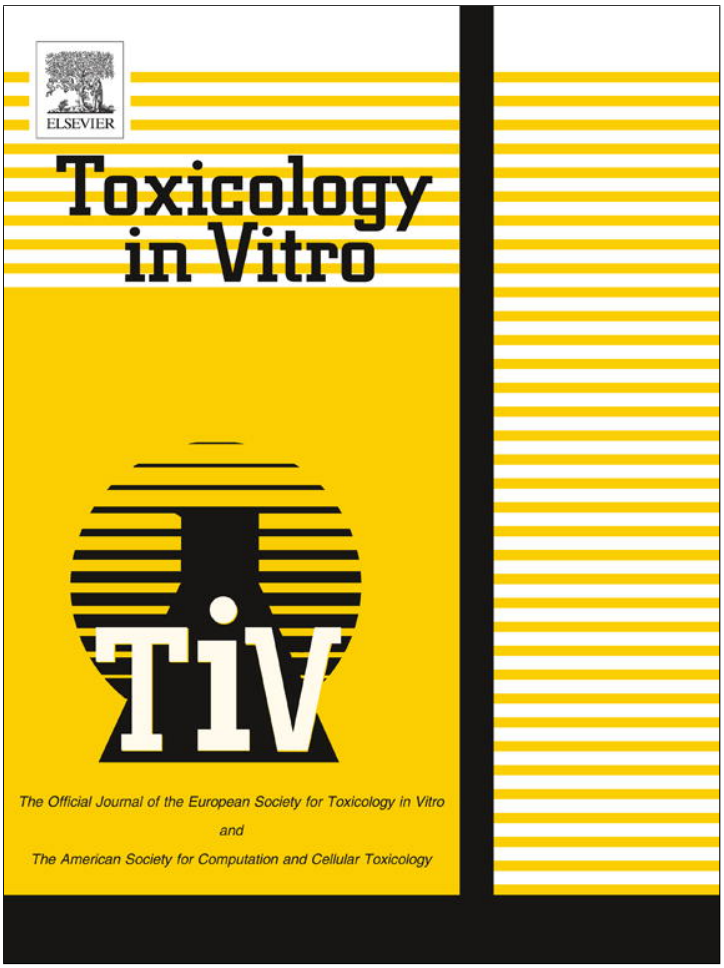

(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Toxicology in Vitro 27 (2013) 852–856

Contents lists available at SciVerse ScienceDirect

Toxicology in Vitro

journal homepage: www.elsevier.com/locate/toxinvit

Prediction of the types of ion channel-targeted conotoxins based on radial basis function network

Lu-Feng Yuan [a], Chen Ding [a], Shou-Hui Guo [a], Hui Ding [a,*], Wei Chen [b,*], Hao Lin [a,*]

[a] Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
[b] Department of Physics, Center for Genomics and Computational Biology, College of Sciences, Hebei United University, Tangshan 063000, China

ARTICLE INFO

Article history:
Received 12 September 2012
Accepted 22 December 2012
Available online 29 December 2012

Keywords:
Ion channel-targeted conotoxins
RBF network
Binomial distribution
Dipeptide

ABSTRACT

Conotoxins are small disulfide-rich peptide toxins, which have the exceptional diversity of sequences. Because conotoxins are able to specifically bind to ion channels and interfere with neurotransmission, they are considered as the excellent pharmacological candidates in drug design. Appropriate type assignment of newly sequenced mature ion channel-targeted conotoxins with computational method is conducive to explore the biological and pharmacological functions of conotoxins. In this paper, we developed a novel method based on binomial distribution and radial basis function network to predict the types of ion-channel targeted conotoxins. We achieved the overall accuracy of 89.3% with average accuracy of 89.7% in the prediction of three types of ion channel-targeted conotoxins in jackknife cross-validation test, indicating that the method is superior to other state-of-the-art methods. In addition, we evaluated the proposed model with an independent dataset including 77 conotoxins. The overall accuracy of 85.7% was achieved, validating that our model is reliable. Moreover, we used the proposed method to annotate 336 function-undefined mature conotoxins in the UniProt Database. The model provides the valuable instructions for theoretical and experimental research on conotoxins.

