Mathematical Biosciences and Engineering

MBE, 18(5): 5943–5958.
DOI: 10.3934/mbe.2021297
Received: 18 May 2021
Accepted: 26 June 2021
Published: 30 June 2021

http://www.aimspress.com/journal/MBE

Research article

Prediction of presynaptic and postsynaptic neurotoxins based on feature extraction

Wen Zhu[1,3,4], Yuxin Guo[1,3,4,*] and Quan Zou[2]

[1] Key Laboratory of Computational Science and Application of Hainan Province, Haikou, China
[2] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China
[3] Key Laboratory of Data Science and Intelligence Education, Hainan Normal University, Ministry of Education, Haikou, China
[4] School of Mathematics and Statistics, Hainan Normal University, Haikou, China

* **Correspondence:** Email: a282029@163.com.

Abstract: A neurotoxin is essentially a protein that mainly acts on the nervous system; it has a selective toxic effect on the central nervous system and neuromuscular nodes, can cause muscle paralysis and respiratory paralysis, and has strong lethality. According to their principle of action, neurotoxins are divided into presynaptic neurotoxins and postsynaptic neurotoxins. Correctly identifying presynaptic and postsynaptic nerve toxins provides important clues for future drug development and the discovery of drug targets. Therefore, a predictive model, Neu_LR, was constructed in this paper. The monoMonokGap method was used to extract the frequency characteristics of presynaptic and postsynaptic neurotoxin sequences and carry out feature selection, then, based on the important features obtained after dimensionality reduction, the prediction model Neu_LR was constructed using a logistic regression algorithm, and ten-fold cross-validation and independent test set validation were used. The final accuracy rates were 99.6078 and 94.1176%, respectively, which proved that the Neu_LR model had good predictive performance and robustness, and could meet the prediction requirements of presynaptic and postsynaptic neurotoxins. The data and source code of the model can be freely download from https: //github.com/gyx123681/.

Keywords: neurotoxin; monoMonokGap; logistic regression; protein classification; Neu_LR

1. Introduction

Neurotoxins are toxic substances that are destructive to nerve tissue, such as AF64A, 6-hydroxydopamine, and kainic acid. In principle, the toxins act on the ion channels in the nerve-muscle junction, destroying cholinergic neurons, inhibiting the release of acetylcholine, blocking nerve-muscle conduction, causing muscle weakness, and thus making the muscle unable to contract. In severe cases, these toxins can cause suffocation and death. Neurotoxins can be classified into presynaptic and postsynaptic types according to their mechanism of action [1]. Presynaptic neurotoxins mainly act on the presynaptic membrane [2]; due to the specificity of enzyme activity, they typically block neuromuscular transmission and inhibit the release of neurotransmitters. The targets of postsynaptic neurotoxins are located in the postsynaptic membrane and can bind to acetylcholine receptors [3]. For example, β-methylamino-L-alanine, also known as BMAA, can damage motor neurons and has been implicated in Parkinson's syndrome. Cobra neurotoxin is a short-chain neurotoxin, the most important lethal component in cobra venom, a mainly postsynaptic neurotoxin. Because cobra venom neurotoxin is nonaddictive and non-drug resistant, it has broad prospects for persons with a drug addiction in detoxification. Therefore, the study of presynaptic and postsynaptic neurotoxins will contribute to the development of medicine, for example, to provide important clues for drug design [4–6].

Neurotoxins are a type of protein, and while their structure and function can be correctly predicted through biochemical experiments, the work is time-consuming and expensive [7–10]. In the genome era, many biological sequences are available [11], giving us a variety of methods to predict protein structure and function [12–15]. The key to correctly predicting protein structure and function is how to analyze these features using computational methods. Therefore, we can use machine learning methods for protein type prediction [16]. Generally, the use of machine learning to predict biological sequences mainly includes the following steps: feature extraction, model construction, and performance evaluation [17–22]. In 2009, a diversity-based method of identifying presynaptic and postsynaptic neurotoxins was proposed. The algorithm is based on the composition of amino acids and pseudo-amino acids [23]. To further improve the accuracy of prediction, Hua Tang et al. proposed a new feature selection technique based on the principle of variance analysis (ANOVA) [7,24]. In this article, we constructed a predictive model, Neu_LR, to correctly identify presynaptic and postsynaptic neurotoxins. The monoMonokGap method was used to extract the frequency characteristics of presynaptic and postsynaptic neurotoxin sequences and carry out feature selection. Then, based on the important features obtained after dimensionality reduction, the logistic regression algorithm is used to construct the prediction model Neu_LR.

2. Material and methods

2.1. The main flow of the article

As the effectiveness of machine learning technology has been continuously verified in recent years, the prediction of protein classification using machine learning-related technology has become a new research category [25,26]. The key to protein classification prediction with the help of machine learning technology lies in data processing and classification algorithms. The general prediction process is to first use the algorithm to extract the features of the protein and then use different classifiers

to predict the protein. Therefore, the effective combination of feature extraction algorithms and classifiers has been extensively studied [27–31].

The research contents of this study include. We first download the presynaptic and postsynaptic neurotoxins from the UniProt database, and then the monoMonokGap feature expression algorithm is used to extract and select features from the data set to get the optimal features. Second, the feature vector obtained by dimensionality reduction was taken as the input, and the model was built using a logistic regression algorithm, and ten-fold cross-validation and independent test set validation was carried out. Figure 1 shows the flow chart of building the model in this paper [32,33].

