RMSCNN: A Random Multi-Scale Convolutional Neural Network for Marine Microbial Bacteriocins Identification

Zhen Cui, Zhan-Heng Chen, Qin-Hu Zhang, Valeriya Gribova, Vladimir Fedorovich Filaretov, and De-Shuang Huang

Abstract—The abuse of traditional antibiotics has led to an increase in the resistance of bacteria and viruses. Similar to the function of antibacterial peptides, bacteriocins are more common as a kind of peptides produced by bacteria that have bactericidal or bacterial effects. More importantly, the marine environment is one of the most abundant resources for extracting marine microbial bacteriocins (MMBs). Identifying bacteriocins from marine microorganisms is a common goal for the development of new drugs. Effective use of MMBs will greatly alleviate the current antibiotic abuse problem. In this work, deep learning is used to identify meaningful MMBs. We propose a random multi-scale convolutional neural network method. In the scale setting, we set a random model to update the scale value randomly. The scale selection method can reduce the contingency caused by artificial setting under certain conditions, thereby making the method more extensive. The results show that the classification performance of the proposed method is better than the state-of-the-art classification methods. In addition, some potential MMBs are predicted, and some different sequence analyses are performed on these candidates. It is worth mentioning that after sequence analysis, the HNH endonucleases of different marine bacteria are considered as potential bacteriocins.

Index Terms—Marine microbial bacteriocins, random, convolutional neural network

1 Introduction

In recent years, antimicrobial resistance has been increasing around the world, which has caused the traditional use of antibiotics to face huge challenges. More and more drugs lose their sensitivity to bacteria and viruses [1]. Such low sensitivity will weaken the resistance of humans, animals, plants and microorganisms. According to statistics, 23,000 patients die from infections caused by antibiotics in the United States every year [2]. Over the past few decades, a variety of researchers have focused on the functional research and clinical use of natural antimicrobial peptides (AMPs) [3], [4]. Compared with conventional drugs, pathogens are hardly resistant to AMPs, which is a great advantage [5].

Bacteriocin is a special kind of AMP, a type of polypeptide or precursor polypeptide with antibacterial activity produced by certain bacteria in the metabolic process through ribosome synthesis [6], [7], [8]. The earliest bacteriocins were discovered about 100 years ago. A bacteriocin named “colicin” was discovered during the wet experiments by researchers [9]. “Colicin” is a heatlabile substance in E.coli culture medium, which can effectively inhibit the activity of E.coli [10]. Therefore, this type of substance is called bacteriocin according to the name of the product. It is worth noting that closely related strains will show a narrow spectrum of activity inhibition [11]. With the development of human science and technology, researchers have gradually expanded their research fields from terrestrial environment to marine environment [12]. More and more marine metagenomic research and microbial protein resources have been obtained through wet experiments, including marine microbial bacteriocins (MMBs). At present, researchers are focusing on the identification of AMP produced by eukaryotes. In the research process, the type of organisms that produce AMP will not be considered. Therefore, this may make it impossible to reveal the general law of antibacterial substances produced by species after being stimulated [13]. Inspired by the research on AMP recognition, we will focus on the identification of MMBs. And the MMBs will

- *Zhen Cui is with the Institute of Machine Learning and Systems Biology, College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China, and also with the Project Management Office of China National Scientific Seafloor Observatory, Tongji University, Shanghai 200092, China. E-mail: cuizhen_tj@163.com.*
- *Zhan-Heng Chen is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China. E-mail: chenzhanheng17@mails.ucas.ac.cn.*
- *Qin-Hu Zhang and De-Shuang Huang are with the Institute of Machine Learning and Systems Biology, College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China. E-mail: 472765713@qq.com, dshuang@tongji.edu.cn.*
- *Valeriya Gribova and Vladimir Fedorovich Filaretov are with the Institute of Automation and Control Processes, Far Eastern Branch of Russian Academy of Sciences, 119991 Moscow, Russia. E-mail: gribova@iacp.dvo.ru, filaretov@inbox.ru.*

Manuscript received 13 March 2021; revised 13 September 2021; accepted 19 October 2021. Date of publication 26 October 2021; date of current version 8 December 2022.
This work was supported in part by the National Key R&D Program of China under Grants 2018AAA0100100 and 2018YFA0902600, in part by the National Natural Science Foundation of China under Grants 61861146002, 61732012, 61772370, 62002266, 61932008, 61772357, and 62073231, in part by the “BAGUI Scholar” Program and the Scientific & Technological Base and Talent Special Program, in part by the GuiKe AD18126015 of the Guangxi Zhuang Autonomous Region of China, in part by China National Scientific Seafloor Observatory.
(Corresponding author: De-Shuang Huang.)
Digital Object Identifier no. 10.1109/TCBB.2021.3122183

inspire the development of new drugs. In immunology, the identification of MMBs can also explore the immune mechanism and immune diversity of marine microorganisms. In addition, the identification of MMBs is of great significance for the ethnic continuation and genetic evolution of marine microorganisms.

Several DNN (Deep Neural Network) methods have been applied to the recognition of AMPs and achieved favorable results. For instance, Veltri *et al.* proposed a neural network method based on CNN (Convolutional Neural Network) and LSTM (Long and Short-Term Memory) to identify AMPs [14]. CNN is first used to learn the features of the sequence, and then the sequence features are input to the LSTM layer to complete the final classification. Hamid *et al.* proposed a word embedding with deep recurrent neural networks to identify AMPs [2]. Unlike the original encoding, word embedding is used to encode protein sequences [15], [16], [17]. Yan *et al.* proposed a deep learning model Deep-AmPEP30 for short sequence AMP recognition [18]. This model is designed to process sequences less than or equal to 30 amino acids in length and has achieved splendid classification results. Several bacteriocin databases and computational models have also been developed by researchers to help identify bacteriocins. In 2010, Hammami *et al.* developed an integrated open database called BACTIBASE to characterize bacterial antimicrobial peptides called bacteriocins [19]. The database mainly collects microbial information through PubMed search based on literature retrieval. In addition, BACTIBASE also integrates some mainstream protein analysis tools. In 2013, van Heel *et al.* developed a search tool BAGEL that can effectively query bacteria based on homology information [20]. One of the important functions of this search tool is genome mining through bacteriocins. In 2014, Mohimani *et al.* developed RiPPquest, a bacteriocin database search tool based on tandem mass spectrometry [21]. In 2015, Weber *et al.* developed a platform called AntiSMASH to realize the genome mining of metabolites [22]. Another function is to realize the mining of bacteriocins. Morton *et al.* designed a software tool called BOA (Bacteriocin Operon Associator), and used it to discover potential bacteriocins through homologous compounds of background genes [23]. Most recently, Mohimani *et al.* improved the automated mass spectrometry of RiPPquest and proposed a software platform MetaRiPPquest based on high-resolution mass spectrometry to identify Ribosomally synthesized and posttranslationally modified peptides (RiPPs) of peptide genomes [24]. However, even if there are several effective software platforms and computational methods to identify bacteriocins, with the development of marine meta-genomic sequencing technology, increasing number of marine microbial protein sequences have been discovered [25]. To discover drug candidates that can replace traditional antibiotics from the huge marine microbial resources, an efficient and reliable calculation method needs to be urgently developed to identify MMBs.

