Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 11, No. 2, 2025

MsaACPpred: Prediction of Anticancer Peptides Based on Multi-Scale Attention Convolutional Network

Yanyu Diao *, Xueyi Liu

School of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei Anhui, 230088, China
*** Corresponding author:** Yanyu Diao (Email: diaoyy0228@163.com)

Abstract: Anticancer peptides (ACPs) constitute a family of biologically active peptides with significant anti-tumor activity. These peptides are composed of 5 to 50 amino acid residues and are generally cationic in nature. Through electrostatic interactions, ACPs are able to target and bind to negatively charged cell membranes on the surface of cancer cells, thereby destroying their structure and inducing apoptosis. Given that the process of recognizing ACPs in the laboratory is highly restricted, which is not only costly but also time-consuming and lengthy, this study proposes a computational method for predicting ACPs based on sequence information. The method is designed with three core modules: feature encoding of peptide sequences, deep learning module layers, and a classification prediction layer. Among them, the deep learning module integrates convolutional neural network (CNN), attention mechanism and bidirectional long short-term memory network (BiLSTM), which enhances the learning and analyzing ability of the model. Ultimately, our model achieves 73.1% accuracy while maintaining 77.8% sensitivity and 69.6% specificity.

Keywords: Anticancer Peptides; CNN; Attention Mechanism; BiLSTM.

1. Introduction

Cancer, a deadly disease, has become a major challenge in human healthcare. Currently, radiotherapy, chemotherapy and targeted therapies constitute the mainstay of cancer treatment [1], and although these drugs are designed to destroy cancer cells, unfortunately, they often also cause damage to normal cells, triggering a series of significant side effects and making treatment unaffordable for many patients. In light of this, the focus of the scientific community has gradually shifted to the study of ACPs. Compared with many conventional cancer therapies, ACPs show several advantages: they are based on natural biological targets and thus may be safer; at the same time, by virtue of their intrinsic cationic properties, ACPs are able to attack cancer cells more selectively and preferentially bind to the anionic membrane structure on the surface of the cancer cells, thus realizing a precise strike [2].

ACPs are widely distributed in many organisms including mammals, amphibians, insects, plants and microorganisms, and serve as a key component of the innate immune defense mechanism of these organisms. In recent years, by virtue of their unique anticancer mechanism of action and potential clinical therapeutic value, ACPs have become a hot spot in the field of antitumor drug development. In this context, computational prediction methods play an increasingly important role in the screening and discovery of ACPs. Researchers are working tirelessly with the aim of developing an efficient and cost-effective strategy to accelerate the discovery and characterization process of novel ACPs.

For example, Hajisharifi et al. [3] combined Chou's pseudo-amino acid composition method with the Ames test, and not only successfully predicted the potential anticancer activity of a group of peptides, but also comprehensively evaluated their safety. The results showed that some of the peptides did not exhibit significant genotoxicity while demonstrating anticancer potential. Vijayakumar et al. [4] introduced ACPP, an online server for prediction and design of anticancer peptides. The ACPP server is centered on support vector machines and feature vectors of protein similarity metrics, and experimental studies revealed that peptides containing apoptotic domains have significant anticancer activity. The ACPred-Fuse method [5] is a comprehensive prediction system that integrates multi-view information from multiple sources and types, which covers amino acid sequences, physicochemical properties, and structural features. By applying a feature selection algorithm, the method is able to filter out the most critical features for anticancer peptide prediction from this rich information. Based on these filtered features, a machine learning model is constructed for accurate prediction of anticancer peptides. In addition, Yu et al. [6] proposed a novel computational method called DeepACP, which utilizes a deep learning algorithm to accurately identify anticancer peptides. The ACP-MHCNN method [7] is a predictive model based on a deep convolutional neural network with multiple heads. This model is able to capture the features of anticancer peptides more accurately by integrating the attention mechanism of multiple heads. DLFF-ACP model [8] is based on deep learning and multi-view feature fusion techniques. The model first extracts candidate features of anticancer peptides from multiple dimensions, and then deeply fuses and learns these features from different perspectives using deep learning algorithms, so as to effectively capture the complex patterns and properties of anticancer peptides. As for ACP-DA [9], its core strategy lies in enriching and balancing the training dataset through data augmentation techniques, thus significantly improving the accuracy of anticancer peptide prediction. There are also many studies that use one-hot or sequential coding methods to encode peptide sequences in the prediction of ACPs. For example, Rao et al. [10] proposed a graph convolutional network-based approach to identify anticancer peptides, where they encoded peptide sequences using one-hot coding in the embedding layer. Lv et al. [11] proposed a computational method called iACP-DRLF, which uses an optical gradient enhancer algorithm with a deep representation of learned features to identify anticancer

