molecules

MDPI

Review

Recent Advances in Conotoxin Classification by Using Machine Learning Methods

Fu-Ying Dao [1], Hui Yang [1], Zhen-Dong Su [1], Wuritu Yang [1,2], Yun Wu [3], Hui Ding [1], Wei Chen [1,4,*], Hua Tang [5,*] and Hao Lin [1,*]

1 Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China; koyee_d@sina.com (F.-Y.D.); huiyang0325@163.com (H.Y.); zhendong__su@163.com (Z.-D.S.); wyang@imu.edu.cn (W.Y.); hding@uestc.edu.cn (H.D.)

2 Development and Planning Department, Inner Mongolia University, Hohhot 010021, China

3 College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China; ywu@xmut.edu.cn

4 Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China

5 Department of Pathophysiology, Southwest Medical University, Luzhou 646000, China

* Correspondence: greatchen@ncst.edu.cn (W.C.); Tanghua771211@aliyun.com (H.T.); hlin@uestc.edu.cn (H.L.)

Received: 17 May 2017; Accepted: 19 June 2017; Published: 25 June 2017

Abstract: Conotoxins are disulfide-rich small peptides, which are invaluable peptides that target ion channel and neuronal receptors. Conotoxins have been demonstrated as potent pharmaceuticals in the treatment of a series of diseases, such as Alzheimer's disease, Parkinson's disease, and epilepsy. In addition, conotoxins are also ideal molecular templates for the development of new drug lead compounds and play important roles in neurobiological research as well. Thus, the accurate identification of conotoxin types will provide key clues for the biological research and clinical medicine. Generally, conotoxin types are confirmed when their sequence, structure, and function are experimentally validated. However, it is time-consuming and costly to acquire the structure and function information by using biochemical experiments. Therefore, it is important to develop computational tools for efficiently and effectively recognizing conotoxin types based on sequence information. In this work, we reviewed the current progress in computational identification of conotoxins in the following aspects: (i) construction of benchmark dataset; (ii) strategies for extracting sequence features; (iii) feature selection techniques; (iv) machine learning methods for classifying conotoxins; (v) the results obtained by these methods and the published tools; and (vi) future perspectives on conotoxin classification. The paper provides the basis for in-depth study of conotoxins and drug therapy research.

Keywords: conotoxin; superfamily; ion channel; machine learning method

1. Introduction

Conotoxins are the group of cysteine-rich neurotoxic peptides isolated from the venom of the marine snails of the genus *Conus*. [1]. Mature conotoxins consist of 10 to 30 residues with $\geq$1 disulfide bonds. By binding to various ion channels, conotoxins possess important biological functions [2]. Conotoxins play key roles in pharmacology and neuroscience as well as new drug development; and have attracted the attention of scientists worldwide [3–9]. Wang et al. [10] found there was apparent synergistic analgesic effects that were produced by ω-conotoxin MVIIA and morphine in rats. Conantokin-R [11] is a highly potent anticonvulsant with a protective index of 17.5 when tested on an audiogenic mouse model of epilepsy.

Over the last few decades, conotoxins have been the subject of pharmacological interest [12], and have been used in the treatment of various diseases such as Alzheimer's disease, Parkinson's disease, epilepsy, chronic pain, and cardiovascular diseases. Conical spirodotoxin, as a non-addictive analgesic, has good prospects. Under the same dose, the effect of conical spirodotoxin is 1000 times higher than that of morphine. Conotoxins have also been characterized by various rapeutic potentials in pre-clinical or clinical trials, such as antinociceptive [13], antiepileptic [14], neuroprotective, and cardioprotective activities [15]. In addition, they also have the potential to cultivate insect-resistant crop varieties and be the candidate of polypeptide pesticide [16,17]. The therapeutic potential of conotoxin is ascribed to their special ion channel-targets in the nervous systems [4]. Thus, they have been regarded as excellent pharmacological probes and potential candidate compounds for drug design for neurological disorders [18].

Based on the N-terminal precursor sequence and disulfide connectivity, uncharted conotoxins may be classified into several superfamilies [19,20]. Currently, conotoxins can be classified into 16 major superfamilies: A, D, I1, I2, I3, J, L, M, O1, O2, O3, P, S, T, V, and Y [4,19–25]. Each superfamily can be further classified into several families based on the cysteine arrangement. For example, A-superfamily conotoxins are classified into α, αA, and κA families; M-superfamily [26,27] includes μ and ψ families; O-superfamily includes δ, μO, ω, κ, and γ families [22,28]. Due to the high specificity and affinity towards ion channels, conotoxins can also be categorized into calcium channel-targeted conotoxins (Ca-conotoxins), sodium channel-targeted conotoxins (Na-conotoxins), and potassium channel-targeted conotoxins (K-conotoxins) [29]. We draw a structural schematic illustration to show conotoxins classifications of superfamily and ion channel-target (Figure 1).

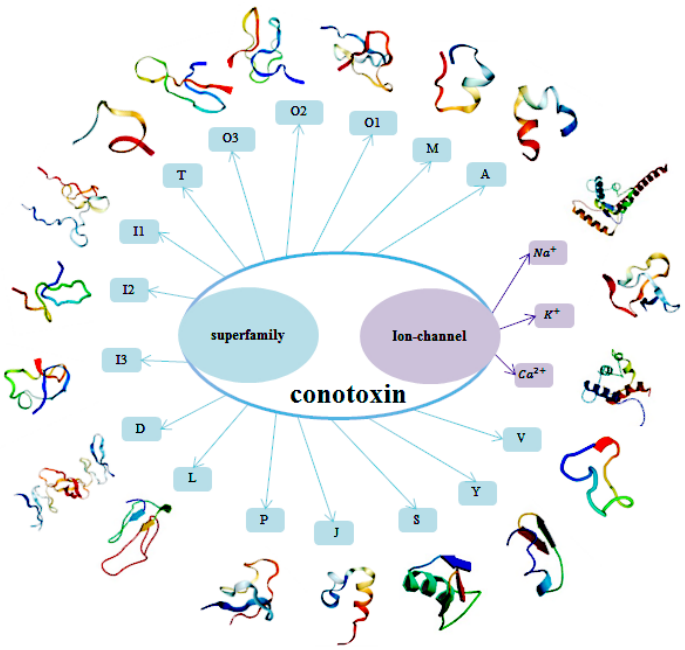

Figure 1. A structural schematic illustration to show the classification of conotoxins in superfamily and ion channel-target. Sixteen major conotoxin superfamilies are A, D, I1, I2, I3, J, L, M, O1, O2, O3, P, S, T, V, and Y. They are also categorized into calcium channel-targeted, sodium channel-targeted, and potassium channel-targeted conotoxins according to their functions.

There are over 100,000 conotoxins in approximately 700 species of cone snails [2]. However, only 8344 conotoxins have been deposited in the Universal Protein Resource (UniProt, 15 May 2017). The functions of most conotoxins are still unknown. With more and more conotoxins being sequenced, determining the function of conotoxins with biochemical experiment-based methods is becoming more and more difficult because of the high cost and long period of wet experiments. Computational

methods have provided opportunities to rapidly and accurately identify the categories of conotoxins and know about some functions of conotoxins while avoiding the disadvantages of biochemical experiments-based methods [30,31].

