Construction and Prediction of Antimicrobial Peptide Predicition Model Based on BERT

Xin-Yu Yu[1]*, Rao Fu[1], Peng-Yu Luo[1], Yue Hong[1], Yue-Han Huang[1]
[1]University of Xiamen,
771702500@qq.com,1090235818@qq.com

Abstract

In daily life, the abuse of antibiotics causes bacteria to develop resistance to antibiotics, which is not conducive to public health. Antimicrobial peptide is a small molecule polypeptide that forms a key component of the biological innate immune system. It kills target bacteria by destroying cell membranes and interfering with DNA. Therefore, antimicrobial peptides are promising alternatives to antibiotics. However, antimicrobial peptides have different lengths and various structures. In order to apply them to various fields, they need to be predicted and identified.The existing antimicrobial peptide identification and prediction methods mainly adopt biological experiments, sequence alignment or machine learning methods. Through deep learning, the accuracy and speed of antimicrobial peptide identification and prediction can be further improved. At present, there are few deep learning network models that specialize in amino acid sequences, and there is no special database to provide samples of non-antimicrobial peptides. Therefore, how to apply the relevant cutting-edge methods of deep learning to the identification and prediction of antimicrobial peptides is a problem worthy of study. At the same time, the accuracy and speed of antimicrobial peptide identification still have room for further improvement.The Bidirectional Encoder Representations from Transformers (BERT) model is a model applied in the field of natural language processing. It builds a network structure with Transformer as the core, and its core mechanism is the self-attention mechanism. We used the BERT model based on the existing protein database, used a large number of protein sequences for pre-training, and combined with the comprehensive data set for fine-tuning, and constructed an antimicrobial peptide prediction model.

Introduction

Since the advent of antibiotics, they have had significant effects in the treatment of various diseases. However, with the emergence of antibiotic abuse, bacterial resistance has gradually increased, which is not conducive to long-term disease treatment. Antimicrobial peptides were artificially induced in the 1980s. (Steiner et al. 1981) Antimicrobial peptides are small molecular peptides that form a key component of the biological innate immune system. The length is generally 9-100 amino acids. It kills target bacteria by damaging cell membranes and interfering with DNA. Studies have shown that antimicrobial peptides are the most promising drugs to replace traditional antibiotics, and some antimicrobial peptides have been used in clinical treatment and other fields.(Boman 2003);(Zelezetsky et al. 2006) At present, there are many prediction methods for antimicrobial peptides, such as biological experiment methods, which need to be designed by professionals, and require rich experience and a large amount of manpower and material resources, and the efficiency is low. The use of deep learning for sequence comparison can speed up the acquisition of data rules and discover the internal correlation of data. Improving the speed and accuracy of predicting antimicrobial peptides by the model can speed up the relevant research process and accelerate its use in various fields. . Therefore, how to improve the identification and prediction speed of antimicrobial peptides and maintain a certain accuracy has become a problem.

On the other hand, the amino acid sequences that make up antimicrobial peptides are similar to text sequences in daily life, and their essence is a sequence composed of symbols. Therefore, deep learning models in the field of natural language processing can be used in the prediction and recognition of antimicrobial peptides. . After BERT was proposed,(Devlin et al. 2018) it has shown excellent performance on multiple natural language processing tasks. Therefore, we applied the BERT model to the task of identifying antimicrobial peptides, and implemented the BERT model for identifying and predicting antimicrobial peptides.

Related work

At present, many researchers have begun to use deep learning algorithms to identify and predict antimicrobial peptides. For example, Marc T et al.(Torrent et al. 2011) used neural networks to predict antimicrobial peptides; Xiao X et al. (Xiao et al. 2013) constructed a secondary classification of antimicrobial peptides. It first uses the amino acid composition (PseAAC) for feature extraction, and then uses fuzzy K nearest neighbors to predict antimicrobial peptides; Fjell CD et al.(Fjell et al. 2009) used QSAR and machine learning techniques to combine antimicrobial peptides to screen ; Veltri D (Veltri, Kamath, and Shehu 2015) constructed an end-to-end model including convolutional neural network and cyclic neural network for the prediction and identifica-

[*] Work done during the deep learning course

tion of antimicrobial peptides; Randou EG et al. (Randou, Veltri, and Shehu 2013) extracted 8 physical and chemical characteristics of peptides and used logic Regression model was used to predict antimicrobial peptides; Lee EY et al. (Lee et al. 2016) used SVM to predict antimicrobial peptides. Yoshida M et al.(Yoshida et al. 2018) used a combination of evolutionary algorithms, neural networks, and in vitro evaluation to optimize the efficacy of antimicrobial peptides , and Michael et al.(Youmans, Spainhour, and Qiu 2019) used Long-Short Term Memory (LSTM) to predict and identify antimicrobial peptides , Dua M et al.(Dua et al. 2018) encapsulated a deep model in a random heuristic search for the screening of antimicrobial peptides . These models have improved the accuracy and speed of antimicrobial peptide identification and prediction, but there is still room for improvement. At the same time, some existing models have the disadvantage of poor mobility.