1. Introduction

Conotoxins are disulfide-rich small peptides and the mature peptide sequences contain only 10–30 amino acids. However, conotoxins have high sequence diversity. Conotoxins have a wide range of targets, including G protein-coupled receptors, nicotinic acetylcholine and neurotensin receptor. Therefore they have widely biological applications, such as the treatment of chronic pain, epilepsy, spasticity and cardiovascular diseases (Han et al., 2008; Watters, 2005). Especially, the majority of conotoxins have high specificity and affinity towards ion channels. They have been deemed as important pharmacological agents in ion channel research and widely used as pharmacological tools for neuroscience research (Han et al., 2008; Terlau and Olivera, 2004; Watters, 2005). According to the types of ion channel activities, conotoxins can be classified as calcium (Ca), sodium (Na) and potassium (K) channel-targeted conotoxins (Terlau and Olivera, 2004). It has been estimated that there are over 100,000 different conotoxins (Daly and Craik, 2009), but only 1703 conotoxins have been published in the Universal Protein Resource (UniProt, Fourth Dec. 2012). And few records can provide function annotation of the types of ion channel-targeted conotoxins. Therefore, identification of the type of a newly sequenced conotoxin is beneficial to study its biological and pharmacological functions.

Unfortunately, the determination of the functions of conotoxins requires long time and high cost for wet experimental approaches. Bioinformatics analysis is a convenient methodology for preliminary function analysis of newly sequenced conotoxins. System classification and function annotation is an important feature and of major interest to the experimental biologists. In the past years, several approaches have been reported to predict conotoxins based on primary sequences. Liu et al. and Peng et al. proposed to use cDNA library to identify different superfamily of conotoxins (Liu et al., 2008; Peng et al., 2008). Mondal et al. developed a support vector machine (SVM) model to predict conotoxin superfamily with pseudo amino acid composition (PseAAC) (Mondal et al., 2006). They correctly predicted 88.1% superfamilies of conotoxins. In our recent work, an IDQD model was proposed to predict conotoxin superfamily and family (Lin and Li, 2007). The overall accuracies of conotoxin superfamily and family were 87.7% and 72%, respectively. Furthermore, the Toxin-AAM method was used to predict conotoxin superfamily with evolutionary information through integrating pairwise sequence comparison with amino acid composition (Zaki et al., 2011a). Subsequently, Zaki et al. developed another method called SVM-Freescore algorithm to improve predictive sensitivity and specificity (Zaki et al., 2011b).

* Corresponding authors. Tel.: +86 28 83208232; fax: +86 28 83208238.
E-mail addresses: hding@uestc.edu.cn (H. Ding), chenwei_imu@yahoo.com.cn (W. Chen), hlin@uestc.edu.cn (H. Lin).

0887-2333/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.tiv.2012.12.024

Recently, Shen and his colleagues used diffusion maps to optimize feature set for the prediction of conotoxin superfamily and improved the overall accuracy to 90.3% (Fan et al., 2011; Yin et al., 2011). Although these methods can achieve some helpful information on conotoxin research, they only indirectly provide possible function information of conotoxins. For example, Delta-conotoxin-like Ac6.1 (Uniprot accession number: P0C8V5) (Gowd et al., 2008) and Omega-conotoxin-like Ai6.2 (Hillyard et al., 2008) (Uniprot accession number: P0CB10), belonging to the conotoxin O1 superfamily, target different types of ion channels. The Delta-conotoxin-like Ac6.1 binds to voltage-gated sodium channels, while the Omega-conotoxin-like Ai6.2 blocks voltage-gated calcium channels. Therefore, it is urgent and necessary to develop a simple and efficient method to predict the types of ion channels-targeted conotoxins.