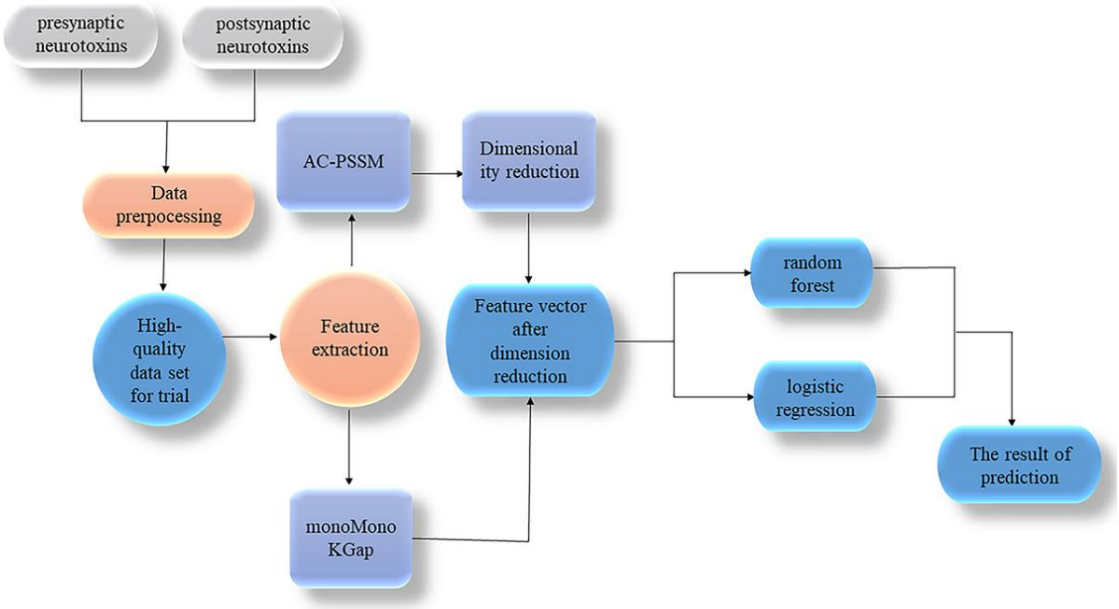

Figure 1. Main flowchart for predicting presynaptic and postsynaptic neurotoxin proteins.

2.2. Dataset

High-quality data sets are the basis for building reliable and accurate models [34,35]. The UniProt database provides the scientific community with a single, centralized, authoritative source of protein sequence and functional information [36]. The data set used in this article is also applied to the research of Hua Tang et al. A total of 91 presynaptic and 165 postsynaptic neurotoxins were downloaded from the UniProt database. Since fuzzy information will reduce the quality of the benchmark data set and will cause the predicted model to become unreliable, we must eliminate unknown residues in the protein sequence (such as "X", "Z", "J", "O" and "B"). Because of the highly similar protein sequences in the data set, the results can be overestimated; therefore, the cut-off value of sequence identity is set to 80%. According to the results of the above screening, our data set contains 90 presynaptic neurotoxins, 165 postsynaptic neurotoxins and a total of 255 types of neurotoxin samples can be expressed by the following formula:

$$S = S_{pr} \cup S_{po}$$

where subset S_{pr} is a collection of 90 presynaptic neurotoxins and S_{po} is a collection of 165 postsynaptic neurotoxins.

Each protein sequence can be expressed by the following formula:

$$R=r_1r_2r_3\cdots r_L$$

R stands for protein sequence, r_i stands for representative residue, and L is the length of the protein sequence. Since some machine learning methods cannot directly learn R, the protein sequences have to be converted to fixed-length vectors [37].

2.3. Feature selection

As the first step in building a biological sequence analysis model, feature extraction is an important part of the correct prediction of protein sequences. Generally, we can use the feature extraction method to convert the input neurotoxin sample order into a fixed-length digital vector and then use MRMD2.0 to reduce the dimensionality of the obtained feature vector as needed. Finally, we use the reduced dimensionality vector as the input vector of the classifier model and classify and process the result [38].

2.3.1. monoMonoKGap

The monoMonoKGap is a feature extraction method, and the best features can be generated from a large number of previously generated features, monoMonoKGap considers kGap in the nucleotide sub-sequence, frequencies of these sub-sequences are treated as prediction features. It can be used for feature extraction of DNA sequences, RNA sequences, and protein sequences [39]. The selection range of kGap in DNA sequences and RNA sequences is 1 to 5, and the selection range of kGap in protein sequences is 1 to 10. When the kGap value is small, the formation of the feature set is small, and the frequency of occurrence of these features retains the partial or short information sequence order, and when the kGap value is moderately large, more features will be generated to retain long sequence information [40]. Specifically, when $\mathrm{kGap}=1$, the sequence can be encoded as frequencies of X_X, Simultaneously generate $4\times1\times4$ dimensional features, or $20\times1\times20$ dimensional features, when $\mathrm{kGap}=2$, generate $4\times2\times4$ dimensional features, or $20\times2\times20$ dimensional features and so on [41]. That is when $\mathrm{kGap}=\mathrm{n}$, the DNA sequence and RNA sequence will produce $4\times4\times n$ features, and the protein sequence will produce $20\times20\times n$ features. The generated feature format is as follows:
when $\mathrm{kGap}=1$, the characteristic structure is X_X;
when $\mathrm{kGap}=2$, the characteristic structure is X_X and $X\_\_X$; and so on.
where X is defined as:

$$X=\begin{cases}\{A,C,G,T\},\\ \{A,C,G,U\}\\ \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}\end{cases}$$

This is followed by the DNA sequence, RNA sequence, and protein sequence.

