IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 31, 2020, accepted August 12, 2020, date of publication September 28, 2020, date of current version October 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3023800

ACP-GCN: The Identification of Anticancer Peptides Based on Graph Convolution Networks

BING RAO[1], LICHAO ZHANG[2], AND GUOYING ZHANG[1]
[1]School of Mechanical Electronic and Information Engineering, China University of Mining and Technology at Beijing, Beijing 100083, China
[2]School of Intelligent Manufacturing and Equipment, Shenzhen Institute of Information Technology, Shenzhen 518172, China

Corresponding authors: Guoying Zhang (zhangguoying1101@163.com) and Lichao Zhang (lczhang5354@szu.edu.cn)

This work was supported by the National Natural Science Foundation under Grant U1704242.

ABSTRACT Anticancer peptide (ACP) is a class of anti-cancer peptide which can inhibit and kill tumor cells. Identification of ACPs is of great significance for the development of new anti-cancer drugs. However, most of computational methods make predictions based on machine learning using hand-crafted features. In this article, we propose a new graph learning based computational model, named ACP-GCN, to automatically and accurately predict ACPs based on graph convolution networks. In this model, we for the first time take the ACP prediction as a graph classification task, where each peptide sample is represented as a graph. The experimental results show that the proposed model outperforms most of state-of-the-art methods, demonstrating that the proposed method can effectively distinguish ACPs from non-ACPs. The excellent predictive ability will rapidly push forward their applications in cancer therapy.

INDEX TERMS Anti-cancer peptides, graph convolution networks, machine learning, prediction methods.

I. INTRODUCTION

Cancer is one of the most deadly diseases in the world, killing millions of people every year. Traditional treatment is to use the conventional chemotherapy. However, its treatment effect is not that good and usually has adverse effect on normal cells. Therefore, there is an urgent need for more effective therapeutic treatment to overcome the shortcomings of traditional chemotherapy. Anticancer peptides (ACPs), with a length of 5 to 30 amino acids, have emerged as a new therapeutic agent, opening a promising perspective for cancer treatment. As compared with conventional chemotherapy, ACPs have multiple attractive advantages, such as high tumor penetration, low cost, high specific, greater efficacy, and selectivity. In recent years, peptide-based therapies are increasingly used to treat various tumor types across different phases of clinical trials. However, the number of experimentally validated ACPs is very limited. How to identify potential ACPs in a large number of proteins is a challenging task.

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

Over the last decade, a series of computational efforts have done for the identification of ACPs. Most of them are motivated by the limit of biological experimental methods, which are highly cost and time consuming for the prediction of ACPs. Recent studies mainly focus on the computational methods like machine learning to identify ACPs. For machine learning based methods, the regular way is to extract features, and train a prediction model based on the features to automatically classify the peptides as ACPs or not [1]–[9]. Researchers have made some achievements in the prediction of anticancer peptides using machine learning methods [10]. In the aspect of feature extraction, a lot of feature extraction methods have been proposed, including amino acid composition [11]–[13], dipeptide composition [14], [15], and binary profile of pattern. For example, Chen *et al.* proposed a sequence-based predictor called iACP by optimizing the g-gap dipeptide components. ACPred-FL can automatically extract and learn sequence information by using several different SVM models [16]–[23]. Moreover, Tyagi *et al.* have designed and discovered a novel anticancer peptide predictor in silico models called AntiCP. Recently, Manavalan *et al.* proposed a machine learning-based model

named MLACP to predict anticancer peptides. As introduced above, most of current methods extract features based on the peptide primary sequences. However, effective feature extraction highly relies on the experience and knowledge of researchers. On the other hand, most of existing hand-craft features capture local information only, ignoring the global information.

Deep learning has recently achieved a remarkable impact on multiple bioinformatics fields, including biological images [24], drug classification [25], protein fold recognition [26], [27] and biomarker discovery [28], [29]. However, most of popular deep neural models, such as convolutional neural networks (CNNs) [30], only work on grid-structured (Euclidean) data, and are not directly applicable to graphs. As well known, the data in most of sequence analysis and classification tasks are non-Euclidean data. For this reason, peptide firstly extracts features from the peptide graphs before applying a CNN. Recently, there has been growing interest in extending deep learning techniques to non-Euclidean data.

In this study, we proposed a novel predictor called ACP-GCN to predict ACPs. In the proposed model, we use one-hot encoding method and graph convolution network (GCN) to predict. To be specific, see the following sections. Firstly, the data sets are described in detail, including training data set and independent test data set. Secondly, the theoretical methods used in the research are introduced in detail, that is, three main steps in ACP-GCN: data set construction; feature embedding; GCN algorithm. Thirdly, the performance evaluation, experimental process and results of this project are described in detail. Finally, the research results in this article are summarized, and the future work is prospected.

