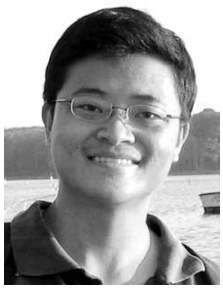
PredCSF: An Integrated Feature-Based Approach for Predicting Conotoxin Superfamily

Yong-Xian Fan[1], Jiangning Song[2,3], Xiangzeng Kong[4] and Hong-Bin Shen[1,*,#]



[1]*Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China;* [2]*Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjian 300308, China;* [3]*Department of Biochemistry and Molecular Biology, Monash University, Clayton, Melbourne, VIC 3800, Australia;* [4]*School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, China*

Abstract: Conotoxins are small disulfide-rich peptides that are invaluable channel-targeted peptides and target neuronal receptors. They show prospects for being potent pharmaceuticals in the treatment of Alzheimer's disease, Parkinson's disease, and epilepsy. Accurate and fast prediction of conotoxin superfamily is very helpful towards the understanding of its biological and pharmacological functions especially in the post-genomic era. In the present study, we have developed a novel approach called PredCSF for predicting the conotoxin superfamily from the amino acid sequence directly based on fusing different kinds of sequential features by using modified one-versus-rest SVMs. The input features to the PredCSF classifiers are composed of physicochemical properties, evolutionary information, predicted secondary structure and amino acid composition, where the most important features are further screened by random forest feature selection to improve the prediction performance. The results show that PredCSF can obtain an overall accuracy of 90.65% based on a benchmark dataset constructed from the most recent database, which consists of 4 main conotoxin superfamilies and 1 class of non-conotoxin class. Systematic experiments also show that combing different features is helpful for enhancing the prediction power when dealing with complex biological problems. PredCSF is expected to be a powerful tool for *in silico* identification of novel conotonxins and is freely available for academic use at http://www.csbio.sjtu.edu.cn/bioinf/PredCSF.

[#]**Author's Profile:** Dr. Hong-Bin Shen is a professor of Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University. His research interests include data mining, protein structure and function prediction, drug discovery and biological network. Dr. Shen has published more than 60 papers and constructed 20 bioinformatics servers in these areas.

Keywords: Conotoxin, physicochemical property, wavelet analysis, random forest, PSSM, PredCSF.

1. INTRODUCTION

Conotoxins are made up of the signal peptide sequence, a propeptide region and the mature peptide sequence that is small and disulfide-rich [1]. They are produced by diverse species of cone snails which are medium-sized to large, intricate predatory snails. The venom of cone snails, such as Magician cone, shows much promise for providing a non-addictive pain reliever 1000 times as powerful as morphine [2]. Many peptides made by the cone snails show prospects for being potent pharmaceuticals, such as compounds of the toxin that can be used in the treatment of Alzheimer's disease, Parkinson's disease, and epilepsy [3-6].

Because there are distinct characteristics such as highly conserved N-terminal precursor sequence, disulfide-crosslinked and similar mode of actions, conotoxins have been grouped into several different superfamilies, namely, A, I, M, O, T, P, S, J, D, V and C superfamily [1, 7]. Each superfamily can further be classified into several families on the basis of the cysteine arrangement. Currently, the genus Conus contains over 700 species, representing a peptide library on the order of 70,000 sequences [8, 9] and we are facing the problem that the majority of conotoxin sequences have not been efficiently classified into their native groups and such a problem is becoming worse with the new conotoxin sequences generated with the new sequencing technologies. Hence, development of systematic computational approaches to classify conotoxin sequences into their superfamilies is an urgent and challenging task. Several studies have been attempted to this task. Modal *et al*. have used pseudo amino acid composition and multi-class support vector machines for dividing 116 conotoxins into their each superfamilies and obtained the overall sensitivity of 87.93% for the four superfamilies [10, 11]. Hao *et al*. have also used pseudo amino acid composition, but the modified Mahalanobis discriminant for this purpose [12]. However, none of these studies has provided the internet based bioinformatics servers for use.

