Hindawi Publishing Corporation
BioMed Research International
Volume 2017, Article ID 3267325, 4 pages
http://dx.doi.org/10.1155/2017/3267325

Research Article

Predicting Presynaptic and Postsynaptic Neurotoxins by Developing Feature Selection Technique

Hua Tang,[1] Yunchun Yang,[2] Chunmei Zhang,[1] Rong Chen,[1] Po Huang,[1] Chenggang Duan,[1] and Ping Zou[1]

[1]*Department of Pathophysiology, Southwest Medical University, Luzhou 646000, China*
[2]*Department of Anesthesiology, The Affiliated Traditional Chinese Medical Hospital of Southwest Medical University, Luzhou 646000, China*

Correspondence should be addressed to Hua Tang; tanghua771211@aliyun.com and Ping Zou; lyzouping@163.com

Received 17 November 2016; Accepted 18 December 2016; Published 12 February 2017

Academic Editor: Ren-Zhi Cao

Presynaptic and postsynaptic neurotoxins are proteins which act at the presynaptic and postsynaptic membrane. Correctly predicting presynaptic and postsynaptic neurotoxins will provide important clues for drug-target discovery and drug design. In this study, we developed a theoretical method to discriminate presynaptic neurotoxins from postsynaptic neurotoxins. A strict and objective benchmark dataset was constructed to train and test our proposed model. The dipeptide composition was used to formulate neurotoxin samples. The analysis of variance (ANOVA) was proposed to find out the optimal feature set which can produce the maximum accuracy. In the jackknife cross-validation test, the overall accuracy of 94.9% was achieved. We believe that the proposed model will provide important information to study neurotoxins.

1. Introduction

Neurotoxins act typically against channels to block or enhance synaptic transmission. According to the mechanism of action, neurotoxins can be classified as presynaptic type and postsynaptic type [1]. The function of presynaptic neurotoxins is to act at the presynaptic membrane [2]. They usually block neuromuscular transmission and inhibit the neurotransmitter release due to their specific enzymatic activities [3]. Postsynaptic neurotoxins can bind to the postsynaptic membrane and acetylcholine receptors [4]. Thus, the study of presynaptic and postsynaptic neurotoxin will give us important clues for drug-target discovery and drug design.

The function and structure of neurotoxins can be correctly measured by biochemical experiments; however, it is time-consuming and costly. The availability of huge amounts of proteins generated in postgenomic age provides us with an important opportunity to design various computational methods for timely and precisely predicting protein functions. Thus, it is important to develop machine learning approach to predict presynaptic and postsynaptic neurotoxins. Recently, Yang and Li developed an increment of diversity-based method to identify presynaptic neurotoxin and postsynaptic neurotoxin [5]. The benchmark dataset including 78 presynaptic neurotoxins and 69 postsynaptic neurotoxins was downloaded from Animal Toxin Database (ATDB) [6]. The overall accuracy was 90.39% in jackknife cross-validation, which is far from satisfactory. Subsequently, Song proposed using bilayer support vector machine (SVM) to improve prediction accuracy based on a new benchmark dataset [7]. Although the overall accuracy was dramatically improved, the sequence identity of the dataset was so high that the results were overestimated.

To overcome the shortcoming mentioned above, in this study, we developed a new method based on feature selection technique to predict presynaptic neurotoxins and postsynaptic neurotoxins. In the following, we will introduce how to construct a new benchmark dataset, to formulate neurotoxin samples using peptide sequences, and to obtain the expected result produced by best feature subset.

2. Materials and Methods

2.1. Benchmark Dataset Construction. A high quality benchmark dataset is the fundamental for building a reliable and accuracy model. The Universal Protein Resource (UniProt) provides the scientific community with a single, centralized, authoritative resource for protein sequences and functional information [8]. Thus, we downloaded presynaptic and postsynaptic neurotoxins from the UniProt. Ambiguous information can reduce the quality of benchmark dataset which makes the prediction model unreliable. Thus, we must exclude the protein sequence which contains ambiguous residues (such as "X," "B," and "Z") and which is the fragment of other proteins. High similar sequences in benchmark dataset will bring about overestimation of results. Thus, the CD-HIT program was used to remove the highly similar sequences by setting the cutoff of sequence identity as 80% [9]. According to above screening procedure, the final benchmark dataset included 256 neurotoxin samples which can be formulated as

$$S = S_{\text{Pre}} \cup S_{\text{Pro}}, \tag{1}$$

where the subset S_{Pre} contains 91 presynaptic neurotoxins and S_{Pro} contains 165 postsynaptic neurotoxins.

