Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 286419, 10 pages
http://dx.doi.org/10.1155/2014/286419

Research Article

iCTX-Type: A Sequence-Based Predictor for Identifying the Types of Conotoxins in Targeting Ion Channels

Hui Ding,[1] En-Ze Deng,[1] Lu-Feng Yuan,[1] Li Liu,[2] Hao Lin,[1,3] Wei Chen,[3,4] and Kuo-Chen Chou[3,5]

[1] *Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China*
[2] *Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China*
[3] *Gordon Life Science Institute, Boston, MA 02478, USA*
[4] *Department of Physics, School of Sciences Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China*
[5] *Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia*

Correspondence should be addressed to Hao Lin; hlin@gordonlifescience.org and Wei Chen; greatchen@heuu.edu.cn

Received 13 March 2014; Revised 22 April 2014; Accepted 7 May 2014; Published 1 June 2014

Academic Editor: Shiwei Duan

Conotoxins are small disulfide-rich neurotoxic peptides, which can bind to ion channels with very high specificity and modulate their activities. Over the last few decades, conotoxins have been the drug candidates for treating chronic pain, epilepsy, spasticity, and cardiovascular diseases. According to their functions and targets, conotoxins are generally categorized into three types: potassium-channel type, sodium-channel type, and calcium-channel types. With the avalanche of peptide sequences generated in the postgenomic age, it is urgent and challenging to develop an automated method for rapidly and accurately identifying the types of conotoxins based on their sequence information alone. To address this challenge, a new predictor, called iCTX-Type, was developed by incorporating the dipeptide occurrence frequencies of a conotoxin sequence into a 400-D (dimensional) general pseudoamino acid composition, followed by the feature optimization procedure to reduce the sample representation from 400-D to 50-D vector. The overall success rate achieved by iCTX-Type via a rigorous cross-validation was over 91%, outperforming its counterpart (RBF network). Besides, iCTX-Type is so far the only predictor in this area with its web-server available, and hence is particularly useful for most experimental scientists to get their desired results without the need to follow the complicated mathematics involved.

1. Introduction

Being peptides consisting of about 10 to 30 amino acid residues, conotoxins are toxins secreted by cone snails for capturing prey and securing themselves. This kind of toxins can bind to various targets, such as G protein-coupled receptors (GPCRs), nicotinic acetylcholine, and neurotensin receptors. In particular, they display extremely high specificity and affinity for ion channels. Ion channels represent a class of membrane spanning protein pores that mediate the flux of ions in a variety of cell types. There are over 300 types of ion channels in a living cell [1]. Many crucial functions in life, such as heartbeat, sensory transduction, and central nervous system response, are controlled by cell signaling via various ion channels. Ion channel dysfunction may lead to a number of diseases, such as epilepsy, arrhythmia, and type II diabetes. These kinds of diseases are primarily treated with the drugs that modulate the ion channels concerned. Ion channels are also the important targets for treating virus diseases (see, e.g., [2–4]). Owing to their importance to human being's life, ion channels have become the 2nd most frequent targets for drug development, just next to GPCRs (G protein-coupled receptors) [5]. The following three kinds of ion channels are usually the targets by conotoxins: potassium (K) channel (Figure 1), sodium (Na) channel (Figure 2), and calcium (Ca) channel (Figure 3). Based on their functions

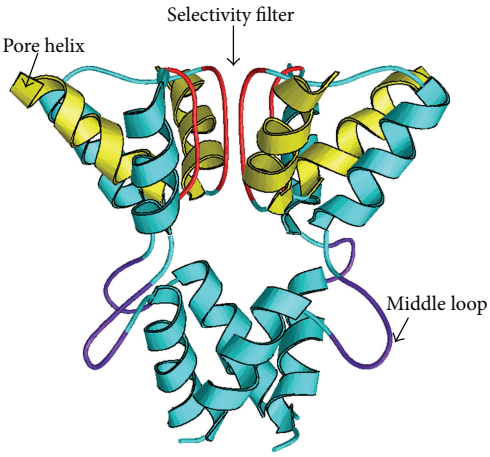

Figure 1: A ribbon drawing to show the human potassium (K) channel. Reproduced from Chou [6] with permission.

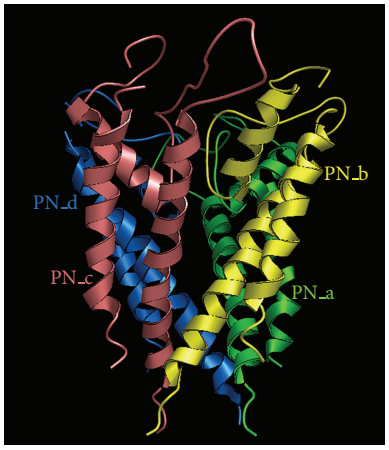

Figure 2: A ribbon drawing to show the human sodium (Na) channel. Reproduced from Chou [6] with permission.

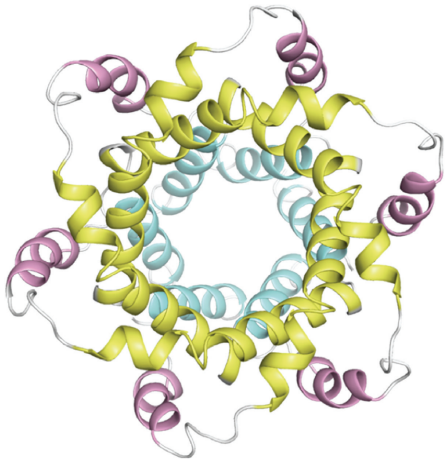

Figure 3: A ribbon drawing to show the calcium (Ca) channel from hepatitis C virus. Reproduced from [4] with permission.

and targeting objects, conotoxins can be classified into the following three types: (i) K-channel-targeting type; (ii) Na-channel-targeting type; and (iii) Ca-channel-targeting type.