For feature selection, to reduce the negative impact of dimensionality and to maintain information features, the AdaBoost classification model can be used to calculate the average impurity reduction. AdaBoost is a popular boosting classification algorithm in data mining. The core idea is to train multiple weak classifiers on the same training set, set these weak classifiers, and finally build a strong classifier [42]. Since Freund and Schapire proposed the AdaBoost algorithm [43–45], the improvement of AdaBoost has mainly involved two aspects: 1) adjusting the weight of the weak classifier in a new way and 2) improving the training method to reduce the error rate of classifier or save the training time.

2.3.2. Profile-based Auto covariance (AC-PSSM)

Each residue in the protein sequence has many physical and chemical properties, so the protein sequence can be regarded as a time series with corresponding properties [46]. PSI-BIAST is run to compare the data set and parameters to generate the outline of each sequence.

AC-PSSM [47] uses PSI-BIAST [48] as a search tool. PSI-BIAST provides a means to detect distant relationships between proteins. It is a protein sequence profile search method used to compare more sensitive protein sequences to other protein sequences. The database used is the nr library.

AC-PSSM can convert PSSMs of different lengths into vectors of fixed length. AC measures the correlation between two residues with the same property, which can be expressed as:

$$\begin{cases} AC(i,lag) = \sum_{j=1}^{L-lag} \left(S_{i,j} - \overline{S}_i\right)\left(S_{i,j+lag} - \overline{S}_i\right) / (L - lag) \\ \overline{S}_i = \sum_{j=1}^{L} S_{i,j} / L \end{cases} \tag{1}$$

where i is one of the residues, L is the length of the protein sequence, $S_{i,j}$ is the PSSM fraction of amino acid i at the jth position, and $\overline{S}_i$ is the average fraction of amino acid i in the entire protein sequence. In this way, the number of ACs can be calculated by $20 \times LAG$, where LAG is the maximum hysteresis, where the value of lag is all integers from 1 to LAG. In this article, we set the value of LAG to the default value of 2, that is, the maximum hysteresis is 2. Figure 2 shows the flowchart for generating AC-PSSM when the hysteresis is 2. That is, when $LAG = 2$, AC-PSSM can generate 40 columns of feature vectors. The first 20 columns represent the characteristics of the lagging item in the PSSM matrix when lag=1, expressed as

$$(A,A,1),(R,R,1),(N,N,1),\cdots,(V,V,1)$$

The last 20 columns represent the characteristics of the two lags in the PSSM matrix when lag=2, expressed as

$$(A,A,2),(R,R,2),(N,N,2),\cdots,(V,V,2)$$

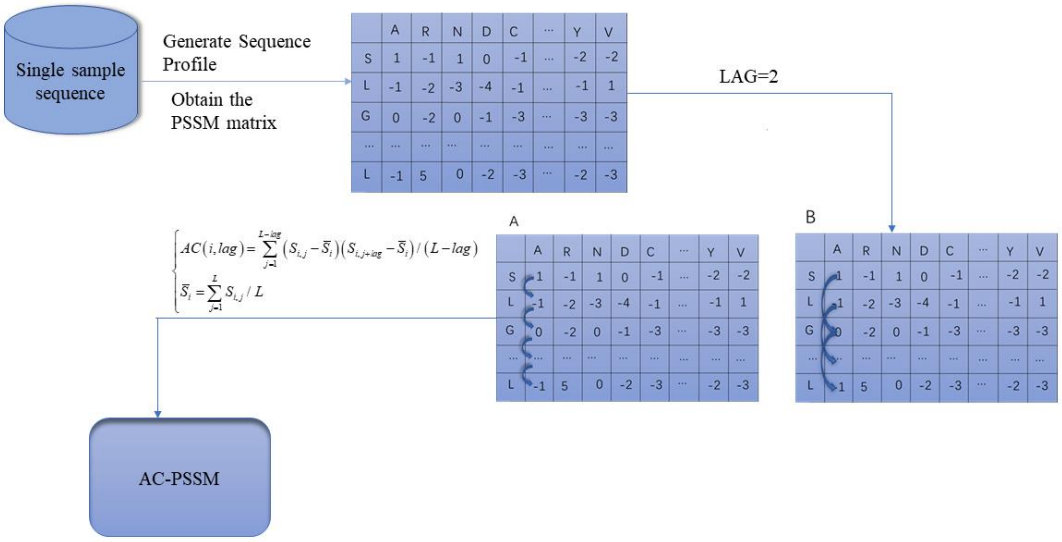

Figure 2. A flowchart of AC-PSSM generation. When $LAG=2$, a 40-column feature vector is generated. (A) First generate the first 20 columns, that is, the features when one item is lagging. (B) Secondly, the characteristics of the last 20 columns that are lagging two items are generated.