Proposed Solution

Bert model

BERT (Bidirectional Encoder Representations from Transformers) is a deep pre-training language model based on Transformer architecture, and its structure is mainly shown in Figure 1.

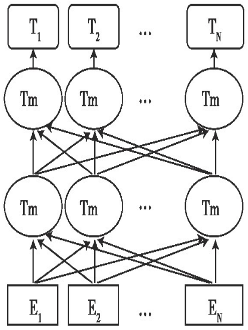

Figure 1: Basic structure of Bert model

Taking the Chinese pre-training model as an example, E1, E2,...EN in Figure 1 indicate that text characters marked with [CLS] and [SEP] are added at the beginning and end. They pass through the 12-layer two-way Transformer (Trm) encoder in turn to obtain the contextual embeddings of text characters. Transformer is an encoder-decoder based on a self-attention mechanism. The input of the bottom Transformer encoder is the sum of character vector, character position vector and sentence fragment vector. Each layer in the model consists of two parts: Multi-head Self-attention and Feed-forward Neural Networks. The former enables the encoder to pay attention to each character when encoding Information about other characters around; the latter is used to enhance the fitting ability of the model. After each layer of the model undergoes an add and normalization (Add & Norm) operation, a new character vector is generated as the input of the encoder of the next layer. The coding vector T1 marked with [CLS] output by the top-level encoder can be regarded as a semantic representation of the entire sentence and used for subsequent text classification tasks.

In addition, in order to enhance the ability of semantic representation, BERT proposed the concepts of masked language model (Masked LM, MLM) and Next Sentence Prediction (NSP). MLM is essentially a cloze task. 15% of the characters in the Chinese corpus will be selected, 80% of which will be replaced with [MASK], 10% will be randomly replaced with another character, and the remaining 10% will remain the original character. The model needs to pass a linear classifier to predict the selected word. In order to be consistent with the following tasks, BERT needs to place the original word or a random word in the predicted word position in a certain proportion, so that the model is more inclined to use context information to predict the selected word. In the next sentence prediction task, the model selects several sentence pairs, among which there is a 50% probability that two sentences are adjacent, and a 50% probability that two sentences are not adjacent. The model can learn the semantic information between words and sentences better through the above two target tasks.

Data acquisition and preprocessing

Language model pre-training has shown excellent performance in NLP tasks. Because antimicrobial peptide sequences are similar to text sequences, language model pre-training can be considered for antimicrobial peptide recognition and prediction. In order to enable the model to capture the long-term dependence and hierarchical relationship of protein sequences after pre-training,(Linzen, Dupoux, and Goldberg 2016);(Gulordava et al. 2018) it is necessary to feed a large number of protein sequences to the model for pre-training. We downloaded 556,603 pieces of data from the UniProt database as pre-training data. By pre-training the above data, the model can be used to capture proteins related to downstream tasks, such as long-term dependencies and hierarchical relationships. Based on the above protein data, we pre-trained a BERT model, and fine-tuned and evaluated the model with reference to the antimicrobial peptide data constructed by others.

Although the antimicrobial peptide sequence has similarities with the text sequence, there are also certain differences. Antimicrobial peptides are not like English texts to divide individual words by spaces, nor do they use dictionary matching algorithms for word segmentation like Chinese texts. In this article, each protein is cut every three amino acids. As a word, the amino acid fragments with the tail of the sequence less than three amino acids in length are individually regarded as a word. In this way, the protein is ”word-divided” so that protein data can be used to predict Train the BERT model.