In this study, a novel method is developed to predict conotoxin superfamilies from the amino acid sequence directly by fusing different kinds of features. We firstly extract the features by using maximal overlap discrete wavelet transform (MODWT) [13] based on 246 physicochemical properties of amino acids [14], then a random forest approach is

*Address correspondence to this author at the Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China; Tel: +86-21-34205320; Fax: +86-21-34204022;
E-mail: hbshen@sjtu.edu.cn

0929-8665/11 $58.00+.00

applied to rank the physicochemical properties' importance and the relatively important features are selected into the first feature subset. To take advantage of the evolutionary information regarding protein sequences and structures, the PSSM (Position-Specific Scoring Matrix) [15] and the predicted SS (secondary structure) [16, 17] are used to generate the second and third feature subsets respectively. The fourth feature subset is constituted based on the amino acid compositions. The four feature subsets are thus combined together and fed into modified multi-class SVMs (support vector machines) for final prediction.

2. MATERIALS AND METHODS

2.1. Datasets

The mature peptides and the corresponding full sequences of conotoxins were extracted from the Swiss-Prot release 57.8 (released on 22-Sep-09). Because the number of entries in some superfamilies like P, S, J, L, D, V and C were less than 10 entries, too few to have statistic significance, these superfamilies were excluded. The I-conotoxin superfamily was not included because there are still some debates about the classification scheme of the I superfamily. For example, some researchers shown that the conotoxins that were previously assigned to the I superfamily should be separated into two different gene superfamilies, namely I1 and I2 [7, 18]. The remaining data set included 403 conotoxin sequences from A, M, O, and T superfamilies. In order to reduce the bias of sequence homology, the redundant sequences with pairwise sequence identity greater than 80% were excluded by using CD-HIT program [19], which make the final data set consisting of 261 entries from four superfamilies, i.e. A (63 entries), M (48 entries), O (95 entries) and T (55 entries). 60 short cysteine rich mature sequences of non-conotonxin sequences were adopted as a negative control data set as constructed by Modal *et al*. [10]. All the samples can be found in the online supporting information.

2.2. Physicochemical Property Information

In order to investigate the importance of the amino acid physicochemical properties in a peptide, we have retrieved all the available physicochemical properties from the APDbase database at http://www.rfdn.org/bioinfo/APDbase/. Currently there are 246 amino acid physicochemical properties in APDbase [14].

Suppose that a protein **P** can be expressed as:

$$\mathbf{P} = R_1R_2R_3R_4R_5R_6R_7\cdots R_L \quad (1)$$

Where R_1 represents the 1st residue of the protein **P**, R_2 the 2nd residue, and so forth. Substituting the 246 amino acid physicochemical properties into Eq. (1), then we can obtain 246 signals, which can be expressed by:

$$\begin{cases} \mathbf{P}_1 = M_1^1M_2^1M_3^1M_4^1M_5^1M_6^1M_7^1\cdots M_L^1 \\ \mathbf{P}_2 = M_1^2M_2^2M_3^2M_4^2M_5^2M_6^2M_7^2\cdots M_L^2 \\ \vdots \\ \mathbf{P}_{246} = M_1^{246}M_2^{246}M_3^{246}M_4^{246}M_5^{246}M_6^{246}M_7^{246}\cdots M_L^{246} \end{cases} \quad (2)$$

Where $\mathbf{P_1}$ denotes the signal generated from the 1st physicochemical property for protein **P** , M_1^1 means the 1st amino acid physicochemical property for the 1st residue, M_2^1 means the 1st amino acid physicochemical property for the 2nd residue and so forth. Fig. (**1**) shows an example of the discrete signal of the CXA14_CONPL peptide using the amino acid physicochemical property of conformational parameter of beta-turn [20].

2.2.1. Maximal Overlap Discrete Wavelet Transform

One of the major steps of designing an effective bioinformatics prediction model is how to encode a protein sequence into a vector by representative features. In this study, a wavelet analysis based approach is used to design feature vector elements that incorporates physicochemical properties of amino acids. The wavelet transform decomposes a signal into several groups (vectors) of coefficients in which different coefficient vectors contain information about the characteristics of the sequence at different scales. Coefficients at coarse scales capture global features of the proteins, whereas coefficients at fine scales contain local details. Therefore, by defining a feature vector in terms of wavelet statistics can make us have the information of physicochemical variations at different scales, which will be useful for the protein's functional and structural classification.