2.3.3. Maximum correlation maximum distance (MRMD2.0)

Feature selection is also known as variable selection or attribute selection, defined as the process of selecting the feature that contributes the most to the predictor variable or output of interest. After extracting features from the sequence, the MRMD2.0 algorithm is used for feature selection.

MRMD2.0 is based on the PageRank algorithm. It is not only a Python-based tool for reducing the dimensionality of data sets but also draws performance curves based on feature dimensions. Accuracy can be sacrificed for fewer features by selecting dimensions from the performance curve [49,50].

2.4. Classification options and tools

The basic idea of classification is to learn the parameters of the classifier through training data, and the goal of classification is to train the parameters of the classifier with the training set with the smallest loss of accuracy [51,52]. Weka (3.8.5) can be used for data mining and prediction models. There are many different classification modes, such as random forest and Bayesian classifiers [27,49,53].

2.4.1. Random forest (RF)

The random forest algorithm is an ensemble algorithm, which is composed of multiple decision tree classifiers, and each subclassifier is a CART classification regression tree; therefore, classification and regression are performed using random forest. This algorithm is highly resistant to overfitting: the

risk of overfitting can be reduced by averaging the decision tree [54,55]. The advantages are its simple implementation, high accuracy, fast training speed, strong anti-overfitting ability, and suitability as a benchmark model. The disadvantage is that the model is prone to overfitting on some sample sets with relatively large noise. When there are many decision trees, the training time and space will be relatively large.

2.4.2. Logistic regression (LR)

Logistic regression is a machine learning method used to solve binary classification problems, and it is a generalized linear regression [56]. A hyperplane can be established to classify samples, which can be described by the following formula:

$$h_\theta(x) = g\left(\theta^T x\right) = \frac{1}{1+e^{-\theta Tx}} \tag{2}$$

where X is the sample $x=\left[x_1, x_2, \cdots, x_n\right]$ is a n-dimensional vector, g is a logistic function, and the general form is $g(z)=\frac{1}{1+e^{-z}}$. The advantages are that the computational cost is not high and it is easy to understand and implement, but the disadvantages are that it is easy to underfit and the classification accuracy may not be high.

2.5. Performance evaluation

In this article, k-fold cross-validation was chosen to test predictions. Specificity (SP), sensitivity (SN), and accuracy (ACC) [57–61] were used to evaluate our proposed method [62]; they can be expressed as:

$$SN = 1 - \frac{N_{po}^{pr}}{N^{pr}} \tag{3}$$

$$SP = 1 - \frac{N_{pr}^{po}}{N^{po}} \tag{4}$$

$$ACC = 1 - \frac{N_{po}^{pr} + N_{pr}^{po}}{N^{pr} + N^{po}} \tag{5}$$

where N^{pr} and N^{po} represent presynaptic neurotoxin and postsynaptic neurotoxin, respectively; N_{po}^{pr} indicates that the presynaptic neurotoxin was incorrectly predicted as a postsynaptic neurotoxin; and N_{pr}^{po} indicates that a postsynaptic neurotoxin was incorrectly predicted as a presynaptic neurotoxin.

3. Results and discussion

Previous studies have shown that feature extraction is very important for predictor variables, and optimized features can improve the accuracy of model prediction [63–67]. Especially in some high-dimensional data, there may be some noise and redundant information, which will have some negative

effects on the prediction.

3.1. Performance comparison of different feature expressions

In this section, the prediction results of monoMonokGap, AC-PSSM, 188D characteristics, Geary-related characteristics, and amino acid composition (AAC) [68,69] characteristics under logistic regression and random forest were compared. The results are shown in Tables 1 and 2 (where the maximum value is indicated in bold). It can be seen from Tables 1 and 2 that the monoMonokGap feature used in this model has the best performance in all indicators. Under the random forest algorithm, when $kGap = 7$, monoMonokGap ($kGap = 7$) performs the best; Under the logistic regression algorithm, when $kGap = 9$, monoMonokGap ($kGap = 9$) achieves the best result. Compared with the two results, $kGap = 9$ had the best performance on LR, among which the values of ACC, AUC, SP, and SN were 99.6078%, 0.996, 0.998, and 0.996, respectively. This also proves the effectiveness of monoMonokGap ($kGap = 9$), so we choose monoMonokGap ($kGap = 9$) as the feature expression method of the model.

Table 1. Different feature representation methods are used to obtain the results of using RF classifier under 5-fold cross-validation.

Method	SN	SP	AUC	ACC (%)
monoMonokGap ($kGap = 1$)	0.910	0.870	0.966	90.9804
monoMonokGap ($kGap = 3$)	0.941	0.912	0.985	94.1176
monoMonokGap ($kGap = 5$)	0.945	0.920	**0.992**	94.5098
monoMonokGap ($kGap = 7$)	**0.957**	**0.936**	0.991	**95.6863**
monoMonokGap ($kGap = 9$)	0.949	0.917	0.992	94.902
AC-PSSM	0.839	0.776	0.868	83.9216
188D	0.710	0.609	0.794	70.9804
Geary	0.533	0.417	0.481	53.3333
AAC	0.686	0.602	0.736	68.6275

To further verify the stability of monoMonokGap, we tested it with an independent test set [70] and compared it with AC-PSSM, 188D characteristics, Geary-related characteristics, AAC and other feature expression methods. The results are shown in Tables 3 and 4. Among them, 80% of the randomly selected data sets are used to train the prediction model, and the remaining 20% are used to test the model.