Although conotoxins are lethally venomous because of blocking the transmission of nerve impulses, they have been widely used to treat chronic pain, epilepsy, spasticity, and cardiovascular diseases. Therefore, conotoxins have been regarded as important pharmacological tools for neuroscience research.

It has been estimated that there are more than 100,000 kinds of conotoxins secreted by over 700 kinds of *Conus* in the world [8]. However, relatively much fewer conotoxins (about 3,000 peptides) have been experimentally confirmed and reported in literature and databases. Moreover, the records about the functions of conotoxins in public databases are no more than 300 items. Hence, developing a computational method to predict the functions of conotoxins has become a challenging task.

In a pioneer work, Mondal et al. [9] proposed a method for predicting conotoxin superfamilies by using the pseudoamino acid composition approach [10, 11]. Subsequently, a series of studies have been reported in predicting conotoxin superfamilies (see, for example, [12–15]). All these methods yielded quite encouraging results, and each of them did play a role in stimulating the development of this area. However, none of these methods can be used to predict the types of conotoxins defined according to their targeting ion-channels. For instance, both delta-conotoxin-like Ac6.1 (UniProt accession number: P0C8V5) [16] and omega-conotoxin-like Ai6.2 [17] (UniProt accession number: P0CB10) belong to the conotoxin O1 superfamily. However, the former targets the voltage-gated sodium channels, while the latter targets the voltage-gated calcium channels.

To deal with this problem, recently, a method was developed [7] to identify conotoxins among the aforementioned three types by using their sequence information alone. However, further work is needed in this regard due to the following reasons. (i) The prediction quality can be further improved. (ii) No web server for the prediction method in [7] was provided, and hence its usage is quite limited, especially for the majority of experimental scientists.

The present study was devoted to develop a new predictor for identifying the conotoxins' types from the above two aspects.

As elaborated in a comprehensive review [18] and conducted by a series of recent publications [19–28], to establish a really useful statistical predictor for a biological system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the biological samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web server for

the predictor that is accessible to the public. In what follows, let us describe how to deal with these procedures one by one.

2. Materials and Methods

2.1. Benchmark Dataset. The sequences of conotoxins and their functions were collected from the UniProt [29]. To ensure its quality, the benchmark dataset was constructed strictly according to the following criteria. (i) Included were only those peptides annotated with "conotoxin" and with the keyword of potassium, calcium, or sodium in their functional ontologies. (ii) Included were only those conotoxins with clear functional annotations based on experiment results. In other words, we excluded those annotated with "uncertain," "predicted," or "inferred from homology" because of lacking confidence. (iii) Excluded were those that were annotated with "immature" due to the incompleteness. (iv) Excluded were also those that contained any invalid amino acid codes, such as "B," "X," and "Z". After going through the above procedures, we obtained 195 conotoxins, of which 37 belonged to the K-channel-targeting type, 86 to the Na-channel-targeting type, and 72 to the Ca-channel-targeting type.

As elaborated in a comprehensive review [18], a benchmark dataset containing many redundant samples with high similarity would lack statistical representativeness. A predictor, if trained and tested by a benchmark dataset with many homologous sequences, might yield misleading results with overestimated accuracy [30]. To remove the homologous sequences from the benchmark dataset, a cutoff threshold of 25% was recommended [31] to exclude those protein/peptide sequences from the benchmark datasets that had ≥25% pairwise sequence identity to any other sample in the same subset. However, in this study we did not use such a stringent criterion because the currently available data did not allow us to do so. Otherwise, the numbers of peptides for some subsets would be very few to have statistical significance. As a compromise, we set the cutoff threshold at 80% and used the CD-HIT software [32] to remove those conotoxin samples that had ≥80% sequence identity to any other in a same subset. After such a screening procedure, we obtained 112 conotoxin samples for the benchmark dataset $\mathbb{S}$, as formulated as follows:

$$\mathbb{S} = \mathbb{S}_{\mathrm{K}} \cup \mathbb{S}_{\mathrm{Na}} \cup \mathbb{S}_{\mathrm{Ca}}, \tag{1}$$

where the subset $\mathbb{S}_{\mathrm{K}}$ contains 24 conotoxin samples of K-channel-targeting type, $\mathbb{S}_{\mathrm{Na}}$ contains 43 samples of Na-channel-targeting type, and $\mathbb{S}_{\mathrm{Ca}}$ contains 45 samples of Ca-channel-targeting type, while the symbol $\cup$ represents the union in the set theory. The codes of 112 conotoxins and their sequences are given in Supporting Information S1 (see Supplementary Material available online at http://dx.doi.org/10.1155/2014/286419).

Likewise, we also constructed an independent dataset $\mathbb{S}^{\mathrm{Ind}}$ as formulated by

$$\mathbb{S}^{\mathrm{Ind}} = \mathbb{S}^{\mathrm{Ind}}_{\mathrm{K}} \cup \mathbb{S}^{\mathrm{Ind}}_{\mathrm{Na}} \cup \mathbb{S}^{\mathrm{Ind}}_{\mathrm{Ca}}, \tag{2}$$

where $\mathbb{S}^{\mathrm{Ind}}_{\mathrm{K}}$ contains 12 K-conotoxins, $\mathbb{S}^{\mathrm{Ind}}_{\mathrm{Na}}$ contains 37 Na-conotoxins, and $\mathbb{S}^{\mathrm{Ind}}_{\mathrm{Ca}}$ contains 21 Ca-conotoxins. None of the samples in the independent dataset occurs in the dataset $\mathbb{S}$ of (1), and their detailed sequences are given in Supporting Information S2.

For simplicity, hereafter, let us use "K-conotoxin," "Na-conotoxin," and "Ca-conotoxin" to represent K-channel-targeting type conotoxin, Na-channel-targeting type conotoxin, and Ca-channel-targeting type conotoxin, respectively.