Machine learning approaches have been widely applied in protein or peptide classification by using amino acid composition, *n*-mer amino acid composition, pseudo amino acid composition, position-specific scoring matrix (PSSM) and so on [32–39]. A process framework of protein or peptide classification with a machine learning approach was shown in Figure 2. These methods were also proposed to identify conotoxins superfamily type. A multi-class support vector machine (SVM) was proposed to predict conotoxin superfamily by using pseudo amino acid composition (PseAAC) [24]. Subsequently, Lin et al. improved the accuracy of classifying conotoxin superfamily by using the modified Mahalanobis discriminant [32]. Inspired by these works, Fan et al. proposed a late-model approach and established a webserver called PredCSF for conotoxin superfamily prediction [33]. Zaki et al. used local alignment partition functions to predict conotoxin superfamilies [34]. Then, they introduced a novel method called Toxin-AAM for classifying conotoxin superfamilies [35]. Yin et al. predicted conotoxin superfamilies by using diffusion maps-based feature selection technique [36]. Laht et al. classified conotoxin superfamilies and families based on profile Hidden Markov Models (pHMMs) [37]. Koua et al. established pHMMs for each of the 48 alignments with the hmmbuild script in the HMMER 3.0 package (Manufacturer, City, US State abbrev. if applicable, Country) and built a webserver called ConoDictor based on the method [38]. Moreover, they defined 50 position-specific scoring matrices (PSSMs) and 47 hidden Markov models to improve accuracy for conotoxin superfamily prediction [39].

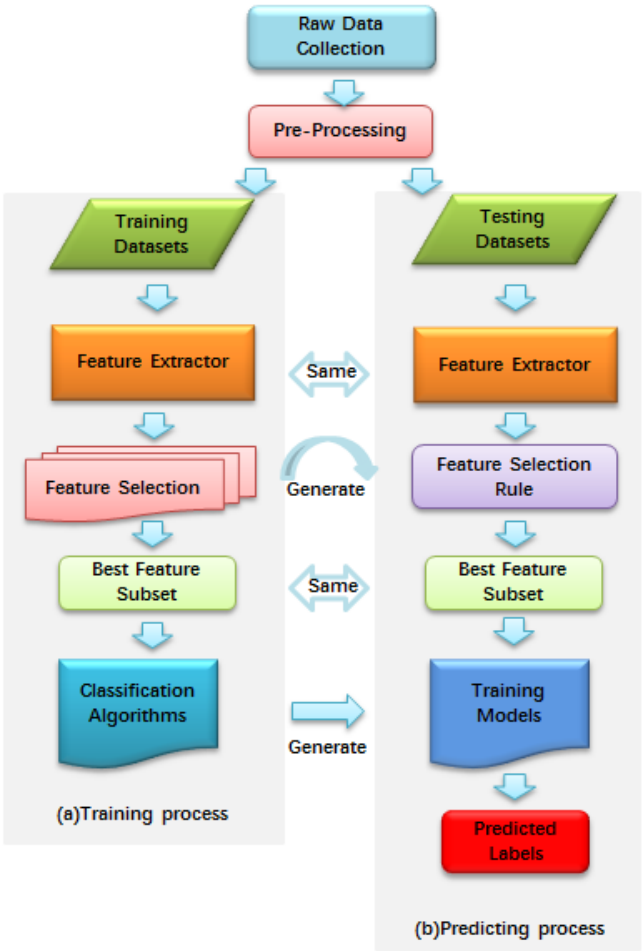

Figure 2. The process framework of conotoxin classification with machine learning methods.

Although these methods and results exemplified above can provide some clues for the study of conotoxins, they only indirectly offer possible function information of conotoxins and they cannot predict the receptor types of the conotoxins. For example, both Delta-conotoxin-like Ac6.1 and Omega-conotoxin-like Ai6.2 belong to the O1 superfamily, but they target different ion channels. The Delta-conotoxin-like Ac6.1 binds to voltage-gated sodium channels, whereas the Omega-conotoxin-like Ai6.2 blocks voltage-gated calcium channels [40]. Thus, it is necessary to develop new computational tools that can recognize the types of ion channel-targeted conotoxins. For the first time, Yuan et al. developed a feature selection technique based on binomial distribution to predict the types of ion channel-targeted conotoxins by using a radial basis function network [41]. Subsequently, they developed a predictor (*iCTX-Type*) to improve prediction accuracies [42]. Zhang et al. applied a hybrid feature in the prediction issue [43]. Wu et al. incorporated new properties of residues into PseAAC to predict the types of conotoxins [44]. Recently, Wang et al. combined the analysis of variance and correlation (AVC) with SVM to reduce redundancy of attributes and improve the prediction accuracy and computation speed [45].

In this review, we summarized recent advances in conotoxin classification by using machine learning methods in the following aspects: (i) benchmark dataset construction; (ii) feature extraction method; (iii) feature selection technique; (iv) classification algorithms; (v) prediction accuracy and web servers establishment; and (vi) prospect of conotoxin prediction with machine learning methods.

2. Benchmark Datasets

2.1. Published Database Resources

Constructing a high quality and reliable benchmark dataset is critical for the protein attribute predictor. Both general databases and special databases play a key role in the construction of bioinformatics benchmark [46–49]. The general databases include the protein knowledgebase (UniProtKB: http://www.uniprot.org) [50], the protein structure data bank (PDB: http://www.rcsb.org/pdb/home/home.do) [51], and the protein database provided by the National Center for Biotechnology information (NCBI) [52]. Researchers used to collect the data from these molecular biology databases.

For the convenience of users, some special databases were constructed. Here, we mainly introduced the ConoServer (http://www.conoserver.org/), which was a specific database for conotoxins [53,54]. The database collected various kinds of information of conotoxins from SwissProt, GenBank, Protein Data Bank and literatures, including peptide sequences, chemical modifications, and their ability to block the ion channels. At present, the ConoServer has managed 2838 nucleic sequences (from 83 *Conus* species), 6255 protein sequences (from 109 *Conus* species) and 176 3D structures (from 35 *Conus* species) until 16 April 2017, provides a convenient overview of current knowledge on conopeptides and furnishes sequence/structure/activity relationships information, which is of particular interest for drug design research.

2.2. Benchmark Dataset Construction

Although the ConoServer contains much information, for the purpose of conotoxin prediction, it is necessary to construct a new benchmark dataset that can be handled by machine learning methods. Generally, a high quality benchmark dataset is constructed in the four following steps. In step 1, samples of conotoxin peptide are acquired from a database with some relevant key words. In step 2, only those proteins with clear functional annotations based on experimental evidence are included. In step 3, the proteins with the annotation information of "immature", "invalid", and "fragment" are excluded. In step 4, redundancy and homology bias are reduced by using the program CD-HIT [55] which has been widely used for clustering and comparing protein or nucleotide sequences.

Based on the strict steps above, some high-quality datasets have been constructed for conotoxin superfamilies. Some superfamilies with relatively less members were not considered in some

studies [24,32]. The first benchmark dataset of superfamily was called S1, which included 116 mature conotoxin sequences including A (25 entries), M (13 entries), O (61 entries) and T (17 entries) superfamilies [24]. At the same time, they also built a negative dataset containing 60 short peptide sequences that did not belong to any of the four superfamilies (A, M, O or T). The second benchmark dataset S2 contains 261 entries consisting of four superfamilies: A (63 samples), M (48 samples), O (95 samples) and T (55 samples) obtained from the SwissProt [33]. In addition, Lath et al. collected 964 sequences from ConoServer [37]. Koua et al. also acquired 933 samples and 967 samples from Conoserver [38,39].

The benchmark dataset of ion channel-targeted conotoxins was also constructed based on the Uniprot. The function type of conotoxins was obtained by searching Gene Ontology. The first benchmark dataset I1 established by Yuan et al. included 112 sequences (24 K-conotoxins, 43 Na-conotoxins, and 45 Ca-conotoxins) [41]. Ding et al. [42], Wu et al. [44] and Wang et al. [45] also established their models based on this dataset. In addition, Zhang et al. built a new dataset called I2 containing 145 samples (26 K-conotoxins, 49 Na-conotoxins and 70 Ca-conotoxins) [43]. The benchmark datasets are provided in Table 1.