2.2. The Dipeptide Composition. One of the most important steps in the prediction problem is to formulate neurotoxin sequences with an effective mathematical expression. Generally, we may formulate a neurotoxin by its entire residue sequence as follows:

$$\mathbf{P} = R_1 R_2 R_3 R_4 \cdots R_L, \tag{2}$$

where R denotes the residue of neurotoxin $\mathbf{P}$ and the subscript L is the number of residues of the neurotoxin $\mathbf{P}$. We may use some straightforward and intuitive tools, such as BLAST or FASTA, to find the similar sequences. However, these tools are only suitable for the query sequences which have high similar sequences in searching dataset. If there are no similar sequences in the training dataset, they cannot work well.

Machine learning approach can overcome such problem and correctly identify presynaptic and postsynaptic neurotoxins. Thus, we must convert neurotoxin sequences into discrete vector. A simplest method used to represent a neurotoxin is its residue composition containing a 20-dimension vector. However, the sequence order information would be completely lost and hence limit the prediction quality [10–13]. Thus, the dipeptide composition was used in this study. Accordingly, each neurotoxin sample in our benchmark dataset can be expressed as a 400-dimension vector and formulated as

$$\mathbf{P} = \left[x_1 \cdots x_u \cdots x_{400}\right]^T, \tag{3}$$

where x_u ($u = 1, 2, \ldots, 400$) is the occurrence frequency of uth dipeptide and given by

$$x_u = \begin{cases} f(AA) & \text{when } u = 1 \\ \vdots & \vdots \\ f(CA) & \text{when } u = 21 \\ \vdots & \vdots \\ f(YY) & \text{when } u = 400, \end{cases} \tag{4}$$

where $A, C, \ldots, W, Y$ are the single letter codes of 20 native amino acids, respectively. x_u can be calculated by

$$x_u = \frac{n_u}{\sum_u n_u}, \tag{5}$$

where n_u denotes the number of the uth dipeptides in the neurotoxin $\mathbf{P}$.

2.3. Support Vector Machine. SVM is a very popular machine learning method and has been widely used in bioinformatics [7, 14–18]. The basic idea of SVM is to transform the input vector into a high-dimension Hilbert space and to determine a separating hyperplane in this space. In this study, we used the LibSVM package 3.18 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) to implement SVM. Because it is more suitable for nonlinear classification, the radial basis function (RBF) defined as $K(\overrightarrow{p_i}, \overrightarrow{p_j}) = \exp(-\gamma \|\overrightarrow{p_i} - \overrightarrow{p_j}\|^2)$ was used as kernel function. In the SVM model construction, a grid search strategy with cross-validation test was used to optimize the regularization parameter C and kernel parameter γ as the following standard:

$$\begin{aligned} 2^{-5} < C < 2^{15} \quad & \text{with step of } 2, \\ 2^{-15} < \gamma < 2^{15} \quad & \text{with step of } 2^{-1}. \end{aligned} \tag{6}$$

2.4. Performance Evaluation. In this study, we used jackknife cross-validation to test the prediction. In the jackknife cross-validation test, each protein sample in the dataset is in turn singled out as an independent test sample and all the rule parameters are calculated based on the remaining proteins without including the one being identified. The performance of our proposed method was estimated by the following three indexes called sensitivity (Sn), specificity (Sp), and overall accuracy (Acc) which can be expressed as

$$\begin{aligned} \text{Sn} &= 1 - \frac{N_{\text{Pro}}^{\text{Pre}}}{N^{\text{Pre}}} \quad 0 \le \text{Sn} \le 1, \\ \text{Sp} &= 1 - \frac{N_{\text{Pre}}^{\text{Pro}}}{N^{\text{Pro}}} \quad 0 \le \text{Sp} \le 1, \\ \text{Acc} &= 1 - \frac{N_{\text{Pro}}^{\text{Pre}} + N_{\text{Pre}}^{\text{Pro}}}{N^{\text{Pre}} + N^{\text{Pro}}} \quad 0 \le \text{Acc} \le 1, \end{aligned} \tag{7}$$