It can be seen from Tables 3 and 4 that, compared with other feature expression methods, the monoMonokGap feature expression method selected in this article has little difference in the results of independent verification set to test and the results of ten-fold cross-validation, which also proves that the feature expression method selected in this paper does not have over-fitting and has good stability.

Table 2. Different feature representation methods are used to obtain the results of using LR classifier under 10-fold cross-validation.

Method	SN	SP	AUC	ACC (%)
monoMonokGap (kGap = 1)	0.929	0.901	0.971	92.9412
monoMonokGap (kGap = 3)	0.925	0.889	0.988	92.549
monoMonokGap (kGap = 5)	0.984	0.986	0.994	98.4314
monoMonokGap (kGap = 7)	0.973	0.980	0.995	97.2549
monoMonokGap (kGap = 9)	**0.996**	**0.998**	**0.996**	**99.6078**
AC-PSSM	0.780	0.663	0.789	78.040
188D	0.631	0.37	0.603	63.1373
Geary	0.627	0.428	0.684	62.7451
AAC	0.667	0.460	0.637	66.6667

Table 3. The performance comparison of different feature algorithms under LR was verified by an independent test set.

Method	SN	SP	AUC	ACC (%)
monoMonokGap (kGap = 1)	0.824	0.828	0.931	82.3592
monoMonokGap (kGap = 3)	0.882	0.835	0.894	88.22353
monoMonokGap (kGap = 5)	0.980	0.989	0.985	98.0392
monoMonokGap (kGap = 7)	0.961	0.979	0.998	96.0784
monoMonokGap (kGap = 9)	0.941	0.968	0.977	94.1176
AC-PSSM	0.686	0.677	0.709	68.6275
188D	0.373	0.658	0.608	37.2549
Geary	0.588	0.397	0.396	58.8235
AAC	0.604	0.445	0.548	60.3774

3.2. *Comparison with other classifiers*

In this section, the performance of the model constructed in this article is compared with RF, Logical Model Tree (LMT), J48, Bayesnet, NaiveBayes, Sequential Minimal Optimization (SMO), and other classifiers. The results of ten-fold cross-validation are shown in Table 5 (the maximum value is indicated in bold). It can be seen from Table 5 that the results of the model constructed in this paper are significantly better than those of other classifiers in terms of various indicators, with the values of

ACC, AUC, SP, and SN being 99.6078%, 0.996, 0.998 and 0.996 respectively. This also proves the validity of the model constructed in this article.

Table 4. The performance comparison of different feature algorithms under RF was verified by an independent test set.

Method	SN	SP	AUC	ACC (%)
monoMonokGap (kGap = 1)	0.804	0.742	0.872	80.3922
monoMonokGap (kGap = 3)	0.804	0.716	0.923	80.3922
monoMonokGap (kGap = 5)	0.902	0.846	0.966	90.1961
monoMonokGap (kGap = 7)	0.882	0.810	0.977	88.2353
monoMonokGap (kGap = 9)	0.922	0.881	0.973	92.1569
AC-PSSM	0.725	0.598	0.697	72.549
188D	0.510	0.707	0.786	50.9804
Geary	0.529	0.390	0.412	52.9412
AAC	0.566	0.501	0.553	56.6038

Table 5. Comparison of performance of different classification algorithms under 10-fold cross-validation.

Classifier	SN	SP	AUC	ACC (%)
Neu_LR	**0.996**	**0.998**	**0.996**	**99.6078**
RF	0969	0.948	0.995	96.8627
LMT	0.918	0.884	0.971	91.7647
J48	0.863	0.804	0.877	86.2745
BayesNet	0.929	0.891	0.987	92.9412
NaiveBayes	0.863	0.779	0.945	86.2745
SMO	0.973	0.965	0.969	97.2549

Table 6. The performance comparison of different classification algorithms was verified by an independent test set.

Classifier	SN	SP	AUC	ACC (%)
Neu_LR	0.941	0.968	0.977	94.1176
RF	0.922	0.881	0.972	92.1569
LMT	0.784	0.756	0.886	78.4314
J48	0.843	0.788	0.824	84.3137
BayesNet	0.902	0.846	0.948	90.1961
NaiveBayes	0.804	0.691	0.920	80.3922
SMO	0.922	0.932	0.927	92.1569

To further verify the robustness of the model constructed in this article, we conducted independent test set verification on it and compared its performance with RF, LMT, J48, Bayesnet, NaiveBayes, SMO, and other classifiers. The results are shown in Table 6. Among them, 80% of the randomly selected data sets are used to train the prediction model, and the remaining 20% are used to test the model.

It can be seen from Table 6 that, compared to other algorithms, the model constructed in this paper achieves the best prediction article, and the values of ACC, AUC, SP, and SN are 94.1176%, 0.977, 0.968, and 0.941, respectively. Moreover, compared with the results of 10-fold cross-validation, the difference is very small, which also proves that the prediction performance of the model constructed in this article does not exist overfitting and has good stability.