2.2. The Dipeptide Mode of Pseudoamino Acid Composition. Given a conotoxin peptide **P** with L amino acids, how do we translate it into a mathematical expression for statistical prediction? This is one of the first important problems to develop a sequence-based predictor for identifying the type of a conotoxin. The most straightforward way to formulate the sample of a conotoxin peptide **P** with L residues is to use its entire amino acid sequence, as can be formulated by

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4 \cdots \mathrm{R}_{\mathrm{L}}, \tag{3}$$

where R_1 represents the 1st residue of the conotoxin peptide and R_2 the 2nd residue of the peptide and so forth. Subsequently, we can utilize various sequence similarity search based tools, such as BLAST [33], to perform statistical prediction. Although this kind of sequence model was very straightforward and intuitive, unfortunately, it failed to work when a query conotoxin peptide did not have significant similarity to any of the peptide sequences in the training dataset. Thus, investigators turned to use vectors to represent the peptide samples. Another reason for them to do so is that the statistical samples in vector format are much easier to be handled than in sequence format by many existing operation engines, such as the correlation angle approach [34], covariance discriminant (CD) [27, 35–37], neural network [38–40], optimization approach [41], support vector machine (SVM) [22, 23, 42, 43], random forest [44, 45], conditional random field [20], nearest neighbor (NN) [46, 47]; K-nearest neighbor (KNN) [30], OET-KNN [48–50], fuzzy K-nearest neighbor [25, 51–55], ML-KNN algorithm [56], and SLLE algorithm [36].

The simplest vector used to represent a peptide or protein sample is its amino acid composition (AAC), as given as follows:

$$\mathbf{P} = [f_1 \quad f_2 \quad \cdots \quad f_{20}]^{\mathbf{T}}, \tag{4}$$

where f_i ($i = 1, 2, \ldots, 20$) is the normalized occurrence frequency of the ith type of native amino acid in the peptide chain and **T** is the transpose operator. The AAC model was used by many in predicting various contributes of proteins (see, e.g., [41, 57–59]). However, as we can see from (4), when using AAC to represent a peptide or protein sample, all its sequence order information would be completely lost and hence limit the prediction quality.

How can we formulate a peptide or protein sequence with a vector yet still keep considerable sequence order information? As reported in many recent publications, in order to incorporate the sequence order information, the pseudoamino acid composition [10, 11] or Chou's PseAAC [60] was proposed. Since the concept of PseAAC was proposed in 2001 [10], it has been penetrating into almost all

the fields of protein attribute predictions (see, e.g., [61–78]). Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA and nucleotides [19, 21, 23, 27, 79], as well as other biological samples (see, e.g., [80–82]). Because it has been widely and increasingly used, in addition to the web server "PseAAC" [83] built in 2008, recently three types of powerful open access software, called "PseAAC-Builder" [84], "propy" [85], and "PseAAC-General" [86], were established: the former two are for generating various modes of Chou's special PseAAC, while the 3rd one is for those of Chou's general PseAAC.

According to a comprehensive review [18], the general PseAAC is formulated by

$$\mathbf{P} = \left[\psi_1 \;\; \psi_2 \;\; \cdots \;\; \psi_u \;\; \cdots \;\; \psi_\Omega\right]^{\mathbf{T}}, \tag{5}$$

where the component ψ_u ($u = 1, 2, \ldots, \Omega$) and the dimension Ω will depend on how to extract the features from the peptide sequences concerned. For the current study, since the conotoxin sequences are not long (about 10–30 residues), we could just consider the sequence order information between two most contiguous amino acid residues. Thus, the dimension of the vector $\mathbf{P}$ in (5) is $\Omega = 20 \times 20 = 400$ and each of the components therein is given by

$$\psi_u = \begin{cases} f(\mathrm{AA}) & \text{when } u = 1 \\ f(\mathrm{AC}) & \text{when } u = 2 \\ \quad\vdots & \quad\vdots \\ f(\mathrm{AY}) & \text{when } u = 20 \\ f(\mathrm{CA}) & \text{when } u = 21 \\ \quad\vdots & \quad\vdots \\ f(\mathrm{YW}) & \text{when } u = 399 \\ f(\mathrm{YY}) & \text{when } u = 400, \end{cases} \tag{6}$$

where A, C, ..., W, Y are, respectively, the single letter codes of 20 native amino acids, $f(\mathrm{AA})$ is the occurrence frequency for the dipeptide AA in the conotoxin sequence (see (3)), and $f(\mathrm{AC})$ is for the dipeptide AC and so forth. The formulation defined by (5)-(6) is actually the dipeptide mode of PseAAC, which can be automatically generated by the PseAAC server [83] for a given peptide or protein sequence.

2.3. Feature Selection. The original raw features usually contain the redundant information and noise that may negatively affect the prediction quality [87]. Using the feature selection techniques to optimize the feature set can not only enhance the prediction accuracy but also provide useful insights for in-depth understanding of the action mechanism of conotoxins. According to the feature selection algorithm [87], the F-score function is defined by

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

where $\overline{f}_i^{\,k}$ is the average frequency of the ith feature in the kth dataset, $\overline{f}_i$ the average frequency of the ith feature in the all datasets concerned, f_{ij}^k is the frequencies of the ith feature of the jth sequence in the kth dataset, and N_k is the number of peptide samples in the kth dataset. The program called "fselect.py" was downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvmtools to calculate F-score defined in (7).