In this work, to better obtain MMBs from marine resources, we collect MMBs data from NCBI database and propose a novel deep learning method to learn the features of protein sequences. Inspired by the multi-scale convolutional neural network, a random model is introduced into the neural network to randomly update the convolution kernel scale [26]. The proposed method inherits the advantages of CNN and has achieved outstanding results in the MMBs classification experiment. And the experimental results reveal that compared with other current classification techniques, the proposed random multi-scale convolutional neural network (RMSCNN) can better classify MMBs.

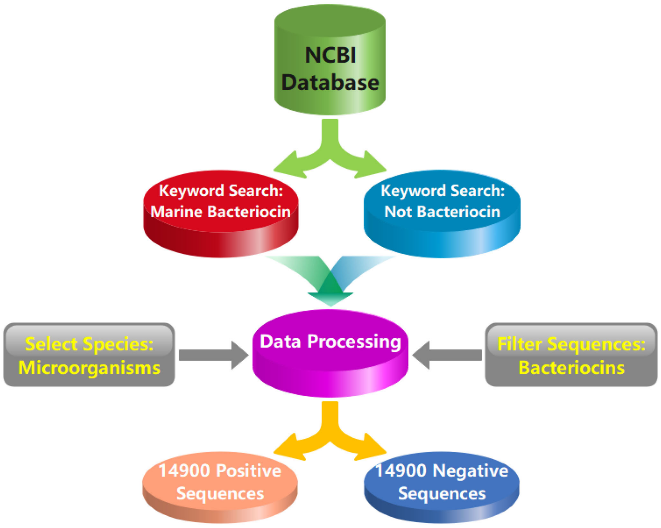
NCBI Database
Keyword Search: Marine Bacteriocin
Keyword Search: Not Bacteriocin
Select Species: Microorganisms
Data Processing
Filter Sequences: Bacteriocins
14900 Positive Sequences
14900 Negative Sequences

Fig. 1. We query the bacteriocins of marine microorganisms from the NCBI database through keyword search, and then further filter to obtain the final dataset.

2 Materials and Methods

2.1 Datasets

In this section, two types of datasets are introduced. One is the AMPs dataset available in the APD vr3 database, which has been used by many researchers [27]. In other words, this is a public and convincing dataset. The second is the MMBs dataset we obtained from the NCBI database through keyword search. Fig. 1 shows the data collection process. These data are essentially derived from a variety of databases, which are finally collected by NCBI database and allowed to be accessed.

For AMPs data, its biological function is mainly reflected in its ability to produce activity against Gram-positive bacteria or Gram-negative bacteria [14]. According to previous studies, sequences less than 10 amino acids in length are filtered out [2], [26], [28]. After the above series of processing, 1778 AMPs can be finally obtained. Considering the division of training set and test set in deep learning, 1778 AMPs are divided into three parts: training set, tuning/evaluation set and test set. Among them, the training set contains 712 AMPs, the tuning/evaluation set contains 354 AMPs, and the test set contains 712 AMPs. Next, negative data as non-AMPs are collected from the UniProt database [29]. Then, sequences less than 10 amino acids in length are still discarded. Finally, to ensure the balance of data category, 1778 non-AMPs are randomly acquired as a negative dataset. Same as the division rule of AMPs sequence, negative data is also divided into 712 training sequences, 354 tuning/evaluation sequences and 712 test sequences.

For the MMBs data, the keyword search “Marine Bacteriocin” is used in the NCBI database and filtered out

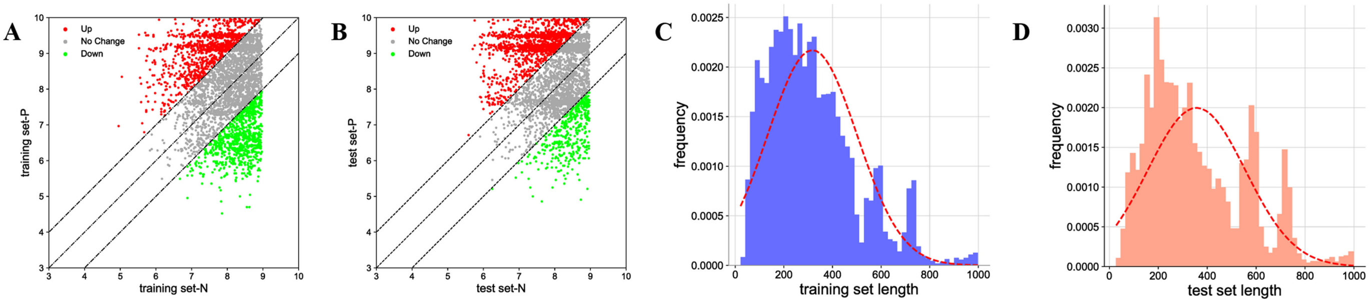

Fig. 2. A. The relationship between the length of the training set positive sequences (training set-P) and the training set negative sequences (training set-N). B. The relationship between the length of the test set positive sequences (test set-P) and the test set negative sequences (test set-N). C. The length of the training set is distributed in frequency, and the sequence with a length between 200 and 400 accounts for the largest proportion. D. The length of the test set is distributed in frequency, and the sequence with a length between 200 and 400 accounts for the largest proportion.

other organisms except marine microorganisms. This step can ensure that the research object is marine microorganisms. Although most bacteriocins are short peptides, excluding longer peptides does not contain meaningful information. Therefore, these sequences with a length of 5 to 1000 amino acids can basically cover most of the feature information of bacteriocins. Finally, 14900 MMBs are collected as positive data. Of course, there are indeed sequences with similar amino acid arrangement in these sequences. For example, there are multiple bacteriocin sequences in a bacterium, but these sequences may differ only by one amino acid. The main reason for this phenomenon is that the MMBs data collected from NCBI database come from different databases. Non-bacteriocin sequences are collected as negative data from the RefSeq database [30]. A total of 14900 sequences ranging in length from 5 to 500 amino acids were obtained by filtering keywords such as antimicrobial peptides, antiviral, antibiotics, bacteriocins, and antifungals. The sequence length distribution of all data is shown in Supplementary Fig. 1.

2.2 Non-redundant MMBs Dataset

Considering that there are many similar sequences in the dataset, the CD-Hit program is used to deal with redundant sequences in the dataset. CD-Hit clusters these sequences by comparing the degree of repetition between them. In this paper, we set the comparison interval of each sequence to account for 80% of the representative sequence. Finally, 8000 MMBs and 8000 non-MMbs were obtained. Specifically, the training set and the test set contain 4000 MMBs and 4000 non-MMBs, respectively. As shown in Figs. 2A and 2B, we construct a scatter plot with the length value, and describe the multiple change line according to the multiple change.

In addition, as shown in Figs. 2C and 2D, the sequence lengths of the training set and the test set are close to the normal distribution. It can be seen from the figure that the sequence length of 500 accounts for the majority. Fig. 3 shows the intersection of the length of the positive and negative sequences in the training set and the test set.

2.3 Evolutionary Tree of AMPs and MMBs

To better show the relationship between AMPs and non-AMPs, MMBs and non-MMBs, the evolutionary tree will be constructed through amino acid sequences. The evolutionary tree calculates the evolutionary relationship betwee organisms through mathematical statistical algorithms [31], [32]. Generally speaking, if there is a common ancestor between species, then there is a root node in the evolutionary tree, otherwise there is no root node. Considering that too many amino acid sequences are not conducive to constructing the evolutionary tree, and the genetic relationship will not be clear. Therefore, we randomly select the same number of amino acid sequences from the positive and negative sequences in the AMPs and non-redundant MMBs data sets to construct the evolutionary tree. Fig. 4 shows the evolutionary tree of AMPs and non-redundant MMBs.