II. METHODOLOGY

A. DATA SETS

In this study, we used the same date set collected in Wei *et al.*'s study [31]. The data collecting procedure is as follows. They collected 3212 ACPs with experimental validation as positive samples. The same number of anti-microbial peptides (AMPs) that are not shown to have anticancer activity are used as negative samples. Afterwards, to avoid the homology bias, they used the CD-HIT program to reduce the similarity of the samples to 0.8 [32]. By doing so, 332 ACPs (positives) and 1023 non-ACPs (negatives) are remaining in the dataset. They collected more other non-ACPs in addition to the negative data set above, ultimately yielding 332 ACPs samples and 2878 non-ACPs samples in the data set. Of this dataset, they used 250 ACPs positive training samples, and the same number of non-ACPs as negative training samples, to construct the training dataset; the remaining 82 ACPs positive samples and 2628 non-ACPs negative samples were yielded as independent dataset. In our study, we used the same training dataset for model training and evaluation.

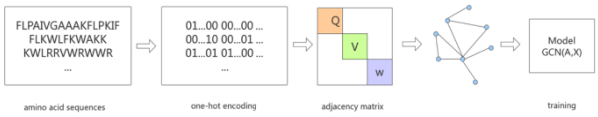

FIGURE 1. Model overview.

B. THE PREDICTIVE MODEL ARCHITECTURE

In this section, we introduce the overall framework of our model. All steps of the model are shown in Figure 1: Step 1, prepare the protein data set and get the amino acid sequence. Step 2, extracted the features of proteins which was represented as amino acid sequence by one-hot encoding method. Step 3, calculate the distance between the samples, and gets the adjacency matrix. Step 4, construct the protein and amino acid graph. Step 5, use GCN model to train datasets. Then minimize the loss and get the classification results and various evaluation indexes. And the cross entropy loss function was used to optimize the classification result.

C. ONE-HOT ENCODING

One-hot encoding is an effective encoding method expressed by binary vectors [33]–[36]. Only one bit is valid at any time, other positions are set to 0. For example, if we want to use one-hot to represent cat, dog and bird, we can use 100 to represent cat, 010 to represent dog, 001 to represent bird. The primary structural information of protein is mainly composed of 20 kinds of common amino acids. 20 kinds of amino acids are represented by a single letter, separately. Set AA is a table of all 20 letter sets involved in our experiment dataset. AA = [A,C,E,F,G,H,I.K,L,M,N,P,Q,R,S,T,V,W,Y]. It is a basic protein encoding method. We encoded the single letter A as 10000000000000000000, and letter C as 01000000000000000000. Consequently, an amino acid sequence CA can be expressed as 01000000000000000000100000000000000000000 by one-hot encoding.

D. GRAPH CONVOLUTION NETWORKS

Graph convolution networks (GCN), Semi-Supervised Classification with Graph Convolution Networks, is a variant of traditional convolution networks, which can be directly used to process graph structure data. The theoretical formula of single layer GCN is as follows:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{1}$$

where $H^{(l)} \in R^{N\times D}$, N is the number of nodes in the graph, each node is represented by a D-dimension feature vector. The input of layer l is $H^{(l)}$, thus the initial input layer is $H^{(0)} = X$. A is an adjacency matrix. A is an adjacency matrix with self-connections, $\tilde{A} = A + I_N$. $\tilde{D}$ is a degree matrix and $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. $W^{(l)}$ is the trainable parameter, $W^{(l)} \in R^{D\times D}$. σ is the corresponding activation function, such as $ReLU(\cdot)$ or $\max(0, \cdot)$. By calculating this formula, we can get the output graph $H^{(l+1)}$. The above is the complete final form of GCN.

E. PROBLEM DEFINITION

In this research, we cast the recognition of therapeutic peptide as a text classification problem. Text classification is a fundamental problem in the field of natural language processing, which has found numerous application scenarios including news filtering, document organization, spam detection, and opinion mining, etc. The models used for text classification include traditional models (e.g., bag-of-word, n-grams, entities in ontologies, etc.) and deep learning-based models (e.g., CNN, RNN, LSTM, etc.). Recently, graph convolutional networks are attracting more and more attention. Its application to text classification achieved state-of-the-art results on a number of benchmark graph datasets. In GCN models, text elements are represented by a graph. The relationship between these elements is embedded in the graph via links between nodes. For peptide recognition, comparison with text classification, a peptide is regarded as a document and an amino acid is regarded as a word. A graph is constructed on the entire peptide dataset. The node represents peptide and amino acid. The edge is built by using the co-occurrence information. Therefore, the peptide classification is converted to node classification of the graph.

