African Journal of Microbiology Research Vol. 6(6), pp. 1354-1358, 16 February, 2012
Available online at http://www.academicjournals.org/AJMR
DOI: 10.5897/AJMR11.1536
ISSN 1996-0808 ©2012 Academic Journals

Full Length Research Paper

Prediction of presynaptic and postsynaptic neurotoxins by bi-layer support vector machine with multi-features

Chaohong Song

College of Science, Huazhong Agricultural University, Wuhan, 430070, China. E-mail: chh_song@mail.hzau.edu.cn.
Tel: +86 189 71212816. Fax: +86 027 87282133.

Accepted 12 January, 2012

Much benefit to biology research and drug design, prediction of neurotoxin gradually became a necessary and popular task in recent year. In this paper, based on multi-feature extraction strategies from primary sequences and support vector machine, a novel Multi-classifier system named bi-layer support vector machine was proposed to predict presynaptic and postsynaptic neurotoxins, and obtained satisfactory results with 98.5% prediction accuracies for presynaptic neurotoxins and 99.18% for postsynaptic neurotoxins, the Matthew's correlation coefficient was 0.9767. The satisfactory results showed that, the current method might play a complementary role to other existing methods for predicting presynaptic and postsynaptic neurotoxins.

Key words: Prediction, bi-layer support vector machine, pseudo amino acid composition, approximate entropy, dipeptide.

INTRODUCTION

There are nearly 3000 species of spiders and 2340 species of snakes living in the world, and hundreds of them are poisonous. Studies have found that, the main components of the venoms of these animals are proteins with various molecular masses that consist mainly of enzymes and toxins (Dolimbek et al., 1998). Although, the roles of the most of these enzymes are still in distinction, but the toxins are the active principles of the venoms, which are expressed in their binding to the elements of the presynaptic or postsynaptic membranes (Afifiyan et al., 1998). Presynaptic acts on nerve endings; it inhibits neurotransmitter by blocking the release of acetylcholine or damaging the cell membrane. Postsynaptic binds specially to the nicotinic acetylcholine receptor resulting in the prevention of nerve transmission, leading to death from asphyxiation. Nearly 100 postsynaptic neurotoxins have been found. Some of these neurotoxins are very important in the research of biological science and medicine design, For example, presynaptic neurotoxins had been used for the treatment of migraine headache and cerebral palsy. Obtaining the information about these neurotoxins provides more information on the function of neurotoxin and makes more application of them. Although, the information about neurotoxins can be obtained by experimental technology, but computer aided prediction is less time consuming and costly, so computer aided prediction of presynaptic and postsynaptic neurotoxins would be very helpful in obtaining these information.

In fact, based on computer aided methods, there are many encouraging results in the field of predicting various toxins. For example, Saha and Raghava (2007) achieved an accuracy of 96.07 and 92.50% for predicting bacterial toxins and non toxins by using support vector machines (SVM). Song (2011) enhanced the predictive accuracy by means of an improved feature extraction and IB1 algorithm fusion method. Using the pseudo amino acid composition, (Lin and Li, 2007) provided a new algorithm of increment of diversity combined with modified Mahalanobis discriminant to predict five conotoxin superfamilies. Directly based on fusing different kinds of sequential features by using modified one-versus-rest SVMs, (Fan et al., 2011) developed a novel approach called PredCSF for predicting the conotoxin superfamily, and obtained an overall accuracy of 90.65%. Zaki et al. (2011) proposed a SVM-Freescore method, which featured an improved sensitivity and specificity by approximately 5.864 and 3.76%, respectively.

For presynaptic and postsynaptic neurotoxins, Yang and Li (2009) used an algorithm of increment of diversity to predict them, and obtained the encouraging prediction accuracies with 90.23% for presynaptic neurotoxins and 89.40% for postsynaptic neurotoxins; their Matthew's correlation coefficient was 0.7963.