peptides.

Inspired by the research results of other scholars, we propose a deep learning-based method, named MsaACPpred, for predicting ACPs. First, we employed both amino acid composition and pseudo amino acid composition to encode the peptide sequences. Subsequently, the high-dimensional representation of the features was further realized using the word2vec model. In order to deeply mine the local features in the data, we designed a multi-scale one-dimensional convolutional neural network and incorporated the attention mechanism, a combination that significantly enhanced the model's ability to capture key information. And then, we introduced a bidirectional LSTM model, which is able to efficiently extract long-distance correlation-dependent information from the multi-scale local features. Finally, we integrated the extracted features through a fully connected network, thus realizing the prediction of ACPs.

2. Materials and Methods

The whole framework diagram of MsaACPpred is shown in Figure 1.

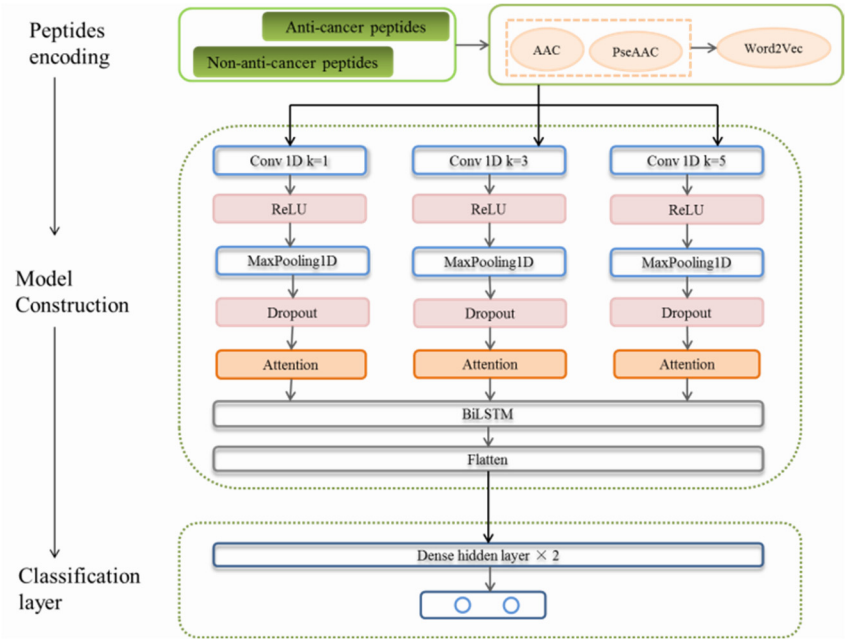

Figure 1. Schematic diagram of MsaACPpred

In the peptide sequence encoding stage, the peptide sequences are encoded by amino acid composition (AAC) and pseudo amino acid composition (PseAAC), and then converted into feature vectors using Word2Vec for subsequent processing. In the model construction stage, first, preliminary feature extraction is performed on the input data by one-dimensional convolutional layers (Conv 1D) combined with the ReLU activation function. The model is designed to include convolutional layers with different convolutional kernel sizes (k=1, k=3, k=5) to capture feature information at different scales. Then, MaxPooling1D layers is utilized to reduce the data dimensionality while retaining important features. To prevent overfitting, Dropout layers is embedded in the model with the dropout rate of 0.3. In addition, the Attention mechanism is introduced to further enhance the model's ability to capture key features [12]. In addition, the model employs a Bidirectional Long and Short-Term Memory Network (Bi-LSTM), which is capable of efficiently dealing with long-term dependencies in sequence data. Subsequently, the multidimensional features are converted to one-dimensional by means of a flatten layer in order to be fed into the fully connected layer for classification. The model contains two hidden fully connected layers for further extraction and combination of features, and finally the classification of ACPs and non-ACPs is achieved through the output layer.