F. PEPTIDE GRAPH CONVOLUTIONAL NETWORK

The peptide and the amino acid are defined as the edges, based on which a graph $G = (V, E)$ is constructed, where V denotes the edges and E denotes the relationship between edges. Let $x_i \in R^m$ be the feature vector associated with node i, where m denotes the dimensionality of the feature space. Let $X \in R^{(n\times m)}$ denotes the matrix containing all n feature vectors. Next, we need to construct an adjacency matrix $A \in R^{(n\times n)}$ that represents the relationship between nodes. A degree matrix $D \in R^{(n\times n)}$ of the adjacency matrix is then obtained by summing over each row of A, i.e., $D_{ii} = \sum_j A_{ij}$, where A_{ij} represents the element of matrix A at position (i, j). Given X as input, the output of the first layer neurons is computed as follows.

$$L^{(1)} = \rho(\widetilde{A}XW_0) \tag{2}$$

where $W_0 \in R^{(m\times k)}$ is a weight matrix, $\widetilde{A} = D^{(-1/2)}AD^{(-1/2)}$ is the normalized symmetric adjacency matrix, and $\rho(\cdot)$ is the activation function, e.g., ReLU $\rho(x) = max(0, x)$. $L^{(1)}$ denotes the feature representation of the first layer neurons in the GCN. In GCN, the feature representation is propagated in a stack of layers all the way to the final output layer. More generally, the output of layer j is computed as follows,

$$L^{(j+1)} = \rho(\widetilde{A}L^{(j)}W_0) \tag{3}$$

where $L^{(0)} = X$.

G. CONSTRUCTION OF THE ADJACENCY MATRIX

As is mentioned before that the nodes represent both peptides and amino acid, the relationship between edges is constructed based on co-occurrence information, including peptides-amino acid co-occurrence and amino acid-amino acid co-occurrence. As such, a large and heterogeneous peptide graph is constructed in which both global peptides-amino acid relationship and local amino acid-amino acid relationship are explicitly modeled, as shown in Figure 1. The whole peptide dataset is used to build the graph, which we call the corpus. The total number of nodes of the graph is the number of peptides plus the number of unique amino acids.

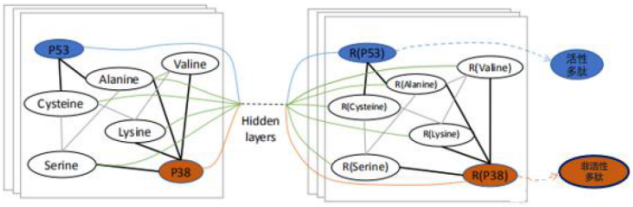

FIGURE 2. Anticancer peptides classification based on graph convolutional neural network.

For modeling an amino acid-amino acid pair, we employ point-wise mutual information (PMI) to calculate weights between two adjacent amino acids. The PMI value of an amino acid-amino acid pair is computed as follows,

$$PMI(i, j) = log\frac{p(i, j)}{p(i)\,p(j)} \tag{4}$$

$$p(i, j) = \frac{\#W(i, j)}{\#W} \tag{5}$$

$$p(i) = \frac{\#W(i)}{\#W} \tag{6}$$

where $\#W(i, j)$ denotes the number of sliding windows in a corpus that contain word i, $\#W(i)$ denotes the number of sliding windows in a corpus that contain word i, and $\#W$ denotes the total number of sliding windows in the corpus. The PMI value implies the semantic correlation of amino acid in a corpus.

For modeling a peptide-amino acid pair, we employ term frequency-inverse document frequency (TF-IDF), where term frequency is the number of times the amino acid appears in the peptide, and inverse document frequency is the logarithmically scaled inverse fraction of the number of peptides that contain amino acids.

As such, the adjacency matrix representing the weight between nodes is constructed as follows,

$$A_{ij} = \begin{cases} PMI(i, j) & \text{i, j are amino acids} \\ TF - IDF_{ij} & \text{i is peptide, j is amino acid} \\ 1 & \text{i = j} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

H. LOSS FUNCTION

The proposed peptide graph convolutional network takes as input the initial data X as in Eq. (1), and propagates through n layers to the output layer $L^{(n)}$, followed by a softmax layer,

$$Z = \text{softmax}\left(L_i^{(n)}\right) = \frac{\exp\left(L_i^{(n)}\right)}{\sum_i \exp\left(L_i^{(n)}\right)} \tag{8}$$

We employ cross entropy to define our loss function,

$$\mathcal{L} = -\sum_{d\in y_D}\sum_{f=1}^{F} Y_{df} \ln Z_{df} \tag{9}$$

where y_D is the set of ground truth labels with the d-th label Y_d, f is the number of classes.