In general, a successful computer aided method is decided by two factors, which are the choice of classifier and the feature extraction method of protein sequence. There are a lot of studies in these fields, as an effective feature extraction method, Pseudo amino acid compositions (PseAA) are usually used to represent a protein sequence with a discrete model yet without completely losing its sequence-order information (Chou, 2001). According to the different composition, there is a variety of pseudo amino acid composition (Chou, 2005; Chou and Cai, 2002; yang and Li, 2009), which had been used for enhancing the prediction quality of protein attributes.

Support vector machine (SVM) is an effective tool for classification and prediction, which has been used in various fields related to protein function prediction (Huang and Shi, 2005; Zhang et al., 2006; Zhou et al., 2008; Shi et al., 2008; Lin et al., 2009), but methods only using a single classifier have some limitations in the prediction (Chou and Shen, 2006a, c). Recently, multi-classifier systems have been proposed to enhance the prediction quality and it obtained satisfactory results (Park and Kanehisa, 2003; Chou and Cai, 2004; Yu et al., 2004; Chou and Shen, 2006a, b, c, 2007; zhou et al., 2008). Using the advantages of SVM and advantages of multi-classifier system, and constructing multiple SVM classifiers to enhance the prediction quality, it will be a fresh idea for neurotoxin prediction.

In this study, a multi-classifier named "bi-layer SVM" was built to further improve the prediction accuracy of presynaptic and postsynaptic neurotoxins, and a relatively good predictive result was obtained.

MATERIALS AND METHODS

Dataset

Presynaptic and postsynaptic neurotoxins used in this study where downloaded from the dataset (Boeckmann et al., 2003), in order to ensure enough protein sequences in experimental data, we selected those presynaptic neurotoxins and postsynaptic neurotoxins with no more than 90% identity. Finally, we got 132 presynaptic neurotoxin sequences and 241 postsynaptic neurotoxin sequences.

Feature extraction

In this paper, besides basic amino acid composition, two kinds of sequence feature were used to construct pseudo amino acid composition of neurotoxin sequences, that is, the approximate entropy of protein sequences, and the dipeptide composition of protein sequences.

The approximate entropy representation of protein sequences

The approximate entropy is generally a measure of system complexity (Pincus, 1991); it has been widely used to deal with physiological signal (Richman and Moorman, 2001) and protein prediction (Song 2011). The algorithm for computing approximate entropy of a protein sequence can be briefly described as follows: first, represent a protein sequence as a time series X_N by replacing every amino acid of a protein sequence by the relevant value of its hydrophobic amino acids, suppose the sequence X_N consisting of *N* components, namely $X_N = \{X_N(1), X_N(2) \cdots X_N(N)\}$, before computing the approximate entropy $ApEn(X_N, m, r)$ of the sequence, we should choose values of two input parameters which are pattern length *m* and the criterion of similarity, *r*. Denote a subsequence with *m* components, beginning at component *i* within X_N by the vector $P_m(i)$. For two subsequence $P_m(i)$ and $P_m(j)$, if the difference between any pair of corresponding components in the subsequences is less than *r*, that is, if

$$| X_N(i+k) - X_N(j+k) | < r \text{ For } 0 \le k \le m$$

We think they are similar.

Suppose $P_m = \{P_m(1), P_m(2), \cdots P_m(N-m+1)\}$ is the set of all subsequences of length *m* within X_N, $n_{im}(r)$ is the number of subsequences in P_m that are similar to $P_m(i)$ for the given similarity criterion *r*, then:

$$C_{im} = \frac{n_{im}(r)}{N-m+1}$$

Where, the quantity $C_{im}(r)$ is the fraction of subsequences of length m, which resembles the subsequence of the same length that begins at interval i.

Finally, for the given similarity criterion r and the length m, we define the approximate entropy of X_N as:

$$ApEn(X_N, m, r) = \ln \frac{C_m(r)}{C_{m+1}(r)}$$

Where, $C_m(r)$ is the mean of all $C_{im}(r)$ values in $P_m(i)$, the quantity $C_m(r)$ expresses the prevalence of repetitive subsequences with length *m* in X_N.