At the same time, we use the data set in Table 1 to fine-tune the model, and also randomly downsample the negative sample set of the training set to ensure sample balance. For the sampled protein sequence, the same segmentation method is used for ”word segmentation”. In addition, in order to provide a predictive model of antimicrobial peptides with strong generalization ability, all the protein sequences of the 4 data sets were merged, and all repetitive sequences

Datasets	Training Set		Test set	
	Positive Sample	Negative Sample	Positive Sample	Negative Sample
Dataset by Veltri, D. et al.	1066	1066	712	712
Dataset by Michael et al.	2087	2536	522	634
Dataset by Xiao, X. et al.	879	2405	920	920
Dataset by Lin, Y. et al.	2617	4371	284	1382

Table 1: Antimicrobial peptide data set composition distribution

were deleted, so as to avoid training on repeated samples, which caused the model to be overly targeted for some samples. Learn. CD-HIT(Meher et al. 2017) is then used to remove sequences with 70% pairwise sequence similarity, and the remaining sequences are used for five-fold cross-validation. Finally, use all the data to train the antimicrobial peptide prediction model.

The construction process of antimicrobial peptide prediction model based on BERT model is shown in Figure 2.

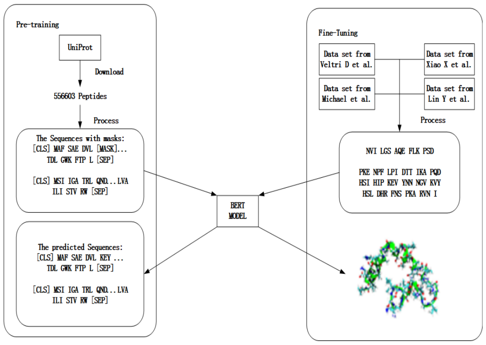

Figure 2: Antimicrobial peptide prediction model construction process

Training method

We use a 12-layer transformer, the hidden layer contains 768 unit nodes and 12 attention heads. In this paper, 15% of the words in each protein sequence in the pre-training data set are randomly masked, and the transformer is trained to predict the masked words, so that the transformer can capture the long-term dependence of the protein. This transformer is trained on a TITAN Xp. The batch size is 32, and the number of training sessions is 10 million. Through a large amount of training, the model can fully learn the long-term dependence and hierarchical relationship of proteins, and can improve the accuracy of the downstream task, that is, the prediction of antimicrobial peptides. Then fine-tuned the model using the processed antimicrobial peptide dataset and compared the results with other models

Experiments

Evaluation metrics

In order to compare the independent test results of this model with other models, we use Sensitivity (Sn), Specificity (Sp), Accuracy (Accuracy, Acc) and Matthews correlation coefficient (Mattews correlation coefficient, abbreviated as MCC) is used as the evaluation index of the model. These four indicators are defined by formulas (1)-(4).

$$Sn = \frac{TP}{TP + FN} \tag{1}$$

$$Sp = \frac{TN}{TN + FP} \tag{2}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

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

TP is the number of predicted correct antimicrobial peptides, that is, the number of true positive samples, FN is the number of predicted incorrect antimicrobial peptides, that is, the number of false negative samples, and TN is the number of predicted correct non-antimicrobial peptides, that is, the number of true negative samples. FP is the number of non-antibacterial peptides that are predicted incorrectly, that is, the number of false positive samples. Comprehensive consideration of the four evaluation indicators of sensitivity, specificity, accuracy and Matthews correlation coefficient can better evaluate the performance of the classification model. Considering the sensitivity and specificity of the model, we can know the model's ability to recognize positive and negative samples respectively. When the positive and negative samples of the test set are balanced, the accuracy rate can better reflect the classification performance of the model, but if the positive and negative samples of the test set are unbalanced, the accuracy rate will lose its reference significance. When the positive and negative samples are unbalanced, MCC is a good indicator when evaluating model performance. It considers both true positives and false positives as well as true negatives and false negatives, and its value is between -1 and +1. When the value is close to 1, the classification performance of the model is better; when the value is close to -1, the model prediction result is opposite to the actual result; when the value is close to 0, the model prediction result is similar to random prediction.

Introduction to the benchmark model

The model we will build will be compared with AMP-Scan(Veltri, Kamath, and Shehu 2018), BiLSM(Yoshida et al. 2018), iAMP-2L(Xiao et al. 2013) and MAMP-Pred(Lin et al. 2019) four models that also do antimicrobial peptide prediction. The source of performance indicators of these four models Yu proposed the literature of the model.

AMPScan is an end-to-end model. The model contains five parts. From the input layer to the output layer, they are embedding layer, convolution layer, maximum pooling layer, LSTM and fully connected layer. The activation method of fully connected layer adopts Sigmoid. Function, the difference from the model in this chapter is that this chapter uses a two-way LSTM.

BiLSTM is an end-to-end model that only uses two-way LSTM. Its structure is relatively simple as a whole. When acquiring the characteristics of a protein sequence, only the global information of the protein is acquired, which may lead to the loss of some local information and key information.