The layer weights W_i for $i = 0, 1, \ldots, n-1$ are determined by training the network on the training set. In the training procedure, we perform 5-fold cross validation. We randomly split the training set into two parts, i.e., 80 percent of the dataset as training examples and 20 percent as validation set. We repeat this procedure for a total of 5 times. We set the size from 5 to 40 in steps of 5 The trained network is applied on the test set to obtained the test results.

I. I. EVALUATION METRICS

For performance evaluation, we used five commonly used metrics including sensitivity (SE), specificity (SP), accuracy (ACC), and Matthew's correlation coefficient (MCC), and AUC (Area Under the Curve), which are widely used in several bioinformatics fields [38]–[53]. The formulas of the first four metrics are as follows:

$$\begin{cases} \mathrm{SE} = \dfrac{TP}{TP+FN} \times 100\% \\ \mathrm{SP} = \dfrac{TN}{TN+FP} \times 100\% \\ \mathrm{ACC} = \dfrac{TP+TN}{TP+TN+FP+FN} \times 100\% \\ MCC = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \times 100\% \end{cases} \tag{10}$$

where TP (true positive) represent the number of real ACPs predicted as real ACPs; TN (true negative) represent the number of non-ACPs predicted as non-ACPs; FP (false positive) represent the number of non-ACPs predicted as ACPs; and FN (false negative) represent the number of real ACPs predicted as non-ACPs. The SE and SP metrics measure the predictive ability of the predictor for the positives and negatives, respectively, while the other two metrics, ACC and MCC, measure the overall predictive performance. Moreover, we also plotted the ROC (receiver operating characteristic) curve to visualize the overall performance of a binary classifier system for comparison purpose. The area under ROC curve is calculated to quantitively evaluate the predictive performance [54], [55]. The value of AUC ranges from 0.5 to 1. The higher the score of AUC achieves, the better the performance of the models is.

III. RESULTS AND DISCUSSION

A. PARAMETER OPTIMIZATION IN THE MODELING PROCESS

Different parameter settings may affect the accuracy of our proposed model. To establish a more discriminative prediction model, we compared the performance under different parameters through five-fold cross validation to search for the best parameters, such as kmer, learning rate, word embeddim and slide size. Specifically, in GCN, we set the k value of kmer from 1 to 8 respectively, and compared the accuracy corresponding to different k values. Experiment results showed that when k is 2, the accuracy reaches the highest value of 73.4%. As for the neural network learning rate, we set the learning rate to 0.001, 0.002, 0.004, 0.006, 0.010, 0.020 and 0.030, respectively. Comparing the performance under different learning rates, the highest accuracy is 77.8% when the learning rate is 0.002. In the process of PMI building, the word embeddim is set to 7 different values, namely, 5, 20, 30, 40, 50, 100 and 200, respectively. The slide size is set from 5 to 40 in steps of 5. We observed that when the word embeddim is 20 or the slide size is 20, the highest accuracy can both arrive 77.8%. Thus, we can determine the best parameters based on the comparison results under different parameters. The specific experimental results are shown in Figure 3.

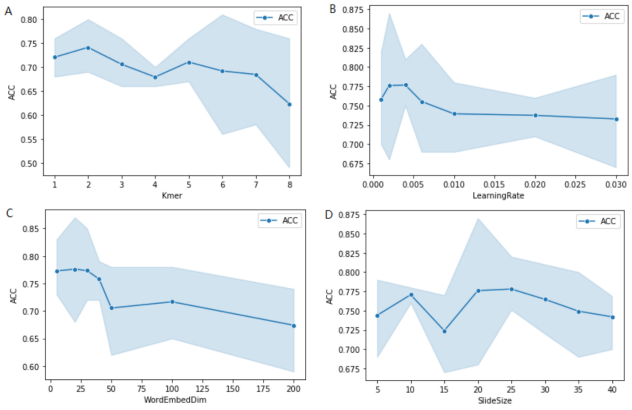

FIGURE 3. Performance comparison with different parameters on training set.