where N^{Pre} and N^{Pro} are the total number of the presynaptic neurotoxins and postsynaptic neurotoxins. $N_{\text{Pro}}^{\text{Pre}}$ is the number of the presynaptic neurotoxins incorrectly predicted as

the postsynaptic neurotoxins and $N_{\text{Pre}}^{\text{Pro}}$ is the number of the postsynaptic neurotoxins incorrectly predicted as presynaptic neurotoxins.

3. Results and Discussion

Many published papers have demonstrated that the optimized features could improve predictive accuracy [19–25]. For high-dimension data, some features are noise or redundant information which has negative contribution to the prediction. Thus, it is very important to develop a feature selection technique to exclude the garbage information. The current study will introduce a new feature selection technique based on the principle of analysis of variance (ANOVA).

Two parameters of feature u can be defined as

$$\begin{aligned} \text{SS}_B(u) &= \sum_{i=\text{Pre,Pro}} N^i \left(\frac{\sum_{j=1}^{N^i} f_{ij}(u)}{N^i} \right. \\ &\quad \left. - \frac{\sum_{i=\text{Pre,Pro}} \sum_{j=1}^{N^i} f_{ij}(u)}{\sum_{i=\text{Pre,Pro}} N^i} \right)^2, \\ \text{SS}_W(u) &= \sum_{i=\text{Pre,Pro}} \sum_{j=1}^{N^i} \left(f_{ij}(u) - \frac{\sum_{i=\text{Pre,Pro}} \sum_{j=1}^{N^i} f_{ij}(u)}{\sum_{i=\text{Pre,Pro}} N^i} \right)^2, \end{aligned} \tag{8}$$

where $f_{ij}(u)$ denotes frequency of the uth feature of the jth sample in the ith group (i = Pre or Pro). N^i denotes number of samples in the ith group (i = Pre or Pro). $\text{SS}_B(u)$ and $\text{SS}_W(u)$ are called sum of squares between groups and sum of squares within groups, respectively. If the sample means within groups are close to each other, $\text{SS}_B(u)$ will be small. If the sample means are close between two groups, $\text{SS}_W(u)$ will be small. Then the sample variance between groups $s_B^2(u)$ and sample variance within groups $s_W^2(u)$ can be given by

$$\begin{aligned} s_B^2(u) &= \frac{\text{SS}_B(u)}{\text{d}f_B}, \\ s_W^2(u) &= \frac{\text{SS}_W(u)}{\text{d}f_W}, \end{aligned} \tag{9}$$

where $\text{d}f_B$ and $\text{d}f_W$ are called degrees of freedom in statistics. In this study, $\text{d}f_B = 1$ and $\text{d}f_W = N^{\text{Pre}} + N^{\text{Pro}} - 2 = 254$, respectively.

According to the statistic theory, the ratio between $s_B^2(u)$ and $s_W^2(u)$ obeys F sampling distribution with $\text{d}f_B$ and $\text{d}f_W$ degrees of freedom under the null hypothesis. Thus, we used ratio $F(u)$ to measure the contribution of each feature defined as follows:

$$F(u) = \frac{s_B^2(u)}{s_W^2(u)}. \tag{10}$$

$F(u)$ reveals how strong the uth feature is related to the group variables. Accordingly, the 400 dipeptides in (3) were ranked according to their $F(u)$. Subsequently, the incremental feature selection (IFS) strategy was proposed to find an optimal of feature subset. In IFS procedure, we firstly examined the performance of the best feature with the highest $F(u)$ by using cross-validation. Subsequently, a new feature with the second highest $F(u)$ was added to form new feature subset which was also inputted into SVM and the accuracy was calculated. This process was repeated until 400 feature subsets were examined. By setting the number of features as abscissa and the Acc as ordinate, the IFS curves were plotted in Figure 1. From the figure, we observed that, in the jackknife cross-validation, the maximum Acc of 94.9% can be obtained by the top 190 features which are regarded as the optimal feature subset.