The larger the F-score is, the more likely it has a better discriminative capability [87]. Accordingly, we ranked the 400 dipeptides in (5) according to their F-scores. Subsequently, based on the ranked dipeptides, we performed the incremental feature selection (IFS) strategy to find an optimal subset of features that yielded the highest predictive accuracy. During the IFS procedure, the feature subset started with one feature with the highest F-score. A new feature subset was composed when one more feature with the second highest F-score was added. By adding these features sequentially from the higher to lower ranks, 400 feature sets would be obtained. The τth feature set can be formulated as

$$S_\tau = \{f_1, f_2, \ldots, f_\tau\}, \quad (1 \le \tau \le 400). \tag{8}$$

For each of the 400 feature sets, a prediction model based on the proposed predictive algorithm was constructed and examined with the jackknife cross-validation on the benchmark dataset. By doing so, we obtained an IFS curve in a 2D (dimensional) Cartesian coordinate system with index τ as the abscissa (or X-coordinate) and the overall accuracy as the ordinate (or Y-coordinate). The optimal feature set is expressed as

$$S_\Theta = \{f_1, f_2, \ldots, f_\Theta\}. \tag{9}$$

with which the IFS curve reached its peak. In other words, in the 2D coordinate system, when $X = \Theta$, the value of the overall accuracy was the maximum. Thus, we used the Θ features to build the final predictor.

2.4. Support Vector Machine (SVM). The classification algorithm used in this work was the support vector machine (SVM). The SVM has been widely used in the realm of bioinformatics (see, e.g., [19, 22, 23, 88–90]). Its basic principle is to transform the input vector into a high-dimension Hilbert space and seek a separating hyperplane with the maximal margin in this space by using the decision function:

$$f\left(\overrightarrow{X}\right) = \mathrm{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \cdot K\left(\overrightarrow{X}, \overrightarrow{X_i}\right) + b\right), \tag{10}$$

where $\overrightarrow{X_i}$ is the ith training vector, the y_i represents the type of the ith training vector, and $K(\overrightarrow{X}, \overrightarrow{X_i})$ is a kernel function which defines an inner product in a high dimensional feature space. Because of its effectiveness and speed in nonlinear classification process, the radial basis kernel function (RBF) $K(\overrightarrow{X_i}, \overrightarrow{X_j}) = \exp(-\gamma \| \overrightarrow{X_i} - \overrightarrow{X_j} \|^2)$ was used in the current work. The original SVM was designed for two-class problems. For multiclass problems, several strategies such as one-versus-rest (OVR), one-versus-one (OVO), and

DAGSVM have been applied to extend the traditional SVM. In the present study, we used the OVO strategy for multiclass prediction. The concrete SVM software (LibSVM) was downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm. A grid search method was used to optimize the regularization parameter C and kernel parameter via the jackknife cross-validation. The search spaces for C and γ are $[2^{15}, 2^{-5}]$ and $[2^{-5}, 2^{-15}]$ with steps of 2^{-1} and 2, respectively. For more details about SVM, see a monograph [91].

3. Results and Discussion

3.1. Test Method and Criteria. In statistical prediction, the independent dataset test, subsampling or K-fold crossover test and jackknife test are the three cross-validation methods often used to check a predictor for its accuracy [92]. However, among the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset [18]. Accordingly, the jackknife test has been increasingly used and widely recognized by investigators to examine the quality of various predictors (see, e.g., [19, 21, 73, 75, 93–95]). Therefore, in this study we also adopted the jackknife test.

In addition to an objective test method, we also need a set of metrics to reasonably measure the test outcome. Here, let us use the criterion proposed in [96, 97] to develop a set of more intuitive and easier-to-understand metrics; that is, the correct rates Λ^{K} in predicting K-conotoxins, Λ^{Na} in predicting Na-conotoxins, and Λ^{Ca} in predicting Ca-conotoxins are defined by

$$\begin{aligned}
\Lambda^{K} &= \frac{N^{K} - N^{K}_{Na} - N^{K}_{Ca}}{N^{K}}, \quad \text{for the K-conotoxins} \\
\Lambda^{Na} &= \frac{N^{Na} - N^{Na}_{K} - N^{Na}_{Ca}}{N^{K}}, \quad \text{for the Na-conotoxins} \\
\Lambda^{Ca} &= \frac{N^{Ca} - N^{Ca}_{K} - N^{Ca}_{Na}}{N^{Ca}}, \quad \text{for the Ca-conotoxins,}
\end{aligned} \tag{11}$$

where N^{K} is the total number of the K-conotoxins investigated, while N^{K}_{Na} is the number of the K-conotoxins incorrectly predicted as the Na-conotoxins, and N^{K}_{Ca} is the number of the K-conotoxins incorrectly predicted as the Ca-conotoxins; N^{Na} is the total number of the Na-conotoxins investigated, while N^{Na}_{K} is the number of the Na-conotoxins incorrectly predicted as the K-conotoxins and N^{Na}_{Ca} is the number of the Na-conotoxins incorrectly predicted as the Ca-conotoxins; and N^{Ca} is the total number of the Ca-conotoxins investigated, while N^{Ca}_{Na} is the number of the Ca-conotoxins incorrectly predicted as the Na-conotoxins and N^{Ca}_{K} is the number of the Ca-conotoxins incorrectly predicted as the K-conotoxins. From (11), it follows that

$$\begin{aligned}
\text{OA} = \Lambda &= 1 - \frac{N^{K}_{Na} + N^{K}_{Ca} + N^{Na}_{K} + N^{Na}_{Ca} + N^{Ca}_{Na} + N^{Ca}_{K}}{N^{K} + N^{Na} + N^{Ca}} \\
\text{AA} &= \frac{\Lambda^{K} + \Lambda^{Na} + \Lambda^{Ca}}{3},
\end{aligned} \tag{12}$$

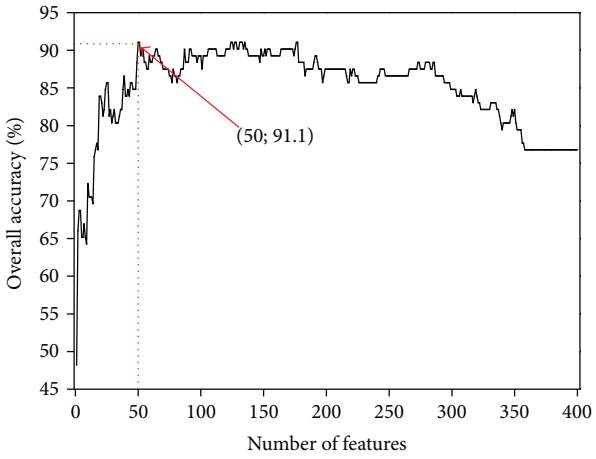

Figure 4: A plot to show the IFS curve, where the abscissa and ordinate axis denote the number of features and the overall accuracy, respectively. As shown in the figure, the value of the overall accuracy reached its peak (91.1%) when the top-ranked 50 dipeptide features were taken into account.