B. PERFORMANCE COMPARISON USING DIFFERENT NEURAL NETWORK METHODS

To verify the effect of different neural network methods on the accuracy of the established model, we compared

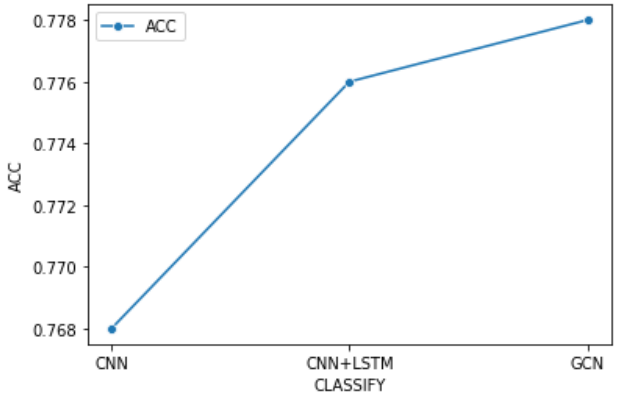

FIGURE 4. Performance comparison of GCN and other commonly-used neural networks methods.

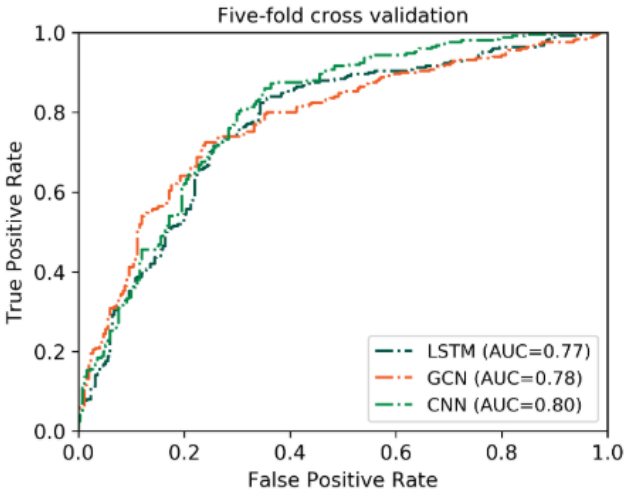

FIGURE 5. The ROC curves of GCN and other commonly-used neural networks methods.

the prediction performance using the GCN method with other two commonly-used neural network methods, CNN and CNN-LSTM. As shown in Figure 4, the GCN method reached a highest accuracy of 77.8%, better than CNN (76.8%) and CNN-LSTM (77.6%). The ROC curves of the prediction models using three different neural networks methods are shown in Figure 5. The GCN obtained an AUC of 0.78, which is 0.02 lower than CNN and 0.01 higher than CNN-LSTM, respectively. Although GCN is slightly lower than CNN in terms of AUC, it is far better than CNN in terms of ACC. Therefore, comprehensively comparing the performance of ACC and AUC, we choose the GCN as our classification method in the modeling process. These results demonstrate that our proposed GCN-based method is better than other commonly used deep learning methods in identification of anti-cancer peptides.

IV. CONCLUSION

In this work, we have established a novel predict model to identify ACPs, called ACP-GCN. It is a powerful bioinformatics tool to identify ACPs. Experiments demonstrated that the proposed method can work better compared to several existing descriptors. Experimental results on both the 10-fold cross validation and independent tests show that this proposed predictor is more effective to discriminate ACPs from non-ACPs. And we found that it can provide a significant improvement of the predictive performance and the excellent predictive ability will accelerate their applications in cancer therapy. In the future work, we will try more computational techniques for more precise and excellent prediction [56]–[58]. Computational methods such as graph neural networks [59], [60] and optimization algorithms [61], [62] would also benefit for the ACPs prediction.

REFERENCES

[1] Q. Zou, "Latest machine learning techniques for biomedicine and bioinformatics," *Current Bioinf.*, vol. 14, no. 3, pp. 176–177, Mar. 2019, doi: 10.2174/1574893614031902201128555.

[2] C. Meng, J. Zhang, X. Ye, F. Guo, and Q. Zou, "Review and comparative analysis of machine learning-based phage virion protein identification methods," *Biochimica et Biophys. Acta (BBA) Proteins Proteomics*, vol. 1868, no. 6, Jun. 2020, Art. no. 140406, doi: 10.1016/j.bbapap.2020.140406.

[3] Q. Zou and Q. Ma, "The application of machine learning to disease diagnosis and treatment," *Math. Biosci.*, vol. 320, Feb. 2020, Art. no. 108305, doi: 10.1016/j.mbs.2019.108305.

[4] T.-H. Zhang and S.-W. Zhang, "Advances in the prediction of protein subcellular locations with machine learning," *Current Bioinf.*, vol. 14, no. 5, pp. 406–421, Jun. 2019, doi: 10.2174/1574893614666181217145156.

[5] K. Qu, F. Guo, X. Liu, Y. Lin, and Q. Zou, "Application of machine learning in microbiology," *Frontiers Microbiol.*, vol. 10, p. 827, Apr. 2019.

[6] K. Patil and U. Chouhan, "Relevance of machine learning techniques and various protein features in protein fold classification: A review," *Current Bioinf.*, vol. 14, no. 8, pp. 688–697, Dec. 2019, doi: 10.2174/1574893614666190204154038.