The maximal overlap discrete wavelet transform (MODWT) is an important improvement of the traditional discrete wavelet transform (DWT) and has been a crucial tool for the analysis of different types of time series data [21]. The MODWT is different from the traditional DWT in that it is a highly redundant and non-orthogonal transform, which makes MODWT can deal with any sample size *N*, while the *J*-th order DWT restricts the sample size to multiples of 2^J. This is quite important for current study because the length of the protein sequences is not a multiple of 2^J. MODWT was implemented using wavelet method for time series analysis (WMTSA) toolkit available at http://www.atmos.washington.edu/~wmtsa/.

As most of the mature peptide sequences have sequence lengths greater than or equal to 9, we thus selected *J*=3. In this study, Daubechies wavelet is used for analysis, which will yield 3 groups (vectors) of coefficients at fine scales and 1 group of coefficients at coarse scales. As shown in Fig. (**1**), D1, D2, and D3 denote three detailed coefficients from *j*=1 to *j*=3 (*j*=1,2,…,*J*); and A3 denotes the approximation coefficient at level *j* = 3 after using MODWT on the original signal. To decrease the dimensionality of the extracted feature vectors, the set of the wavelet coefficients are statistically analyzed. The following statistical features calculated from the approximation coefficients and detailed coefficients are used: (i) mean of the approximation coefficient A3 (mean of the detail coefficients is not adopted because it always equal to zero), (ii) standard deviation of the approximation A3 and three detailed coefficients D1, D2 and D3. So, a single mature peptide can be characterized as a (1+1+3)×246=1230 dimensional feature vector as described as follows:

$$F_{MODWT} = [f_1^{1,1}, f_2^{1,2}, f_3^{1,3}, f_4^{1,4}, f_5^{1,5}, f_6^{2,1}, f_7^{2,2}, f_8^{2,3}, f_9^{2,4}, f_{10}^{2,5}, \cdots, f_{1230}^{246,5}] \quad (3)$$

where the subscript 1 stands for the 1st feature, and the superscript (1,1) stands for the 1st statistics of the 1st amino acid physicochemical property for $f_1^{1,1}$, the subscript 2 stands for

the 2nd feature, and the superscript (1,2) means for the 2nd statistics of the 1st amino acid physicochemical property for $f_2^{1,2}$ and so forth.

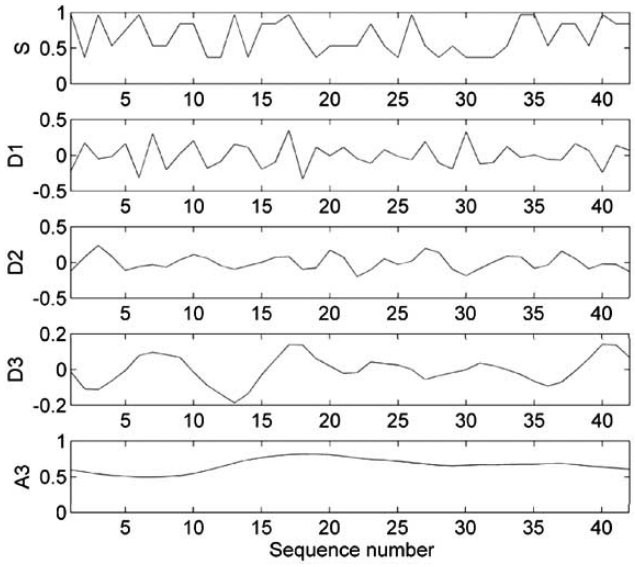


Figure 1. MODWT wavelet coefficients plot of the discrete signal of the CXA14_CONPL peptide sequence based on the amino acid physicochemical property of conformational parameter. S is the original discrete signal; D1, D2, and D3 denote three detailed coefficients from j=1 to j=3 (j=1,2,…,J), respectively; A3 denotes the approximation coefficient at level j = 3 after the wavelet analysis of MODWT on S.

As shown in Eq. (3), we can represent a single mature peptide sequence by a 1230-D vector for a single mature peptide based on the MODWT wavelet analysis from the 246 amino acid physicochemical properties. Now the question is which physicochemical properties (or wavelet coefficient statistics) are more important for the current classification task. It is necessary to select the most important features to reduce the initially high dimension feature vector to a much lower space to avoid "dimension disaster" or "over-fitting", which is an critical issue as discussed by Smialowski in [22]. Random forests algorithm is applied to handle this problem and analyze the importance of each of the 1230-D features, which has been demonstrated very effective for this purpose [23-25]. The random forest is constructed by fusing multiple decision trees so that each tree depends on the values of a random vector sampled independently from the initial feature space and with the identical distribution for all trees in the forest. According to the criteria that selected different features need to be the tree nodes in the decision tree, the random forest is able to evaluate the importance of different features. We ranked the 1230 features according to the importance degree outputted from the random forest and an optimal threshold is then used to determine whether a feature should be excluded or not. Based on the cross-validation, 18 of the 1230 coefficients are selected with the optimized random forest threshold 1.1, which is illustrated in Fig. (**2**). As a result, we obtained the first feature subset of 18 feature elements.