3.3. *Comparison with state-of-the-art predictors*

This section compares the model constructed in this paper with other existing methods. The comparison results are shown in Table 7, where the results of ID [23] and ANOVA [7] are obtained directly from literature. It can be seen from Table 7 that the model Neu_LR constructed in this paper has the best performance in all indicators, among which ACC, SP, and SN reach the maximum value of 99.6078%, 0.998, and 0.996 respectively, and the effect is better than the other two methods, which also proves the effectiveness of the model Neu_LR constructed in this paper [71].

Table 7. Comparison of different methods for predicting presynaptic and postsynaptic neurotoxins.

Method	SN	SP	ACC (%)
ID [23]	88.46	91.30	89.80
ANOVA [7]	94.51	95.15	94.92
Neu_LR	0.996	0.998	99.6078

4. Conclusions

The correct understanding of presynaptic and postsynaptic neurotoxins is an essential first step in the discovery of drug targets and drug design. And protein prediction mainly involves two aspects, feature extraction, and selection of classification algorithms. Therefore, the prediction model Neu_LR was constructed in this article. The monoMonokGap method was used to extract the frequency characteristics of presynaptic and postsynaptic neurotoxin sequences, and the feature selection was carried out. Then, based on the important features obtained after dimension-reduction, the logistic regression algorithm was used to construct the prediction model Neu_LR. In this paper, we use 10-fold cross-validation and independent test set validation to judge whether the Neu_LR model is good or not. In the 10-fold cross-validation, we achieved 99.6078% accuracy, and in the independent test set validation, we achieved 94.1176% accuracy, which shows that our model is feasible and effective.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (Grant Nos 61863010, 11926205, 11926412, and 61873076), National Key R&D Program of China (No.2020YFB2104400) and Natural Science Foundation of Hainan, China (Grant Nos. 119MS036 and 120RC588), and Hainan Normal University 2020 Graduate Student Innovation Research Project

(hsyx2020-41), the Special Science Foundation of Quzhou (2020D003).

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. F. Afifiyan, A. Armugam, P. Gopalakrishnakone, N. H. Tan, C. H. Tan, K. Jeyaseelan, Four new postsynaptic neurotoxins from Naja naja sputatrix venom: cDNA cloning, protein expression, and phylogenetic analysis, *Toxicon*, **36** (1998), 1871–1885.
2. A. J. Alexandrou, R. S. Duncan, A. Sullivan, J. C. Hancox, D. J. Leishman, H. J. Witchel, et al., Mechanism of hERG K^+ channel blockade by the fluoroquinolone antibiotic moxifloxacin, *Brit. J. Pharmacol.*,**147** (2006), 905–916.
3. J. P. Forder, M. Tymianski, Postsynaptic mechanisms of excitotoxicity: Involvement of postsynaptic density proteins, radicals, and oxidant molecules, *Neuroscience*, **158** (2009), 293–300.
4. F. Li, M. Luo, W. Zhou, J. Li, X. Jin, Z. Xu, et al., Single cell RNA and immune repertoire profiling of COVID-19 patients reveal novel neutralizing antibody, *Protein Cell*, (2020), 1–5.
5. R. Su, X. Liu, L. Wei, Q. Zou, Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response, *Methods*, **166** (2019), 91–102.
6. R. Su, X. Liu, G. Xiao, L. Wei, Meta-GDBP: A high-level stacked regression model to improve anticancer drug response prediction, *Briefings Bioinf.*, **21** (2020), 996–1005.
7. H. Tang, Y. Yang, C. Zhang, R. Chen, P. Huang, C. Duan, et al., Predicting presynaptic and postsynaptic neurotoxins by developing feature selection technique, *BioMed. Res. Int.*, **2017** (2017), 3267325.
8. Y. Ding, J. Tang, F. Guo, Identification of drug-target interactions via dual laplacian regularized least squares with multiple kernel fusion, *Knowl.-Based Syst.*, **204** (2020), 106254.
9. Y. Ding, J. Tang, F. Guo, Identification of drug-side effect association via multiple information integration with centered kernel alignment, *Neurocomputing*, **325** (2019), 211–224.
10. Z. Hong, X. Zeng, L. Wei, X. Liu, Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism, *Bioinformatics*, **36** (2020), 1037–1043.
11. Y. Shen, J. Tang, F. Guo, Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC, *J. Theor. Biol.*, **462** (2019), 230–239.
12. D. Zhang, H. D. Chen, H. Zulfiqar, S. Yuan, Q. Huang, Z. Zhang, et al., iBLP: An XGBoost-Based Predictor for Identifying Bioluminescent Proteins, *Comput. Math. Methods Med.*, **2021** (2021).
13. X. J. Zhu, C. Q. Feng, H. Y. Lai, W. Chen, L. Hao, Predicting protein structural classes for low-similarity sequences by evaluating different features, *Knowl.-Based Syst.*, **163** (2019), 787–793.
14. J. X. Tan, S. H. Li, Z. M. Zhang, C. Chen, W. Chen, H. Tang, et al., Identification of hormone binding proteins based on machine learning methods, *Math. Biosci. Eng.*, **16** (2019), 2466–2480.
15. Z. Guo, P. Wang, Z. Liu, Y. Zhao, Discrimination of thermophilic proteins and non-thermophilic proteins using feature dimension reduction, *Front. Bioeng. Biotechnol.*, **8** (2020), 584807.