2.4 Architecture of Proposed RMSCNN

Deep learning methods have achieved great success in bioinformatics [33], [34]. One of the important reasons for using deep learning is because it can identify potentially complex patterns that can represent different MMBs from a large amount of irregular sequence data [35], [36]. We design the proposed DNN method based on the currently popular deep learning frameworks Keras (http://www.keras.io) and Tensorflow deep learning library. Fig. 5 shows the framework of our proposed method. In detail, the convolutional layer and the maximum pooling layer are the main structures of DNN [37], [38]. The main function of these two layers is to extract the feature pattern of the sequence and achieve the effect of dimensionality reduction on the extracted sequence features [39]. The convolutional layer used is a multi-scale convolutional layer. The advantage of the multi-scale convolutional layer is that it can capture sequence feature patterns at different scales [40]. On the one hand, the global information of these training sequences can be captured, and on the other hand, the local information of these training sequences can also be captured.

According to previous research, the general multi-scale convolutional layer adopts a certain regular and fixed scale value when designing the scale. At present, this type of multi-scale has a wide range of applications in image segmentation, target detection and other fields [33], [41], [42], [43]. However, protein sequence data is complex and changeable, and the distribution of each amino acid is almost irregular, so the traditional multi-scale cannot mine its potential features more accurately. As shown in Supplementary Fig. 2, available

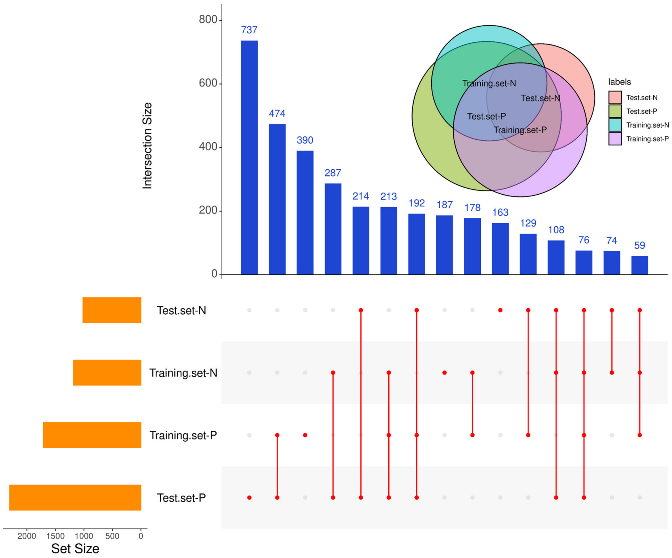

Fig. 3. The intersection of the length of the positive and negative sequences in the training set and the test set. Points connected by lines represent the intersection of multiple sets. The bar graph on the bottom left represents the number of each set; the bottom right is the intersection of each set, where the red dots indicate yes, and the gray dots indicate none.

online, after the AMP sequences are encoded, they show a complex distribution. Regular and fixed scales are used, which may lead to high contingency of the method and is not conducive to the feature learning of AMPs. Specifically, a step set with an initial value of 0 is created, and the maximum step size is limited to 100. Then, a random model randomly selects a step from the set as the length of the sliding window each time, and ends the loop when the selected step reaches 100. It is worth noting that the random model is a positive growth, that is, the next step value is greater than the previous step. Through this mechanism, the regular selection of scale values is transformed into random behavior, which greatly reduces the contingency of the method [44]. Fig. 5 shows the sequence processing flow and the structure of the proposed method from a global perspective. The source code and data are available at: https://github.com/cuizhensdws.

As shown in Fig. 5, sequences of different lengths are uniformly encoded into numerical vectors of fixed size, and the length is set to 500. If the length of some sequences is less than 500, then 0 elements will be filled in. The 20 amino acids that make up a protein will be replaced by numbers from 1 to 20 in order [14]. If a character other than 20 amino acids appears in the sequence, such as ‘X’, the character is represented by the number 0. As shown in Fig. 5, an embedding layer is used to receive the input encoded sequences. Then, these encoded sequences continue to be fed into a 1D Conv layer. The role of the embedding layer is to convert the index similar to the amino acid discrete symbol into a fixed-size vector [45]. It is worth mentioning that the advantage of the embedding layer is that the semantically similar symbols in the vector space can be generated, and the symbols can become closer and more logically related [23]. Of course, during the training process, the weights of the embedding layer will also be constantly updated.

After the sequence flows from the embedding layer, a multi-scale convolutional network is used, which contains different convolution kernel lengths. Next, the random model is embedded to generate a set of sliding window lengths. Each time a convolution operation is performed, the length of the sliding window will be randomly selected from this set. Then, the values passed from the filter are downsampled by a max pooling layer with a pooling size of 500. The overfitting of the DNN method can be reduced by this step. Then a max pooling layer continues to be introduced, where the pooling size is 1. Finally, the output from the max pooling layer is fed to a fully connected layer to perform the final classification. In the final output layer, the sigmoid function is used to control the range of predicted values between $[0, 1]$. The predicted value of 0.5 is considered as the classification limit, a value greater than 0.5 can be considered as a marine microbial bacteriocin, otherwise it is not. To measure the difference between the predicted value and the actual value output by the neural network, the binary cross entropy loss function is used to represent the starting point of back propagation [46]. The binary cross entropy loss function is as follows:

$$Loss = -\frac{1}{S}\sum_{i=1}^{S} y_i \times \log\left(p(y_i)\right) + (1 - y_i) \times \log\left(1 - p(y_i)\right) \tag{1}$$

where y is the label, $p(y)$ is a probability to indicate that S samples are all predicted to be positive.

2.5 Experimental Setup and Runtime Performance

In this work, experiments are conducted on an AMD 4800H laptop with an eight core 2.9Ghz processor and 16GB of RAM. Specifically, the operating system is Windows 10, IDE is PyCharm Community, GPU is NVIDIA GeForce GTX1650, Cuda version is 10.1, CUDNN version is 7.6, Python version is 3.7, and Tensorflow version is 2.2. During the experiments, the time required for each epoch is less than three minutes. However, if GPU acceleration is not used, the time for each epoch will be far more than three minutes, depending on the configuration of the computer used in the experiments.

3 Results

3.1 Model Evaluation

To evaluate the performance of the DNN method, multiple commonly used evaluation indicators are used to measure the classification effect of the proposed method. We use accuracy (ACC), recall, F1-score, precision and the area under the curve (AUC) to evaluate classification per-formance. The calculations of the above indicators are based on the number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The value range of the AUC is $[0.5, 1]$. If the value is equal to 0.5, this can be considered a random guess. If the value is equal to 1, this can be considered as the case where all predicted categories are correct [47]. It is worth noting that if the value is less than 0.5, then the classifier can be considered invalid. In this paper, the evaluation indicators used are all derived from the ‘sklearn metrics’ module. The formulas for these