Table 1: List of the 50 optimal features or dipeptides derived according to (7)–(9) as elaborated in the Section 2.3.

AA	AS	CC	CH	CS	DH	DN	EN	GA	GH
GL	GT	GY	HA	HL	HS	IY	KD	KK	KM
KP	LN	LV	MC	MY	ND	NQ	NS	PI	QK
QT	RC	RD	RF	RN	RT	RW	SC	SG	TE
TF	TT	VV	WG	WI	YD	YH	YL	YT	YY

where OA stands for the overall accuracy and AA for the average accuracy.

3.2. The Optimal Features. As mentioned above, it would be no good for a sample vector to contain either too few or too many features. This is because the former would limit the prediction quality due to lack of information, while the latter would generate a lot of noise due to redundancy. Therefore, we should find a set of optimal features, for which there is minimal redundancy among themselves but maximal relevancy to the target to be predicted. In the present study, such an optimal feature-set is none but (9).

Shown in Figure 4 is the IFS curve for the value of OA against the number of the counted features, as described in Section 2.3. As can be seen from there, the value of OA reached its peak of 91.1% when the top-ranked 50 dipeptides (Table 1) were taken into account.

The predictor thus obtained via the aforementioned procedures is called "iCTX-Type," where "i" stands for "identify" and "CTX" for "conotoxin."

A comparison of the current predictor iCTX-Type with the one in [7] (i.e., to the best of our knowledge, it is the only existing predictor in this area) is given in Table 2, from which we can see the following. (i) For four of the five metrics defined in (10)-(11), iCTX-Type yielded higher scores than the method in [7]. Particularly, iCTX-Type achieved

Table 2: Comparison of the current method with the one in [7] by the jackknife test on the same benchmark dataset (Supporting Information S1) according to the metrics defined in (11)-(12).

Method	Number of features counted	Λ^{K} (%)	Λ^{Na} (%)	Λ^{Ca} (%)	AA (%)	OA (%)
RBF network[a]	70	91.7	88.4	88.9	89.7	89.3
iCTX-Type[b]	50	83.3	97.8	89.8	90.3	91.1

[a]See [7].
[b]This paper.

higher overall accuracy (OA) and average accuracy (AA). (ii) Compared with the method of [7] using 70 features, only 50 features were used in the present method (Table 1), indicating that the iCTX-Type is more efficient in excluding redundancy and noise as well as in capturing the core features.

To further verify the performance of the current predictor, iCTX-Type was also used to identify the samples in the independent dataset $\mathbb{S}^{Ind}$ (see Supporting Information S2), and the success rates (see (11)) thus obtained were 91.7%, 91.9%, and 90.5% for K-, Na-, and Ca-conotoxins, respectively. These results are fully consistent with those obtained by the jackknife test as given in Table 2, furtherindicating that the new predictor iCTX-Type is quite promising and holds a high potential to become a useful tool for in-depth studying ion channel-targeted conotoxins.

To enhance the value of its practical applications [98], a web server for the new iCTX-Type predictor was established as described below.

3.3. Web-Server Guide. For the convenience of the vast majority of experimental scientists, below a step-by-step guide is provided for how to use the web server to get the desired results without the need to follow the mathematic equations that were presented in this paper just for the integrity in developing the predictor.

Step 1. Open the web server at http://lin.uestc.edu.cn/server/iCTX-Type and you will see the top page of iCTX-Type on your computer screen, as shown in Figure 5. Click on the Read Me button to see a brief introduction about the predictor and the caveat when using it.

Step 2. Either type or copy/paste the query peptide sequences into the input box at the center of Figure 5. The input sequence should be in the FASTA format. A sequence in FASTA format consists of a single initial line beginning with a greater-than symbol ">" in the first column, followed by lines of sequence data. The words right after the ">" symbol in the single initial line are optional and only used for the purpose of identification and description. All lines should be no longer than 120 characters and usually do not exceed 80 characters. The sequence ends if another line starting with a ">" appears; this indicates the start of another sample sequence. Example sequences in FASTA format can be seen by clicking on the Example button right above the input box.

Step 3. Click on the Submit button to see the predicted result. For instance, when using the three peptide sequences as an input and clicking the Submit button, you will see the following shown on the screen of your computer: the outcome for the 1st query example is "Ca-conotoxin"; the outcome for the 2nd query sample is "K-conotoxin"; the outcome for the 3rd query sample is "Na-conotoxin." All these results are fully consistent with the experimental observations. It takes only a few seconds for the above computation before the predicted result appears on your computer screen; the more number of query sequences, the longer time it usually needs.

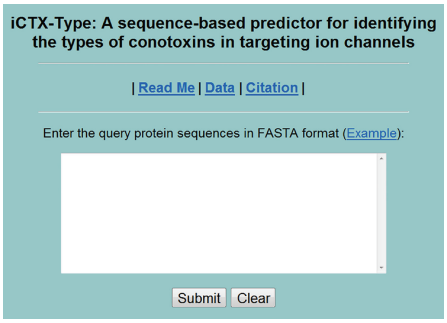

Figure 5: A screenshot to show the top page of the iCTX-Type web server. Its website address is http://lin.uestc.edu.cn/server/iCTX-Type.

Step 4. Click on the Data button to download the benchmark datasets used to train and test the iCTX-Type predictor.

Step 5. Click on the Citation button to find the relevant papers that document the detailed development and algorithm of iCTX-Type.

Caveats. The input query sequences must be formed by the single-letter codes of the 20 native amino acids; any other characters such as "B," "X," "U," and "Z" are invalid and should not be part of the peptide sequence.