Table 1. The benchmark datasets of conotoxin superfamily and ion channel-targeted conotoxin.

	Superfamily				Total Number	Reference
	A	**M**	**O**	**T**		
S1	25	13	16	17	116	[24,32,34,35]
S2	63	48	95	55	216	[33,36]
	Type of Ion Channel				**Total Number**	**Reference**
	K-Conotoxin	**Na-Conotoxin**	**Ca-Conotoxin**			
I1	24	43	45		112	[41,42,44,45]
I2	26	49	70		145	[43]

3. Conotoxin Sample Description Methods

In the process of protein classification with machine learning methods, the second step is to represent protein samples. Two strategies may be adopted: the continuous model and the discrete model. In the continuous model, the BLAST or FASTA programs are used to search homology. For a highly similar sequence (sequence identity $\geq$40%) in the searching dataset, its predictive results are always good. Thus, the similarity-based method is straightforward and intuitive. However, if a query protein has no similar sequence in the training dataset, these methods cannot work. Therefore, various discrete models were recommended [24,32–36,41–45,56]. The way to formulate conotoxin samples with discrete models is provided below.

3.1. Amino Acid Compositions and Dipeptide Compositions

The amino acid compositions (AAC) and dipeptide compositions are the most widely used features to formulate the protein samples, and can be formulated as:

$$X_{20} = [x_1 \cdots x_i \cdots x_{20}]^T, \quad (1)$$

$$Y_{400} = [y_1 \cdots y_i \cdots y_{400}]^T, \quad (2)$$

where x_i (i = 1,2,... , 20) and y_i (i = 1, 2,..., 400) are, respectively, the absolute occurrence frequencies of 20 native amino acids and 400 dipeptides, which, respectively, describe the sequence composition and neighborhood information of residues.

Based on the two kinds of parameters above, Lin et al. [32] developed a method to predict conotoxin superfamilies by combining the increment of diversity with modified Mahalanobis

discriminant. Recently, the 400 dipeptide compositions were also used to represent a conotoxin sequence by Wang et al. [45].

3.2. Pseudo Amino Acid Composition

The pseudo amino acid composition (PseAAC) is a widely used strategy for peptide sample description in protein classification [57,58]. PseAAC can not only include amino acid composition, but also the correlation of physicochemical properties between two residues [59]. Its merits have been demonstrated in a series of studies [24,44,57,58].

Mondal et al. constructed a model by using Type-I PseAAC to formulate samples for predicting superfamilies of conotoxins [24]. The Type-I PseAAC is also called parallel correlation PseAAC, which contains 20 + λ components. The number '20' reflects the occurrence frequency of one of the 20 native amino acids in a protein P and λ reflects the rank of correlation and is a non-negative integer. In the discrete descriptor, an arbitrary conotoxin (P) can be expressed by a 20 + λ-dimensional vector and is defined as follows:

$$P = [x_1 \cdots x_{20} x_{20+1} \cdots x_{20+\lambda}]^T, \quad (3)$$

where

$$x_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \theta_j}, & (1 \le u \le 20) \\ \frac{\omega \theta_{u-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \theta_j}, & (20+1 \le u \le 20+\lambda) \end{cases}, \quad (4)$$

where f_i is denoted as the normalized frequency of the 20 residues in a conotoxin. ω is weight factor for sequence order effect and was previously defined as 0.7 [24]. θ_j is the j-tier sequence correlation factor and calculated as:

$$\theta_j = \frac{1}{L-j} \sum_{i=1}^{L-j} F(R_i, R_{i+1}), \ (j < L), \quad (5)$$

where θ_j is the j-th tire correlation factor that reflects the sequence order correlation between all the j-th most contiguous residues along a protein sequence. In addition, the correlation function is given by:

$$F(R_i, R_j) = \frac{1}{k} \left\{ [H_1(R_j) - H_1 R_i]^2 + [H_2(R_j) - H_2 R_i]^2 + \cdots + [H_k(R_j) - H_k R_i]^2 \right\}, \quad (6)$$

where k is the number of factors and $H_l(R_i)$ is the l-th physiochemical properties of the residue R_i:

$$H_l(R_i) = \frac{H_l^0(i) - \sum_{i=1}^{20} (H_l^0(i)/20)}{\sqrt{\frac{\sum_{i=1}^{20} [H_l^0(i) - \sum_{i=1}^{20} (H_l^0(i)/20)]^2}{20}}}, \quad (7)$$

where $H_l^0(i)$ is the l-th original value of the i-th residue. The numerical indices 1, 2, 3,$\cdots$, 20, respectively, represent the 20 native amino acids: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y. The five factors of polarity index, secondary structure factor, molecular size, relative amino acid composition in various proteins and electrostatic charge were used in the model of Mondal et al. [24].

Wu et al. [44] used the Type-II PseAAC, which is also called series correlation PseAAC, to formulate their samples. In the descriptor, an arbitrary conotoxin (P) is expressed as a vector containing (20^2 + 3λ) components:

$$P = [x_1 \cdots x_{400} \cdots x_{400+3\lambda}]^T, \quad (8)$$

where $x_1 \cdots x_{400}$ denote the frequencies of 20^2 dipeptides. The '3' is the number of amino acid properties, namely, rigidity, flexibility, and irreplaceability; λ reflects the rank of correlation, which is the same as that in Type-I PseAAC:

$$x_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{400} f_u + \omega \sum_{j=1}^{3\lambda} \tau_j}, & (1 \le u \le 400) \\ \frac{\omega \tau_u}{\sum_{i=1}^{400} f_u + \omega \sum_{j=1}^{3\lambda} \tau_j}, & (400+1 \le u \le 400+3\lambda) \end{cases}, \quad (9)$$

where ω is weight factor for sequence order effect; and f_u was the normalized frequency of the 400 dipeptides in conotoxin (P);

$$f_u = \frac{n_u}{\sum_u n_u}, \quad (10)$$

where n_u denotes the number of occurrences of u-th dipeptide in conotoxin (P); τ_u in Equation (9) is the correlation factor of the physicochemical properties between residues:

$$\begin{cases} \tau_1 = \frac{1}{L-1} \sum_{K=1}^{L-1} H_{k,k+1}^1 \\ \tau_2 = \frac{1}{L-1} \sum_{K=1}^{L-1} H_{k,k+1}^2 \\ \quad \vdots \\ \tau_n = \frac{1}{L-1} \sum_{K=1}^{L-1} H_{k,k+1}^n \\ \tau_{n+1} = \frac{1}{L-2} \sum_{K=1}^{L-2} H_{k,k+2}^1 \\ \tau_{n+2} = \frac{1}{L-2} \sum_{K=1}^{L-2} H_{k,k+2}^2 \\ \quad \vdots \\ \tau_{n+n} = \frac{1}{L-2} \sum_{K=1}^{L-2} H_{k,k+2}^n \\ \quad \vdots \\ \tau_{n\lambda} = \frac{1}{L-\lambda} \sum_{K=1}^{L-\lambda} H_{k,k+\lambda}^n \end{cases} \quad (\lambda < L), \quad (11)$$

where $H_{k,k+\lambda}^n$ is the correlation function:

$$H_{k,k+\lambda}^n = h^n(R_k) \cdot h^n(R_{k+\lambda}), \quad (12)$$

where $h^n(R_k)$ is the n-th kind of the physicochemical values of the residue R_k. The values should be converted to standard type:

$$h^n(R_k) = \frac{h_0^n(R_k) - \langle h_0^n(R_k) \rangle}{SD\langle h_0^n(R_k) \rangle}, \quad (13)$$

where $h_0^n(R_k)$ is the original physicochemical value of the k-th residue.