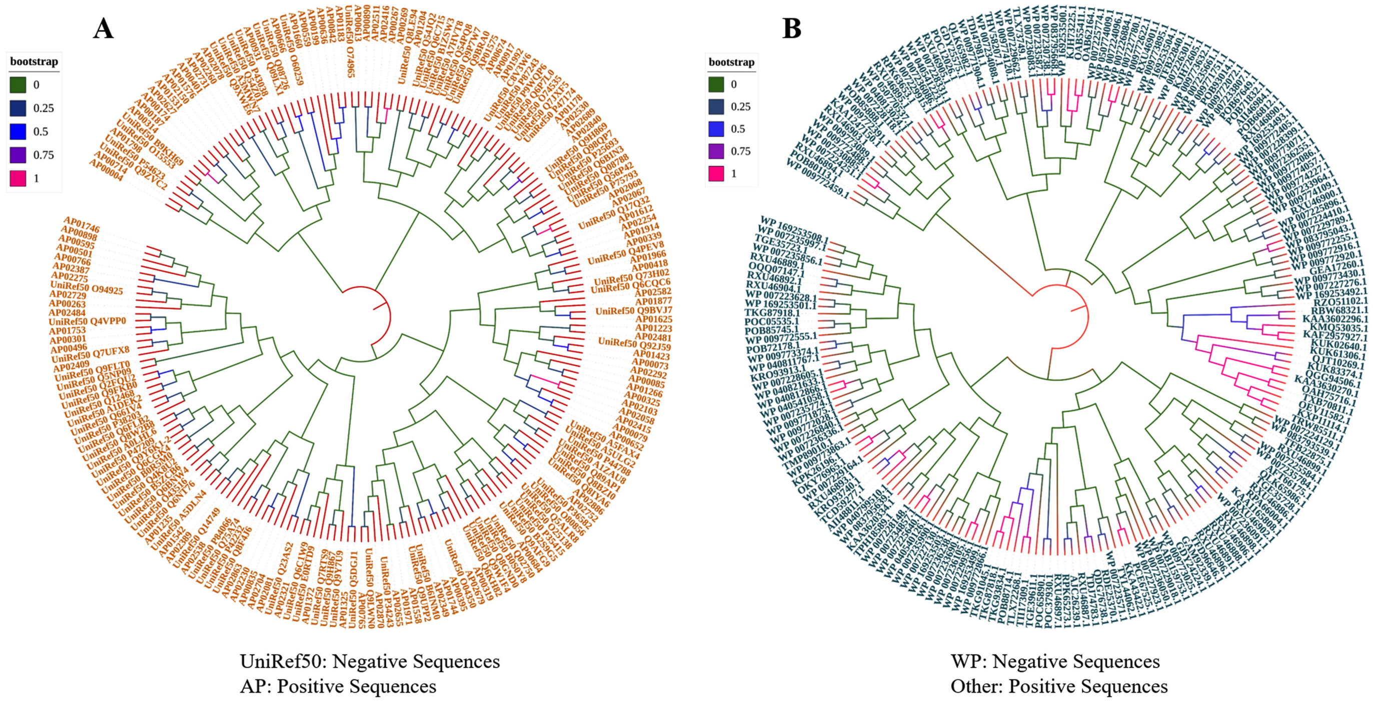

Fig. 4. 100 AMP and non-AMP sequences were randomly selected to construct an evolutionary tree. 100 MMB and non-MMB sequences were randomly selected to construct an evolutionary tree. A. The evolutionary tree between partial AMP sequences and non-AMP sequences. B. The evolutionary tree between partial MMB sequences and non-MMB sequences.

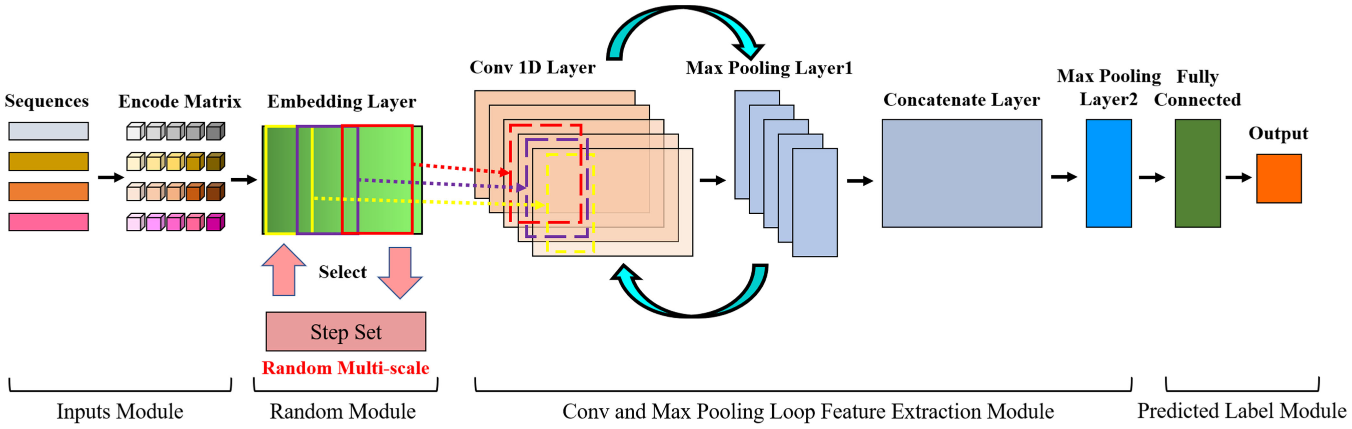

Fig. 5. The framework of our proposed method RMSCNN. We use a random model, the scale of each convolution kernel is randomly generated by the model, and finally these random convolution kernel scales are randomly selected. Each time of convolution, the max pooling is performed once, and the operation is performed in a loop until the random ends.

indicators are as follows:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{4}$$

$$F1{-}Measure = \left(1 + \beta^2\right) \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall} \times 100\% \tag{5}$$

$$AUC = \frac{\sum_{ins_i \in positiveclass} rank_{ins_i} - \frac{M \times (M+1)}{2}}{M \times N} \times 100\% \tag{6}$$

where β is a hyperparameter, it is generally set to 1. M and N represent the number of positive samples and negative samples, respectively, $rank_{ins_i}$ represents the serial number of the i-th sample.

3.2 Comparing With Baseline Methods

In this paper, we mainly compare two types of methods, one is machine learning and the other is neural network. Machine learning methods include Decision Tree, Random Forest, Extra Trees, SVM and AdaBoost [48], [49]. Neural network methods mainly include CNN, DNN, multi-scale CNN and the proposed method. These methods are currently popular classifiers and have been widely used in many fields. Decision trees, random forests and extra trees can be considered as tree-like methods. Although SVM is an effective classifier, its performance is often related to the setting of the kernel function. Commonly used kernel functions include Linear kernel function, Poly kernel function,

TABLE 1
Performance of Different Methods on the AMPs Test Dataset

Methods	ACC	Recall	F1-Measure	Precision	AUC
Decision Tree	0.6654	0.6400	0.6450	0.6450	0.6416
Random Forest	0.6617	0.6450	0.6450	0.6500	0.6451
Extra Trees	0.6736	0.6350	0.6350	0.6350	0.6338
SVM	0.6437	0.6450	0.6350	0.6650	0.7507
AdaBoost	0.6564	0.6550	0.6600	0.6550	0.7513
CNN	0.9003	0.8876	0.8990	0.9107	0.9553
DNN (Conv and LSTM)	0.8906	**0.9179**	0.8949	0.8642	0.9499
Multi-Scale CNN	0.9010	0.9031	0.9012	0.8993	0.9606
RMSCNN	**0.9122**	0.8876	**0.9100**	**0.9335**	**0.9661**