K-means clustering and dipeptide feature representation

K-means clustering is one of clustering analysis methods. Given observations X_i $i = 1, 2 \cdots n$, where X_i is an *m*-dimensional real vector, the aim of *k*-means clustering is to partition these *n* observations into *k* clusters S_j $j = 1, 2 \cdots k$ ($k \le n$) so as to minimize the within-cluster sum of squares:

$$\arg\min_S \sum_{i=1}^{k} \sum_{X_j \in S_i} \| X_j - u_i \|^2$$

Where u_i is the mean of points in S_i.

In simple terms, the steps of K-means clustering can be describes

Table 1. Selection of the feature and parameters for each classifier.

Parameter	Classifier	Feature	Parameter
	SVM1	$(f_1, f_2 \ldots f_{20})$	γ=2, c=4
The first layer	SVM2	$(f_1, f_2 \ldots f_{20}, 0.12e_1, \ldots 0.12e_{12})$	γ=1, c=2
	SVM3	$(e_1, \ldots e_{12}, 0.02d_1, \ldots 0.02d_{16})$	γ=1, c=2
The second layer	SVM4	(r_1, r_2, r_3)	γ=1, c=2

Table 2. Comparison of prediction performance for presynaptic and postsynaptic neurotoxins.

Parameter	Presynaptic neurotoxins				Postsynaptic neurotoxins			
	Sn (%)	Sp (%)	Ac (%)	Mcc	Sn (%)	Sp (%)	Ac (%)	Mcc
Our method	100	96.97	98.50	0.9796	98.37	100	99.18	0.9796
Diversity increment[a]	88.46	92.00	90.23	0.7963	91.30	87.50	89.40	0.7963

[a]Comes from Yang and Li (2009).

as the following: first, randomly select k observations as the initial center of each cluster, then each one of other observations are assign to the center with the closest distance measure. Afterwards, compute a new center for each cluster by averaging the feature vectors of all observations assigned to it, repeat this process until each center keeps unchangeable.

The dipeptide components are important parameters for protein structure and function, it has been widely used to protein bioinformatics (Lin et al., 2007, 2008; Lin and Ding, 2011; Lin and Li, 2011). In this study, we first extracted dipeptides features of a neurotoxin sequence according to the hydrophilicity value of the corresponding residue, and obtained 400 vectors of dipeptides. In order to be easy to calculate, we sorted these 400 vectors into k clusters by k-means clustering. Where the distance measured in k-means clustering, we chose Euclidean distance, and the selection of k was by the histogram: firstly, we obtained the histogram of these 400 vectors by the soft linkage, from the histogram, we could estimate the value scope of k, and for all possible k, we selected the k whose prediction effect is the best one, here $k = 16$.

Bi-layer support vector machines (SVM) classifier

SVM is a popular algorithm for pattern recognition and protein predicting protein (Chen et al., 2010; Lin and Chen, 2011). But the single classification prediction efficiency is always affected by noise and complex datasets, in order to reduce these effects, in this study; we built a bi-layer SVM classifier, which consists of four SVM classifiers, three in the first layer, denoted as SVM1, SVM2 and SVM3, and one in the second layer, denoted as SVM4. The sequence feature used in SVM1 is only the occurrence frequencies of the 20 amino acids (denoted them as $f_i \ (i = 1, 2, \cdots, 20)$). In SVM2 and SVM3, we selected the pseudo amino acid composition as the sequence feature, which were constructed by the occurrence frequencies of the 20 amino acids combined with approximate entropy (denoted them as $e_i \ (i = 1, 2, \cdots, 12)$) with the weight 0.12 in SVM2, and the approximate entropy combined with 16 dipeptide composition (denoted them as $d_i \ (i = 1, 2, \cdots, 16)$) with the weight 0.02, respectively. Training parameters are γ =2, c =4 in SVM1, γ =1, c =2 in SVM2 and SVM3, respectively. The value of these parameters is taking from the corresponding values which have the best effect in the training prediction. Then, the corresponding dimensions of input vector of these three SVM classifiers are 20, 32, and 28, respectively. The second layer was trained with the output (denoted them as $r_i \ (i = 1, 2, 3)$) generated by 3 classifiers in the first layer (here we also set, γ=1, c=2) (Table 1). The classifiers used here are OSU-SVM (http://www.ece.osu.edu/~maj/osu_svm).