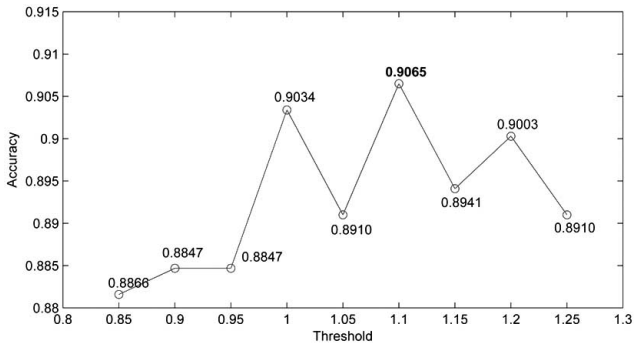


Figure 2. Threshold optimization for selecting important wavelet coefficients based on random forest.

2.3. Position-Specific Scoring Matrix (PSSM) Information

Numerous studies have shown that evolutionary information contained in protein multiple sequence alignments is important for improving the predictive performance [26, 27]. To incorporate this information, we generated the PSSM matrix of $L \times 20$ using PSI-BLAST [28] to search the non-redundant (NR) protein sequence database through three iterations with 0.001 as the *E*-value cutoff for multiple sequence alignment against the sequence of the conotoxin.

$$\mathrm{SEQ}_{\mathrm{PSSM}} = \begin{bmatrix} \mathrm{E}_{1\to1} & \mathrm{E}_{1\to2} & \cdots & \mathrm{E}_{1\to20} \\ \mathrm{E}_{2\to1} & \mathrm{E}_{2\to2} & \cdots & \mathrm{E}_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{E}_{L\to1} & \mathrm{E}_{L\to2} & \cdots & \mathrm{E}_{L\to20} \end{bmatrix} \quad (4)$$

where $\mathrm{E}_{i\to j}$ can partially stands for the score of the amino acid residue in the i-th position of the sequence being converted into the amino acid type j during the evolution process.

To make the $\mathrm{SEQ}_{\mathrm{PSSM}}$ a same-sized matrix, one feasible method is to represent a peptide by

$$\overline{\mathrm{PEP}}_{\mathrm{PSSM}} = \begin{bmatrix} \overline{\mathrm{E}}_1 & \overline{\mathrm{E}}_2 & \cdots & \overline{\mathrm{E}}_{20} \end{bmatrix} \quad (5)$$

where $\overline{\mathrm{E}}_j = \frac{1}{\mathrm{L}} \sum_{i=1}^{\mathrm{L}} \mathrm{E}_{i\to j} \quad (j = 1,2,\cdots,20)$ of Eq. (**5**). In this way, we then obtain another subset of features consisting of 20 elements.

2.4. Secondary Structure (SS) Information

Given a sequence, its secondary structure can be predicted by means of various secondary structure prediction tools. In the present study, based on the information thus obtained by using PSIPRED program, we can represent the protein **P** of L residues by a matrix (L rows and 3 columns):

$$\mathrm{S} = \begin{bmatrix} \alpha_1 & \beta_1 & c_1 \\ \alpha_2 & \beta_2 & c_2 \\ \vdots & \vdots & \vdots \\ \alpha_L & \beta_L & c_L \end{bmatrix} \quad (6)$$

where α_i is the probability of the *i*-th residue being α-helix, β_i the β-sheet, and c_i the coiled-coil. According to PSIPRED, $\alpha_i + \beta_i + c_i = 1$. Based on Eq. (**6**), we have the secondary structure content ratios for the whole protein chain **P**, as formulated by

$$\begin{cases} \Gamma_\alpha = \sum_{i=1}^{L} \alpha_i / L \\ \Gamma_\beta = \sum_{i=1}^{L} \beta_i / L \\ \Gamma_C = \sum_{i=1}^{L} c_i / L \end{cases} \quad (7)$$

where Γ_α, Γ_β and Γ_C are the ratios of the α-helix, β-sheet, and coiled-coil residues for the protein **P** and then Γ_α, Γ_β and Γ_C are then used as three global features for **P**, which gives us another 3 features.