The feature vector structure of iAMP-2L includes the following content: 1. Count the occurrence frequency of each amino acid in the sequence. This part contains the overall information of the sequence, but the sequence information of the amino acids is missing; 2. Calculate the two positions at a fixed relative position. The product of five physicochemical properties (molecular mass, hydrophobicity, etc.) of an amino acid. This method obtains the sequence information of amino acids to a certain extent. The classifier of iAMP-2L uses fuzzy K-nearest neighbors, which is a variant of the K-nearest neighbor algorithm. The main difference from the K-nearest neighbor algorithm is that an index is added to the Euclidean distance between the target sample and the neighbor sample, which is called Is the fuzzy coefficient. Adding this method can further enhance or weaken the influence of distance on the classification results. Compared with the classic K nearest neighbor algorithm, it is more flexible, but it also introduces a hyperparameter that requires manual selection.

MAMP-Pred obtains sequence features through the SVM-Prot 188D algorithm based on 8 physical and chemical properties and the Co-Pse-AAC algorithm based on 5 physical and chemical properties. These two algorithms are commonly used algorithms for protein feature extraction and have good results.

Evaluation and comparison with the benchmark model

In each data set, the BERT-based model performs better than other models. The ACC value of our model is more than 1% higher than other models. In particular, on ACC, the BERT-based model is about 3% higher than iAMP-2L. The results show that the BERT-based model performs well on most data sets. In the models iAMP-2l and MAMP, protein features are manually constructed by feature engineering. The difference is that iAMP-2l uses fuzzy K-nearest neighbors, while MAMP uses SVM to predict whether the protein is an antimicrobial peptide. Both iAMP-2l and MAMP use traditional machine learning methods to predict and identify antimicrobial peptides. The construction of features depends on the experimenter's settings, and the performance of the model depends on the experimenter's experience to a certain extent. Both AMPScan and Bi-LSTM use end-to-end networks to predict and identify antimicrobial peptides, and obtain features through adaptive learning of the network model. The performance of the model does not depend on the experience of the experimenter. The number of end-to-end network training is relatively small, and the number of samples containing positive and negative sample labels is small, and the characteristics of the sequence cannot be fully captured, and performance is lost. By pre-training a large amount of unlabeled data, the BERT model can further capture the characteristics of the sequence and fine-tune specific data sets to identify and predict antimicrobial peptides.

Datasets	Model	Sn(%)	Sp(%)	Acc(%)	MCC
Dataset by Veltri, D. et al.	AMPScan	89.89	92.13	91.01	0.8204
	BERT	90.62	93.07	91.84	0.8376
Dataset by Michael et al.	Bi-LSTM	-	-	94.98	0.899
	BERT	92.7	96.87	95.27	0.9023
Dataset by Xiao, X. et al.	iAMP-2L	97.72	86.84	92.23	0.8446
	BERT	97.98	92.6	95.53	0.9895
Dataset by Lin, Y. et al.	MAMP-Pred	83.1	84.4	84.16	-
	BERT	79.42	86.21	85.32	0.586

Table 2: comparison with the benchmark model

In summary, the BERT model can effectively capture the long-term dependence of the sequence through long-term pre-training, and effectively improve the performance of the model. The fine-tuning process is end-to-end, avoiding the dependence on the expert domain, and avoiding the influence of the pros and cons of the feature extraction algorithm on the performance of the model.

Five-fold cross validation

We merge the four data sets in Table 1 to remove duplicate sample data, and use CD-HIT to remove sequences with 70% sequence pair similarity to reduce data redundancy. At the same time, the BERT-based model was cross-validated 5 times, and the cross-validation results are shown in Table 3. The cross-validated average Sn, Sp, Acc, and MCC were 87.68%, 85.82%, 85.98%, 0.6261, respectively. This numerical result is low compared to the training and independent testing of the model on a single data set. It may be because the four data sets are constructed differently, especially the negative samples are constructed differently, which leads to the emergence of the data set. The pollution makes the model appear to be degraded in terms of performance value. In addition, the accuracy of the model in identifying positive samples is higher than the accuracy of identifying negative samples. This may be because the construction process of positive samples is relatively similar. During training, there are fewer false positive samples, which makes the model in The ability to identify positive samples is stronger.

