Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 3981478, 5 pages
http://dx.doi.org/10.1155/2016/3981478

Research Article

Identifying the Types of Ion Channel-Targeted Conotoxins by Incorporating New Properties of Residues into Pseudo Amino Acid Composition

Yun Wu,[1] Yufei Zheng,[1] and Hua Tang[2]

[1]*College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China*
[2]*Department of Pathophysiology, Southwest Medical University, Luzhou 646000, China*

Correspondence should be addressed to Hua Tang; tanghua771211@aliyun.com

Received 13 July 2016; Accepted 31 July 2016

Academic Editor: Ren-Zhi Cao

Conotoxins are a kind of neurotoxin which can specifically interact with potassium, sodium type, and calcium channels. They have become potential drug candidates to treat diseases such as chronic pain, epilepsy, and cardiovascular diseases. Thus, correctly identifying the types of ion channel-targeted conotoxins will provide important clue to understand their function and find potential drugs. Based on this consideration, we developed a new computational method to rapidly and accurately predict the types of ion-targeted conotoxins. Three kinds of new properties of residues were proposed to use in pseudo amino acid composition to formulate conotoxins samples. The support vector machine was utilized as classifier. A feature selection technique based on F-score was used to optimize features. Jackknife cross-validated results showed that the overall accuracy of 94.6% was achieved, which is higher than other published results, demonstrating that the proposed method is superior to published methods. Hence the current method may play a complementary role to other existing methods for recognizing the types of ion-target conotoxins.

1. Introduction

The marine cone snail can secrete venom for predation and defense. A key component of venom is called conotoxin which is a kind of disulfide-rich neurotoxic peptide with 10–30 residues long. The high diversity of their sequences makes it difficult to systemically study them. It has been reported that there are over 100,000 conotoxins existing in approximately 700 species of cone snails [1]. Conotoxins can target G protein-coupled receptors (GPCRs) [2], nicotinic acetylcholine, and neurotensin receptors. Particularly, they interact with ion channels with extremely high specificity and affinity [3]. Thus, they have been regarded as important drug candidates to treat chronic pain, epilepsy, spasticity, and cardiovascular diseases [4, 5].

With more and more conotoxins being discovered, biochemical experiments-based method to investigate the function of conotoxins becomes more and more difficult because of high cost and long period of wet experiment. Using computational method to predict the function of conotoxins provides us with a convenient way to perform systemic analysis of conotoxins. In 2006, Mondal et al. combined support vector machine (SVM) with pseudo amino acid composition (PseAAC) to predict the superfamily of conotoxins [6]. Subsequently, Lin and Li developed a novel method called increment of diversity (ID) to describe dipeptide sequence and used quadratic discriminant (QD) to predict superfamily and family of conotoxins [7]. Zaki et al. used sequence alignment which was also used by Zou et al. [8] combined with amino acid composition to predict superfamily of conotoxins by use of SVM [9]. They further provide a SVM-Freescore method to improve accuracy [10]. Recently, Yin et al. developed a method called dHKNN to predict superfamily of conotoxins and achieved the overall accuracy of 90.3% by using hidden Markov model to select best features [11, 12]. Lisacek et al. used profile Hidden Markov Models (pHMMs) and position-specific scoring matrix (PSSM) to improve accuracy for conotoxin superfamily prediction [13–15].

Although the methods and results mentioned above can give some guide to study conotoxins, they did not provide

more information for the prediction of conotoxins' function. A case shows that two conotoxins (delta-conotoxin-like Ac6.1 and omega-conotoxin-like Ai6.2) belong to the same superfamily; however, they can target different ion channels [16]. Thus, it is necessary to develop new bioinformatics tools to identify the function of conotoxins. In 2007, Saha and Raghava proposed a method based on SVM and PSI-BLAST to predict the function of neurotoxins [17]. Soli et al. developed a statistical-based model to predict the activity of scorpion toxins by using motifs and secondary structure information [18]. Recently, Yuan et al. developed a feature selection technique based on binomial distribution to predict the types of ion channel-targeted conotoxins by using radial basis function network [19]. Subsequently, they improved the accuracy by using SVM with optimal dipeptide composition [20]. However, the prediction accuracy can be further improved.

