. no. 510.
- [35] K. Chen, Y. Jiang, L. Du, and L. Kurgan, "Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs," *J. Comput. Chem.*, vol. 30, no. 1, pp. 163–172, 2009.
- [36] K. Chen, L. Kurgan, and M. Rahbari, "Prediction of protein crystallization using collocation of amino acid pairs," *Biochem. Biophys. Res. Commun.*, vol. 355, no. 3, pp. 764–769, 2007.
- [37] K. Chen, L. A. Kurgan, and J. Ruan, "Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs," *BMC Struct. Biol.*, vol. 7, no. 1, 2007, Art. no. 25.
- [38] K. Chen, L. A. Kurgan, and J. Ruan, "Prediction of protein structural class using novel evolutionary collocation-based sequence representation," *J. Comput. Chem.*, vol. 29, no. 10, pp. 1596–1604, 2008.
- [39] S. P. Wang, Q. Zhang, J. Lu, and Y.-D. Cai, "Analysis and prediction of nitrated tyrosine sites with the mRMR method and support vector machine algorithm," *Current Bioinf.*, vol. 13, no. 1, pp. 3–13, 2018.
- [40] B. Liu, F. Yang, D.-S. Huang, and K.-C. Chou, "iPromoter-2L: A two-layer predictor for identifying promoters and their types by multi-window-based PseKNC," *Bioinformatics*, vol. 34, no. 1, pp. 33–40, Jan. 2018.
- [41] H.-Y. Lai, Z.-Y. Zhang, Z.-D. Su, W. Su, H. Ding, W. Chen, and H. Lin, "iProEP: A computational predictor for predicting promoter," *Mol. Therapy Nucleic Acids*, vol. 17, pp. 337–346, Jun. 2019.
- [42] F.-Y. Dao, H. Lv, F. Wang, C.-Q. Feng, H. Ding, W. Chen, and H. Lin, "Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique," *Bioinformatics*, vol. 35, no. 12, pp. 2075–2083, Jun. 2019.
- [43] J. Song, Y. Wang, F. Li, T. Akutsu, N. D. Rawlings, G. I. Webb, and K.-C. Chou, "iProt-Sub: A comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites," *Briefings Bioinf.*, vol. 20, no. 2, pp. 638–658, Mar. 2019.
- [44] M. Zhang, F. Li, T. T. Marquez-Lago, A. Leier, C. Fan, C. K. Kwok, K.-C. Chou, J. Song, and C. Jia, "MULTiPLY: A novel multi-layer predictor for discovering general and specific types of promoters," *Bioinformatics*, vol. 35, no. 17, pp. 2957–2965, Sep. 2019.
- [45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
- [46] B. Liu, R. Long, and K.-C. Chou, "iDHS-EL: Identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework," *Bioinformatics*, vol. 32, no. 16, pp. 2411–2418, 2016.
- [47] B. Liu, F. Yang, and K. C. Chou, "2L-piRNA: A two-layer ensemble classifier for identifying PIWI-interacting RNAs and their function," *Mol. Therapy-Nucleic Acids*, vol. 7, pp. 267–277, Jun. 2017.
- [48] B. Liu, S. Wang, R. Long, and K.-C. Chou, "iRSpot-EL: Identify recombination spots with an ensemble learning approach," *Bioinformatics*, vol. 33, no. 1, pp. 35–41, Jan. 2017.
- [49] Q. Zou, J. Guo, Y. Ju, M. Wu, X. Zeng, and Z. Hong, "Improving tRNAscan-SE annotation results via ensemble classifiers," *Mol. Inform.*, vol. 34, nos. 11–12, pp. 761–770, Nov. 2015.
- [50] Q. Zou, Z. Wang, X. Guan, B. Liu, Y. Wu, and Z. Lin, "An approach for identifying cytokines based on a novel ensemble classifier," *Biomed. Res. Int.*, vol. 2013, Jul. 2013, Art. no. 686090.