Evaluation of the performance

In order to compare with other prediction performance, we also adopt the sensitivity (*Sn*), specificity (*Sp*), Matthew's correlation coefficient (*Mcc*) and accuracy (*Acc*) as appraisal targets to estimate the performance of our method, these appraisal targets can be calculated by the following formulae, respectively:

$$Sn = TP / (TP + FN)$$

$$Sp = TP / (TP + FP)$$

$$MCC = (TP \times TN - FP \times FN) / \sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}$$

$$Acc = (Sn + Sp) / 2$$

Where, *TP* denotes the numbers of the correctly recognized positives, *FN* denotes the numbers of the positives recognized as negatives, *FP* denotes the numbers of the negatives recognized as positives, and *TN* denotes the numbers of correctly recognized negatives (yang and Li, 2009).

RESULTS AND DISCUSSION

In this paper, we provided rough comparison of the performance between our method and the other method (Table 2). From Table 2, we could see that, for presynaptic neurotoxins, by 10 fold cross validation, the results of the sensitivity, specificity and MCC value were appreciably improved, and the increments were 11.54,

Table 3. Prediction performance for presynaptic and postsynaptic neurotoxins by single-layer SVM.

Parameter	Presynaptic neurotoxins				Postsynaptic neurotoxins			
	Sn (%)	Sp (%)	Ac (%)	Mcc	Sn (%)	Sp (%)	Ac (%)	Mcc
Feature two	77.94	80.30	79.12	0.6441	89.03	87.55	88.29	0.9796
Feature three	83.20	78.79	80.99	0.7099	88.71	91.29	90.00	0.7099

4.97 and 18.33%, respectively. For postsynaptic neurotoxins, these three appraisal targets were also improved, and with 7.07, 12.5 and 18.33% increments, respectively. These satisfactory results were enough to show that our method was effective for predicting presynaptic and postsynaptic neurotoxins.

In order to further study the prediction performance of bi-layer support vector machine, we predicted presynaptic and postsynaptic neurotoxins only by single-layer SVM based on the feature which was the same as the ones which was used in SVM2 and SVM3, the results were listed in Table 3. From the Table 3, we could see that the prediction performance were not satisfactory, but based on the same feature by bi-layer SVM, we could obtain fairly good results, which might be that bi-layer SVM built in this paper could make full use of multiple features information, and could take better advantage of the sequence information of a protein than that of the single-layer SVM based on individual feature.

The successful prediction showed that bi-layer support vector machine could make full use of multiple features information and fairly improved the sensitivity, specificity and MCC value; it was quite suitable to predict presynaptic and postsynaptic neurotoxins. It was also evidence that bi-layer SVM was a promising classifier; we hoped this method would be helpful for the analysis of possible functions of new neurotoxins.

ACKNOWLEDGEMENT

It is partly supported by the Fundamental Research Funds for the Central Universities (2010JC003).

REFERENCES

Afifiyan F, Armugam A, Gopalakrishnakone P, Tan NH, Tan CH, Jeyaseelan K (1998). Four new postsynaptic neurotoxins from naji naja sputatrix venom: cDNA cloning, protein expression, and phylogenetic analysis. Toxicon., 36: 1871–1885.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res., 31: 365–370.

Chen W, Lin H (2010). Prediction of midbody, centrosome and kinetochore proteins based on gene ontology information, Biochem. Biophys. Res. Commun., 401: 382–384.