[7] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers Genet.*, vol. 9, p. 515, Nov. 2018, doi: 10.3389/fgene.2018.00515.

[8] 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.

[9] K. Yan, X. Fang, Y. Xu, and B. Liu, "Protein fold recognition based on multi-view modeling," *Bioinformatics*, vol. 35, no. 17, pp. 2982–2990, Sep. 2019.

[10] H.-Y. Lai, C.-Q. Feng, Z.-Y. Zhang, H. Tang, W. Chen, and H. Lin, "A brief survey of machine learning application in cancerlectin identification," *Current Gene Therapy*, vol. 18, no. 5, pp. 257–267, Nov. 2018, doi: 10.2174/1566523218666180913112751.

[11] Z. Lv, S. Jin, H. Ding, and Q. Zou, "A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features," *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 215, Sep. 2019.

[12] W. Yang, X.-J. Zhu, J. Huang, H. Ding, and H. Lin, "A brief survey of machine learning methods in protein sub-golgi localization," *Current Bioinf.*, vol. 14, no. 3, pp. 234–240, Mar. 2019.

[13] 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, Jul. 2019.

[14] H. Tang, Y.-W. Zhao, P. Zou, C.-M. Zhang, R. Chen, P. Huang, and H. Lin, "HBPred: A tool to identify growth hormone-binding proteins," *Int. J. Biol. Sci.*, vol. 14, no. 8, pp. 957–964, 2018, doi: 10.7150/ijbs.24174.

[15] 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, doi: 10.3934/mbe.2019123.

[16] C. Meng, F. Guo, and Q. Zou, "CWLy-SVM: A support vector machine-based tool for identifying cell wall lytic enzymes," *Comput. Biol. Chem.*, vol. 87, Aug. 2020, Art. no. 107304, doi: 10.1016/j.compbiolchem.2020.107304.

[17] L. Chao, L. Wei, and Q. Zou, "SecProMTB: A SVM-based classifier for secretory proteins of mycobacterium tuberculosis with imbalanced data set," *Proteomics*, vol. 19, Aug. 2019, Art. no. e1900007.

[18] C. Meng, S. Jin, L. Wang, F. Guo, and Q. Zou, "AOPs-SVM: A sequence-based classifier of antioxidant proteins using a support vector machine," *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 224, Sep. 2019.

[19] H. Bu, J. Hao, J. Guan, and S. Zhou, "Predicting enhancers from multiple cell lines and tissues across different developmental stages based on SVM method," *Current Bioinf.*, vol. 13, no. 6, pp. 655–660, Nov. 2018, doi: 10.2174/1574893613666180726163429.

[20] X. Du, X. Li, W. Li, Y. Yan, and Y. Zhang, "Identification and analysis of cancer diagnosis using probabilistic classification vector machines with feature selection," *Current Bioinf.*, vol. 13, no. 6, pp. 625–632, Nov. 2018, doi: 10.2174/1574893612666170405125637.

[21] Y. Wang, F. Shi, L. Cao, N. Dey, Q. Wu, A. S. Ashour, R. S. Sherratt, V. Rajinikanth, and L. Wu, "Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images," *Current Bioinf.*, vol. 14, no. 4, pp. 282–294, Apr. 2019, doi: 10.2174/1574893614666190304125221.

[22] M.-L. Liu, W. Su, Z.-X. Guan, D. Zhang, W. Chen, L. Liu, and H. Ding, "An overview on predicting protein subchloroplast localization by using machine learning methods," *Current Protein Peptide Sci.*, vol. 21, pp. 1–13, Jan. 2020, doi: 10.2174/1389203721666200117153412.

[23] S.-H. Li, J. Zhang, Y.-W. Zhao, F.-Y. Dao, H. Ding, W. Chen, and H. Tang, "IPhoPred: A predictor for identifying phosphorylation sites in human protein," *IEEE Access*, vol. 7, pp. 177517–177528, 2019.

[24] B. Wu, H. Zhang, L. Lin, H. Wang, Y. Gao, L. Zhao, Y.-P.-P. Chen, R. Chen, and L. Gu, "A similarity searching system for biological phenotype images using deep convolutional encoder-decoder architecture," *Current Bioinf.*, vol. 14, no. 7, pp. 628–639, Sep. 2019, doi: 10.2174/1574893614666190204150109.

[25] L. Yu, X. Sun, S. W. Tian, X. Y. Shi, and Y. L. Yan, "Drug and non-drug classification based on deep learning with various feature selection strategies," *Current Bioinf.*, vol. 13, no. 3, pp. 253–259, 2018, doi: 10.2174/1574893612666170125124538.