To the best of our knowledge, there is no computational system for predicting the types of ion channel-targeted conotoxins. In this work, we developed a computational model based on radial basis function network (RBF network) to predict the types of ion channel-targeted conotoxins, providing a useful tool for further physiological and pharmacological studies. A non-redundant benchmark dataset including 112 mature conotoxins was established to train and test the performance of the proposed model. The binomial distribution was used to reduce feature redundancy for optimal feature set. Jackknife cross-validation was used to evaluate the accuracy of the proposed method. An overall accuracy of 89.3% was obtained. Furthermore, we used an independent dataset to evaluate the proposed model and obtained the overall accuracy of 85.7%. Finally, we used the proposed model to predict 336 function-undefined mature conotoxins. Other researchers may use the model and relevant results to study the functions of conotoxins.

2. Materials and methods

2.1. Datasets

The raw datasets adopted in this research were extracted from the UniProt (Magrane and Consortium, 2011). For the purpose of obtaining a reliable benchmark dataset, the following steps were used to construct high quality datasets. Firstly, conotoxins with ambiguous existence annotations, such as 'uncertain', 'predicted' and 'inferred from homology' were excluded because they lack in confidence. Secondly, only those peptides with experimental confirmed functional annotation for targeting ion channel were included because they can provide correct and validated information. Thirdly, we only chose the mature conotoxin sequences since the mature peptides have biological functions. Finally, the sequences containing nonstandard letters, such as 'B', 'X' or 'Z', were excluded because their meanings are ambiguous. After the above strict screening procedure, we obtained 195 sequences including 37 potassium ion channel-targeted conotoxins (K-conotoxins), 86 sodium ion channel-targeted conotoxins (Na-conotoxins) and 72 calcium ion channel-targeted conotoxins (Ca-conotoxins).

In order to include as many sequences as possible without increase sequence homology bias, the CD-HIT program (Li and Godzik, 2006) was used to prune the data. By setting the cutoff of sequence identity to 80%, 112 sequences were remained in the final datasets including 24 K-conotoxins, 43 Na-conotoxins and 45 Ca-conotoxins.

For further estimating the performance of the method, we also collected 77 ion channel-targeted conotoxins which are independent from training set. Among these conotoxins, 12 cases target potassium ion channel; 41 cases target sodium ion channel; 24 cases target calcium ion channel. The conotoxin sequences with unclearly function annotation were regarded as function-undefined conotoxins. We totally collected 336 function-undefined conotoxins.

2.2. Features extraction

A large amount of feature extraction methods has been put forward (Daly and Craik, 2009; Han et al., 2008; Terlau and Olivera, 2004; Yin et al., 2011; Zaki et al., 2011a,b). Because dipeptides can reflect the order of amino acids and encapsulate the global information for each protein sequence, it has been widely used in protein bioinformatics. In this work, we also extracted features from dipeptide compositions. The dipeptide compositions can provide total 400 (20×20) dimensions of vectors and can be calculated by the following equation:

$$F_{ab} = m_{ab}/M_b \tag{1}$$

where m_{ab} represents the occurrence number of the a-th dipeptide in b-th sequence, M_b represents the total number of dipeptides in the b-th sequence.

2.3. Binomial distribution

Feature selection is an important issue in pattern recognition, not only for the insight gained from determining relevant modeling variables, but also for the improved understandability, scalability, possibility, and accuracy of the resulting models. Thus, the optimized parameters could improve predictive accuracy. So far, many feature selection techniques have been proposed to optimize feature set (Feng and Luo, 2008; Li et al., 2012), such as, principal component analysis (PCA) (Rocchi et al., 2004), minimal-redundancy-maximal-relevance (mRMR) (Peng et al., 2005), diffusion maps (Yin et al., 2011) and the analysis of variance (ANOVA) (Lin and Ding, 2011). Here, a novel method based on binomial distribution was used to perform feature selection (Feng and Luo, 2008).