Both Type-I and Type-II PseAAC can not only describe the information of the constituent elements of the conotoxin sequence, but also reflect the long-range correlation information of residues' physicochemical properties. Therefore, PseAAC can usually produce better prediction accuracy compared with the traditional peptide frequency. Because Type-II PseAAC considers the contributions of each kind of physicochemical property, it exhibits a better prediction performance as shown in Ref. [44]

3.3. Hybrid Features

Instead of using a single discrete model, different features were used to describe conotoxin samples. Recently, the 246 physicochemical properties of residues obtained from APDbase [60] were used to formulate protein samples [33,36]:

$$\begin{cases} P_1 = R_1^1 R_2^1 R_3^1 R_4^1 \cdots R_L^1 \\ P_2 = R_1^2 R_2^2 R_3^2 R_4^2 \cdots R_L^2 \\ \quad \vdots \\ P_{246} = R_1^{246} R_2^{246} R_3^{246} R_4^{246} \cdots R_L^{246} \end{cases}. \quad (14)$$

By using the maximal overlap discrete wavelet transform (MODWT) to construct the eigenvectors [61], a conotoxin sample can thus be represented by a 1230-dimensional feature vector ((1 + 1 + 3) × 246 = 1230):

$$F_{MODWT} = \left[f_1^{1,1}, f_2^{1,2}, f_3^{1,3}, f_4^{1,4}, f_5^{1,5}, f_6^{2,1}, f_7^{2,2}, f_8^{2,3}, f_9^{2,4}, f_{10}^{2,5}, \cdots f_{1230}^{246,5}, \right]. \quad (15)$$

In addition, three characteristics were also incorporated in their model: 20D features of evolutionary information, 3D secondary structural (SS) information, and 20D AAC. Therefore, the final feature set to formulate conotoxin sample was a (1230 + 20 + 3 + 20)1273D vector.

Compared with the above two methods, the method combines with several models to represent protein samples. Thus, the bias caused by a single discrete model can be significantly reduced.

4. Feature Selection Techniques

Feature selection is important in pattern recognition for the insight gained from determining relevant modeling variables. By feature selection, generalization ability of prediction model will improve, information redundancy or noise will be excluded; and the dimension disaster will be resolved [62]. It can significantly increase the comprehensibility of classifier models and often build a better model [63]. The ultimate goal of feature selection is to find the best feature subset that can produce the maximum accuracy and to establish a robust prediction model. Currently, many feature selection techniques have been developed to optimize a feature set, such as principal component analysis (PCA) [64], minimal-redundancy-maximal-relevance (mRMR) [65], maximum-relevancy-maximum-distance (MRMD) [66], diffusion maps [36] and the analysis of variance (ANOVA) [67]. The following feature selection techniques have been used in conotoxin prediction.

4.1. Binomial Distribution

Binomial distribution is a discrete probability and can deal with the experiments that have two types of results. Thus, Yuan et al. [41] proposed using the binomial distribution to perform feature selection in order to improve the accuracy of conotoxin prediction. In their model, the confidence level (*CL*) of each feature was calculated by:

$$CL_{ij} = 1 - \sum_{n=n_{ij}}^{N_i} \frac{N_i!}{n!(N_i - n)!} p_j^n (1 - p_j)^{N_i - n}, \quad (16)$$

where CL_{ij} is the confidence level of the i-th dipeptide in the j-th type; N_i represents the total number of the i-th dipeptide in the dataset; n_{ij} represents the occurrence number of the i-th dipeptide in the j-th type and the sum is taken from n_{ij} to N_i; the probability p_j is the relative frequency of Type j in the database; the confidence level of peptide i in benchmark dataset is defined as follows:

$$CL_i = max\{CL_{i\ k}, CL_{i\ Na}, CL_{i\ Ca}\}, \quad (17)$$

All features can be ranked in descending order according to their *CL*s. According to the principle of feature selection, the CL_i reveals the degree that the i-th feature is related to the group variables. The larger *CL* the feature is, the higher its contribution to the classification. The binomial distribution-based technique is a powerful statistical method that can extract the over-represented motifs; however, it needs more computational resources.

4.2. Relief Algorithm

Zhang et al. [43] proposed another feature selection technique called relief algorithm in conotoxin classification. The relevance between the features and class labels can be depicted by this algorithm [68].

Based on the ability of the feature to distinguish the near samples, the weighted features can be formulated by [69]:

$$W_P^{i+1} = W_P^i - \frac{diff(Y, x_i, H(x_i))}{m} + \frac{diff(S, x_i, M(x_i))}{m}, \tag{18}$$

$$diff(*, x, y) = \begin{cases} \|x - y\|, x \neq y, \\ 0, \ x \neq y, \end{cases} \tag{19}$$

where W_P^i and W_P^{i+1} denote the current and next weighting values, respectively. p stands for a given feature; x_i denotes the i-th sample sequence; $H(x_i)$ represents the nearest neighbor samples from the same class label against x_i; $M(x_i)$ represents the nearest neighbor samples from the different class labels against x_i; Y and S are, respectively, the sample sets with the same and different class labels against x_i; m denotes the number of random samples; the function of diff$(*, x, y)$ is used to calculate the distance between the random samples.

The algorithm is not dependent on heuristics, runs in low-order polynomial time, and is noise-tolerant and robust to feature interactions; however, it does not discriminate between redundant features.

4.3. F-Score Algorithm

Ding et al. [42] and Wu et al. [44] used the F-score to sort the features for conotoxin classification:

$$F(i) = \frac{\sum_{k=1}^{3}\left(\overline{x}_i^k - \overline{x}_i\right)^2}{\sum_{k=1}^{3}(1/(N_k - 1))\sum_{j=1}^{N_k}\left(x_{ij}^k - \overline{x}_i^k\right)^2}, \tag{20}$$

where $\overline{x}_i^k$ is the average frequency of the i-th feature in the k-th dataset; $\overline{x}_i$ the average frequency of the i-th feature in all of the datasets concerned; x_{ij}^k is the frequency of the i-th feature of the j-th sequence in the k-th dataset; N_k is the number of peptide samples in the k-th dataset. The larger the F value is, the better the predictive capability the feature has. A python script fselect.py downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/ can be used to perform the F-score calculation and rank the features.

The F-score is a simple but effective technique for evaluating the discriminative power of each feature in the feature set. It has strict mathematical definition but does not take the true negatives into account.

4.4. Diffusion Map Reduction

For conotoxin superfamily classification, Yin et al. [36] proposed using the diffusion maps to project the data into diffusion space. Diffusion maps algorithm can effectively reduce the data dimensionality while keeping the original data structure [70]. Thus, high-dimensional data can be projected into a low-dimensional space based on diffusion maps, while the intrinsic properties are kept almost invariant. Compared with other methods, the diffusion maps algorithm is robust to noise perturbation and is computationally inexpensive. It has strict mathematical definition but does not take the true negatives into account [71].

It is assumed that there is a dataset Ω with N observations and each of them has p attributes. If a weighted graph on the dataset is defined, the margin between two observations x and y is defined as:

$$w(x, y) = exp\left(-\frac{(x - y)^2}{\varepsilon}\right), \tag{21}$$

where $(x - y)^2$ is the application dependent dissimilarity between x and y; the degree of node x is defined as:

$$d(x) = \sum_{z \in \Omega} w(x, z). \tag{22}$$

Next, a Markov random walk over the weighted graph can be constructed. The transition probability from x to y in one-step is:

$$p(x, y) = \frac{w(x, z)}{d(x)}, \tag{23}$$

where $w(x, z)$ and $d(x)$ are, respectively, defined in Equations (21) and (22).