In the process of training the MsaACPpred model, we used the AdamW optimizer and set the initial learning rate to 0.01. For the activation function, we chose ReLU, and selected the cross-entropy function as the loss function. The dimension of the bidirectional LSTM hidden layer is 4, and the dimensions of the two hidden layers in the fully connected network are 128 and 32, respectively.

(1) Dataset

We collected a total of 1722 experimental data from the published literature [13], covering two main categories: one is the rigorously experimentally validated ACPs, totaling 861, which were derived from the CancerPPD database, ensuring the reliability and expertise of the data; and the second is the corresponding 861 non-ACPs, which served as negative samples, derived from antimicrobial peptides. In order to validate the predictive performance of the model, we divided the entire dataset into two subsets. One of them, as the main body of training and evaluation, this subset is used to implement a 5-fold cross-validation strategy to ensure the stability and generalization ability of the model under different data combinations; at the same time, this process also promotes the continuous optimization of model training. The other subset, which serves as an independent test set, is independent of the training process and is used to ultimately assess the predictive accuracy of the model. The dataset is now fully open to the scientific community and is freely available to researchers by visiting the AntiCP 2.0 server (https://webs.iiitd.edu.in/raghava/anticp2/).

(2) Peptide sequence encoding

Before inputting the peptide sequence into the deep

learning model, it first needs to be encoded numerically. In this paper, we use the compositional properties of peptides to encode the sequences, and this process includes two major features, amino acid composition (AAC) and pseudo amino acid composition (PseAAC). These two features can be obtained with the help of Pfeature web server [14]. Thus, the final feature dimension obtained is 40, which contains 20 AAC features and 20 PseAAC features.

In order to further transform these features into a form suitable for modeling, we employ word2vec to encode the features as vectors of dimension D, forming an encoding length of D × 40. word2vec, as a state-of-the-art technique, is capable of mapping discrete words into a continuous vector space. By deeply learning the distributed representation of words in the corpus, the word2vec model is able to ensure that words with similar contexts present similar mappings in the vector space, which improves the accuracy and effectiveness of the encoding.

(3) Multi-Scale Convolutional Neural Network and Attention

Traditional 2D convolutional neural networks are commonly used to process image data and their basic structure is two-dimensional [15]. Unlike two-dimensional convolutional neural networks, the convolutional kernel of 1D-CNN moves in only one direction, making it particularly suitable for processing one-dimensional sequence data. The core function of the Attention mechanism is to filter out the information that is more critical to the current task goal from a large amount of information, which is achieved by assigning a weight vector to each feature that indicates the importance of the association. Attention mechanisms have been widely used in the processing of sequence data, enabling the model to selectively focus on different parts of the input sequence, assigning different weights to various parts of the input sequence as a way of highlighting information that is more critical to the task. The 1D-CNN with attention captures local features in the input data through convolutional operations. The convolutional kernel slides over the input data and is able to recognize sequential features at different scales, which helps the model to understand the structure and pattern of the text. Meanwhile, the attention mechanism is used to enhance the output of the CNN convolutional layer. It allows the model to focus on specific parts of the input sequence and pay more attention to the information that is important for solving the task.

(4) Bidirectional Long Short-Term Memory

The long short-term memory (LSTM), as a variant of recurrent neural network (RNN), possesses the ability to learn and memorize long-term dependencies, and is able to capture complex associations between different features in the feature vector, while effectively mitigating the problem of gradient vanishing or gradient explosion. The bidirectional long short-term memory network (BiLSTM) is composed of a combination of two independent LSTM networks, which process the input sequences from two opposite directions. This design enables the model to capture information before and after the current time step simultaneously, which greatly enriches the feature extraction. The BiLSTM layer consists of a stack of two network layers, the forward LSTM and the backward LSTM, which breaks the limitation of the traditional LSTM model that only relies on the previous timing information to predict the output of the next moment. Figure 2 illustrates the network structure of BiLSTM, which consists of several basic units, each of which contains four layers: input layer, forward layer, backward layer, and output layer. The forward propagation layer focuses on extracting the forward features of the input vectors, while the backward propagation layer works on capturing the reverse features of the input sequence. Ultimately, the output layer fuses the outputs of the forward and backward propagation layers, enabling BiLSTM to capture contextual information more comprehensively.