Three types of ion channel-targeted conotoxin dataset may contain four hundreds kinds of dipeptides. Each kind of dipeptide in one type may be a stochastic event. Then, the probability of the i-th dipeptide occurring in the j-th class (j = K-conotoxin, Na-conotoxin and Ca-conotoxin) can be defined by:

$$CL_{ij} = 1 - \sum_{n=n_{ij}}^{N_i} \frac{N_i!}{n!(N_i-n)!} p_j^n (1-p_j)^{N_i-n} \tag{2}$$

where CL_{ij} is also called the confidence level (CL) of the i-th dipeptide in the j-th type; N_i represents the total number of the i-th dipeptide in the dataset. n_{ij} represents the occurrence number of the i-th dipeptide in the j-th type. The sum is taken from n_{ij} to N_i. The probability p_j is the relative frequency of class j in the database and defined as:

$$p_j = \sum_{i=1}^{400} n_{ij} \Big/ \sum_{i=1}^{400} N_i \tag{3}$$

where $\sum_{i=1}^{400} N_i$ and $\sum_{i=1}^{400} n_{ij}$ are the total occurrence number of all dipeptides in the dataset and in the j-th type conotoxin, respectively.

In the three types of conotoxins, three CLs ($CL_{i\,K}$, $CL_{i\,Na}$ and $CL_{i\,Ca}$) of the i-th dipeptide may be calculated according to Eq. (2). Then we may define the confidence level of dipeptide i in benchmark dataset as follows:

$$CL_i = \max\{CL_{i\,\mathrm{K}},\ CL_{\mathrm{i\,Na}},\ CL_{iCa}\} \tag{4}$$

If there are m dipeptides whose CL_i is larger than a given cutoff, CL_o, the frequencies of these dipeptides are selected as optimized features and expressed as follows:

$$F_m = [f_1, f_2, \cdots, f_i, \cdots, f_m]^T \quad (5)$$

If CL_o is set to zero, all the 400 dipeptides are selected. If $CL_o > 1$, no dipeptides are selected. Based on confidence level (Eq. (2)), high-dimensional data can be projected into low-dimensional space. The parameter m or CL_o can be chosen through cross-validation.

2.4. RBF network

Artificial Neural Network (ANN) is one of the most successful technologies in the last two decades and is widely applied in various fields, including data mining and bioinformatics (Anthony et al., 1995; Lin et al., 2005). The RBF network is a special type of ANNs. Due to its faster training procedure and better approximation capabilities compared with other network types, such as simple network configuration, it has been widely used in protein prediction fields (Chen et al., 2011; Ou et al., 2008, 2010). Based on its universal approximation capability, the RBF network can approximate any nonlinear function with sufficient neurons in the hidden layer.

A typical RBF network is composed of three layers: an input layer, a hidden layer with a non-linear RBF activation function and a linear output layer. The input layer consists of many nodes and connects the network with external environment. The second layer is the only hidden one in the network and its role is to carry out a nonlinear transformation from the input space to the hidden space. And the hidden space will have a high dimension. The output layer supplies the response to the activations of the hidden nodes. The hidden layer is connected with output layer by the weight, ω. The RBF network is modeled by the following relation:

$$\hat{y}_k = \sum_{i=1}^{m} \omega_{ik} R_i(x) \quad (k = 1, 2, \cdots, p) \quad (6)$$

where $R_i(x)$ represents the RBF.

The gaussian function is the most widely used basis function in nonlinear transformation and it can be defined as follows:

$$R_i(x) = \exp(-\|x - c_i\|^2 / 2\sigma_i^2), \quad (i = 1, 2, \cdots, m) \quad (7)$$

where $\|x - c_i\|$ represents Euclidean norm; c_i, σ_i and R_i are the center, the width and the output of the i-th hidden unit, respectively.

The software used in this work was Weka (Waikato Environment for Knowledge Analysis) (Hall et al., 2009) developed at the University of Waikato, New Zealand.