Then, a transition matrix R of size N × N can be built, and each element of R is calculated by Equation (23). R is the transition matrix for a Markov random walk and can be used to calculate the transition probability matrix R_t, where each entry in R_t represents the probability going from x to y in t steps. Based on R_t, the stationary distribution of the random walk $\varphi_0(x)$ can be calculated:

$$\varphi_0(x) = \frac{d(x)}{\sum_{z \in \Omega} d(z)}. \tag{24}$$

The next step is to define the diffusion distance between two points at the scale t as:

$$D_t^2(x, y) = \sum_z \frac{(R_t(x, z) - R_t(y, z))^2}{\varphi_0(z)}. \tag{25}$$

The diffusion map at scale t can project the data x from the original space into the m-dimensional diffusion space by taking the first m eigenvectors as follows:

$$x \to \left[\frac{\lambda_1}{1 - \lambda_1} \psi_1(x), \frac{\lambda_2}{1 - \lambda_2} \psi_2(x), \cdots, \frac{\lambda_m}{1 - \lambda_m} \psi_m(x) \right], \tag{26}$$

where λ_j and ψ_j are, respectively, the eigenvalue and right eigenvector of R_t; m is the final reduced dimension.

4.5. Analysis of Variance

In order to select optimal features from the 400D dipeptide compositions, Wang et al. classified the ion channel-targeted conotoxins with the analysis of variance (ANOVA) method [45]. The variance-based analysis is used to calculate the ratio of the variance among groups and the variance within the group for each attribute [72,73]. It has a good foundation of statistics and can test the feature difference between groups intuitively. The formula expressions are as follows:

$$F(u) = \frac{S_b^2(u)}{S_w^2(u)}. \tag{27}$$

The F value represents the u-th dipeptide, and $S_b^2(u)$ is the variance between groups, $S_w^2(u)$ is the variance within groups. The calculation methods are shown in Equations (28) and (29), respectively:

$$S_b^2(u) = \frac{SS_b(u)}{K - 1}, \tag{28}$$

$$S_w^2(u) = \frac{SS_w(u)}{N - K}, \tag{29}$$

where K is the total of classes; N is the total of samples; $SS_b(u)$ is the sum of the squares between the groups; and $SS_w(u)$ is the sum of squares within the groups.

4.6. Feature Selection Process

Picking out informative features can overcome the high-dimensional disaster, reduce information redundancy, exclude noise, and improve the accuracy and robust of the proposed models. Obviously,

the most objective and strict method to select the best feature subset is to examine the performance of all the feature combinations. However, the computation time is too long. Taking a 20-dimensional feature vector as an example, there are 1,048,575 possible combinations. Thus, feature selection techniques described above are developed to economize the computational time and source.

The incremental feature selection (IFS) is a popular strategy to determine the optimal feature subset. The selection process is described as follows. At first, all features are ranked according to a score obtained from one of the feature selection techniques described above. Subsequently, the feature subset is built from the first feature in the ranked feature set. Furthermore, a new feature subset is built when the second feature is added. This process is repeated from the first feature to the last feature until all candidate features are added. For each feature subset, the machine learning methods are used to investigate their performance with cross-validation [57,74–77]. The highest accuracy is produced by the best feature subset, which is selected to build the final prediction model. The machine learning methods in conotoxin prediction is discussed below.

5. Prediction Algorithms

The four key steps for conotoxin classification are to select a highly efficient and powerful machine learning method to make a predictive decision. In the prediction, the classification function or classification model was constructed with a machine learning method for predicting the input conotoxin to a given category.

5.1. Support Vector Machine

Support Vector Machine (SVM) was originally developed by Vapnik et al. [78]. As SVM is always suitable for small sample, SVM has been widely used to deal with many pattern recognition problems [42,79–87], and also some hierarchical classification [88]. As shown in Table 1, the number of train data is from 13 to 95 for each type; thus, several works used SVM to predict conotoxin types [24,32–39,42–45].

The basic idea of SVM is to transform the input vector into a high-dimensional Hilbert space and seek a separating hyperplane in this space. Gaussian Radial Basis Function (RBF) kernel function ($K_{Gaussion}(x_i, x_j) = e^{\frac{\|x_i - x_j\|^2}{2\sigma}}$) is a widely used kernel function because of its high performance in non-line classification.

In order to reduce the programming burden of researchers, some software packages including LIBSVM, mySVM and SVMLight [89,90] have been developed and can be freely downloaded from the internet. LIBSVM is the most popular software to implement SVM and can be downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvm/. A grid search strategy with cross-validation test is always utilized to obtain the best values for the regularization parameter C and kernel parameter g.

5.2. Profile Hidden Markov Models

Profile Hidden Markov Models (pHMMs) are statistical models for capturing position-specific information [91]. The pHMMs provide a formal probabilistic framework for sequence comparison [92] and leverage the information contained in a sequence alignment to improve detection of distantly related sequences [93,94]. More recently, Hidden Markov Models have been extended to pairwise Markov models and triplet Markov models for the consideration of more complex data structures [95] and the model of non-stationary data [96]. Obviously, compared to the classic tool BLAST [97], pHMMs can more accurately detect remote homologs and provide more information by using a statistical representation of a multiple sequence alignment. Recently, a pHMM has been built for each subset using *hmmbuild* from the HMMER package and can be acquired from http://www.hmmer.org/. In addition, more packages about pHMMs can be acquired [91].

5.3. K-Local Hyperplane Distance Nearest Neighbor Algorithm

The *K*-local hyperplane distance nearest neighbor algorithm (HKNN) [98] was introduced to overcome the generalization problems of the well-known *K*-nearest neighbor algorithm (KNN). Unlike SVM, it can be used to establish a nonlinear decision surface directly in the original sample space with a local linear manifold. With the HKNN method, the closest *K* neighbors should be firstly found to test the samples for each class. Then, these neighbors are used to build the local linear manifolds of the classes. Finally, the query is allocated to the class that is associated with the closest manifold.

Suppose there are *C* classes in the training set. Let $V_i^k(x_q) = \{x_1^i, x_2^i, \cdots, x_k^i\}$ represent the set of *K* nearest samples of the tested sample $x_q \in \aleph^m$ in the training set belonging to the *i*-th class. Here, the dimension of the sample space *m* is assumed to be larger than or equal to *K*. The local affine hull of each class is defined in terms of the closest *K* sample vectors as:

$$LH_i^k(x_q) = \left\{ v \mid v = u_i + \sum_{i=1}^{l_i} \beta_j^i z_j^i, \beta_j^i \in \aleph^m \right\}, \; i = 1, \cdots, C, \tag{30}$$

where $u_i = \frac{1}{K}\sum_{j=1}^{K} x_j^i$, z_j^i are the linearly independent vectors obtained from the difference vectors $\{x_1^i - u_i, x_2^i - u_i, \cdots, x_K^i - u_i\}$; l_i is the number of linearly independent difference vectors and $l_i \leq K-1$.

In order to classify a query x_q, the minimum distances between the query vector and the local manifolds should be computed as follows:

$$\mathrm{dis}\left(x_q, LH_i^K(x_q)\right) = \min_{v \in LH_i^K(x_q)} \|x_q - v\| = \min_{\beta_i \in \aleph^{l_i}} \|x_q - u_i - Z_i\beta_i\|, \; i = 1, \cdots, C. \tag{31}$$

Thus, the query is assigned to the class whose manifold is the closest to x_q. The details about HKNN can be obtained from the results reported by Yin et al. [36].