TABLE 2
Performance of Different Methods on the MMBs Test Dataset

Methods	ACC	Recall	F1-Measure	Precision	AUC
Decision Tree	0.6595	0.6595	0.6595	0.6594	0.6594
Random Forest	0.6771	0.6769	0.6756	0.6801	0.6770
Extra Trees	0.6853	0.6855	0.6834	0.6894	0.6851
SVM	0.6498	0.6514	0.6068	0.7695	0.7522
AdaBoost	0.6625	0.6625	0.6623	0.6630	0.7594
CNN	0.9041	0.8911	0.9028	0.9148	0.9582
DNN (Conv and LSTM)	0.8996	0.8581	0.8907	0.9356	0.9593
Multi-Scale CNN	0.9141	**0.9322**	0.9156	0.8996	0.9704
RMSCNN	**0.9304**	0.9189	**0.9434**	**0.9663**	**0.9774**

RBF kernel function, Sigmoid kernel function and Precomputed kernel function [50], [51]. In this paper, the Poly kernel function is used to achieve the best experimental results. The AdaBoost method is a boosting algorithm that combines multiple weak learning algorithms [52]. First, the method tries to find a weak classifier, and then multiple weak classifiers can be obtained through repeated learning, and finally all the weak classifiers are integrated together to obtain the final strong classifier. In this work, the number of weak classifiers is set to 100 to achieve the best results.

CNN is one of the most successful methods used in deep learning [53]. In this paper, a simple CNN method with only one convolutional layer is used, and the default parameters are set in the method. In the DNN method that combines the CNN and LSTM layers, it contains a convolutional layer, a max pooling layer and an LSTM layer. The LSTM layer receives the sequence patterns sent from the CNN layer and filters the undesired information by setting a forget gate, and finally controls the prediction result range to $[0, 1]$ through a sigmoid activation function [54], [55].

3.3 Model Performance

Table 1 shows the classification results of different methods on the AMP dataset. As can be seen from Table 1, the values of all neural network methods on ACC and AUC have reached more than 90%, and the values on other indicators can also reach more than 85%. This can indicate that when applying neural network methods to process AMP data, the overall performance is better than traditional machine learning methods. Of course, due to the introduction of the random model, it reduces the contingency of scale selection, so the proposed method achieves the best in all indicators except Recall. Compared with multi-scale CNN, the proposed method improves ACC and AUC values by 1.12% and 0.55%, respectively. However, for the Recall value, the DNN method reached the highest value, followed by the multi-scale CNN.

Table 2 shows the classification results of different methods on the MMB dataset. It can be seen from Table 2 that the performance of traditional machine learning methods on the MMB dataset is close to that of the AMP dataset. In addition, the values of all neural network methods on ACC and AUC have reached more than 90%, and the values on other indicators can also reach more than 85%. Compared with multi-scale CNN, the proposed method improves the ACC value by 1.63%. Surprisingly, for the Recall value, multi-scale CNN reached the highest value, followed by RMSCNN, which increased by 1.33%. Table 3 shows the classification results of neural network methods on the non-redundant MMBs test dataset. Compared with these methods, RMSCNN still achieved high prediction values. Fig. 6 shows the comparison of the top 200 predicted label values between our method and multi-scale CNN.

It is worth noting that the performance of multiple machine learning methods on these two datasets is not satisfactory. They are lower than the neural network method in all evaluation indicators. Therefore, to explore the specific situation of classification, taking the SVM method as an example, we visualize its classification results on two datasets. As shown in Supplementary Fig. 3, available online, the blue points and the red points represent positive data and negative data, respectively, and the data points with black outlines represent test points. The green and yellow areas indicate different division interfaces. It can be seen from Supplementary Fig. 3, available online, that the SVM method can hardly find an effective dividing line to distinguish between positive and negative data. In addition, the features of positive and negative data are particularly similar may be one of the important reasons for poor classification performance.

TABLE 3
Performance of Neural Network Methods on the Non-redundant MMBs Test Dataset

Methods	ACC	Recall	F1-Measure	Precision	AUC
CNN	0.8965	0.9032	0.9064	0.9023	0.9476
DNN (Conv and LSTM)	0.9063	0.9133	0.9160	0.9188	0.9667
Multi-Scale CNN	0.9009	**0.9225**	0.9031	0.8843	0.9610
RMSCNN	**0.9195**	0.9197	**0.9195**	**0.9193**	**0.9788**

3.4 Prediction of Potential MMBs

Table 4 shows the MMBs prediction results. The top ten sequences with the largest predicted probability values are listed in the table. We retrieve the details of each sequence through NCBI. Among them, four sequences did not find detailed information. Interestingly, these retrieved sequences are all HNH endonucleases, which come from different microorganisms, such as Staphylococcus carnosus, Ligilactobacillus agilis, Phocaeicola, Fictibacillus and Bacillus [56]. It

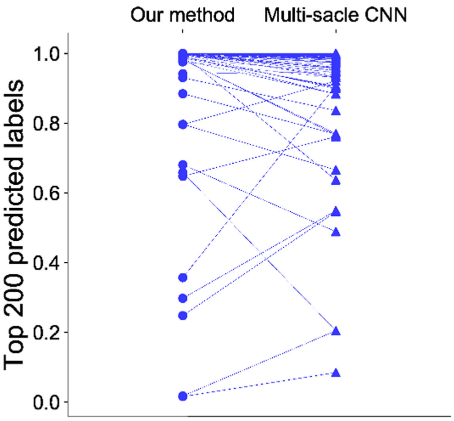

Fig. 6. In our method and multi-scale CNN, the top 200 sequence prediction labels (the original labels of the first 200 sequences are all 1) are selected for comparison. Overall, the prediction value of our method is higher than that of multi-scale CNN.

can be seen from Table 4 that the proposed method is determined after prediction that these sequences are highly consistent with the characteristics of MMB, which means that HNH endonuclease may have a function similar to antibiotics. It is worth noting that endonuclease is a hydrolase, which can be divided into DNase I and DNase II enzymes that break down DNA [57]. The main function of endonuclease is to protect the host (bacteria) from bacteriophages [58], [59]. Its mechanism of action is to cut the DNA of the virus at a specific site and cut the gene into small gene fragments, thereby achieving the purpose of protecting the host and decomposing the virus. Fig. 7 shows the secondary structure of HNH endonuclease. All in all, the function of bacteriocins is to inhibit or kill bacteria, while the function of HNH endonuclease is to protect the host from phage by cutting the viral genes. They have similar immune mechanisms.

To explore the biological connection of these sequences, a sequence comparison tree is constructed through CLUSTALW (https://www.genome.jp/tools-bin/clustalw) [60]. As shown in Fig. 8, the graph contains six nodes, and the distance of each node indicates the distance of the kinship.