2.5. The criteria definitions

In statistical prediction, independent dataset test, sub-sampling test (such as 5-fold or 10-fold cross validation) and jackknife test are often used to examine the power of a predictor (Lu et al., 2009; Shen and Chou, 2009; Zhou and Cai, 2006). Here, the jackknife test is applied to evaluate the performance of the proposed methods. In the jackknife cross-validation, each sequence in the training dataset is selected in turn as an independent testing sample and all the rule-parameters are calculated without including the one being identified.

The predictive capability of the algorithm is estimated by the three parameters: sensitivity (Sn), overall accuracy (Ac) and average accuracy (AA) which are defined as follows:

$$Sn_i = TP_i / (TP_i + FN_i) \quad (8)$$

$$OA = \sum_{i=1}^{\mu} TP_i / N \quad (9)$$

$$AA = \sum_{i=1}^{\mu} Sn_i / \mu \quad (10)$$

where TP_i represents the number of the correctly recognized i-th type of conotoxins; FN_i represents the number of the i-th types of conotoxins recognized as other types of conotoxins; μ represents the number of types (here $\mu = 3$); N represents the total number of sequences (here $N = 112$).

3. Results

3.1. Improvement of accuracy by feature selection

Generally, high confidence level of selecting dipeptides allows relatively small feature sets and more reliable information for prediction. The minimum number of variables to fit the data can increase the robustness of prediction (Park et al., 2005). However, the number of these features is too small to afford enough information, which results in the poor predictive accuracy. For example, if we select 99.9% as the cutoff of confidence level, we can achieve nine kinds of dipeptides. But the obtained overall accuracy is only 64.29% in jackknife cross-validation. In contrast, the dipeptide sets with low confidence contains too many components. The larger the dimension of the vectors is, the more information the representation bears. However, if the input vector contains too many components, it would reduce the cluster-tolerant capacity so as to lower the cross-validated accuracy. For instance, 349 dipeptides with >50% of confidence level can only produce the overall accuracy of 61.61% in jackknife cross-validation. Therefore, appropriate dipeptide sets allow the higher prediction accuracy.

By changing the cutoff of confidence level, we can obtain a series of dipeptide sets. Each of the dipeptide sets was input into RBF network to investigate its prediction performance through jackknife cross-validation. Accordingly, we plotted a three-dimensional curve for confidence level, feature dimension and overall accuracy shown in Fig. 1. The results show that the peak of the curve appears at 70 dipeptides with the confidence level of 96.658%. The maximum overall accuracy is 89.3% with the average accuracy of 89.7%. In the set of 70 kinds of dipeptides, 27, 24 and 19 kinds of dipeptides were selected from K-conotoxin, Na-conotoxin and Ca-conotoxin, respectively. Subsequently, we examined the predictive results for three types of conotoxins in details. The results are recorded in Table 1. As can be seen from Table 1, the sensitivities for K-conotoxin, Na-conotoxin and Ca-conotoxin are 91.7%, 88.4% and 88.9%, respectively. Our results demonstrate that the proposed feature selection technique can effectively improve the prediction performance.

3.2. Comparison with other methods

Since there is no other published work to predict the types of ion channels-targeted conotoxins, we cannot provide the

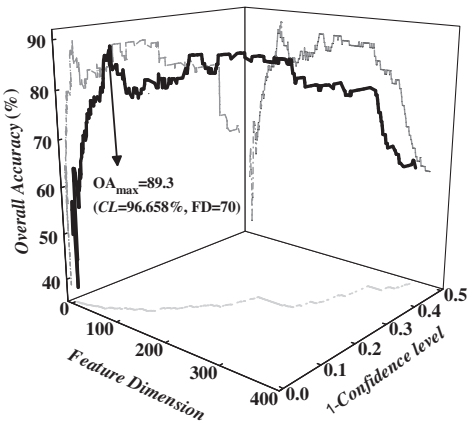

Fig. 1. The 3D curve for confidence level, feature dimension and overall accuracy.

Table 1
The jackknife test results of different methods and different features.