Thus, the present study aimed to develop a new prediction method to improve the prediction quality of conotoxins' types. We incorporated three kinds of new properties of residues into PseAAC for formulating conotoxins samples. Subsequently, we used SVM to perform classification. After feature selection, we found that the accuracy was dramatically improved in jackknife cross-validation. In the following section, we will introduce the process of model construction in detail.

2. Materials and Methods

2.1. Benchmark Dataset. The benchmark dataset extracted from the UniProt [21] was constructed by Lin's group [19, 20]. The dataset is reliable and objective because (i) the conotoxins with ambiguous annotations have been excluded, (ii) the function of all conotoxins in benchmark dataset has been experimentally confirmed, and (iii) high similar sequences (cutoff = 80%) have been pruned by using CD-HIT program. The benchmark dataset contains 112 mature conotoxins peptide sequences including 24 potassium ion channel-targeted conotoxins (K-conotoxins), 43 sodium ion channel-targeted conotoxins (Na-conotoxins), and 45 calcium ion channel-targeted conotoxins (Ca-conotoxins). All calculations and model construction in the following section are based on the data.

2.2. Feature Extraction. A key point in protein prediction is how to extract important information from peptide sequences. In the past studies, the amino acid composition has been widely used in protein prediction. To consider the correlation of residues, the dipeptide composition was used in prediction model. Chou proposed a very popular and elegant descriptor called PseAAC which describes not only the correlation of physicochemical properties of residues but also the amino acid composition [22]. Furthermore, recently some web servers or stand-alone tools have been proposed to generate different modes of PseAAC, such as PseKNC [23], PseKNC-General [24], Pse-in-One [25], repRNA [26], and repDNA [27]. The authors should introduce these tools. In this study, we proposed three kinds of new properties, that is, rigidity, flexibility, and irreplaceability. The flexibility and rigidity of residues correlate with the protein structure and function. The irreplaceability of residues can reflect the evolution of life. The values of three properties for 20 residues [28] have been listed in Table 1. In the following, we will describe how to formulate conotoxins with PseAAC [22].

Table 1: The values of rigidity, flexibility, and irreplaceability of 20 residues.

Residues	Rigidity	Flexibility	Irreplaceability
G	−1.097	−2.746	0.56
A	−1.338	−3.102	0.52
V	−1.641	−1.339	0.54
L	−1.741	0.424	0.58
I	−1.741	0.424	0.65
F	2.877	−0.466	0.86
W	5.913	−1.000	1.82
Y	2.714	−0.672	0.98
D	−0.204	0.424	0.77
H	2.269	−0.223	0.94
N	−0.204	0.424	0.79
E	−0.365	2.009	0.76
K	−1.822	3.950	0.81
Q	−0.365	2.009	0.86
M	−1.741	2.484	1.25
R	1.169	3.06	0.6
S	−1.511	0.957	0.64
T	−1.641	−1.339	0.56
C	−1.511	0.957	1.12
P	1.979	−2.404	0.61

Consider a conotoxin $\mathbf{P} = R_1R_2R_3R_4 \cdots R_L$, where R_1, R_2, and R_L denote the 1st, 2nd, and Lth residue of the conotoxin sample $\mathbf{P}$; it can be defined by a $400 + 3\lambda$-dimensional vector as shown by

$$\mathbf{P} = \left[x_1 \cdots x_{400} \cdots x_{400+3\lambda}\right]^T, \tag{1}$$

where

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{400} f_u + w \sum_{j=1}^{3\lambda} \tau_j} & (1 \le u \le 400) \\ \dfrac{\omega\tau_u}{\sum_{i=1}^{400} f_u + \omega \sum_{j=1}^{n\lambda} \tau_j} & (400 + 1 \le u \le 400 + 3\lambda), \end{cases} \tag{2}$$

where f_u is the normalized frequency of the 400 dipeptides in conotoxin $\mathbf{P}$ and can be defined as

$$f_u = \frac{n_u}{\sum_u n_u}, \tag{3}$$

where n_u denotes the number of occurrences of uth dipeptide in conotoxin $\mathbf{P}$.