TABLE 4
We List the Top Ten Peptides Predicted From the MMBs Data Based on the Predicted Probability. Except for the Details of the Four Sequences, the Remaining Six are HNH Endonucleases

Rank	Name	Predictive value	Details
1	>WP_007225938.1	1	Not found
2	>WP_007224760.1	1	Not found
3	**>WP_015901688.1**	**1**	**HNH endonuclease**
4	**>WP_191997391.1**	**1**	**HNH endonuclase]**
5	**>WP_191996029.1**	**1**	**HNH endonuclease**
6	**>WP_191764297.1**	**1**	**HNH endonuclease**
7	**>WP_191755117.1**	**1**	**HNH endonuclease**
8	>WP_007225148.1	0.9999999	Not found
9	>WP_007226964.1	0.99999976	Not found
10	**>WP_191779918.1**	**0.99999964**	**HNH endonuclease**

The higher the score on the same branch, the closer the relationship between the sequences. Considering that there are four sequences without finding the details, they are no longer considered. The function of these six sequences is similar, so there may be some similar conservative sequences in these sequences. A sequence similarity comparison visualization map created by ESPript is shown in Supplementary Fig. 4, available online. The yellow window indicates that the sequence is relatively similar, and the red window indicates completely conserved sequence [61]. It can be seen from Supplementary Fig. 4, available online, that there are multiple sets of similar sequences in these endonucleases, as well as a set of completely conservative sequences. And the sequences of WP_191997391.1 and WP_191996029.1 are highly correlated. Both sequences are from Ligilactobacillus agilis [62]. It means that the physical and chemical properties of the two sequences are similar. The secondary structure of a protein is an important factor in determining its function, mainly including alpha helix and beta helix, and the transmembrane tendency of protein also affects the structure of the protein [63]. Supplementary Figs. 5, 6, and 7, available online, how their amino acid profiles. These pictures can be obtained in Supplementary data. It can be seen

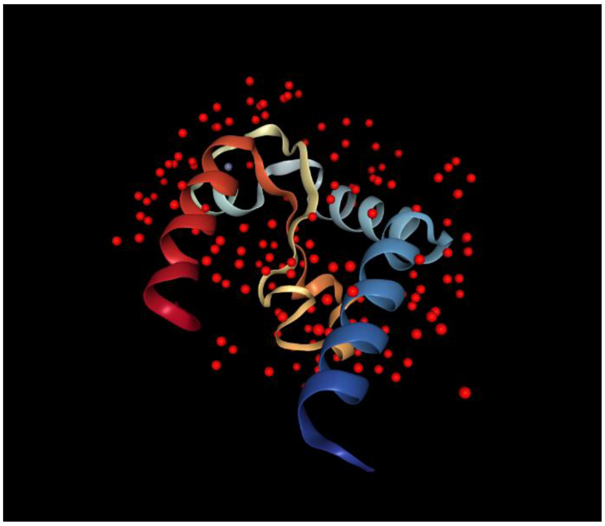

Fig. 7. The secondary structure of HNH endonuclease is obtained by using the X-ray diffraction method. The 3D model is the deep-sea thermophilic phage GVE2 HNH endonuclease. The surrounding points are water molecules and hydrogen ions.

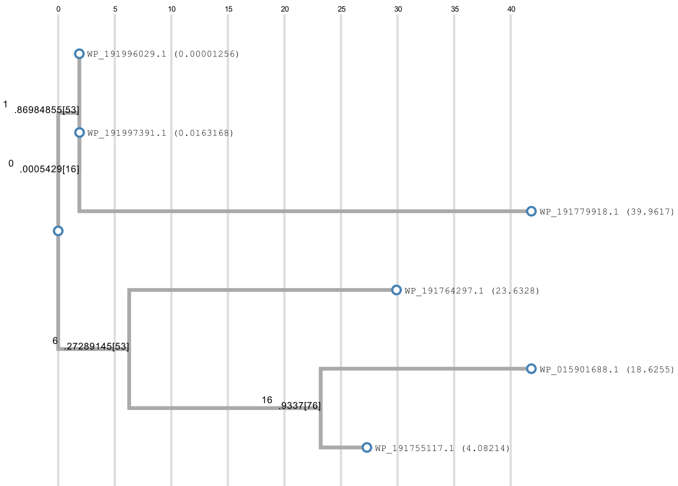

Fig. 8. The secondary structure of HNH endonuclease is obtained by using the X-ray diffraction method. The 3D model is the deep-sea thermophilic phage GVE2 HNH endonuclease. The surrounding points are water molecules and hydrogen ions.

that the structure and transmembrane trend of the two sequences are almost the same, so it can be proved that their functions are similar.

4 Discussions

During the analysis, some interesting results are discovered. First of all, most of the predicted proteins are HNH endonucleases. The function of endonucleases is to protect the host from phage [64]. And MMBs kill the bacteria that are homologous to it to achieve the purpose of immunity. In other words, endonucleases and MMBs have different immune mechanisms but similar immune effects. Second, in multiple analysis experiments, we notice that two endonucleases are homologous, so we can assume that they have the same immune mechanism. Considering that the structure of the protein determines its function, and the transmembrane trend determines its structural characteristics, to verify our hypothesis, the protein secondary structure profile and transmembrane trend profile of the two sequences are constructed separately. In the end, the protein profiles of the two sequences are almost identical. The above findings will help wet-laboratory researchers to explore new immune molecules from endonucleases and provide new perspectives for the development of new drug candidates.

5 Conclusion

In this paper, a novel CNN classifier is proposed to identify marine microbial bacteriocins. Compared with other advanced methods, the proposed method achieves the best performance on multiple evaluation indicators. Not only that, to the best of our knowledge, this is the first time that a neural network method has been used to classify bacteriocins related to marine microorganisms. Considering that biomolecular sequences are often disorderly arranged in their primary structure, the vectors constructed according to their sequences are usually disordered, and random multi-scale selection is beneficial to capture the potential features of these sequences. In addition, due to the embedding of the random model, the randomized selection of the convolution kernel size also increases the generalization ability of the method.

The richness of marine resources is far greater than that of land. There are abundant types and numbers of organisms in the ocean, and future resources will come from the ocean. In the future, our research will include not only AMP and MMB, but also other therapeutic peptides, such as anti-cancer peptides, anti-inflammatory peptides, surface-binding peptides, and cell penetrating peptides. The identification of these therapeutic peptides has great medical significance.

References

[1] M. Magana *et al.*, "The value of antimicrobial peptides in the age of resistance," *Lancet Infect. Dis.*, vol. 20, no. 9, pp. e216–e230, 2020.

[2] M.-N. Hamid and I. Friedberg, "Identifying antimicrobial peptides using word embedding with deep recurrent neural networks," *Bioinformatics*, vol. 35, no. 12, pp. 2009–2016, 2018.

[3] H. G. Boman, "Antibacterial peptides: Basic facts and emerging concepts," *J. Intern. Med.*, vol. 254, no. 3, pp. 197–215, 2003.

[4] T. Ganz, "Defensins: Antimicrobial peptides of innate immunity," *Nat. Rev. Immunol.*, vol. 3, no. 9, pp. 710–720, 2003.

[5] E. Guaní-Guerra, T. Santos-Mendoza, S. O. Lugo-Reyes, and L. M. Terán, "Antimicrobial peptides: General overview and clinical implications in human health and disease," *Clin. Immunol.*, vol. 135, no. 1, pp. 1–11, 2010.

[6] A. Guder, I. Wiedemann, and H.-G. Sahl, "Posttranslationally modified bacteriocins—The lantibiotics," *Peptide Sci*, vol. 55, no. 1, pp. 62–73, 2000.

[7] J. M. Willey and W. A. van der Donk, "Lantibiotics: Peptides of diverse structure and function," *Annu. Rev. Microbiol.*, vol. 61, no. 1, pp. 477–501, 2007.

[8] A. Correia and A. Weimann, "Protein antibiotics: Mind your language," *Nat. Rev. Microbiol.*, vol. 19, no. 1, 2021, Art. no. 7.

[9] F. Desriac, D. Defer, N. Bourgougnon, B. Brillet, P. Le Chevalier, and Y. Fleury, "Bacteriocin as weapons in the marine animal-associated bacteria warfare: Inventory and potential applications as an aquaculture probiotic," *Mar. Drugs*, vol. 8, no. 4, pp. 1153–1177, 2010.