[26] 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.

[27] 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.

[28] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring MicroRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 4, pp. 905–915, Jul. 2017, doi: 10.1109/tcbb.2016.2550432.

[29] X. Zeng, L. Liu, L. Lù, and Q. Zou, "Prediction of potential disease-associated microRNAs using structural perturbation method," *Bioinformatics*, vol. 34, no. 14, pp. 2425–2432, Jul. 2018, doi: 10.1093/bioinformatics/bty112.

[30] F. Ren, C. Yang, Q. Qiu, N. Zeng, C. Cai, C. Hou, and Q. Zou, "Exploiting discriminative regions of brain slices based on 2D CNNs for Alzheimer's disease classification," *IEEE Access*, vol. 7, pp. 181423–181433, 2019.

[31] O. A. Brown, M. Canatelli-Mallat, G. M. Console, G. Camihort, G. Luna, E. Spinedi, and R. G. Goya, "IGF-1 gene therapy as a potentially useful therapy for spontaneous prolactinomas in senile rats," *Current Gene Therapy*, vol. 18, no. 4, pp. 240–245, Oct. 2018, doi: 10.2174/1566523218666180905170020.

[32] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: An empirical study," *Briefings Bioinf.*, vol. 21, no. 1, pp. 1–10, Sep. 2018, doi: 10.1093/bib/bby090.

[33] Q. Zou, P. Xing, L. Wei, and B. Liu, "Gene2vec: Gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA," *RNA*, vol. 25, no. 2, pp. 205–218, Feb. 2019.

[34] K. Liu and W. Chen, "IMRM: A platform for simultaneously identifying multiple kinds of RNA modifications," *Bioinformatics*, vol. 36, no. 11, pp. 3336–3342, Mar. 2020, doi: 10.1093/bioinformatics/btaa155.

[35] H. Lv, F.-Y. Dao, D. Zhang, Z.-X. Guan, H. Yang, W. Su, M.-L. Liu, H. Ding, W. Chen, and H. Lin, "IDNA-MS: An integrated computational tool for detecting DNA modification sites in multiple genomes," *iScience*, vol. 23, no. 4, Apr. 2020, Art. no. 100991, doi: 10.1016/j.isci.2020.100991.

[36] B. Liu and K. Li, "IPromoter-2L2.0: Identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features," *Mol. Therapy Nucleic Acids*, vol. 18, pp. 80–87, Dec. 2019.

[37] C. M. Abreu, R. Prakash, P. J. Romanienko, I. Roig, S. Keeney, and M. Jasin, "Shu complex SWS1-SWSAP1 promotes early steps in mouse meiotic recombination," *Nature Commun.*, vol. 9, no. 1, p. 3961, Oct. 2018, doi: 10.1038/s41467-018-06384-x.

[38] L. Wei, H. Chen, and R. Su, "M6APred-EL: A sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning," *Mol. Therapy Nucleic Acids*, vol. 12, pp. 635–644, Sep. 2018.

[39] 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.

[40] L. Wei, P. Xing, R. Su, G. Shi, Z. S. Ma, and Q. Zou, "CPPred-RF: A sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency," *J. Proteome Res.*, vol. 16, no. 5, pp. 2044–2053, May 2017, doi: 10.1021/acs.jproteome.7b00019.

[41] L. Wei, S. Wan, J. Guo, and K. K. Wong, "A novel hierarchical selective ensemble classifier with bioinformatics application," *Artif. Intell. Med.*, vol. 83, pp. 82–90, Nov. 2017.

[42] R. Su, H. Wu, B. Xu, X. Liu, and L. Wei, "Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 4, pp. 1231–1239, Jul. 2019.

[43] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, "Meta-4mCpred: A sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation," *Mol. Therapy Nucleic Acids*, vol. 16, pp. 733–744, Jun. 2019.

[44] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, "MAHTPred: A sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation," *Bioinformatics*, vol. 35, no. 16, pp. 2757–2765, Aug. 2019.

[45] W. Chen, P. Feng, and F. Nie, "IATP: A sequence based method for identifying anti-tubercular peptides," *Medicinal Chem.*, vol. 16, no. 5, pp. 620–625, Aug. 2020, doi: 10.2174/1573406415666191002152441.

[46] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "DUNet: A deformable network for retinal vessel segmentation," *Knowl.-Based Syst.*, vol. 178, pp. 149–162, Aug. 2019.

[47] B. Rao, C. Zhou, G. Zhang, R. Su, and L. Wei, "ACPred-fuse: Fusing multi-view information improves the prediction of anticancer peptides," *Briefings Bioinf.*, to be published, doi: 10.1093/bib/bbz088.