In (2), ω is weight factor for sequence order effect. τ_j is called the j-tier sequence correlation factor computed by the following formula:

$$\begin{aligned}
\tau_1 &= \frac{1}{L-1}\sum_{i=1}^{L-1} H_{i,i+1}^1,\\
\tau_2 &= \frac{1}{L-1}\sum_{i=1}^{L-1} H_{i,i+1}^2,\\
\tau_3 &= \frac{1}{L-1}\sum_{i=1}^{L-1} H_{i,i+1}^3,\\
&\vdots\\
\tau_{3\lambda} &= \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^3 \qquad (\lambda < L),
\end{aligned} \tag{4}$$

where $H_{i,i+\lambda}^n$ ($n = 1, 2, 3$ denotes rigidity, flexibility, and irreplaceability) is called the correlation function and can be given by

$$H_{i,i+\lambda}^n = h^n\left(R_i\right) \cdot h^n\left(R_{i+\lambda}\right), \tag{5}$$

where $h^n(R_i)$ is the nth kind of the physicochemical values of the amino acid R_i. The values should be converted to standard type by

$$h^n\left(R_i\right) \Longleftarrow \frac{h_0^n\left(R_i\right) - \left\langle h_0^n\left(R_i\right)\right\rangle}{\mathrm{SD}\left\langle h_0^n\left(R_i\right)\right\rangle}, \tag{6}$$

where $h_0^n(R_i)$ is the original physicochemical values of the ith amino acid.

For the purpose of finding the best feature subset which can produce the maximum accuracy, we performed feature selection by using the algorithm called F-score which can be defined as

$$F\left(i\right) = \frac{\sum_{k=1}^{3}\left(\overline{x}_i^k - \overline{x}_i\right)^2}{\sum_{k=1}^{3}\left(1/\left(N_k - 1\right)\right)\sum_{j=1}^{N_k}\left(x_{ij}^k - \overline{x}_i^k\right)^2}, \tag{7}$$

where $\overline{x}_i$ and $\overline{x}_i^k$ are the average values of the ith feature in whole dataset and the kth dataset; x_{ij}^k is the value of the ith feature of the jth conotoxin in the kth dataset; and N_k is the numbers of conotoxin in the kth dataset. We noticed that the larger the $F(i)$ value is, the better the predictive capability the ith feature has. We used a python script fselect.py downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/ to calculate F-score.

2.3. Support Vector Machine. SVM is a very popular machine learning method which is very suitable for small sample classification [29–31] and regressions [32, 33]. Its basic idea is to map the original samples into a high-dimensional space and search for the best hyperplane in this space which can separate different samples. In this study, the LibSVM soft package was used to implement SVM. The radial basis function (RBF) usually exhibits excellent performance in nonlinear classification [34]. Thus the RBF kernel function was used in the current work. We utilized grid search method to find out the best values of the regularization parameter C and kernel parameter γ via jackknife cross-validation. The search spaces for C and γ are $[2^{15}, 2^{-5}]$ and $[2^{-5}, 2^{-15}]$ with steps being 2^{-1} and 2, respectively.

2.4. The Evaluation of Model Performance. We used jackknife cross-validation to evaluate the performance of proposed method. Three metrics, namely, sensitivity (Sn), overall accuracy (OA), and average accuracy (AA) as defined in [19, 20], were used to quantitatively estimate the accuracy of the model:

$$\begin{aligned}
\mathrm{Sn}_k &= \frac{m_k}{N_k},\\
\mathrm{OA} &= \frac{\sum_{k=1}^{3} m_k}{\sum_{k=1}^{3} N_k},\\
\mathrm{AA} &= \frac{\mathrm{Sn}_k}{3},
\end{aligned} \tag{8}$$

where N_k is the total number of the kth types of conotoxins and m_k denotes the number of the kth types of conotoxins which was correctly recognized.