Methods	Sn (%)			AA (%)	OA (%)
	K	Na	Ca		
RBF network (dipeptide,70-D)	91.7	88.4	88.9	89.7	89.3
Random forest (dipeptide, 143-D)	58.3	74.4	86.7	73.1	75.9
naïve Bayes (dipeptide,184-D)	91.7	90.7	77.8	86.7	85.7
SVM (dipeptide,180-D)	83.3	83.7	93.3	86.8	87.5
RBF network (PseAAC, $\omega = 0.35$, $\lambda = 7$)	58.3	72.1	71.1	66.7	67.0
Random forest (PseAAC, $\omega = 0.25$, $\lambda = 1$)	58.3	76.7	71.1	68.7	70.5
naïve Bayes (PseAAC, $\omega = 0.30$, $\lambda = 6$)	54.2	74.4	48.9	59.2	59.8
SVM (PseAAC, $\omega = 0.20$, $\lambda = 6$)	58.3	83.7	60.0	67.3	68.8

comparison analysis with published results to confirm that the predictive model proposed here is superior to other methods. For the purpose of comparison, we compared the performance of the method with the performances of the random forest, naïve Bayes and SVM algorithm. Firstly, we repeated the feature selection process in which binomial distribution was adopted to optimize dipeptides. Secondly, each feature set was input into the three algorithms. Finally, the maximum accuracies of three algorithms were selected for comparison. Results listed in Table 1 show that the overall accuracies of random forest, naïve Bayes and SVM are 75.9%, 85.7% and 87.5%, respectively, which are lower than the accuracy obtained through RBF network. Besides, the number of features utilized in RBF network is only 70. The number is dramatically less than those used through other three algorithms (143-D, 184-D and 180-D, respectively), indicating that the RBF network-based model is more robust. But, we should note that the SVM can achieve maximum sensitivity (93.3%) for Ca-conotoxin and that naïve Bayes can achieve maximum sensitivity (90.7%) for Na-conotoxin.

The PseAAC is a popular parameter and has been widely applied in protein classification and prediction (Chen et al., 2006; Chou, 2001; Xiao et al., 2006). Because it has successfully improved prediction quality in diverse protein predictions (Chou, 2011; Hayat et al., 2012; Huang et al., 2011; Xiao et al., 2011), we also examined the accuracies of RBF network, random forest, naïve Bayes and SVM by use of PseAAC. The parameters of PseAAC, ω and λ, are optimized by jackknife cross-validation. Comparison indicates that our method is better than the other prediction methods of ion channel-targeted conotoxins.

3.3. Evaluation on independent dataset

Moreover, for the purpose of further evaluating the performance of the proposed method, we used 77 independent conotoxins with experimentally-confirmed function to examine the method. As a result, 66 conotoxins were correctly predicted by the proposed method. The predicted successful rates of K^+, Na^+ and Ca^+ channel-targeted conotoxins are 91.7%, 80.5% and 91.7%, respectively. These results further demonstrate the excellent performance of our model.

3.4. Prediction of function-undefined conotoxins

The above investigation proves that the proposed method has the ability to predict the types of ion channel-targeted conotoxins with a high accuracy. Thus, we utilized the present model to predict 336 function-undefined mature conotoxins derived from the UniProt Database. The numbers and percentages of the predicted types of conotoxins are shown in Table 2. According to Table 2, 30.06%, 30.06% and 39.88% conotoxins are predicted to Na, K and Ca ion channel-targeted conotoxins, respectively. These predictions redound to further experimental research and can be freely accessed from our web server (http://cobi.uestc.edu.cn/people/hlin/data/conotoxin/).

Table 2
The predicted results of 336 function-undefined conotoxins.