2.5. Amino Acid Composition

We also took into account the global information of mature peptides, represented by the amino acid composition (AAC). AAC can be represented by a 20-dimensional vector, where each element denotes each amino acid's occurrence in the whole sequence.

Finally, according to above steps, four different feature subsets are constructed, i.e. the wavelet coefficients from the physicochemical properties, the evolutional information, the predicted secondary structure information and the amino acid composition. By combining them together, a mature peptide sequence is then can be represented by a 18+20+3+20=61-D vector.

2.6. Modified One-Versus-Rest SVMs

SVM is a kind of learning machine based on statistical learning theory. SVM is fascinating to biological sequence analysis because of its robustness and generalization ability. For its implementation, we used the LIBSVM package (Version 2.89) [29]. However, SVM was originally designed for binary classification, whereas prediction of conotoxin superfamilies is a multiclass classification problem. To solve this problem, one-versus-rest (o-v-r) or one-versus-one (o-v-o) approach can employed to decompose multiclass into a series of binary SVMs. This method includes the construction of each binary SVM classifier. In present study, we have constructed five SVM classifiers, i.e. SVM-A specifically for the A type superfamily, SVM-M for the M superfamily, SVM-O for the O superfamily, SVM-T for the T superfamily and SVM-N for the negative control dataset. The kernel parameter and regularization parameter are optimized for each SVM classifier as shown in Table **1**. In response to a query sequence, each classifier will output a probability value of current input belonging to the corresponding class. The query sequence will be assigned to the class with the highest propensity value.

The entire predictor thus developed is called PredCSF (Predicting Conotoxin SuperFamily). To help to better understand the architecture of PredCSF, a flowchart showing how to predict conotoxin superfamily from primary sequence by PredCSF is given in Fig. (**3**).

2.7. Assessment of Predictive Ability

The predictive ability of the present approach is evaluated with several measures, namely, Sensitivity (S_n), Specificity (S_p), the Mathews correlation coefficient (MCC), and the overall accuracy (Acc). They are defined as follows:

$$S_n = \frac{TP}{TP + FN}$$

$$S_p = \frac{TP}{TP + FP}$$

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

$$Acc = \frac{\sum_i TP_i}{\sum_i (TP_i + FN_i)}$$

where TP, FP, TN, and FN denote the number of true positives, false positives, true negatives, and false negatives, respectively. TP_i and FN_i denote the number of true positives and false negatives of class *i*, respectively.

3. RESULTS AND DISCUSSION

We followed the procedure shown in Fig. (**3**) to predict the conotoxin superfamilies and the results are presented in

Table 1. The Parameters of Five SVM Classifiers Built in This Study

SVM Classifiers [a]	SVM Kernel	Regularization Parameter	Kernel Parameter
SVM-A	Gaussian	8	0.25
SVM-M	Gaussian	32	1
SVM-O	Gaussian	4	1
SVM-T	Gaussian	32	0.125
SVM-N	Gaussian	4	0.5

[a]SVM-A, SVM-M, SVM-O, SVM-T and SVM-N stand for the one-versus-rest classifiers built for four different conotoxin superfamilies plus the negative data set.

Table **2**. As shown in Table **2**, PredCSF correctly identified the conotoxin superfamilies with the overall accuracy Acc 90.65% through the jackknife test which is considered as the most stringent and objective test [30]. It can be seen that the sensitivity in predicting the A, M, O, T conotoxin superfamilies and the non-conotoxin superfamily are 84.13%, 93.75%, 93.68%, 94.55% and 86.67%, respectively.