Chou KC, (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Struct. Funct. Genet., 43: 246–255.

Chou KC (2005). Using amphiphilic pseudo-amino acid composition to predict enzyme subfamily classes. Bioinform., 21: 10–19.

Chou KC, Cai YD (2002). Using functional domain composition and support vector machines for prediction of protein subcellular location. J. Biol. Chem., 277: 45765–45769.

Chou KC, Cai YD (2004) Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. J. Cell Biochem., 91: 1197–1203.

Chou KC, Shen HB (2006a) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic k-nearest neighbor classifiers. J. Proteome Res., 5: 1888–1897.

Chou KC, Shen HB (2006b). Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. Biochem. Biophys. Res. Commun., 347: 150–157.

Chou KC, Shen HB (2006c). Predicting protein subcellular location by fusing multiple classifiers. J. Cell Biochem., 99: 517–527.

Chou KC, Shen HB (2007). Large-scale plant protein subcellular location prediction. J. Cell Biochem., 100: 665–678

Dolimbek BZ, Atassi MZ, Sa]ikhov SI (1998). Presynaptic and postsynaptic neurotoxins investigation of the structures of the immune recognition sections, Chem. Nat. Compd., 34: 15-28.

Fan YX, Song JN, Kong XZ and Shen HB (2011). PredCSF: An Integrated feature-based approach for predicting conotoxin superfamily, Protein Peptide Lett., 18: 261-267.

Lin H (2008). The modified Mahalanobis Discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J. Theor. Biol., 252: 350–356.

Lin H, Chen W (2011). Prediction of thermophilic proteins using feature selection technique, J. Microbiol. Methods, 84: 67–70.

Lin H, Ding H (2011). Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition, J. Theor. Biol., 269: 64-69.

Lin H, Li QZ (2007). Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem. Biophys. Res. Commun., 354: 548–551.

Lin H, Li QZ (2011). Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J. Comput. Chem., 28: 1463-1466

Lin H, Wang H, Ding H, Chen YL and Li QZ (2009), Prediction of Subcellular Localization of Apoptosis Protein Using Chou's Pseudo Amino Acid Composition. Acta Biotheor., 57: 321–330.

Huang J, Shi F (2005). Support vector machine for predicting apoptosis proteins types, Acta biotheor., 53: 39-47.

Park KJ, KanehisaM (2003). Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinform., 19: 1656–1663.

Pincus SM (1991). Approximate entropy as a measure of system complexity. Proc. Natl. Acad Sci USA., 88: 2297~2301.

Richman JS, Moorman JR (2001). Physiological time-series analysis using approximate entropy and sample entropy. Am. J. Physiol. Heart Circ. Physiol., 278: 2039-2049.

Saha S, Raghava GP (2007). BTXpred: Prediction of bacterial toxins. In Silico Biol., 7: 405-412.

Shi F, Chen QJ, Li NN (2008). Hilbert Huang transform for predicting proteins subcellular location. J. Biomed. Sci. Eng., 1: 59–63.

Song CH (2011). Prediction of bacterial toxins by an improved feature extraction and IB1 algorithm fusion. Afr. J. Microbiol. Res., 5: 1479-1483.

Yang L, Li QZ (2009). Prediction of presynaptic and postsynaptic neurotoxins by the increment of diversity. Toxicol. *in vitro*., 23: 346-348.

Yu CS, Lin CJ, Huang JK (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci., 13: 1402–1406.

Zaki N, Wolfsheimer S, Nuel G and Khuri S (2011). Conotoxin protein classification using free scores of words and support vector machines. BMC Bioinform., 12: 217.

Zhang ZH, Wang ZH, Zhang ZR, Wang YX (2006). A novel method for apoptosis protein subcellular localization prediction combing encoding based on grouped weight and support vector machine. FEBS Lett., 580: 6169–6174.

Zhou XB, Chen C, Li ZC, Zou XY (2008). Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine. Amino Acids 35:383–388.