	Predicted number	Percentage (%)
K-conotoxin	101	30.06
Na-conotoxin	101	30.06
Ca-conotoxin	134	39.88

4. Discussion

The number of newly identified conotoxins is growing fast, while their functional characterization is lagging (Tan et al., 2003). Although a significant experimental effort can provide systematic functional study of one individual conotoxin or even small groups of toxins, systematic functional study is extremely time-consuming and costly. It is increasingly difficult to study them with wet-experimental method alone. Bioinformatics offers a promising and efficient methodology for the analysis of the *in silico* possible functions of new conotoxins and the acceleration of the *in vitro* screening of potential conotoxin candidates. Therefore, we constructed a prediction model to identify the types of ion channel-targeted conotoxins. Through optimizing features with binomial distribution theory and utilizing RBF network in the prediction, high prediction accuracy was obtained. The model may become the important tool for analysis of these pharmacologically important peptides. And the predicted results on the function-undefined conotoxins may improve relevant research in the assistance of selection and design of critical laboratory experiments.

In pattern recognition, feature selection technique plays an important role in improving the accuracy of the model. In this study, we adopted the binomial distribution to optimize dipeptides. Although the statistical method can pick out over-represented dipeptides, it cannot directly provide the correlation information between two dipeptides. Hybridizing different parameters to represent sequence is an important and popular strategy for improving predictive accuracy. However, for the dimensions of different kinds of features are different, our proposed strategy is limited in dealing with hybrid parameters selection. In the future, we will develop an approach of cascade prediction combined with binomial distribution to further improve the performance of the model.

In summary, we introduced a novel method to identify the types of ion channel-targeted conotoxins only using primary sequence information. The binomial distribution can fully utilize the important features of different types of conotoxins. High predicted accuracies demonstrate that the proposed model is an effective tool to predict K^+, Na^+ and Ca^+ channel-targeted conotoxins. Thus, this method will become a useful tool for conotoxin analysis and further experimental research.

Conflict of interest

The authors have no conflict to disclose.

Acknowledgements

This work was supported by the National Nature Scientific Foundation of China (Nos. 61202256 and 61100092), the Fundamental Research Funds for the Central Universities (ZYGX2012J113) and the Scientific Research Startup Foundation of UESTC.

References

Anthony, M.L., Rose, V.S., Nicholson, J.K., Lindon, J.C., 1995. Classification of toxin-induced changes in 1H NMR spectra of urine using an artificial neural network. J. Pharm. Biomed. Anal. 13, 205–211.