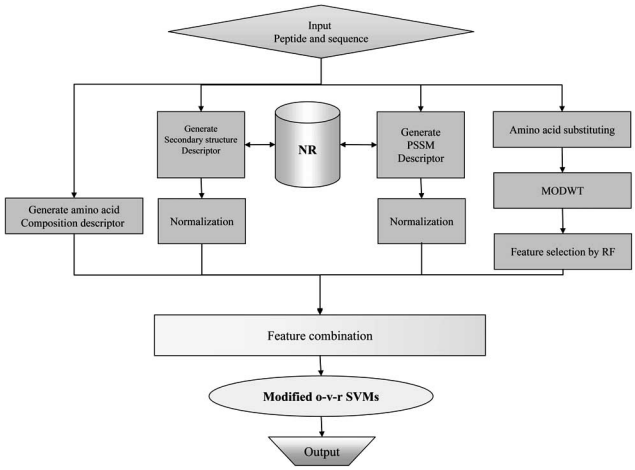


Figure 3. A flowchart diagram of showing how PredCSF predictor works.

We have employed several different sequential encoding schemes and it would be interesting to further investigate their independent contributions. We thus performed extensive experiments on seven different combinations of encoding schemes in order to evaluate their corresponding contributions. The jackknife test results are thus reported in Table **3**. These results show that using the PSSM feature subset alone gives the best sensitivity, specificity and the overall accuracy, in comparison with other single feature subsets. This indicates that the evolutionary information in the PSSM profile is the most important determinant for predicting the contoxin superfamily. Furthermore, when gradually incorporating the SS, AAC and PCP feature subsets with the PSSM subset, i. e. adopting the PSSM+SS, PSSM+SS+AAC and PSSM+SS+AAC+PCP sequence encoding schemes, MCC of 88.47%, 89.41% and 90.65% are obtained respectively. It is interesting to note that the predictive performance improves with the increasing complexity of the added features. As a result, the sequence encoding scheme PSSM+SS+AAC+PCP achieves the best overall accuracy.

In order to fairly compare the proposed prediction model with previously published methods, the same benchmark dataset constructed by Mondal *et al*. [10] is employed, which consists of 5 classes, i.e. 25 A- conotoxins, 13 M-conotoxins, 61 O-conotoxins, 16 T-conotoxins and 60 non-conotoxin proteins. Table **4** shows the rigorous jackknife test of different algorithms. As can be seen, the overall accuracy by PredCSF is 90.34%, which is about 2-18% respectively higher than the other methods based on the same datasets when using the jackknife test to evaluate the performance, indicating the proposed prediction scheme is really promising.

Table 2. The Predictive Performances of the PredCSF Algorithm by the Jackknife Test for 321 Mature Peptides

	A (%)	M (%)	O (%)	T (%)	N (%)
S_n	53/63=84.13	45/48=93.75	89/95=93.68	52/55=94.55	52/60=86.67
S_p	53/58=91.38	45/47=95.74	89/99=89.90	52/56=92.86	52/61=85.25
MCC	84.84	93.83	88.22	92.38	82.69

A, M, O and T denote the four conotoxin superfamilies, and N denotes the negative data set.

Table 3. The Predictive Performance Comparison Using Different Feature Subsets

Features [a]	S_n(%) [b]					S_p(%)					Overall Acc(%)
	A	M	O	T	N	A	M	O	T	N	
SS	52.38	8.33	91.58	94.55	68.33	66.00	44.44	71.31	60.47	75.93	67.60
AAC	63.49	64.58	82.11	56.36	68.33	67.80	58.49	70.91	67.39	77.36	68.85
PCP	77.78	91.67	83.16	90.91	68.33	84.48	89.80	74.53	92.59	75.93	81.93
PSSM	77.78	93.75	91.58	96.36	80.00	87.50	90.00	87.00	92.98	82.76	87.85
PSSM+SS	77.78	93.75	93.68	96.36	80.00	87.50	90.00	89.00	91.38	84.21	88.47
PSSM+SS+AAC	80.95	95.83	94.74	92.73	81.67	89.47	90.20	89.11	94.44	84.48	89.41
PSSM+SS+AAC+PCP	84.13	93.75	93.68	94.55	86.67	91.38	95.74	89.90	92.86	85.25	**90.65**

[a]SS denotes the predicted secondary structure, AAC the amino acid composition, PCP the wavelet features from the physicochemical properties, PSSM the position specific scoring matrix. [b]A, M, O and T denote the four conotoxin superfamilies, and N denotes the negative data set.