[10] H. M. Behrens, A. Six, D. Walker, and C. Kleanthous, "The therapeutic potential of bacteriocins as protein antibiotics," *Emerg. Top. Life Sci.*, vol. 1, no. 1, pp. 65–74, 2017.

[11] A. Radaic, M. B. de Jesus, and Y. L. Kapila, "Bacterial anti-microbial peptides and nano-sized drug delivery systems: The state of the art toward improved bacteriocins," *J. Controlled Release*, vol. 321, pp. 100–118, 2020.

[12] C. Ao, Y. Zhang, D. Li, Y. Zhao, and Q. Zhao, "Progress in the development of antimicrobial peptide prediction tools," *Curr. Protein Peptide Sci.*, vol. 22, no. 3, pp. 211–216, 2021.

[13] C. Wang, J. Wu, L. Xu, and Q. Zou, "NonClasGP-Pred: Robust and efficient prediction of non-classically secreted proteins by integrating subset-specific optimal models of imbalanced data," *Microbial Genomic*, vol. 6, no. 12, 2020, Art. no. mgen000483.

[14] D. Veltri, U. Kamath, and A. Shehu, "Deep learning improves antimicrobial peptide recognition," *Bioinformatics*, vol. 34, no. 16, pp. 2740–2747, 2018.

[15] L. Wang, Z. You, D. Huang, and F. Zhou, "Combining high speed ELM learning with a deep convolutional neural network feature encoding for predicting Protein-RNA interactions," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 3, pp. 972–980, May/Jun. 2020.

[16] L. Zhu, S. Deng, Z. You, and D.-Shuang Huang, "Identifying spurious interactions in the protein-protein interaction networks using local similarity preserving embedding," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 2, pp. 345–352, 2017.

[17] H.-C. Yi *et al.*, "A deep learning framework for robust and accurate prediction of ncRNA-Protein interactions using evolutionary information," *Mol. Ther. - Nucleic Acids*, vol. 11, pp. 337–344, 2018.

[18] J. Yan *et al.*, "Deep-AmPEP30: Improve short antimicrobial peptides prediction with deep learning," *Mol. Ther. Nucleic Acids*, vol. 20, pp. 882–894, 2020.

[19] R. Hammami, A. Zouhir, C. Le Lay, J. B. Hamida, and I. Fliss, "BACTIBASE second release: A database and tool platform for bacteriocin characterization," *BMC Microbiol*, vol. 10, no. 1, 2010, Art. no. 22.

[20] A. J. van Heel, A. de Jong, M. Montalbán-López, J. Kok, and O. P. Kuipers, "BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides," *Nucleic Acids Res.*, vol. 41, no. W1, pp. W448–W453, 2013.

[21] H. Mohimani *et al.*, "Automated genome mining of ribosomal peptide natural products," *ACS Chem. Biol.*, vol. 9, no. 7, pp. 1545–1551, 2014.

[22] T. Weber *et al.*, "antiSMASH 3.0—A comprehensive resource for the genome mining of biosynthetic gene clusters," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W237–W243, 2015.

[23] J. T. Morton, S. D. Freed, S. W. Lee, and I. Friedberg, "A large scale prediction of bacteriocin gene blocks suggests a wide functional spectrum for bacteriocins," *BMC Bioinf.*, vol. 16, no. 1, 2015, Art. no. 381.

[24] H. Mohimani *et al.*, "MetaRiPPquest: A peptidogenomics approach for the discovery of ribosomally synthesized and post-translationally modified peptides," *BioRxiv*, p. 227504, 2017.

[25] Y. Huang, Q. Zou, and X. J. Shen, "Construction of baculovirus expression vector of miRNAs and its expression in insect cells," *Mol. Gene., Microbiol. Virol.*, vol. 27, no. 2, pp. 85–90, 2012.

[26] X. Su *et al.*, "Antimicrobial peptide identification using multi-scale convolutional network," *BMC Bioinform.*, vol. 20, no. 1, 2019, Art. no. 730.

[27] G. Wang, X. Li, and Z. Wang, "APD3: The antimicrobial peptide database as a tool for research and education," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1087–D1093, 2015.

[28] L. Aguilera-Mendoza *et al.*, "Overlap and diversity in antimicrobial peptide databases: Compiling a non-redundant set of sequences," *Bioinformatics*, vol. 31, no. 15, pp. 2553–2559, 2015.

[29] M. Magrane and U. Consortium, "UniProt knowledgebase: A hub of integrated protein data," *Database*, vol. 2011, vol. 2011, Art. no. bar009.

[30] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Res.*, vol. 35, no. suppl_1, pp. D61–D65, 2006.

[31] S. Kumar, M. Nei, J. Dudley, and K. Tamura, "MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences," *Brief. Bioinf.*, vol. 9, no. 4, pp. 299–306, 2008.

[32] B. Liu, K. Li, D.-S. Huang, and K.-Chen Chou, "iEnhancer-EL: Identifying enhancers and their strength with ensemble learning approach," *Bioinformatics*, vol. 34, no. 22, pp. 3835–3842, 2018.

[33] Q. Zhang, Z. Shen, and D.-S. Huang, "Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network," *Sci. Rep.*, vol. 9, no. 1, 2019, Art. no. 8484.

[34] Y. He, Z. Shen, Q. Zhang, S. Wang, and D. -Shuang Huang, "A survey on deep learning in DNA/RNA motif mining," *Brief. Bioinform.*, vol. 22, no. 4, Jul. 2021, Art. no. bbaa229.

[35] Q. Zhang, Z. Shen, and D. S. Huang, "Predicting in-vitro transcription factor binding sites using DNA sequence + shape," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 2, pp. 667–676, Mar./Apr. 2021.

[36] W. He *et al.*, "Sc-ncDNAPred: A sequence-based predictor for identifying non-coding DNA in saccharomyces cerevisiae," *Front. Microbiol.*, vol. 9, 2018, Art. no. 2174.

[37] Q. Zhang *et al.*, "Weakly-supervised convolutional neural network architecture for predicting protein-DNA binding," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 2, pp. 679–689, Mar./Apr. 2020.

[38] W. Xu, L. Zhu, and D. S. Huang, "DCDE: An efficient deep convolutional divergence encoding method for human promoter recognition," *IEEE Trans. NanoBiosci.*, vol. 18, no. 2, pp. 136–145, Apr. 2019.

[39] L. Li *et al.*, "Network analysis of the hot spring microbiome sketches out possible niche differentiations among ecological guilds," *Ecological Model.*, vol. 431, 2020, Art. no. 109147.

[40] Z. Shen, S. P. Deng, and D. S. Huang, "RNA-Protein binding sites prediction via multi scale convolutional gated recurrent unit networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 5, pp. 1741–1750, Sep./Oct. 2020.

[41] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks." in *Proc. Int. Joint Conf. Neural Netw.*, 2011, pp. 2809–2813.

[42] G. Chuai *et al.*, "DeepCRISPR: Optimized CRISPR guide RNA design by deep learning," *Genome Biol*, vol. 19, no. 1, 2018, Art. no. 80.

[43] L. Yuan *et al.*, "Nonconvex penalty based low-rank representation and sparse regression for eQTL mapping," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 5, pp. 1154–1164, Sep./Oct. 2017.