16. L. Cheng, Y. Hu, J. Sun, M. Zhou, Q. Jiang, DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function, *Bioinformatics*, **34** (2018), 1953–1956.
17. B. Liu, X. Wang, Q. Zou, Q. Dong, Q. Chen, Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation, *Mol. Inf.*, **32** (2013), 775–782.
18. X. Zeng, Y. Zhong, W. Lin, Q. Zou, Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods, *Briefings Bioinf.*, **21** (2020), 1425–1436.
19. S. Jin, X. Zeng, F. Xia, W. Huang, X. Liu, Application of deep learning methods in biological networks, *Briefings Bioinf.*, **22** (2021), 1902–1917.
20. B. Liu, X. Gao, H. Zhang, BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches, *Nucleic Acids Res.*, **47** (2019), e127.
21. J. Shao, K. Yan, B. Liu, FoldRec-C2C: protein fold recognition by combining cluster-to-cluster model and protein similarity network, *Briefings Bioinf.*, **22** (2021).
22. L. Yu, M. Wang, Y. Yang, F. Xu, X. Zhang, F. Xie, et al., Predicting therapeutic drugs for hepatocellular carcinoma based on tissue-specific pathways, *PLoS Comput. Biol.*, **17** (2021), e1008696.
23. Y. Lei, Q. Li, Prediction of presynaptic and postsynaptic neurotoxins by the increment of diversity, *Toxicol. Vitro*, **23** (2009), 346–348.
24. X. Zhao, Q. Jiao, H. Li, Y. Wu, H. Wang, S. Huang, et al., ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles, *BMC Bioinf.*, **21** (2020), 43.
25. R. Su, J. Hu, Q. Zou, B. Manavalan, L. Wei, Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools, *Briefings Bioinf.*, **21** (2020), 408–420.
26. L. Yu, D. Zhou, L. Gao, Y. Zha, Prediction of drug response in multilayer networks based on fusion of multiomics data, *Methods*, 2020.
27. J. Zhang, Y. Ju, H. Lu, P. Xuan, Q. Zou, Accurate identification of cancerlectins through hybrid machine learning technology, *Int. J. Genomics*, **2016** (2016).
28. X. Zeng, W. Lin, M. Guo, Q. Zou, A comprehensive overview and evaluation of circular RNA detection tools, *Plos Comput. Biol.*, **13** (2017), e1005420.
29. J. Shao, B. Liu, ProtFold-DFG: protein fold recognition by combining Directed Fusion Graph and PageRank algorithm, *Briefings Bioinf.*, **22** (2021).
30. Y. Shang, L. Gao, Q. Zou, L. Yu, Prediction of drug-target interactions based on multi-layer network representation learning, *Neurocomputing*, **434** (2021), 80–89.
31. X. Pan, H. Li, T. Zeng, Z. Li, L. Chen, T. Huang, et al., Identification of protein subcellular localization with network and functional embeddings, *Front. Genet.*, **11** (2021), 626500.
32. L. Wei, S. Wan, J. Guo, K. Wong, A novel hierarchical selective ensemble classifier with bioinformatics application, *Artif. Intell. Med.*, **83** (2017), 82–90.
33. W. Yu, Z. Jiang, J. Wang, R. Tao, Using feature selection technique for drug-target interaction networks prediction, *Current Med. Chem.*, **18** (2011), 5687–5693.
34. W. Su, M. L. Liu, Y. H. Yang, J. Wang, S. Li, H. Lv, et al., PPD: A Manually Curated Database for Experimentally Verified Prokaryotic Promoters, *J. Mol. Biol.*, **433** (2021), 166860.