- [51] X. Ru, P. Cao, L. Li, and Q. Zou, "Selecting essential MicroRNAs using a novel voting method," *Mol. Therapy Nucleic Acids*, vol. 18, pp. 16–23, Dec. 2019.
- [52] Q. Liao and Q. Zhang, "Local coordinate based graph-regularized NMF for image representation," *Signal Process.*, vol. 124, pp. 103–114, Jul. 2016.
- [53] J. Zhang and B. Liu, "PSFM-DBT: Identifying DNA-binding proteins by combing position specific frequency matrix and distance-bigram transformation," *Int. J. Mol. Sci.*, vol. 18, no. 9, p. 1856, 2017.
- [54] B. Liu, F. Weng, D.-S. Huang, and K.-C. Chou, "iRO-3wPseKNC: Identify DNA replication origins by three-window-based PseKNC," *Bioinformatics*, vol. 34, no. 18, pp. 3086–3093, Sep. 2018.
- [55] H. Lv, Z.-M. Zhang, S.-H. Li, J.-X. Tan, W. Chen, and H. Lin, "Evaluation of different computational methods on 5-methylcytosine sites identification," *Briefings Bioinf.*, to be published.
- [56] Y. Xie, X. Luo, Y. Li, L. Chen, W. Ma, J. Huang, J. Cui, Y. Zhao, Y. Xue, Z. Zuo, and J. Ren, "DeepNitro: Prediction of protein nitration and nitrosylation sites by deep learning," *Genomics Proteomics Bioinf.*, vol. 16, no. 4, pp. 294–306, Aug. 2018.
- [57] L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, and F. Guo, "Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier," *Artif. Intell. Med.*, vol. 83, pp. 67–74, Nov. 2017.
- [58] L. Wei, S. Wan, J. Guo, and K. K. L. Wong, "A novel hierarchical selective ensemble classifier with bioinformatics application," *Artif. Intell. Med.*, vol. 83, pp. 82–90, Nov. 2017.
- [59] F. Li, C. Li, T. T. Marquez-Lago, A. Leier, T. Akutsu, A. W. Purcell, A. I. Smith, T. Lithgow, R. J. Daly, J. Song, and K.-C. Chou, "Quokka: A comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome," *Bioinformatics*, vol. 34, no. 24, pp. 4223–4231, Dec. 2018.
- [60] F. Li, Y. Wang, C. Li, T. T. Marquez-Lago, A. Leier, N. D. Rawlings, G. Haffari, J. Revote, T. Akutsu, K.-C. Chou, A. W. Purcell, R. N. Pike, G. I. Webb, A. I. Smith, T. Lithgow, R. J. Daly, J. C. Whisstock, and J. Song, "Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: A comprehensive revisit and benchmarking of existing methods," *Briefings Bioinf.*, to be published.
- [61] F. Li, Y. Zhang, A. W. Purcell, G. I. Webb, K.-C. Chou, T. Lithgow, C. Li, and J. Song, "Positive-unlabelled learning of glycosylation sites in the human proteome," *BMC Bioinf.*, vol. 20, no. 1, p. 112, Mar. 2019.
- [62] S. Gerstberger, M. Hafner, and T. Tuschl, "A census of human RNA-binding proteins," *Nature Rev. Genet.*, vol. 15, no. 12, pp. 829–845, Nov. 2014.
- [63] R. D. Finn, "The PFAM protein families database," *Nucleic Acids Res.*, vol. 38, pp. D211–D222, Jan. 2010.
- [64] Y. Liu, X. Wang, and B. Liu, "A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction," *Briefings Bioinf.*, vol. 20, no. 1, pp. 330–346, 2019.
- [65] F. Li, C. Fan, T. T. Marquez-Lago, A. Leier, J. Revote, C. Jia, Y. Zhu, A. I. Smith, G. I. Webb, Q. Liu, L. Wei, J. Li, and J. Song, "PRISM: A comprehensive 3D structure database for post-translational modifications and mutations with functional impact," *Briefings Bioinf.*, to be published.



**DONGHUA WANG** received the bachelor's and master's degrees in medicine from the Harbin Medical School, in 1989 and 1992, respectively. From 2004 to 2019, he was the Chief Physician with the General Hospital of the Provincial Agricultural Reclamation. He is currently the Vice Chairman of the association, the Executive Director of the Provincial Medical Association, the Executive Director of the Provincial Association of doctors, and the Executive Director of the Association of Youth Prosperity Leaders of the Nongnongken. His research is mainly focused on general surgery. He is a General Member of the Provincial Medical Association, and a member of the Cancer Committee of the Provincial Medical Association.



**JUN ZHANG** received the B.S. degree from the Henan University of Science and Technology, China, in 2016, and the M.S. degree from the Harbin Institute of Technology, Shenzhen, China, in 2018, where he is currently pursuing the Ph.D. degree in computer science and technology. His research interests include bioinformatics, natural language processing, and machine learning.



**QING LIAO** received the Ph.D. degree in computer science and engineering from the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, in 2016, under the supervision by Prof. Q. Zhang. She is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. Her research interests include artificial intelligence and bioinformatics.



**BIN LIU** received the Ph.D. degree from the Harbin Institute of Technology, China, in 2010. From 2010 to 2012, he was a Postdoctoral Researcher with The Ohio State University, USA. He was with the Harbin Institute of Technology, Shenzhen, China, from 2012 to 2019, as a Professor. He is currently a Full Professor with the Beijing Institute of Technology. His research interests include bioinformatics, machine learning, and natural language processing. He was focused on exploring the language models of biological sequences and proposing computational predictors for some important tasks in bioinformatics based on natural language processing techniques.



**XIN GAO** is currently pursuing the master's degree in computer science and technology with the Harbin Institute of Technology, Shenzhen, China. His research interests include bioinformatics and machine learning.