Table 4. The Predictive Performance Comparison between PredCSF and Other Methods

Method	S_n(%) [a]					S_p(%)					Overall Acc(%)
	A	M	O	T	N	A	M	O	T	N	
ISort predictor[10]	76.00	69.23	70.49	88.24	68.33	79.17	60.00	68.25	78.95	74.55	72.16
Least Hamming distance[10]	80.00	53.85	77.05	82.35	71.67	66.67	53.85	72.31	82.35	84.31	74.43
Least Euclidean distance[10]	76.00	53.85	73.77	82.35	73.33	70.37	77.78	71.43	73.68	75.86	73.30
One-versus-rest SVMs[10]	84.00	84.62	81.97	76.47	76.67	95.45	100	96.15	92.86	88.46	80.11
Multi-class SVMs[10]	84.00	92.31	86.89	94.12	88.33	95.45	80.00	86.89	94.12	86.89	88.07
IDQD[12]	96.0	92.3	82.0	94.1	89.2	92.3	100	89.3	100	91.7	88.3
PredCSF	88.00	76.92	93.44	94.12	90.00	100	76.92	87.69	100	90.00	**90.34**

[a] A, M, O and T denote the four conotoxin superfamilies, and N denotes the negative data set.

4. CONCLUSIONS

In this paper, an integrated multiple feature-based approach is proposed for the prediction of the conotoxin superfamily by combining different features extracted directly from the amino acid sequence. Specifically, a wavelet-based method is used to analyze the sequential signal obtained from 246 amino acid physicochemical properties, followed by selecting the most important features through the use of a random forest algorithm. These features are further combined with the evolutionary information, predicted secondary structure and amino acid composition. Systematic experiments suggest that the prediction power of PredCSF is in truth enhanced by fusion of these different types of sequence descriptors. As an implementation of our approach, PredCSF has been made freely available for academic use at http://www.csbio.sjtu.edu.cn/bioinf/PredCSF, which is anticipated to become a powerful tool in the area of *in silico* identification of the conotoxin superfamilies.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 60704047), A Foundation for the Author of National Excellent Doctoral Dissertation of PR China (Grant No. 2011048), Science and Technology Commission of Shanghai Municipality (Grant No. 08ZR1410600, 08JC1410600), sponsored by Shanghai Pujiang Program, Innovation Program of Shanghai Municipal Education Commission (10ZZ17), Shanghai Jiao Tong University Innovation Fund for Postgraduates and Shanghai Leading Academic Discipline Project (Grant No. S30201). JS would like to thank the National Health and Medical Research Council of Australia (NHMRC) for financially supporting this research via the NHMRC Peter Doherty Fellowship and the Hundred Talents Program of the Chinese Academy of Sciences (CAS).

REFERENCES

[1] Terlau, H.; Olivera, B. Conus venoms: a rich source of novel ion channel-targeted peptides. *Physiol. Rev.*, **2004**, *84*(1), 41-68.

[2] Mehdiratta, R.; Saberwal, G. Bio-business in brief: the case of conotoxins. *Curr. Sci.*, **2007**, *92*(1), 39-45.

[3] Olivera, B.; Cruz, L. Conotoxins, in retrospect. *Toxicon*, **2001**, *39*(1), 7-14.

[4] Jimenez, E.C.; Donevan, S.; Walker, C.; Zhou, L.M.; Nielsen, J.; Cruz, L.J.; Armstrong, H.; White, H.S.; Olivera, B.M. Conantokin-L, a new NMDA receptor antagonist: determinants for anticonvulsant potency. *Epilepsy Res.*, **2002**, *51*(1-2), 73-80.

[5] Livett, B.; Gayler, K.; Khalil, Z. Drugs from the sea: conopeptides as potential therapeutics. *Curr. Med. Chem.*, **2004**, *11*(13), 1715-1723.

[6] Twede, V.; Miljanich, G.; Olivera, B.; Bulaj, G. Neuroprotective and cardioprotective conopeptides: an emerging class of drug leads. *Curr. Opin. Drug Discov. Devel.*, **2009**, *12*(2), 231-239.

[7] Mondal, S.; Babu, R.M.; Bhavna, R.; Ramakumar, S. I-conotoxin superfamily revisited. *J. Pept. Sci.*, **2006**, *12*(11), 679-685.

[8] Halai, R.; Craik, D. Conotoxins: natural product drug leads. *Nat. Prod. Rep.*, **2009**, *26*(4), 526-536.