35. Z. Y. Liang, H. Lai, H. Yang, C. Zhang, H. Yang, H. Wei, et al., Pro54DB: a database for experimentally verified sigma-54 promoters, *Bioinformatics*, **33** (2017), 467–469.
36. The UniProt Consortium, The universal protein resource (UniProt) in 2010, *Nucleic Acids Res.*, **38** (2010), D142–148.
37. B. Liu, BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches, *Briefings Bioinf.*, **20** (2019), 1280–1294.
38. B. Manavalan, S. Basith, T. H. Shin, D. Y. Lee, L. Wei, G. Lee, 4mCpred-EL: an ensemble learning framework for identification of DNA N4-methylcytosine sites in the mouse genome, *Cells*, **8** (2019), 1332.
39. M. Mandal, A. Mukhopadhyay, U. Maulik, Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou's PseAAC, *Med. Biol. Engi. Comput.*, **53** (2015), 331–344.
40. R. Muhammod, S. Ahmed, D. M. Farid, S. Shatabda, A. Sharma, A. Dehzangi, et al., PyFeat: a Python-based effective feature generation tool for DNA, RNA and protein sequences, *Bioinformatics*, **35** (2019), 3831–3833.
41. L. Dou, X. Li, H. Ding, L. Xu, H. Xiang, Prediction of m5C modifications in RNA sequences by combining multiple sequence features, *Mol. Ther. Nucleic Acids*, **21** (2020), 332–342.
42. E. Teimoury, M. R. Gholamian, B. Masoum, M. Ghanavati, An optimized clustering algorithm based on K-means using Honey Bee Mating algorithm, *Sensors*, **16** (2016), 1–19.
43. Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, *J. Comput. Syst. Sci.*, **55** (1997), 119–139.
44. Y. Freund, R. E. Schapire, Experiments with a new boosting algorithm, in *icml*, **96** (1996), 148–156.
45. L. Cai, X. Ren, X. Fu, L. Peng, M. Gao, X. Zeng, iEnhancer-XG: Interpretable sequence-based enhancers and their strength predictor, *Bioinformatics*, **37** (2021). 1060–1067.
46. Q. Dong, S. Zhou, J. Guan, A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation, *Bioinformatics*, **25** (2009), 2655–2662.
47. B. Liu, H. Wu, K. C. Chou, Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences, *Nat. Sci.*, **9** (2017).
48. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25** (1997), 3389–3402.
49. L. Yu, Y. Shi, Q. Zou, S. Wang, L. Zheng, L. Gao, Exploring drug treatment patterns based on the action of drug and multilayer network model, *Int. J. Mol. Sci.*, **21** (2020), 5014.
50. Z. Tao, Y. Li, Z. Teng, Y. Zhao, A method for identifying vesicle transport proteins based on LibSVM and MRMD, *Comput. Math. Methods Med.*, **2020** (2020), 1–9.
51. I. M. Javed, F. Ibrahima, S. B. Belhaouari, A. M. Said, Efficient feature selection and classification of protein sequence data in bioinformatics, *Sci. World J.*, **2014** (2014), 314–319.
52. L. Xu, G. Liang, L. Wang, C. Liao, A novel hybrid sequence-based model for identifying anticancer peptides, *Genes*, **9** (2018), 158.
53. Y. H. Zhang, H. Li, T. Zeng, L. Chen, Z. Li, T. Huang, et al., Identifying transcriptomic signatures and rules for SARS-CoV-2 infection, *Front. Cell Dev. Biol.*, **8** (2021), 627302.

54. X. Zhou, T. Liu, D. Yan, X. Shi, X. Jin, An action-based Markov chain modeling approach for predicting the window operating behavior in office spaces, in *Building Simulation*, **14** (2021), 301–315.
55. Y. H. Zhang, T. Zeng, L. Chen, T. Huang, Y. Cai, Determining protein-protein functional associations by functional rules based on gene ontology and KEGG pathway, *Biochim. Biophys. Acta (BBA)-Proteins Proteomics*, **1869** (2021), 140621.
56. H. Yang, Y. Luo, X. Ren, M. Wu, X. He, B. Peng, et al., Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators, *Inf. Fusion*, **75** (2021), 140–149.
57. H. Wang, Y. Ding, J. Tang, F. Guo, Identification of membrane protein types via multivariate information fusion with Hilbert-Schmidt Independence Criterion, *Neurocomputing*, **383** (2020), 257–269.
58. Y. Ding, J. Tang, F. Guo, Identification of drug-target interactions via fuzzy bipartite local model, *Neural Comput. Appl.*, **32** (2020), 10303–10319.
59. R. Su, H. Wu, X. Bo, X. Liu, L. Wei, Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **16** (2018), 1231.
60. L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, F. Guo, Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier, *Artif. Intell. Med.*, **83** (2017), 67–74.
61. X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, F. Cheng, deepDR: a network-based deep learning approach to in silico drug repositioning, *Bioinformatics*, **35** (2019), 5191–5198.
62. R. W. Snow, C. A. Guerra, A. M. Noor, H. Y. Myint, S. I. Hay, The global distribution of clinical episodes of Plasmodium falciparum malaria, *Nature*, **434** (2005), 214–217.
63. H. Wang, J. Tang, Y. Ding, F. Guo, Exploring associations of non-coding RNAs in human diseases via three-matrix factorization with hypergraph-regular terms on center kernel alignment, *Briefings Bioinf.*, 2021.
64. J. Li, Y. Pu, J. Tang, Q. Zou, F. Guo, DeepATT: a hybrid category attention neural network for identifying functional effects of DNA sequences, *Briefings Bioinf.*, **22** (2021).
65. Y. Shen, Y. Ding, J. Tang, Q. Zou, F. Guo, Critical evaluation of web-based prediction tools for human protein subcellular localization, *Briefings Bioinf.*, **21** (2020), 1628–1640.
66. X. Fu, L. Cai, X. Zeng, Q. Zou, StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency, *Bioinformatics*, **36** (2020), 3028–3034.
67. L. Yu, F. Xu, L. Gao, Predict new therapeutic drugs for hepatocellular carcinoma based on gene mutation and expression, *Front. Bioeng. Biotechnol.*, **8** (2020).
68. L. Cai, L. Wang, X. Fu, C. Xia, X. Zeng, Q. Zou, ITP-Pred: an interpretable method for predicting, therapeutic peptides with fused features low-dimension representation, *Briefings Bioinf.*, 2020.
69. Z. Chen, P. Zhao, F. Li, A. Leier, T. T. Marquez-Lago, Y. Wang, et al., iFeature: a python package and web server for features extraction and selection from protein and peptide sequences, *Bioinformatics*, **34** (2018), 2499–2502.
70. L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, Q. Zou, Improved and promising identification of human microRNAs by incorporating a high-quality negative set, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **11** (2013), 192–201.

71. L. Wei, H. Chen, S. Ran, M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning, *Mol. Ther. Nucleic Acids*, **12** (2018), 635–644.