4. Conclusion

It is anticipated that iCTX-Type may become a useful high throughput tool for both basic research and drug development, particularly for in-depth investigation into the mechanisms of ion-channels and developing new drugs to treat chronic pain, epilepsy, spasticity, and cardiovascular diseases, among others.

It is instructive to point out that since the binding of conotoxins to ion-channel is highly selective and specific, the information obtained by iCTX-Type in identifying the

types of conotoxins may be also very useful for designing ion channel inhibitors according to the Chou's distorted key theory as elaborated in [99] and briefed in a Wikipedia article at http://en.wikipedia.org/wiki/Chou's_distorted_key_theory_for_peptide_drugs.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors wish to thank the anonymous reviewers for their constructive comments, which were very helpful for strengthening the presentation of this study. This work was supported by the National Nature Scientific Foundation of China (nos. 61202256, 61301260, and 61100092), the Nature Scientific Foundation of Hebei Province (no. C2013209105), and the Fundamental Research Funds for the Central Universities (nos. ZYGX2012J113 and ZYGX2013J102).

References

[1] I. S. Gabashvili, B. H. Sokolowski, C. C. Morton, and A. B. Giersch, "Ion channel gene expression in the inner ear," *Journal of the Association for Research in Otolaryngology*, vol. 8, no. 3, pp. 305–328, 2007.

[2] J. R. Schnell and J. J. Chou, "Structure and mechanism of the M2 proton channel of influenza A virus," *Nature*, vol. 451, no. 7178, pp. 591–595, 2008.

[3] R. M. Pielak, J. R. Schnell, and J. J. Chou, "Mechanism of drug inhibition and drug resistance of influenza A M2 channel," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 18, pp. 7379–7384, 2009.

[4] B. OuYang, S. Xie, M. J. Berardi et al., "Unusual architecture of the p7 channel from hepatitis C virus," *Nature*, vol. 498, no. 7455, pp. 521–525, 2013.

[5] X. Xiao, J. L. Min, and P. Wang, "Predict drug-protein interaction in cellular networking," *Current Topics in Medicinal Chemistry*, vol. 13, no. 14, pp. 1707–1712, 2013.

[6] K.-C. Chou, "Insights from modeling three-dimensional structures of the human potassium and sodium channels," *Journal of Proteome Research*, vol. 3, no. 4, pp. 856–861, 2004.

[7] 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.

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

[9] S. Mondal, R. Bhavna, R. Mohan 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.

[10] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins*, vol. 43, no. 3, pp. 246–255, 2001, Erratum in: *Proteins*, vol. 44, no. 1, article 60, 2001.

[11] K. C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[12] 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.

[13] 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.

[14] 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*, vol. 1824, no. 3, pp. 488–492, 2012.

[15] 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*, vol. 1834, no. 4, pp. 717–724, 2013.

[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] D. R. Hillyard, M. J. Mcintosh, R. M. Jones et al., "O-superfamily conotoxin peptides," Patent number JP2003533178, 2008.

[18] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[19] S. H. Guo, E. Z. Deng, L. Q. Xu, H. Ding, H. Lin, and W. Chen, "iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition," *Bioinformatics*, 2014.

[20] Y. Xu, J. Ding, and L. Y. Wu, "iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition," *PLoS ONE*, vol. 8, no. 2, Article ID e55844, 2013.

[21] W. R. Qiu and X. Xiao, "iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components," *International Journal of Molecular Sciences*, vol. 15, no. 2, pp. 1746–1766, 2014.

[22] B. Liu, D. Zhang, R. Xu et al., "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

[23] W. Chen, P. M. Feng, H. Lin, and K. C. Chou, "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Research*, vol. 41, no. 6, article e68, 2013.

[24] J. L. Xiao, X. Min, and K.-C. Chou, "iEzy-Drug: a web server for identifying the interaction between enzymes and drugs in cellular networking," *BioMed Research International*, vol. 2013, Article ID 701317, 13 pages, 2013.

[25] X. Xiao, J. L. Min, and P. Wang, "iCDI-PseFpt: identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints," *Journal of Theoretical Biology C*, vol. 337, pp. 71–79, 2013.

[26] Y. Xu, X. J. Shao, L. Y. Wu, N. Y. Deng, and K. C. Chou, "iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins," *PeerJ*, vol. 1, article e171, 2013.

[27] W. Chen, H. Lin, P. M. Feng, C. Ding, and Y. C. Zuo, "iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties," *PLoS ONE*, vol. 7, no. 10, Article ID e47843, 2012.

[28] Y. Xu, X. Wen, X. J. Shao, and N. Y. Deng, "iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition," *International Journal of Molecular Sciences*, vol. 15, no. 5, pp. 7594–7610, 2014.

[29] T. U. Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 40, pp. D71–D75, 2012.

[30] K. C. Chou and H. B. Shen, "Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers," *Journal of Proteome Research*, vol. 5, no. 8, pp. 1888–1897, 2006.

[31] K. C. Chou and H. B. Shen, "Recent progress in protein subcellular location prediction," *Analytical Biochemistry*, vol. 370, no. 1, pp. 1–16, 2007.

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

[33] J. C. Wootton and S. Federhen, "Statistics of local complexity in amino acid sequences and sequence databases," *Computers and Chemistry*, vol. 17, no. 2, pp. 149–163, 1993.

[34] J. J. Chou, "A formulation for correlating properties of peptides and its application to predicting human immunodeficiency virus protease-cleavable sites in proteins," *Biopolymers*, vol. 33, no. 9, pp. 1405–1414, 1993.

[35] K. C. Chou, "Prediction of G-protein-coupled receptor classes," *Journal of Proteome Research*, vol. 4, no. 4, pp. 1413–1418, 2005.