Table 1: Comparison of prediction performance for presynaptic and postsynaptic neurotoxins.

	Sn	Sp	Acc
ID [5]	88.46	91.30	89.80
Bilayer SVM [7]	100.00	98.37	98.93
Our method	94.51	95.15	94.92

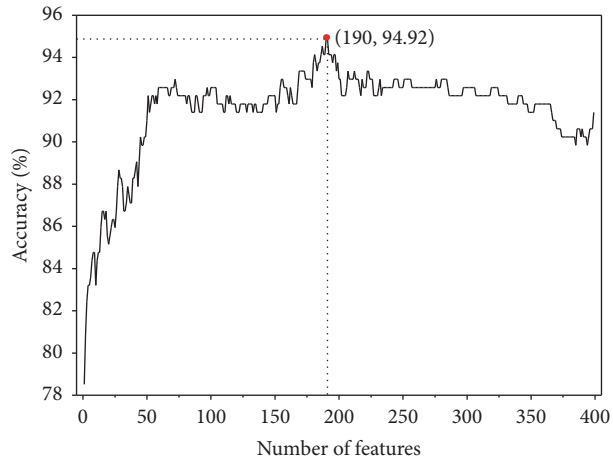

Figure 1: A plot to show the feature selection results. The maximum accuracy is 94.92% by using the top 190 features.

It is very important to compare the performance of different methods. However, it is not feasible because the benchmark datasets are different. Thus, we made a rough comparison and recorded the results in Table 1. Yang and Li proposed ID-based method to predict presynaptic and postsynaptic neurotoxins on a benchmark dataset with the sequence identity of <80% [5]. Thus, our method is superior to Yang's method. Song developed bilayer support vector machine to improve the accuracy [7]. We noticed that the sequence identity of the benchmark dataset reaches 90% which results in the overestimation of the method. Thus, our proposed model is more objective and real.

4. Conclusions

The knowledge for neurotoxin is conductive to the development of drug design and drug-target discovery. Thus, the aim

of the study is to develop a computational method to predict presynaptic and postsynaptic neurotoxins. A new feature selection technique was proposed to optimize features and to improve prediction accuracy. The feature selection technique can also be used in other bioinformatics fields.

Competing Interests

The authors declare that there is no conflict of interests.

Acknowledgments

This work was supported by the Applied Basic Research Program of Sichuan Province (14JC0121) and the Scientific Research Foundation of the Education Department of Sichuan Province (11ZB122).

References

[1] F. Afifiyan, A. Armugam, P. Gopalakrishnakone, N. H. Tan, C. H. Tan, and K. Jeyaseelan, "Four new postsynaptic neurotoxins from Naja naja sputatrix venom: cDNA cloning, protein expression, and phylogenetic analysis," *Toxicon*, vol. 36, no. 12, pp. 1871–1885, 1998.

[2] O. Rossetto, M. Rigoni, and C. Montecucco, "Different mechanism of blockade of neuroexocytosis by presynaptic neurotoxins," *Toxicology Letters*, vol. 149, no. 1–3, pp. 91–101, 2004.

[3] X. Wang, K. L. Engisch, Y. Li, M. J. Pinter, T. C. Cope, and M. M. Rich, "Decreased synaptic activity shifts the calcium dependence of release at the mammalian neuromuscular junction in vivo," *Journal of Neuroscience*, vol. 24, no. 47, pp. 10687–10692, 2004.

[4] J. P. Forder and M. Tymianski, "Postsynaptic mechanisms of excitotoxicity: involvement of postsynaptic density proteins, radicals, and oxidant molecules," *Neuroscience*, vol. 158, no. 1, pp. 293–300, 2009.

[5] L. Yang and Q. Li, "Prediction of presynaptic and postsynaptic neurotoxins by the increment of diversity," *Toxicology in Vitro*, vol. 23, no. 2, pp. 346–348, 2009.

[6] Q.-Y. He, Q.-Z. He, X.-C. Deng et al., "ATDB: a uni-database platform for animal toxins," *Nucleic Acids Research*, vol. 36, no. 1, pp. D293–D297, 2008.