3. Results and Discussion

As we can see from (2), the results of the proposed method depend on two parameters λ and ω, where λ represents the long-range sequence order effect and ω is called weight factor which reflects the weight imposed between the local and global effects. Generally speaking, the greater λ is, the more global sequence order information it contains. However, if λ is too large, it would cause the high-dimensional disaster as mentioned above. Therefore, our searching for the optimal values of the three parameters was carried out in the following regions:

$$\begin{aligned}
1 \le \lambda \le 10 \quad &\text{with step } \Delta = 1\\
0.1 \le w \le 1.0 \quad &\text{with step } \Delta = 0.1.
\end{aligned} \tag{9}$$

From (9), a total of $10 \times 10 = 100$ individual combinations needed to be considered for finding the optimal parameter combination. This was actually a routine but tedious process to optimize the model via a 2-dimensional grid search. We used the jackknife cross-validation approach to deal with the parameter optimization. The results show that when $\lambda = 6$ and $\omega = 0.2$, the accuracy reaches to maximum value. We noticed that the current model contains 418 features which is still so large that the high-dimensional and overfitting problems will appear.

Therefore, we must select the key features from the 418 components. These key features can produce the maximum Acc. The best feature subset will be obtained by

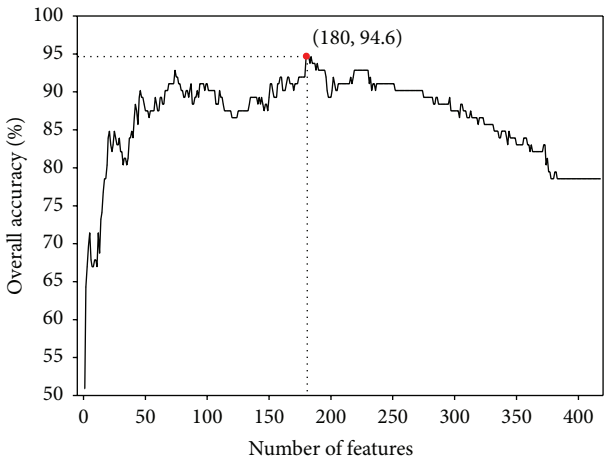

Figure 1: A plot to show the feature selection results. When the top 180 features were used to perform prediction, the overall success rate reached its peak of 94.6%.

Table 2: Comparison of the current method with published methods.

Methods	Sn (%)			AA (%)	OA (%)
	K	Na	Ca		
RBF network [19]	91.7	88.3	88.9	89.7	89.3
iCTX-Type [20]	83.3	97.8	89.8	90.31	91.1
Our method	91.7	95.3	95.6	94.2	94.6

investigating all the combinations of features. However, it is time-consuming and even beyond computational capability for most computers to examine all possible combinations. Based on this reason, we used F-score defined in (7) to perform feature selection. At first, all 418 features were ranked according to their F-scores from large to small. Secondly, the SVM was used to classify three samples and calculate the accuracy based on the feature with maximum F-score. Thirdly, a new feature subset was produced by adding the feature with the second highest F value to the former feature subset. We repeated the process until all combinations were investigated and the accuracies were calculated.

We plotted the accuracies with feature dimension in Figure 1 and noticed that the maximum accuracy is 94.6% when 180 best features were used. The detailed results were recorded in Table 1. Other published results were also listed in Table 2. We noticed that Sns of Na- and Ca-conotoxins of our method are 95.3% and 95.6%, respectively, which are higher than those of RBF network-based method [19]. The Sns of K- and Ca-conotoxins of our method are 91.7% and 95.6%, respectively, which are higher than those of iCTX-Type [20]. Thus, in summary, our proposed method is superior to other published methods.