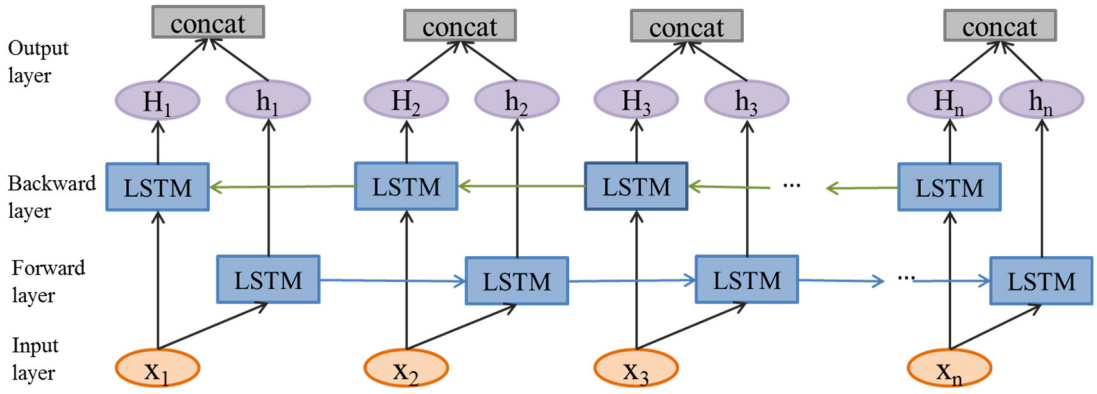

Figure 2. BiLSTM network structure

In this study, the BiLSTM was employed, aiming to capture sequence dependencies in peptide sequences more accurately. BiLSTM can comprehensively consider the contextual information before and after each moment in the sequence, providing a complete sequence perspective for anticancer peptide identification. Through BiLSTM, the model is able to deeply understand the associations between various parts of the peptide sequence, especially their temporal and contextual dependencies, and thus more accurately capture features and patterns, significantly improving the prediction performance of anticancer peptides.

(5) Performance metrics

In this paper, we use four commonly used metrics for model performance evaluation, including accuracy (Acc), sensitivity (Sn), specificity (Sp), and Matthew's correlation coefficient (MCC). Their formulas are as follows:

$$\mathrm{Acc} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\mathrm{Sn} = \frac{TP}{TP+FN} \quad (2)$$

$$\mathrm{Sp} = \frac{TN}{TN+FN} \quad (3)$$

$$\mathrm{MCC} = \frac{TP \times \mathrm{TN} - \mathrm{TP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (4)$$

where TP denotes true positive, signifying the number of correctly predicted true ACPs; TN stands for true negative, indicating the number of correctly identified non-ACPs; FP represents false positive, referring to the instances where non-ACPs were incorrectly predicted as true ACPs; and FN signifies false negative, which is the number of true ACPs mistakenly predicted as non-ACPs. The metrics Sn and Sp assess a model's predictive proficiency in positive and negative cases, respectively, as referenced in [35, 36]. Additionally, Acc and MCC are employed to evaluate the overall efficacy of the model. Notably, higher scores in all these metrics indicate better performance by the models.

(6) Five-fold cross-validation method

Five-fold cross-validation is a method for evaluating the performance of machine learning models with the following procedure: first, the original dataset is uniformly divided into five subsets of similar size. Subsequently, five iterations are performed, and in each iteration, a different subset is selected as the test set, while the remaining four subsets are combined and used as the training set. The model is trained using the training set data and then the performance of the model is judged with the help of the test set data. During each iteration, various performance evaluation metrics of the model on the test set are recorded. After five iterations are completed, the final performance evaluation of the model is concluded by calculating the average value of these performance metrics. This average value not only reflects the overall performance level of the model, but is also an important basis for judging the stability and generalization ability of the model. The advantage of the five-fold cross-validation is that it can provide a comprehensive assessment of the model performance and reduce the performance assessment bias caused by improper data set division.