[7] C. Song, "Prediction of presynaptic and postsynaptic neurotoxins by bi-layer support vector machine with multi-features," *African Journal of Microbiology Research*, vol. 6, no. 6, pp. 1354–1358, 2012.

[8] A. Bairoch, R. Apweiler, C. H. Wu et al., "The Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 33, pp. D154–D159, 2004.

[9] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[10] H. Tang, P. Zou, C. Zhang, R. Chen, W. Chen, and H. Lin, "Identification of apolipoprotein using feature selection technique," *Scientific Reports*, vol. 6, Article ID 30441, 2016.

[11] C. Zhang, H. Tang, W. Li, H. Lin, W. Chen, and K. Chou, "iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition," *Oncotarget*, vol. 7, no. 43, pp. 69783–69793, 2016.

[12] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, and K.-C. Chou, "IHyd-PseCp: identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC," *Oncotarget*, vol. 7, no. 28, pp. 44310–44321, 2016.

[13] J. Zhang, P. Sun, X. Zhao, and Z. Ma, "PECM: prediction of extracellular matrix proteins using the concept of Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 363, pp. 412–418, 2014.

[14] H. Lin, E.-Z. Deng, H. Ding, W. Chen, and K.-C. Chou, "iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition," *Nucleic Acids Research*, vol. 42, no. 21, pp. 12961–12972, 2014.

[15] H. Tang, W. Chen, and H. Lin, "Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique," *Molecular BioSystems*, vol. 12, no. 4, pp. 1269–1275, 2016.

[16] R. Cao, D. Bhattacharya, J. Hou, and J. Cheng, "DeepQA: improving the estimation of single protein model quality with deep belief networks," *BMC Bioinformatics*, vol. 17, article no. 495, 2016.

[17] S. Colic, R. G. Wither, M. Lang, L. Zhang, J. H. Eubanks, and B. L. Bardakjian, "Prediction of antiepileptic drug treatment outcomes using machine learning," *Journal of Neural Engineering*, vol. 14, no. 1, Article ID 016002, 2017.

[18] Y. Bao, M. Hayashida, and T. Akutsu, "LBSizeCleav: improved support vector machine (SVM)-based prediction of Dicer cleavage sites using loop/bulge length," *BMC Bioinformatics*, vol. 17, article 487, 2016.

[19] R. Yang, C. Zhang, R. Gao, and L. Zhang, "A novel feature extraction method with feature selection to identify golgi-resident protein types from imbalanced data," *International Journal of Molecular Sciences*, vol. 17, no. 2, article 218, 2016.

[20] P. Tao, T. Liu, X. Li, and L. Chen, "Prediction of protein structural class using tri-gram probabilities of position-specific scoring matrix and recursive feature elimination," *Amino Acids*, vol. 47, no. 3, pp. 461–468, 2015.

[21] M. Mandal, A. Mukhopadhyay, and U. Maulik, "Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou's PseAAC," *Medical & Biological Engineering & Computing*, vol. 53, no. 4, pp. 331–344, 2015.

[22] M. J. Iqbal, I. Faye, B. B. Samir, and A. Md Said, "Efficient feature selection and classification of protein sequence data in bioinformatics," *Scientific World Journal*, vol. 2014, Article ID 173869, 12 pages, 2014.

[23] A. Srivastava, S. Ghosh, N. Anantharaman, and V. K. Jayaraman, "Hybrid biogeography based simultaneous feature selection and MHC class I peptide binding prediction using support vector machines and random forests," *Journal of Immunological Methods*, vol. 387, no. 1-2, pp. 284–292, 2013.

[24] M. Bhattacharyya, L. Feuerbach, T. Bhadra, T. Lengauer, and S. Bandyopadhyay, "MicroRNA transcription start site prediction with multi-objective feature selection," *Statistical Applications in Genetics and Molecular Biology*, vol. 11, no. 1, article no. 6, 2012.

[25] W. Yu, Z. Jiang, J. Wang, and R. Tao, "Using feature selection technique for drug-target interaction networks prediction," *Current Medicinal Chemistry*, vol. 18, no. 36, pp. 5687–5693, 2011.