Chen, C., Tian, Y.X., Zou, X.Y., Cai, P.X., Mo, J.Y., 2006. Using pseudo-amino acid composition and support vector machine to predict protein structural class. J. Theor. Biol. 243, 444–448.
Chen, S.A., Ou, Y.Y., Lee, T.Y., Gromiha, M.M., 2011. Prediction of transporter targets using efficient RBF networks with PSSM profiles and biochemical properties. Bioinformatics 27, 2062–2067.
Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43, 246–255.
Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247.
Daly, N.L., Craik, D.J., 2009. Structural studies of conotoxins. IUBMB Life 61, 144–150.
Fan, Y.X., Song, J., Shen, H.B., Kong, X., 2011. PredCSF: an integrated feature-based approach for predicting conotoxin superfamily. Protein Pept. Lett. 18, 261–267.
Feng, Y., Luo, L., 2008. Use of tetrapeptide signals for protein secondary-structure prediction. Amino Acids 35, 607–614.
Gowd, K.H., Dewan, K.K., Iengar, P., Krishnan, K.S., Balaram, P., 2008. Probing peptide libraries from Conus achatinus using mass spectrometry and cDNA sequencing: identification of delta and omega-conotoxins. J Mass Spectrom. 43, 791–805.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11, 10–18.
Han, T.S., Teichert, R.W., Olivera, B.M., Bulaj, G., 2008. Conus venoms – a rich source of peptide-based therapeutics. Curr. Pharm. Des. 14, 2462–2479.
Hayat, M., Khan, A., Yeasin, M., 2012. Prediction of membrane proteins using split amino acid and ensemble classification. Amino Acids 42, 2447–2460.
Hillyard, D.R., Mcintosh, M.J., Jones, R.M., Cartier, E.G., Watkins, M., Olivera, B.M., Layer, R.T. 2008. O-superfamily conotoxin peptides. Patent number JP2003533178.
Huang, T., Chen, L., Cai, Y.D., Chou, K.C., 2011. Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property. PLoS One 6, e25297.
Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.
Li, B.Q., Hu, L.L., Niu, S., Cai, Y.D., Chou, K.C., 2012. Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches. J. Proteomics 75, 1654–1665.
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, Q.Z., 2007. Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem. Biophys. Res. Commun. 354, 548–551.
Lin, C.T., Lin, K.L., Yang, C.H., Chung, I.F., Huang, C.D., Yang, Y.S., 2005. Protein metal binding residue prediction based on neural networks. Int. J. Neural Syst. 15, 71–84.
Liu, L., Wu, X., Yuan, D., Chi, C., Wang, C., 2008. Identification of a novel S-superfamily conotoxin from vermivorous Conus caracteristicus. Toxicon 51, 1331–1337.
Lu, L., Niu, B., Zhao, J., Liu, L., Lu, W.C., Liu, X.J., Li, Y.X., Cai, Y.D., 2009. GalNAc-transferase specificity prediction based on feature selection method. Peptides 30, 359–364.
Magrane, M., Consortium, U., 2011. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) r9, r9.
Mondal, S., Bhavna, R., Mohan, B.R., Ramakumar, S., 2006. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J. Theor. Biol. 243, 252–260.
Ou, Y.Y., Gromiha, M.M., Chen, S.A., Suwa, M., 2008. TMBETADISC-RBF: Discrimination of beta-barrel membrane proteins using RBF networks and PSSM profiles. Comput. Biol. Chem. 32, 227–231.
Ou, Y.Y., Chen, S.A., Gromiha, M.M., 2010. Classification of transporters using efficient radial basis function networks with position-specific scoring matrices and biochemical properties. Proteins 78, 1789–1797.
Park, K.J., Gromiha, M.M., Horton, P., Suwa, M., 2005. Discrimination of outer membrane proteins using support vector machines. Bioinformatics 21, 4223–4229.
Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1226–1238.
Peng, C., Liu, L., Shao, X., Chi, C., Wang, C., 2008. Identification of a novel class of conotoxins defined as V-conotoxins with a unique cysteine pattern and signal peptide sequence. Peptides 29, 985–991.
Rocchi, L., Chiari, L., Cappello, A., 2004. Feature selection of stabilometric parameters based on principal component analysis. Med. Biol. Eng. Comput. 42, 71–79.
Shen, H.B., Chou, K.C., 2009. QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. J. Proteome. Res. 8, 1577–1584.
Tan, P.T., Khan, A.M., Brusic, V., 2003. Bioinformatics for venom and toxin sciences. Brief Bioinform. 4, 53–62.
Terlau, H., Olivera, B.M., 2004. Conus venoms: a rich source of novel ion channel-targeted peptides. Physiol. Rev. 84, 41–68.
Watters, M.R., 2005. Tropical marine neurotoxins: venoms to drugs. Semin. Neurol. 25, 278–289.
Xiao, X., Shao, S.H., Huang, Z.D., Chou, K.C., 2006. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J. Comput. Chem. 27, 478–482.
Xiao, X., Wang, P., Chou, K.C., 2011. GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. Mol. Biosyst. 7, 911–919.
Yin, J.B., Fan, Y.X., Shen, H.B., 2011. Conotoxin superfamily prediction using diffusion maps dimensionality reduction and subspace classifier. Curr. Protein Pept. Sci. 12, 580–588.
Zaki, N., Sibai, F., Campbell, P., 2011a. Conotoxin protein classification using pairwise comparison and amino acid composition: toxin-AAM. In: Proceedings of the 13th annual conference on Genetic and evolutionary computation. ACM, Dublin, Ireland, pp. 323–330.
Zaki, N., Wolfsheimer, S., Nuel, G., Khuri, S., 2011b. Conotoxin protein classification using free scores of words and support vector machines. BMC Bioinformatics 12, 217.
Zhou, G.P., Cai, Y.D., 2006. Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. Proteins 63, 681–684.