3. The Application Value of Traditional Embroidery in Modern Fashion Design

Traditional embroidery art still has wide application value in contemporary fashion design. For modern fashion design, it still has a very strong sound. First of all, we need to break the integrity of the design. Traditional embroidery art is often lack of abstractness, and the pattern image is very complete, but it has the characteristics of rigidity. It is not conducive to the long-term promotion of traditional ideas, and modern fashion design in the process of integration must break through the limitations of traditional embroidery, the integrity of the continuous to be broken, in order to innovate. For fashion designers, we can divide the patterns properly and pick up the relevant composition methods to make them more in line with modern people's clothing aesthetics. For example, when designing a dress, you can combine embroidery art to form a hollow way, so as to decorate the waist line and make the design effect of the whole curve more prominent. The second is to strengthen the aesthetic level of clothes and make them more in line with people's understanding of beauty. Traditional embroidery art is rooted in traditional culture, so the pattern is very exquisite and meticulous, which makes embroidery art unique and coquettish. It can enhance the aesthetic feeling of modern fashion design, so as to optimize contemporary fashion design. At present, many designers will actively try embroidery techniques in the process of design, so as to enhance the artistic value of clothing design, and even properly use embroidery techniques in foreign high-definition clothing, so as to further improve the quality of clothing.

4. Results

(1) Ablation experiments

The method we propose in this paper, MsaACPpred, uses convolutional neural networks, long short-term memory networks, and attention mechanisms, and each of these different modules plays a unique role in characterizing the object of study. For example, convolutional neural networks excel at refining local features, bidirectional long and short-term memory networks excel at capturing long-distance dependencies between words, and the attention mechanism highlights key associations between words. In this paper, we employ a five-fold cross-validation method to evaluate the performance of our MsaACPpred prediction model. In addition, we explored the contributions of each component to the prediction of ACPs by removing the corresponding components from the model. As shown in Table 1, removing these components leads to a significant decrease in model performance.

Table 1. Five-fold cross-validation results of ablation experiments

ablation experiments	Acc(%)	Sn(%)	Sp(%)	MCC
delete Multi-scale CNN	64.61	66.87	62.35	0.29
delete Attention mechanism	66.08	67.17	64.94	0.32
delete Bi-LSTM	65.9	61.8	69.9	0.32
MsaACPpred	72.5	75.6	69.2	0.45

In order to verify the effectiveness of combining different types of features, the five-fold cross-validation method was used to evaluate the features, and the related experimental results are detailed in Table 2. When using a combination of AAC and PseAAC features, the constructed model demonstrated a sensitivity of 75.6%, a specificity of 69.2%, an accuracy of 72.5%, and an MCC value of 0.45. Through comparative analysis, the results show that models incorporating multiple types of features can significantly improve performance compared to models relying only on a single feature representation.

Table 2. Five-fold cross validation results of the hybrid feature sets

Features	Acc(%)	Sn(%)	Sp(%)	MCC
ACC	64.2	69.4	59.0	0.28
PseAAC	68.9	67.5	70.3	0.38
ACC+PseAAC	72.5	75.6	69.2	0.45

(2) Comparison with Other Predictors

Table 3. Comparison MsaACPpred with other ACPs predictors on the same independent dataset

Methods	Acc(%)	Sn(%)	Sp(%)
AntiCP	50.6	100	12
ACPred	53.5	85.6	21.4
ACPred-FL	44.8	67.1	22.5
ACPpred-Fuse	68.9	69.2	68.6
PEPred-Suite	53.5	33.1	73.8
iACP	55.1	77.9	33.2
MsaACPpred	73.7	77.8	69.6

We also compare the results of the MsaACPpred method with existing methods on an independent test set. The prediction results on the independent test set are taken from reference [13]. The results are shown in Table 3, where ACPred ranks first in terms of Sn, but its Sp is only 21.4%. Compared to other methods, MsaACPpred has an overall better performance.