[44] H. Zhang, L. Zhu, and D. S. Huang, "DiscMLA: An efficient discriminative motif learning algorithm over high-throughput datasets," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 6, pp. 1810–1820, 2018.

[45] Z. Shen, S. P. Deng, and D. S. Huang, "Capsule network for predicting RNA-Protein binding preferences using hybrid feature," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 5, pp. 1483–1492, Sep./Oct. 2020.

[46] W. Lee, D.-S. Huang, and K. Han, "Constructing cancer patient-specific and group-specific gene networks with multi-omics data," *BMC Med. Genomic.*, vol. 13, no. 6, 2020, Art. no. 81.

[47] L. Zhu, H.-B. Zhang, and D.-S. Huang, "Direct AUC optimization of regulatory motifs," *Bioinformatics*, vol. 33, no. 14, pp. i243–i251, 2017.

[48] F. Yang and Q. Zou, "mAML: An automated machine learning pipeline with a microbiome repository for human disease classification," *Database*, vol. 2020, 2020, Art. no. baaa050.

[49] K. Qu *et al.*, "Application of machine learning in microbiology," *Front. Microbiol.*, vol. 10, 2019, Art. no. 827.

[50] F. Friedrichs and C. Igel, "Evolutionary tuning of multiple SVM parameters," *Neurocomputing*, vol. 64, pp. 107–117, 2005.

[51] D.-S. Huang *et al.*, "Prediction of protein-protein interactions based on protein-protein correlation using least squares regression," *Curr. Protein Peptide Sci.*, vol. 15, no. 6, pp. 553–560, 2014.

[52] R. E. Schapire, "Explaining adaboost," in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, B. Schölkopf, Z. Luo, and V. Vovk, Eds., Berlin, Germany: Springer, 2013, pp. 37–52.

[53] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4509–4522, Sep. 2017.

[54] X. Shi *et al.*, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[55] Z. Shen, Q. Zhang, K. Han, and D. -S. Huang, "A deep learning model for RNA-Protein binding preference prediction based on hierarchical LSTM and attention network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jul. 7, 2020, doi: 10.1109/TCBB.2020.3007544.

[56] N. Quiles-Puchalt *et al.*, "Staphylococcal pathogenicity island DNA packaging system involving <em>cos</em>-site packaging and phage-encoded HNH endonucleases," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 16, pp. 6016–6021, 2014.

[57] C. J. Evans and R. J. Aguilera, "DNase II: Genes, enzymes and function," *Gene*, vol. 322, pp. 1–15, 2003.

[58] S. Alguwaizani *et al.*, "Predicting interactions between virus and host proteins using repeat patterns and composition of amino acids," *J. Healthcare Eng.*, vol. 2018, 2018, Art. no. 1391265.

[59] B. Kim *et al.*, "An improved method for predicting interactions between virus and human proteins," *J. Bioinf. Comput. Biol.*, vol. 15, no. 01, 2016, Art. no. 1650024.

[60] M. A. Larkin *et al.*, "Clustal w and clustal x version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[61] P. Gouet, X. Robert, and E. Courcelle, "ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3320–3323, 2003.

[62] L. Zhang *et al.*, "Biochemical characterization of a thermostable HNH endonuclease from deep-sea thermophilic bacteriophage GVE2," *Appl. Microbiol. Biotechnol.*, vol. 100, no. 18, pp. 8003–8012, 2016.

[63] K. G. Fleming and D. M. Engelman, "Specificity in transmembrane helix–helix interactions can define a hierarchy of stability for sequence variants," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 25, pp. 14340–14344, 2001.

[64] A. Pingoud and A. Jeltsch, "Structure and function of type II restriction endonucleases," *Nucleic Acids Res*, vol. 29, no. 18, pp. 3705–3727, 2001.

Zhen Cui is currently working toward the PhD degree with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include bioinformatics, machine learning, and deep learning.

Zhan-Heng Chen received the PhD degree from the University of Chinese Academy of Sciences, China. He has authored more than 28 research publications in journals, including *Briefings in Bioinformatics*, *Molecular Therapy-Nucleic Acids*, *BMC Genomics*, *BMC Systems Biology*, *Frontiers in Genetics*, *International Journal of Molecular Sciences*, *iScience*, *Journal of Cellular and Molecular Medicine* and international conferences, including ICIBM and ICIC. His research interests include data mining, natural language processing, bioinformatics, machine learning, and pattern recognition.

Qin-Hu Zhang received the PhD degree in computer science and technology from Tongji University, China, in 2019. He is currently a postdoctor with Tongji University. His research interests include bioinformatics, machine learning, and deep learning.

Valeriya Gribova is with the Institute of Automation and Control Processes, Far Eastern Branch of Russian Academy of Sciences, Russia. She is currently an expert with Analytic Center, Government of Russian Federation, the vice-president of the Russian Association of Artificial Intelligence, the member of ITHEA, the member of the Expert Council of Russian Foundation for Basic Research, and an expert of Russian Science Foundation. She was the recipient of the ITHEA award for Outstanding Achievement in the Field of Information Theory and Application (2009), the commendation certificate of the Far East Branch of the Russian Academy of Sciences (2001, 2006, and 2012), and the commendation certificate of the Ministry of Education and Science of Primorsky Krai (2011). Her research interests include artificial intelligence and decision making, user interface, multiagent systems, program models and systems, and specialized program models and systems.

Vladimir F. Filaretov was born in 1948. In 1966, he finished school with an honors (gold) medal. He received the graduation degree (Hons.) in automatic systems from Moscow State Technical University, the candidate of sciences degree in engineering in 1976, the doctor of sciences degree in automatic control in 1990, and the professor's degree in 1992. In 1995, he was elected the member of an Russian and in 1996 the member of International Engineering Academy. He is currently the head of Department of Automation and Control of Far Eastern Federal University and the head of Robotic Laboratory of the Institute of Automatics and Control Process of Russian Academy of Sciences, the president of Far Eastern Branch Russian Engineering Academy, and the vice president of Russian Engineering Academy. His research interests include the creation of industrial and underwater robots and manipulators and also other dynamic systems, allowing to automate technical devices and technological processes.

De-Shuang Huang (Fellow, IEEE) received the BSc degree in in electronic engineering from the Institute of Electronic Engineering, Hefei, China, in 1986, the MSc degree in electronic engineering from the National Defense University of Science and Technology, Changsha, China, in 1989, and the PhD degree in electronic engineering, from Xidian University, Xian, China, in 1993. From 1993 to 1997, he was a postdoctoral student with the Beijing Institute of Technology and National Key Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China. In September, 2000, he was with the Institute of Intelligent Machines, Chinese Academy of Sciences as the recipient of "Hundred Talents Program of CAS". In September 2011, he was with Tongji University as chaired professor. From September 2000 to March 2001, he was a research associate with Hong Kong Polytechnic University. From August to September 2003, he was with George Washington University as a visiting professor, Washington, DC, USA. From July to December 2004, he was the university fellow with Hong Kong Baptist University. From March 2005 to March 2006, he was the research fellow with the Chinese University of Hong Kong. From March to July 2006, he was a visiting professor with the Queen's University of Belfast, U.K. In 2007, 2008, and 2009, he was a visiting professor with Inha University, Korea, respectively. He is currently the director of the Institute of Machines Learning and Systems Biology, Tongji University. He has authored or coauthored 220 journal papers. His research interest includes bioinformatics, pattern recognition, and machine learning. He is currently an IAPR fellow.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.