[36] M. Wang, J. Yang, Z. J. Xu, and K.C. Chou, "SLLE for predicting membrane protein types," *Journal of Theoretical Biology*, vol. 232, no. 1, pp. 7–15, 2005.

[37] X. Xiao, P. Wang, and K.-C. Chou, "Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image," *Journal of Theoretical Biology*, vol. 254, no. 3, pp. 691–696, 2008.

[38] K. Y. Feng, Y. D. Cai, and K. C. Chou, "Boosting classifier for predicting protein domain structural class," *Biochemical and Biophysical Research Communications*, vol. 334, no. 1, pp. 213–217, 2005.

[39] Y. D. Cai and K. C. Chou, "Artificial neural network model for predicting α-turn types," *Analytical Biochemistry*, vol. 268, no. 2, pp. 407–409, 1999.

[40] T. B. Thompson, C. Zheng, and K.-C. Chou, "Neural network prediction of the HIV-1 protease cleavage sites," *Journal of Theoretical Biology*, vol. 177, no. 4, pp. 369–379, 1995.

[41] C. T. Zhang and K. C. Chou, "An optimization approach to predicting protein structural class from amino acid composition," *Protein Science*, vol. 1, no. 3, pp. 401–408, 1992.

[42] P. M. Feng, W. Chen, and H. Lin, "iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition," *Analytical Biochemistry*, vol. 442, no. 1, pp. 118–125, 2013.

[43] X. Xiao, P. Wang, and K. C. Chou, "iNR-physchem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix," *PLoS ONE*, vol. 7, no. 2, Article ID e30869, 2012.

[44] W. Z. Lin, J. A. Fang, X. Xiao, and K. C. Chou, "iDNA-prot: identification of DNA binding proteins using random forest with grey model," *PLoS ONE*, vol. 6, no. 9, Article ID e24756, 2011.

[45] K. K. Kandaswamy, K.-C. Chou, T. Martinetz et al., "AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties," *Journal of Theoretical Biology*, vol. 270, no. 1, pp. 56–62, 2011.

[46] Y. D. Cai and K. C. Chou, "Predicting subcellular localization of proteins in a hybridization space," *Bioinformatics*, vol. 20, no. 7, pp. 1151–1156, 2004.

[47] K. C. Chou and Y. D. Cai, "Prediction of protease types in a hybridization space," *Biochemical and Biophysical Research Communications*, vol. 339, no. 3, pp. 1015–1020, 2006.

[48] H. Shen and K. C. Chou, "Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types," *Biochemical and Biophysical Research Communications*, vol. 334, no. 1, pp. 288–292, 2005.

[49] K. C. Chou and H. B. Shen, "Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites," *Journal of Proteome Research*, vol. 6, no. 5, pp. 1728–1734, 2007.

[50] H. B. Shen and K. C. Chou, "A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0," *Analytical Biochemistry*, vol. 394, no. 2, pp. 269–274, 2009.

[51] T. L. Zhang, Y. S. Ding, and K. C. Chou, "Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern," *Journal of Theoretical Biology*, vol. 250, no. 1, pp. 186–193, 2008.

[52] X. Xiao, P. Wang, and K. C. Chou, "GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions," *Molecular BioSystems*, vol. 7, no. 3, pp. 911–919, 2011.

[53] H. B. Shen, J. Yang, and K. C. Chou, "Fuzzy KNN for predicting membrane protein types from pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 240, no. 1, pp. 9–13, 2006.

[54] X. Xiao, J. L. Min, and P. Wang, "iGPCR-Drug: a web server for predicting interaction between GPCRs and drugs in cellular networking," *PLoS ONE*, vol. 8, no. 8, Article ID e72234, 2013.

[55] X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, and K.-C. Chou, "iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Analytical Biochemistry*, vol. 436, no. 2, pp. 168–177, 2013.

[56] K. C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Molecular Biosystems*, vol. 9, no. 6, pp. 1092–1100, 2013.

[57] H. Nakashima, K. Nishikawa, and T. Ooi, "The folding type of a protein is relevant to the amino acid composition," *Journal of Biochemistry*, vol. 99, no. 1, pp. 153–162, 1986.

[58] J. Cedano, P. Aloy, J. A. Pérez-Pons, and E. Querol, "Relation between amino acid composition and cellular location of proteins," *Journal of Molecular Biology*, vol. 266, no. 3, pp. 594–600, 1997.

[59] G.-P. Zhou, "An intriguing controversy over protein structural class prediction," *Protein Journal*, vol. 17, no. 8, pp. 729–738, 1998.

[60] S.-X. Lin and J. Lapointe, "Theoretical and experimental biology in one," *Journal of Biomedical Science and Engineering (JBiSE)*, vol. 6, pp. 435–442, 2013.

[61] X.-B. Zhou, C. Chen, Z.-C. Li, and X.-Y. Zou, "Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes," *Journal of Theoretical Biology*, vol. 248, no. 3, pp. 546–551, 2007.

[62] L. Nanni and A. Lumini, "Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization," *Amino Acids*, vol. 34, no. 4, pp. 653–660, 2008.

[63] D. N. Georgiou, T. E. Karakasidis, J. J. Nieto, and A. Torres, "Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 257, no. 1, pp. 17–26, 2009.

[64] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203–209, 2010.

[65] H. Mohabatkar, "Prediction of cyclin proteins using chou's pseudo amino acid composition," *Protein and Peptide Letters*, vol. 17, no. 10, pp. 1207–1214, 2010.

[66] S. S. Sahu and G. Panda, "A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction," *Computational Biology and Chemistry*, vol. 34, no. 5-6, pp. 320–327, 2010.

[67] H. Mohabatkar, M. Mohammad Beigi, and A. Esmaeili, "Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine," *Journal of Theoretical Biology*, vol. 281, no. 1, pp. 18–23, 2011.

[68] M. Mohammad Beigi, M. Behjati, and H. Mohabatkar, "Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach," *Journal of Structural and Functional Genomics*, vol. 12, no. 4, pp. 191–197, 2011.