5. Conclusion

In this study, we propose a method called MsaACPpred, which is a deep learning-based method for predicting ACPs that uses one-dimensional convolutional neural networks, attention mechanisms, and bidirectional long and short-term memory networks. This method can not only speed up the screening of anticancer peptides, but also reduce the experimental cost and provide new ideas for the identification of anticancer drugs.

In the future, we will continue to collect rich and diverse data and actively explore new models to continuously improve the performance of our models and accurately identify anticancer peptides that work against specific cancers.

Acknowledgments

The authors gratefully acknowledge the financial support from the Key Scientific Research Project of Anhui Provincial Research Preparation Plan in 2023 (No. 2023AH051806); the School-level Scientific Research Projects of Anhui Xinhua University (No. 2024zr016); the Innovation Training Program Project for College Students in Anhui Province (No. S202412216206); Anhui Province Quality Engineering Project (No. 2022xsxx089).

References

[1] Cheng L, Zhao H, Wang P, et al. Computational Methods for Identifying Similar Diseases[J]. Mol Ther Nucleic Acids, 2019,18:590–604.

[2] Ge R, Feng G, Jing X, et al. EnACP: An Ensemble Learning Model for Identification of Anticancer Peptides[J]. Frontiers in Genetics,2020,11:760.

[3] Hajisharifi Z, Piryaiee M, Beigi MM, et al. Predicting anticancer peptides with chou's pseudo amino acid composition and investigating their mutagenicity via Ames test [J]. J Theor Biol,2014,341:34–40.

[4] Saravanan Vijayakumar, Lakshmi PTV.ACPP: A Web Server for Prediction and Design of Anti-cancer Peptides[J]. International Journal of Peptide Research &Therapeutics, 2015, 21 (1):99–106.

[5] Rao B, Zhou C, Zhang G, et al. ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides [J]. Briefings in Bioinformatics, 2020, 21(5): 1846–1855.

[6] Yu L, Jing R, Liu F, et al. DeepACP: A Novel Computational Approach for Accurate Identification of Anticancer Peptides by Deep Learning Algorithm[J]. Mol Ther Nucleic Acids,2020,22: 862–870.

[7] Sajid Ahmed, Rafsanjani Muhammod, Zahid Hossain Khan, et al. ACP-MHCNN: an accurate multi-headed deep-convolutional neural network to predict anticancer peptides[J]. Scientific Reports,2021,11(1): 1–15.

[8] Cao R, Wang M, Bin Y, et al. DLFF-ACP: prediction of ACPs based on deep learning and multi-view features fusion[J]. PeerJ, 2021,9: e11906.

[9] Chen X, Zhang W, Yang X, et al. ACP-DA: Improving the Prediction of Anticancer Peptides Using Data Augmentation[J]. Frontiers in Genetics, 2021,12: 1-9.

[10] Rao B, Zhang L, Zhang G. ACP-GCN: The Identification of Anticancer Peptides Based on Graph Convolution Networks[J]. IEEE Access, 2020, 8: 176005–176011.

[11] Lv Z, Cui F, Zou Q, et al. Anticancer peptides prediction with deep representation learning features[J]. Briefings in Bioinformatics, 2021,22(5): bbab008.

[12] Yuan Q, Chen K, Yu Y, et al. Prediction of anticancer peptides based on an ensemble model of deep learning and machine learning using ordinal positional encoding[J]. Briefings in Bioinformatics, 2023, 24(1):1–10.

[13] Agrawal P, Bhagat D, Mahalwal M, et al. AntiCP 2.0: an updated model for predicting anticancer peptides[J]. Briefings in Bioinformatics, 2020,22(3): bbaa153.

[14] Pande A, Patiyal S, Lathwal A, et al. Computing wide range of protein/peptide features from their sequence and structure[J]. bioRxiv, 2019.DOI:10.1101/599126.

[15] Pfahringer B , Reutemann P , Witten I H ,et al.The WEKA data mining software: an update[J].Acm Sigkdd Explorations Newsletter, 2009, 11(1):10-18.