[48] R. Su, J. Hu, Q. Zou, B. Manavalan, and L. Wei, "Empirical comparison and analysis of Web-based cell-penetrating peptide prediction tools," *Briefings Bioinf.*, vol. 21, no. 2, pp. 408–420, Mar. 2020, doi: 10.1093/bib/bby124.

[49] R. Su, X. Liu, and L. Wei, "MinE-RFE: Determine the optimal subset from RFE by minimizing the subset-accuracy–defined energy," *Briefings Bioinf.*, vol. 21, no. 2, pp. 687–698, Mar. 2020.

[50] R. Su, X. Liu, L. Wei, and Q. Zou, "Deep-resp-forest: A deep forest model to predict anti-cancer drug response," *Methods*, vol. 166, pp. 91–102, Aug. 2019.

[51] R. Su, X. Liu, G. Xiao, and L. Wei, "Meta-GDBP: A high-level stacked regression model to improve anticancer drug response prediction," *Briefings Bioinf.*, vol. 21, no. 3, pp. 996–1005, May 2020, doi: 10.1093/bib/bbz022.

[52] R. Su, H. Wu, X. Liu, and L. Wei, "Predicting drug-induced hepatotoxicity based on biological feature maps and diverse classification strategies," *Briefings Bioinf.*, to be published, doi: 10.1093/bib/bbz165.

[53] Y.-J. Tang, Y.-H. Pang, and B. Liu, "IDP-Seq2Seq: Identification of intrinsically disordered regions based on sequence to sequence learning," *Bioinformatics*, to be published, doi: 10.1093/bioinformatics/btaa667.

[54] H. Yang, W. Yang, F.-Y. Dao, H. Lv, H. Ding, W. Chen, and H. Lin, "A comparison and assessment of computational method for identifying recombination hotspots in saccharomyces cerevisiae," *Briefings Bioinf.*, to be published, doi: 10.1093/bib/bbz123.

[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.*, vol. 21, no. 3, pp. 982–995, May 2020, doi: 10.1093/bib/bbz048.

[56] Z. Lv, C. Ao, and Q. Zou, "Protein function prediction: From traditional classifier to deep learning," *Proteomics*, vol. 19, no. 14, Jul. 2019, Art. no. 1900119, doi: 10.1002/pmic.201900119.

[57] L. Peng, M. Peng, B. Liao, G. Huang, W. Li, and D. Xie, "The advances and challenges of deep learning application in biological big data processing," *Current Bioinf.*, vol. 13, no. 4, pp. 352–359, Jul. 2018, doi: 10.2174/1574893612666170707095707.

[58] 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, doi: 10.1016/j.knosys.2018.10.007.

[59] X. Zeng, W. Wang, G. Deng, J. Bing, and Q. Zou, "Prediction of potential disease-associated MicroRNAs by using neural networks," *Mol. Therapy Nucleic Acids*, vol. 16, pp. 566–575, Jun. 2019, doi: 10.1016/j.omtn.2019.04.010.

[60] X. Liu, Z. Hong, J. Liu, Y. Lin, A. Rodríguez-Patón, Q. Zou, and X. Zeng, "Computational methods for identifying the critical nodes in biological networks," *Briefings Bioinf.*, vol. 21, no. 2, pp. 486–497, Mar. 2020, doi: 10.1093/bib/bbz011.

[61] H. Xu, W. Zeng, D. Zhang, and X. Zeng, "MOEA/HD: A multiobjective evolutionary algorithm based on hierarchical decomposition," *IEEE Trans. Cybern.*, vol. 49, no. 2, pp. 517–526, Feb. 2019, doi: 10.1109/TCYB.2017.2779450.

[62] T. Song, A. Rodriguez-Paton, P. Zheng, and X. Zeng, "Spiking neural p systems with colored spikes," *IEEE Trans. Cognit. Develop. Syst.*, vol. 10, no. 4, pp. 1106–1115, Dec. 2018.

BING RAO received the master's degree from the Harbin Institute of Technology, China. He is currently pursuing the Ph.D. degree with the School of Mechanical Electronic and Information Engineering, China University of Mining and Technology at Beijing. His research interests include bioinformatics and machine learning.

LICHAO ZHANG is currently a Lecturer with the School of Intelligent Manufacturing and Equipment, Shenzhen Institute of Information Technology. Her research interests include machine learning and bioinformatics.

GUOYING ZHANG received the Ph.D. degree from the Beijing Institute of Technology, China. She is currently a Professor with the School of Mechanical Electronic and Information Engineering, China University of Mining and Technology at Beijing. Her research interests include pattern recognition and artificial intelligence.