[69] S. Mei, "Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization," *Journal of Theoretical Biology*, vol. 293, pp. 121–130, 2012.

[70] L. Nanni, S. Brahnam, and A. Lumini, "Wavelet images and Chou's pseudo amino acid composition for protein classification," *Amino Acids*, vol. 43, no. 2, pp. 657–665, 2012.

[71] L. Nanni, A. Lumini, D. Gupta, and A. Garg, "Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's Pseudo amino acid composition and on evolutionary information," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 2, pp. 467–475, 2012.

[72] M. K. Gupta, R. Niyogi, and M. Misra, "An alignment-free method to find similarity among protein sequences via the general form of Chou's pseudo amino acid composition," *SAR and QSAR in Environmental Research*, vol. 24, no. 7, pp. 597–609, 2013.

[73] Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani, and H. Mohabatkar, "Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test," *Journal of Theoretical Biology*, vol. 341, pp. 34–40, 2014.

[74] C. Huang and J. Yuan, "Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites," *Biosystems*, vol. 113, no. 1, pp. 50–57, 2013.

[75] C. Huang and J. Q. Yuan, "Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions," *Journal of Theoretical Biology*, vol. 335, pp. 205–212, 2013.

[76] H. Mohabatkar, M. Mohammad Beigi, K. Abdolahi, and S. Mohsenzadeh, "Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach," *Medicinal Chemistry*, vol. 9, no. 1, pp. 133–137, 2013.

[77] A. N. Sarangi, M. Lohani, and R. Aggarwal, "Prediction of essential proteins in prokaryotes by incorporating various physico-chemical features into the general form of Chou's pseudo amino acid composition," *Protein and Peptide Letters*, vol. 20, no. 7, pp. 781–795, 2013.

[78] S. Wan, M. W. Mak, and S. Y. Kung, "GOASVM: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 323, pp. 40–48, 2013.

[79] W. Chen, T. Y. Lei, D. C. Jin, and H. Lin, "PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition," *Analytical Biochemistry*, vol. 456, pp. 53–60, 2014.

[80] B.-Q. Li, T. Huang, L. Liu, Y.-D. Cai, and K.-C. Chou, "Identification of colorectal cancer related genes with mrmr and shortest path in protein-protein interaction network," *PLoS ONE*, vol. 7, no. 4, Article ID e33393, 2012.

[81] T. Huang, J. Wang, Y.-D. Cai, H. Yu, and K.-C. Chou, "Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma," *PLoS ONE*, vol. 7, no. 4, Article ID e34460, 2012.

[82] Y. Jiang, T. Huang, L. Chen, Y. F. Gao, Y. Cai, and K.-C. Chou, "Signal propagation in protein interaction network during colorectal cancer progression," *BioMed Research International*, vol. 2013, Article ID 287019, 9 pages, 2013.

[83] H.-B. Shen and K.-C. Chou, "PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition," *Analytical Biochemistry*, vol. 373, no. 2, pp. 386–388, 2008.

[84] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions," *Analytical Biochemistry*, vol. 425, no. 2, pp. 117–119, 2012.

[85] D. S. Cao, Q. S. Xu, and Y. Z. Liang, "propy: a tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, no. 7, pp. 960–962, 2013.

[86] P. Du, S. Gu, and Y. Jiao, "PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets," *International Journal of Molecular Sciences*, vol. 15, no. 3, pp. 3495–3506, 2014.

[87] L. C. Chen YW, "Combining SVMs with various feature selection strategies," in *Feature Extraction*, I. Guyon, N. Nikravesh, S. Gunn, and L. Zadeh, Eds., pp. 315–324, Springer, Berlin, Germany, 2006.

[88] H. Lin, H. Ding, F.-B. Guo, and J. Huang, "Prediction of subcellular location of mycobacterial protein using feature selection techniques," *Molecular Diversity*, vol. 14, no. 4, pp. 667–671, 2010.

[89] K.-C. Chou and Y.-D. Cai, "Using functional domain composition and support vector machines for prediction of protein subcellular location," *The Journal of Biological Chemistry*, vol. 277, no. 48, pp. 45765–45769, 2002.

[90] Y.-D. Cai, G.-P. Zhou, and K.-C. Chou, "Support vector machines for predicting membrane protein types by using functional domain composition," *Biophysical Journal*, vol. 84, no. 5, pp. 3257–3263, 2003.

[91] N. Cristianini and J. Shawe-Taylor, *An Introduction of Support Vector Machines and Other Kernel-Based Learning Methodds*, Cambridge University Press, Cambridge, UK, 2000.

[92] K. C. Chou and C. T. Zhang, "Review: prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 4, pp. 275–349, 1995.

[93] G. P. Zhou and N. Assa-Munt, "Some insights into protein structural class prediction," *Proteins: Structure, Function and Genetics*, vol. 44, no. 1, pp. 57–59, 2001.

[94] K.-C. Chou, Z.-C. Wu, and X. Xiao, "ILoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Molecular BioSystems*, vol. 8, no. 2, pp. 629–641, 2012.

[95] K.-C. Chou, Z.-C. Wu, and X. Xiao, "iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins," *PLoS ONE*, vol. 6, no. 3, Article ID e18258, 2011.

[96] K.-C. Chou, "Using subsite coupling to predict signal peptides," *Protein Engineering*, vol. 14, no. 2, pp. 75–79, 2001.

[97] K. C. Chou, "Prediction of signal peptides using scaled window," *Peptides*, vol. 22, no. 12, pp. 1973–1979, 2001.

[98] K. C. Chou and H. B. Shen, "Review: recent advances in developing web-servers for predicting protein attributes," *Natural Science*, vol. 1, no. 2, pp. 63–92, 2009.

[99] K. C. Chou, "Review: prediction of human immunodeficiency virus protease cleavage sites in proteins," *Analytical Biochemistry*, vol. 233, no. 1, pp. 1–14, 1996.