[9] Jacob, R.B.; McDougal, O.M. The M-superfamily of conotoxins: a review. *Cellular and Molecular Life Sciences*, **2010**, *67*(1), 17-27.

[10] Mondal, S.; Bhavna, R.; Babu, R.M.; Ramakumar, S. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. *J. Theor. Biol.*, **2006**, *243*(2), 252-260.

[11] Du, P.; Li, Y. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. *BMC Bioinformatics*, **2006**, *7*(1), 518.

[12] Lin, H.; Li, Q.Z. Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. *Biochem. Biophys. Res. Commun.*, **2007**, *354*(2), 548-551.

[13] Buendia, F.; Tarquis, A.M.; Buendia, G.; Andina, D. Feature Extraction Via Multiresolution MODWT Analysis in a Rainfall Forecast System. *Wmsci 2008: 12th World Multi-Conference on Systemics, Cybernetics and Informatics, Vol Viii, Proceedings*, **2008**, 69-73, p. 69-73, 178.

[14] Mathura, V.S.; Kolippakkam, D. APDbase: amino acid physicochemical properties Database. *Bioinformation*, **2005**, *1*(1), 2-4.

[15] Schaffer, A.A.; Aravind, L.; Madden, T.L.; Shavirin, S.; Spouge, J.L.; Wolf, Y.I.; Koonin, E.V.; Altschul, S.F. Improving the accuracy of PSI-BLAST protein database searches with composition-

based statistics and other refinements. *Nucleic Acids Res.*, **2001**, *29*(14), 2994-3005.

[16] Song, J.; Burrage, K.; Yuan, Z.; Huber, T. Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information. *BMC Bioinformatics*, **2006**, *7*, 124.

[17] Song, J.; Yuan, Z.; Tan, H., Huber, T.; Burrage, K. Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure. *Bioinformatics*, **2007**, *23*(23), 3147-3154.

[18] Buczek, O.; Yoshikami, D.; Watkins, M.; Bulaj, G.; Jimenez, E.C.; Olivera, B.M. Characterization of D-amino-acid-containing excitatory conotoxins and redefinition of the I-conotoxin superfamily. *FEBS J.*, **2005**, *272*(16), 4178-4188.

[19] Li, W.; Jaroszewski, L.; Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **2001**, *17*(3), 282-283.

[20] Beghin-Dirkx. Une methode statistique simple de prediction des conformations proteiques. *Physiol. Biochim.*, **1975**, *83*, 167-168.

[21] Percival, D.B.; Walden, A.T. Wavelet *Methods for Time Series Analysis*. Cambridge University Press: Cambridge, **2000**.

[22] Smialowski, P.; Frishman, D.; Kramer, S. Pitfalls of supervised feature selection. *Bioinformatics*, **2010**, *26*(3), 440-443.

[23] Breiman, L. Random forests. *Machine Learning*, **2001**, *45*(1), 5-32.

[24] Pan, X.Y.; Shen, H.B. Robust prediction of B-factor profile from sequence using two-stage SVR based on random forest feature selection. *Protein Pept. Lett.*, **2009**, *16*(12), 1447-1454.

[25] Tian, W.; Zhang, L.V.; Tasan, M.; Gibbons, F.D.; King, O.D.; Park, J.; Wunderlich, Z.; Cherry, J.M.; Roth, F.P. Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function. *Genome Biol.*, **2008**, *9*(Suppl 1), S7.

[26] Shen, H.B.; Chou, K.C. QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. *J. Proteome Res.*, **2009**, *8*(3), 1577-1584.

[27] Mishra, N.K.; Raghava, G.P. Prediction of FAD interacting residues in a protein from its primary sequence using evolutionary information. *BMC Bioinformatics*, **2010**, *11*(Suppl 1), S48.

[28] Altschul, S. Hot papers - Bioinformatics - Gapped BLAST and PSI-BLAST: a new generation of protein database search programs by S.F. Altschul, T.L. Madden, A.A. Schaffer, J.H. Zhang, Z. Zhang, W. Miller, D.J. Lipman - Comments. *Scientist*, **1999**, *13*(8), 15-15.

[29] Chang, C.; Lin, C. LIBSVM: a library for support vector machines. **2001**, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[30] Chou, K.C.; Zhang, C. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, **1995**, *30*(4), 275-349.

Received: April 26, 2010 **Revised: September 26, 2010** **Accepted: September 28, 2010**