5.4. Mahalanobis Discriminant

The Mahalanobis distant is a measure of the distance between a point *P* and a distribution *D*, introduced by P. C. Mahalanobis in 1936 [99]. If *P* is at the mean of *D*, this distance is zero and the mean grows as *P* moves away.

Mahalanobis Discriminant has been widely used in cluster analysis and data classification [100]. Due to the imbalance of data samples, based on Bayes theory, the modified Mahalanobis Discriminant was deduced [101].

The modified Mahalanobis Discriminant (*MD*) between test sequence *x* and training set *x* can be calculated as:

$$MD = \left(x - \overline{x}_§\right)^T \sum\nolimits_§^{-1} \left(x - \overline{x}_§\right) + \log\left|\sum\nolimits_§\right|, \tag{32}$$

where $\overline{x}_§$ and $\sum_§$ are, respectively, the group mean and covariance matrix for §-th training set; $\left(x - \overline{x}_§\right)^T \sum_§^{-1} \left(x - \overline{x}_§\right)$ denotes the square of Mahalanobis distance between test sequence *x* and training set $x_§$.

MD is unitless and scale-invariant, and takes into account the correlations of the features. According to the principle of similarity, the smaller the *MD* between test sequence *x* and training set $x_§$ is, the higher the probability that sequence *x* belongs to class § is.

5.5. Radial Basis Function Network

The RBF network is a special type of Artificial Neural Network (ANN). Due to its faster training procedure and better approximation capabilities, it has been widely used in protein prediction fields [102], and noncoding RNA identification [103]. The RBF network can approximate any nonlinear function with sufficient neurons in the hidden layer. A typical RBF network is composed of three

layers: an input layer, a hidden layer and a linear output layer. Training is usually carried out in two stages: (1) fixed width and centers, and (2) fixed weights. This can be demonstrated by considering the different properties of nonlinear hidden neurons versus linear output neurons.

Generally, the RBF network is modeled by the following relation:

$$\hat{y}_k = \sum_{i=1}^{m} \omega_{ik} R_i(x), \ (k = 1, 2, \cdots, p), \tag{33}$$

where $R_i(x)$ represents the RBF and is expressed as:

$$R_i(x) = \exp\left(-\frac{\|x - c_i\|^2}{2\sigma_i^2}\right), \ (i = 1, 2, \cdots, m), \tag{34}$$

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

The WEKA [104] soft package (version, Manufacturer, City, US State abbrev. if applicable, Country) is used to execute the RBF network with default parameters.

5.6. Random Forest Algorithm

The Random Forest (RF) algorithm is also a popular learning algorithm and has been successfully employed in dealing with various biological prediction problems [105–108]. The principle of RF is based on the training of multiple decision trees. It just needs two parameters: one is the number of building decision trees t, another is the number of input features to be considered when each node of the decision tree splits m. By establishing many tree predictors, the type of a new sample can be determined. The results obtained from many experiments have shown that combining multiple trees generated in randomly selected subspaces can significantly improve the prediction accuracy. The algorithm can produce a high accuracy classifier and handle a large number of input variables with fast learning process. For an unbalanced dataset, it can balance the random error. For more detailed information about the RF algorithm, readers can refer to the http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

6. Prediction Accuracy

In this section, we listed the commonly-used metrics for the performance evaluation of proposed models and introduced the published results.

6.1. Commonly-Used Evaluation Metrics

A jackknife test can yield a unique result for a given benchmark dataset and has been wildly applied in various predictions [109,110]. A set of metrics, namely, sensitivity (Sn), average accuracy (AA) (or called average sensitivity) and overall accuracy (OA) are commonly used to quantitatively estimate the accuracy of the models and respectively calculated as:

$$Sn = \frac{TP}{TP + FN}, \tag{35}$$

$$AA = \frac{\sum Sn}{\mu}, \tag{36}$$

$$OA = \frac{TP}{TP + TN + FP + FN}, \tag{37}$$

where TP, FP, TN, and FN, respectively, denote the number of true positives, false positives, true negatives, and false negatives; μ is the type of samples.

The receiver operating characteristic (ROC) curve [111] shows the predictive capability. The ROC curve can also present the model behavior of the true positive rate (sensitivity) against the false positive rate (1-specificity) in a visual way. The area under the ROC (auROC) is calculated to quantitatively and objectively measure the performance of the proposed method. A perfect classifier gives auROC = 1 and the random performance gives auROC = 0.5.

6.2. Published Results

Based on the benchmark dataset S1, by using Multi-class SVMs combined with PseAAC, Mondal et al. [24] achieved an overall accuracy of 88.1%, which was higher than those obtained by other methods like BLAST and ISort (Intimate Sorting) predictor. In order to improve the accuracy, Lin et al. [32] proposed a new algorithm that combined increment of diversity with modified Mahalanobis discriminant. The algorithm can reduce the dimension of inputting vector, extract important classify information, and improve the calculation efficiency. The average sensitivity and specificity respectively reached 88% and 91% in the jackknife cross-validation test. Zaki et al. developed a scoring system called SVM-Free score based on local alignment partition functions [34] and increased the average sensitivity and specificity to 97.42% and 99.17%, respectively. Based on SVM-Free score method, a soft package was constructed and could be freely downloaded from http://faculty.uaeu.ac.ae/nzaki/SVM-Freescore.htm. Based on this work, a novel method called Toxin-AAM [35] was introduced with evolutionary information and amino acid composition and the average sensitivity reached 94.5% in jackknife cross-validation test.

Based on the benchmark dataset S2, Fan et al. [33] proposed a novel method called *PredCSF* for predicting the conotoxin superfamily by using modified one-versus-rest SVMs. *PredCSF* can realize an overall accuracy of 90.65% in jackknife cross-validation tests. A user-friendly webserver was established and could be freely accessible at http://www.csbio.sjtu.edu.cn/bioinf/PredCSF/. Yin et al. [36] proposed an improved HKNN version called dHKNN algorithm for predicting conotoxin superfamily by considering the local density information in the diffusion space. The overall accuracy of 91.90% was obtained by the jackknife cross-validation test on the benchmark dataset S2. The results indicated that the proposed dHKNN was more promising.

Based on the dataset of ConoServer, Laht et al. [37] acquired 964 sequences and built 62 profile Hidden Markov Models (pHMMs) for the classification of all the described conopeptide superfamilies and families based on the primary sequences. As a result, the mature peptide models realized an accuracy of 96% and the propeptide and signal peptide models got an accuracy of 100%. Koua et al. [38] constructed pHMMs for each of the 48 alignments using the HMMbuild script from the HMMER 3.0 package (version, Manufacturer, City, US State abbrev. if applicable, Country) [112] based on 933 samples of conotoxin superfamily. The model obtained promising discriminative abilities with the sensitivity of ~95% and specificity of ~99%. Based on the model, they published the webserver *ConoDoctor* to predict the conotoxin superfamily, and the package could be freely downloaded from http://conco.ebc.ee. For further improving the accuracy, they established 50 position-specific scoring matrices and 47 hidden Markov models based on 967 sequences from ConoServer [39] and realized the sensitivity of 99.42% and specificity of 92.81%, respectively. Although the accuracies of these models are high, the benchmark datasets used in these models are not objective. Many redundant sequences are included in these datasets. Moreover, the some samples lack biochemical experimental proofs.