4. Conclusion

In this paper, we designed a new method based on three kinds of new properties to predict three kinds of ion channel-targeted conotoxins. By using feature selection technique, prediction accuracy was dramatically improved. Comparison with published methods demonstrated the advantage of our method. The properties of residues used in this paper can also be used in other fields of protein classification. In the future, we will construct a free webserver based on the proposed method for the convenience of the vast majority of experimental scientists.

Competing Interests

The authors declare that they have no competing financial interests.

Acknowledgments

This work was supported by the Applied Basic Research Program of Sichuan Province (LZ-LY-45) and the Scientific Research Foundation of the Education Department of Sichuan Province (11ZB122).

References

[1] N. L. Daly and D. J. Craik, "Structural studies of conotoxins," *IUBMB Life*, vol. 61, no. 2, pp. 144–150, 2009.

[2] Z. Liao, Y. Ju, and Q. Zou, "Prediction of G protein-coupled receptors with SVM-prot features and random forest," *Scientifica*, vol. 2016, Article ID 8309253, 10 pages, 2016.

[3] H. Terlau and B. M. Olivera, "Conus venoms: a rich source of novel ion channel-targeted peptides," *Physiological Reviews*, vol. 84, no. 1, pp. 41–68, 2004.

[4] T. S. Han, R. W. Teichert, B. M. Olivera, and G. Bulaj, "Conus venoms—a rich source of peptide-based therapeutics," *Current Pharmaceutical Design*, vol. 14, no. 24, pp. 2462–2479, 2008.

[5] M. R. Watters, "Tropical marine neurotoxins: venoms to drugs," *Seminars in Neurology*, vol. 25, no. 3, pp. 278–289, 2005.

[6] S. Mondal, R. Bhavna, R. M. Babu, and S. Ramakumar, "Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification," *Journal of Theoretical Biology*, vol. 243, no. 2, pp. 252–260, 2006.

[7] H. Lin and Q.-Z. Li, "Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant," *Biochemical and Biophysical Research Communications*, vol. 354, no. 2, pp. 548–551, 2007.

[8] Q. Zou, Q. Hu, M. Guo, and G. Wang, "HAlign: fast multiple similar DNA/RNA sequence alignment based on the centre star strategy," *Bioinformatics*, vol. 31, no. 15, pp. 2475–2481, 2015.

[9] N. Zaki, F. Sibai, and P. Campbell, "Conotoxin protein classification using pairwise comparison and amino acid composition," in *Proceedings of the 13th Annual Genetic and Evolutionary Computation Conference (GECCO '11)*, pp. 323–330, ACM, Dublin, Ireland, July 2011.

[10] N. Zaki, S. Wolfsheimer, G. Nuel, and S. Khuri, "Conotoxin protein classification using free scores of words and support vector machines," *BMC Bioinformatics*, vol. 12, article 217, 2011.

[11] Y.-X. Fan, J. Song, X. Kong, and H.-B. Shen, "PredCSf: an integrated feature-based approach for predicting conotoxin superfamily," *Protein and Peptide Letters*, vol. 18, no. 3, pp. 261–267, 2011.

[12] J.-B. Yin, Y.-X. Fan, and H.-B. Shen, "Conotoxin superfamily prediction using diffusion maps dimensionality reduction and subspace classifier," *Current Protein and Peptide Science*, vol. 12, no. 6, pp. 580–588, 2011.

[13] D. Koua, A. Brauer, S. Laht et al., "ConoDictor: a tool for prediction of conopeptide superfamilies," *Nucleic Acids Research*, vol. 40, no. 1, pp. W238–W241, 2012.

[14] D. Koua, S. Laht, L. Kaplinski et al., "Position-specific scoring matrix and hidden Markov model complement each other for the prediction of conopeptide superfamilies," *Biochimica et Biophysica Acta (BBA)—Proteins and Proteomics*, vol. 1834, no. 4, pp. 717–724, 2013.