	Sn(%)	Sp(%)	Acc(%)	MCC
1	87.35	86.52	86.98	0.6245
2	87.21	86.04	85.96	0.6312
3	89.87	85.12	85.22	0.6327
4	85.76	86.12	86.14	0.6210
5	88.21	85.32	85.62	0.6212

Table 3: The result of five-fold cross validation

Although the performance value of the five-fold cross-validation is lower than the performance values of the three independent tests, this does not fully explain that the model trained on the combined data set cannot be used, because the five-fold cross-validation and independent testing use The training set and the test set are different and cannot be directly compared, and Sp, Sn and Acc are all higher than 85%, indicating that the prediction results of the model still have a certain reference value.

Conclusion

We built an antimicrobial peptide recognition model based on BERT, which is pre-trained on the data provided by UniProt, and then fine-tuned on a specific antimicrobial peptide data set. Experimental results show that the generalization ability of this model is better than other models in identifying antimicrobial peptides. In addition, 4 data sets of antimicrobial peptides were sorted and used to train a model for predicting antimicrobial peptides.

The final training model uses all the antimicrobial peptide data sets, but the results of the five-fold cross verification show that the performance indicators are not particularly good. The possible reason is that the current method of constructing the data set is flawed. Therefore, a method for constructing a complete antimicrobial peptide data set or a complete antimicrobial peptide data set is needed. If there is no such method, then the true performance test method of the model should be obtained through testing and verification by experimenters, but this is inefficient.

References

Boman, H. 2003. Antibacterial peptides: basic facts and emerging concepts. *Journal of internal medicine* 254(3): 197–215.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Dua, M.; Veltri, D.; Bishop, B.; and Shehu, A. 2018. Guiding Exploration of Antimicrobial Peptide Space with a Deep Neural Network. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2082–2087. IEEE.

Fjell, C. D.; Jenssen, H.; Hilpert, K.; Cheung, W. A.; Pante, N.; Hancock, R. E.; and Cherkasov, A. 2009. Identification of novel antibacterial peptides by chemoinformatics and machine learning. *Journal of medicinal chemistry* 52(7): 2006–2015.

Gulordava, K.; Bojanowski, P.; Grave, E.; Linzen, T.; and Baroni, M. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138* .

Lee, E. Y.; Fulan, B. M.; Wong, G. C.; and Ferguson, A. L. 2016. Mapping membrane activity in undiscovered peptide sequence space using machine learning. *Proceedings of the National Academy of Sciences* 113(48): 13588–13593.

Lin, Y.; Cai, Y.; Liu, J.; Lin, C.; and Liu, X. 2019. An advanced approach to identify antimicrobial peptides and their function types for penaeus through machine learning strategies. *BMC bioinformatics* 20(8): 291.

Linzen, T.; Dupoux, E.; and Goldberg, Y. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4: 521–535.

Meher, P. K.; Sahu, T. K.; Saini, V.; and Rao, A. R. 2017. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Scientific reports* 7(1): 1–12.

Randou, E. G.; Veltri, D.; and Shehu, A. 2013. Binary response models for recognition of antimicrobial peptides. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, 76–85.

Steiner, H.; Hultmark, D.; Engström, Å.; Bennich, H.; and Boman, H. 1981. Sequence and specificity of two antibacterial proteins involved in insect immunity. *Nature* 292(5820): 246–248.

Torrent, M.; Andreu, D.; Nogués, V. M.; and Boix, E. 2011. Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. *PloS one* 6(2): e16968.

Veltri, D.; Kamath, U.; and Shehu, A. 2015. Improving recognition of antimicrobial peptides and target selectivity through machine learning and genetic programming. *IEEE/ACM transactions on computational biology and bioinformatics* 14(2): 300–313.

Veltri, D.; Kamath, U.; and Shehu, A. 2018. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* 34(16): 2740–2747.

Xiao, X.; Wang, P.; Lin, W.-Z.; Jia, J.-H.; and Chou, K.-C. 2013. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Analytical biochemistry* 436(2): 168–177.

Yoshida, M.; Hinkley, T.; Tsuda, S.; Abul-Haija, Y. M.; McBurney, R. T.; Kulikov, V.; Mathieson, J. S.; Reyes, S. G.; Castro, M. D.; and Cronin, L. 2018. Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. *Chem* 4(3): 533–543.

Youmans, M.; Spainhour, J. C.; and Qiu, P. 2019. Classification of Antibacterial Peptides using Long Short-Term Memory Recurrent Neural Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* .

Zelezetsky, I.; Pontillo, A.; Puzzi, L.; Antcheva, N.; Segat, L.; Pacor, S.; Crovella, S.; and Tossi, A. 2006. Evolution of the Primate Cathelicidin Correlation between Structural Variations and Antimicrobial Activity. *Journal of Biological Chemistry* 281(29): 19861–19871.