Based on the benchmark dataset I1, Yuan et al. [41] predicted the types of ion channel-targeted conotoxins by using binomial distribution and radial basis function network, and achieved an average accuracy of 89.7% and overall accuracy of 85.7% in the prediction of three types of ion channel-targeted conotoxins in the jackknife cross-validation test. The model provides the valuable instructions for theoretical and experimental studies on conotoxins. For further improving the accuracy, Ding et al. [42] used the SVM to classify three kinds of samples based on the feature selection technique *F*-score. The average sensitivity and the overall accuracy respectively reached 90.3% and 91.1%, which are higher than those of the RBF network-based method [41]. For the convenience of the vast majority

of experimental scientists, they provided the webserver *iCTX-Type*, and user guide details could be obtained from http://lin.uestc.edu.cn/server/iCTX-Type. By incorporating new properties of residues into pseudo amino acid composition, Wu et al. [44] achieved a higher overall accuracy of 94.6%. Recently, an overall accuracy of 91.98% with an average accuracy of 92.17% was obtained by the AVC-SVM model proposed by Wang et al. [45].

Based on the benchmark dataset I2, Zhang et al. [43] proposed a random forest based predictor called ICTCPred for the prediction of the types of ion channel-targeted conotoxins and yielded the satisfactory performance with an average accuracy of 91.0%.

The detailed results obtained by these theoretical methods are provided in Table 2. The published webservers are listed in Table 3. Although many methods have been proposed to predict superfamily types and ion channel-target types of conotoxins, only a few tools were established based on the proposed methods. PsedCSF provides not only a free webserver, but also a stand-alone soft package. The establishment of ConoDictor is based on the cone snail genome project for health. The project website also provides a database called ConoDB, which collects the peptide sequences from cone snails stored in NCBI and Uniprot. The iCTX-Type is the only webserver for the prediction of the types of ion channel-target conotoxins.

Table 2. A list of published results for conotoxin superfamilies and ion channel-targeted conotoxin classifications.

Superfamily Prediction								Reference
Dataset	**Methods**	**A**	**M**	**O**	**T**	**AA**	**OA**	
S1	Multi-class SVMs	0.840	0.923	0.869	0.941	0.893	0.881	[24]
	IDQD	0.960	0.923	0.820	0.940	0.911	0.883	[32]
	SVM-Freescore	0.960	0.984	0.984	1	0.982	0.974	[34]
	Toxin-AAM	0.957	0.966	0.891	0.966	0.945	0.966	[35]
S2	PredCFS	0.960	0.984	0.984	1	0.982	0.903	[33]
	dHKNN	0.957	0.966	0.891	0.966	0.945	0.919	[36]
Type of Ion Channel-Targeted Prediction								**Reference**
Dataset	**Methods**	**K-Conotoxin**	**Na-Conotoxin**		**Ca-Conotoxin**	**AA**	**OA**	
I1	RBF network	0.917	0.884		0.889	0.897	0.893	[41]
	iCTX-Type	0.833	0.978		0.898	0.903	0.911	[42]
	Fscore-SVM	0.917	0.953		0.953	0.942	0.946	[44]
	AVC-SVM	0.931	0.942		0.892	0.922	0.920	[45]
I2	ICTCPred	1	0.919		1	0.973	0.957	[43]

Table 3. A list of the published prediction tools for conotoxin classification.

Name	Prediction Type	URL	Reference
PredCSF	Superfamily	http://www.csbio.sjtu.edu.cn/bioinf/PredCSF/	[33]
ConoDictor	Superfamily	http://conco.ebc.ee	[38]
iCTX-Type	ion channel-target	http://lin.uestc.edu.cn/server/iCTX-Type	[42]

7. Conclusions

Conotoxins have a wide application prospect in the fields of neuroscience development and neuroscience research and play different physiological functions and therapeutic potentials. Accurate identification of conotoxin types will provide vital clues in revealing the physiological mechanism and pharmacological therapeutic of conotoxins. It is necessary to develop computational tools for both basic research and drug development, particularly for in-depth investigation into the mechanisms of conotoxins and the development of new drugs to treat chronic pain, epilepsy, spasticity, and cardiovascular diseases.

Similarly, the computational-based methods can be also applied to other disulfide-rich venom peptides that target the same ion channels and receptors as conotoxins and show similar pharmacological, biochemical and structural properties, such as spider venoms, centipede or snake venoms.

Although encouraging results have been obtained in conotoxin superfamily and ion channel-target type prediction, further improvements should be made. At first, the prediction accuracy should be further improved. Different types of PseAAC can be applied in the field to formulate conotoxin samples for improving the accuracy. A PSSM (position-specific scoring matrix) produced by similarity search can also be used as an important feature in prediction. Moreover, different feature selection techniques, such as minimal-redundancy-maximal-relevance (mRMR) and principal component analysis (PCA), can also be used to reduce the feature dimension and extract key features. Furthermore, with machine learning approaches, more methods such as deep learning, deep forest, and random forest can also be used to obtain higher accuracies. In addition to superfamily and target type prediction, the signal peptide cleavage sites, the position of two disulfide bonds, and the transition from *L* to *D*-residues are also required to be computationally identified by using machine learning methods. We hope that more and more scholars devote themselves to this field.

Acknowledgments: This work was supported by the Applied Basic Research Program of Sichuan Province (Nos. 2015JY0100 and LZ-LY-45), the Fundamental Research Funds for the Central Universities of China (Nos. ZYGX2015J144; ZYGX2015Z006; ZYGX2016J118; ZYGX2016J125; ZYGX2016J126), the Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028), the Outstanding Youth Foundation of North China University of Science and Technology (No. JP201502), the China Postdoctoral Science Foundation (No.2015M582533) and the Scientific Research Foundation of the Education Department of Sichuan Province (11ZB122).

Author Contributions: H.L., W.C. and H.T. conceived and designed the experiments; F.Y.D. performed the experiments; F.Y.D, H.Y. andZ.D.S. analyzed the data; W.Y., Y.W. and H.D. contributed reagents/materials/analysis tools; F.Y.D., H.Y., Z.D.S., W.C. and H.L. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kohn, A.J. The ecology of *Conus* in hawaii. *Ecol. Monogr.* **1959**, *29*, 47–90. [CrossRef]
2. Daly, N.L.; Craik, D.J. Structural studies of conotoxins. *IUBMB Life* **2009**, *61*, 144–150. [CrossRef] [PubMed]
3. Adams, D.J. Conotoxins and their potential pharmaceutical applications. *Drug Dev.* **1999**, *46*, 219–234. [CrossRef]
4. Terlau, H.; Olivera, B.M. *Conus* venoms: A rich source of novel ion channel-targeted peptides. *Phys. Rev.* **2004**, *84*, 41–68. [CrossRef] [PubMed]
5. Craik, D.J.; Adams, D.J. Chemical modification of conotoxins to improve stability and activity. *ACS Chem. Biol.* **2007**, *2*, 457–468. [CrossRef] [PubMed]
6. Livett, B.G.; Gayler, K.R.; Khalil, Z. Drugs from the sea: Conopeptides as potential therapeutics. *Curr. Med. Chem.* **2004**, *11*, 1715–1723. [CrossRef] [PubMed]
7. Aguilar, M.B.; Lopez-Vera, E.; Heimer de la Cotera, E.P.; Falcon, A.; Olivera, B.M.; Maillo, M. I-conotoxins in vermivorous species of the west atlantic: Peptide sr11a from *Conus* spurius. *Peptides* **2007**, *28*, 18–23. [CrossRef] [PubMed]
8. Vincler, M.; McIntosh, J.M. Targeting the alpha9alpha10 nicotinic acetylcholine receptor to treat severe pain. *Expert Opin. Ther. Targets* **2007**, *11*, 891–897. [CrossRef] [PubMed]
9. Twede, V.D.; Miljanich, G.; Olivera, B.M.; Bulaj, G. Neuroprotective and cardioprotective conopeptides: An emerging class of drug leads. *Curr. Opin. Drug Discov. Dev.* **2009**, *12*, 231–239.
10. Wang, Y.X.; Pettus, M.; Gao, D.; Phillips, C.; Scott Bowersox, S. Effects of intrathecal administration of ziconotide, a selective neuronal n-type calcium channel blocker, on mechanical allodynia and heat hyperalgesia in a rat model of postoperative pain. *Pain* **2000**, *84*, 151–158. [CrossRef]
11. Feng, W.H.; Zan, J.B.; Zhu, Y.P. Advances in study of structures and functions of conantokins. *Zhejiang Da Xue Xue Bao Yi Xue Ban J. Zhejiang Univ. Med. Sci.* **2007**, *36*, 204–208.