[15] S. Laht, D. Koua, L. Kaplinski, F. Lisacek, R. Stöcklin, and M. Remm, "Identification and classification of conopeptides using profile Hidden Markov Models," *Biochimica et Biophysica Acta (BBA)—Proteins and Proteomics*, vol. 1824, no. 3, pp. 488–492, 2012.

[16] K. H. Gowd, K. K. Dewan, P. Iengar, K. S. Krishnan, and P. Balaram, "Probing peptide libraries from *Conus achatinus* using mass spectrometry and cDNA sequencing: identification of δ and ω-conotoxins," *Journal of Mass Spectrometry*, vol. 43, no. 6, pp. 791–805, 2008.

[17] S. Saha and G. P. S. Raghava, "Prediction of neurotoxins based on their function and source," *In Silico Biology*, vol. 7, no. 4-5, pp. 369–387, 2007.

[18] R. Soli, B. Kaabi, M. Barhoumi, M. El-Ayeb, and N. Srairi-Abid, "Bioinformatic characterizations and prediction of K^+ and Na^+ ion channels effector toxins," *BMC Pharmacology*, vol. 9, article 4, 2009.

[19] L.-F. Yuan, C. Ding, S.-H. Guo, H. Ding, W. Chen, and H. Lin, "Prediction of the types of ion channel-targeted conotoxins based on radial basis function network," *Toxicology in Vitro*, vol. 27, no. 2, pp. 852–856, 2013.

[20] H. Ding, E.-Z. Deng, L.-F. Yuan et al., "ICTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels," *BioMed Research International*, vol. 2014, Article ID 286419, 10 pages, 2014.

[21] M. Magrane and UniProt Consortium, "UniProt Knowledgebase: a hub of integrated protein data," *Database*, vol. 2011, Article ID bar009, 2011.

[22] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Structure, Function and Genetics*, vol. 43, no. 3, pp. 246–255, 2001.

[23] W. Chen, T.-Y. Lei, D.-C. Jin, H. Lin, and K.-C. Chou, "PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition," *Analytical Biochemistry*, vol. 456, no. 1, pp. 53–60, 2014.

[24] W. Chen, X. Zhang, J. Brooker, H. Lin, L. Zhang, and K.-C. Chou, "PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions," *Bioinformatics*, vol. 31, no. 1, pp. 119–120, 2015.

[25] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K. Chou, "Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Research*, vol. 43, no. W1, pp. W65–W71, 2015.

[26] B. Liu, F. Liu, L. Fang, X. Wang, and K.-C. Chou, "repRNA: a web server for generating various feature vectors of RNA sequences," *Molecular Genetics and Genomics*, vol. 291, no. 1, pp. 473–481, 2016.

[27] B. Liu, F. Liu, L. Fang, X. Wang, and K.-C. Chou, "RepDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects," *Bioinformatics*, vol. 31, no. 8, pp. 1307–1309, 2015.

[28] 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.

[29] P.-P. Zhu, W.-C. Li, Z.-J. Zhong et al., "Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition," *Molecular BioSystems*, vol. 11, no. 2, pp. 558–563, 2015.

[30] R. Wang, Y. Xu, and B. Liu, "Recombination spot identification Based on gapped k-mers," *Scientific Reports*, vol. 6, Article ID 23934, 2016.

[31] D. Li, Y. Ju, and Q. Zou, "Protein Folds Prediction with Hierarchical Structured SVM," *Current Proteomics*, vol. 13, no. 2, pp. 79–85, 2016.

[32] R. Cao, Z. Wang, and J. Cheng, "Designing and evaluating the MULTICOM protein local and global model quality prediction methods in the CASP10 experiment," *BMC Structural Biology*, vol. 14, no. 1, article 13, 2014.

[33] R. Cao, Z. Wang, Y. Wang, and J. Cheng, "SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines," *BMC Bioinformatics*, vol. 15, no. 1, article 120, 2014.

[34] J. Chen, X. Wang, and B. Liu, "IMiRNA-SSF: improving the identification of MicroRNA precursors by combining negative sets with different distributions," *Scientific Reports*, vol. 6, article 19062, 2016.