12. Olivera, B.M.; Teichert, R.W. Diversity of the neurotoxic *Conus* peptides: A model for concerted pharmacological discovery. *Mol. Interv.* **2007**, *7*, 251–260. [CrossRef] [PubMed]
13. Miljanich, G.P. Ziconotide: Neuronal calcium channel blocker for treating severe chronic pain. *Curr. Med. Chem.* **2004**, *11*, 3029–3040. [CrossRef] [PubMed]
14. Barton, M.E.; White, H.S.; Wilcox, K.S. The effect of cgx-1007 and ci-1041, novel nmda receptor antagonists, on nmda receptor-mediated epscs. *Epilepsy Res.* **2004**, *59*, 13–24. [CrossRef] [PubMed]
15. Han, T.S.; Teichert, R.W.; Olivera, B.M.; Bulaj, G. *Conus* venoms—A rich source of peptide-based therapeutics. *Curr. Pharm. Des.* **2008**, *14*, 2462–2479. [CrossRef] [PubMed]
16. Pallaghy, P.K.; Alewood, D.; Alewood, P.F.; Norton, R.S. Solution structure of robustoxin, the lethal neurotoxin from the funnel-web spider atrax robustus. *FEBS Lett.* **1997**, *419*, 191–196. [CrossRef]
17. Savarin, P.; Guenneugues, M.; Gilquin, B.; Lamthanh, H.; Gasparini, S.; Zinn-Justin, S.; Menez, A. Three-dimensional structure of kappa-conotoxin pviia, a novel potassium channel-blocking toxin from cone snails. *Biochemistry* **1998**, *37*, 5407–5416. [CrossRef] [PubMed]
18. Botana, L.M. Seafood and freshwater toxins. *Phytochemistry* **2000**, *60*, 549–550.
19. Kaas, Q.; Westermann, J.C.; Craik, D.J. Conopeptide characterization and classifications: An analysis using conoserver. *Toxicon Off. J. Int. Soc. Toxinol.* **2010**, *55*, 1491–1509. [CrossRef] [PubMed]
20. Jones, R.M.; Bulaj, G. Conotoxins—New vistas for peptide therapeutics. *Curr. Pharm. Des.* **2000**, *6*, 1249–1285. [CrossRef] [PubMed]
21. Mouhat, S.; Jouirou, B.; Mosbah, A.; De Waard, M.; Sabatier, J.M. Diversity of folds in animal toxins acting on ion channels. *Biochem. J.* **2004**, *378*, 717–726. [CrossRef] [PubMed]
22. McIntosh, J.M.; Jones, R.M. Cone venom—From accidental stings to deliberate injection. *Toxicon Off. J. Int. Soc. Toxinol.* **2001**, *39*, 1447–1451. [CrossRef]
23. Rajendra, W.; Armugam, A.; Jeyaseelan, K. Toxins in anti-nociception and anti-inflammation. *Toxicon Off. J. Int. Soc. Toxinol.* **2004**, *44*, 1–17. [CrossRef] [PubMed]
24. Mondal, S.; Bhavna, R.; Mohan Babu, R.; Ramakumar, S. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. *J. Theor. Biol.* **2006**, *243*, 252–260. [CrossRef] [PubMed]
25. Akondi, K.B.; Muttenthaler, M.; Dutertre, S.; Kaas, Q.; Craik, D.J.; Lewis, R.J.; Alewood, P.F. Discovery, synthesis, and structure-activity relationships of conotoxins. *Chem. Rev.* **2014**, *114*, 5815–5847. [CrossRef] [PubMed]
26. Jacob, R.B.; McDougal, O.M. The m-superfamily of conotoxins: A review. *Cell. Mol. Life Sci. CMLS* **2010**, *67*, 17–27. [CrossRef] [PubMed]
27. Corpuz, G.P.; Jacobsen, R.B.; Jimenez, E.C.; Watkins, M.; Walker, C.; Colledge, C.; Garrett, J.E.; McDougal, O.; Li, W.; Gray, W.R.; et al. Definition of the m-conotoxin superfamily: Characterization of novel peptides from molluscivorous *Conus* venoms. *Biochemistry* **2005**, *44*, 8176–8186. [CrossRef] [PubMed]
28. Baldomero, M.; Olivera, B.M. *Conus* venom peptides, receptor and ion channel targets, and drug design: 50 million years of neuropharmacology. *Mol. Biol. Cell* **1997**, *8*, 2101–2109.
29. Lewis, R.J. Conotoxins as selective inhibitors of neuronal ion channels, receptors and transporters. *IUBMB Life* **2004**, *56*, 89–93. [CrossRef] [PubMed]
30. Yu, R.; Craik, D.J.; Kaas, Q. Blockade of neuronal alpha7-nachr by alpha-conotoxin imi explained by computational scanning and energy calculations. *PLoS Comput. Biol.* **2011**, *7*, e1002011. [CrossRef] [PubMed]
31. Patel, D.; Mahdavi, S.; Kuyucak, S. Computational study of binding of mu-conotoxin giiia to bacterial sodium channels navab and navrh. *Biochemistry* **2016**, *55*, 1929–1938. [CrossRef] [PubMed]
32. 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*, 548–551. [CrossRef] [PubMed]
33. Fan, Y.X.; Song, J.; Shen, H.B.; Kong, X. Predcsf: An integrated feature-based approach for predicting conotoxin superfamily. *Protein Pept. Lett.* **2011**, *18*, 261–267. [CrossRef] [PubMed]
34. Zaki, N.; Wolfsheimer, S.; Nuel, G.; Khuri, S. Conotoxin protein classification using free scores of words and support vector machines. *BMC Bioinform.* **2011**, *12*, 217. [CrossRef] [PubMed]
35. Nazar Zaki, F.S. Conotoxin protein classification using pairwise comparison and amino acid composition. In Proceedings of the Genetic & Evolutionary Computation Conference, Dublin, Ireand, 2011; Volume 5540, pp. 323–330.

36. Yin, J.B.; Fan, Y.X.; Shen, H.B. Conotoxin superfamily prediction using diffusion maps dimensionality reduction and subspace classifier. *Curr. Protein Pept. Sci.* **2011**, *12*, 580–588. [CrossRef] [PubMed]
37. Laht, S.; Koua, D.; Kaplinski, L.; Lisacek, F.; Stocklin, R.; Remm, M. Identification and classification of conopeptides using profile hidden markov models. *Biochim. Biophys. Acta* **2012**, *1824*, 488–492. [CrossRef] [PubMed]
38. Koua, D.; Brauer, A.; Laht, S.; Kaplinski, L.; Favreau, P.; Remm, M.; Lisacek, F.; Stocklin, R. Conodictor: A tool for prediction of conopeptide superfamilies. *Nucleic Acids Res.* **2012**, *40*, W238–W241. [CrossRef] [PubMed]
39. Koua, D.; Laht, S.; Kaplinski, L.; Stocklin, R.; Remm, M.; Favreau, P.; Lisacek, F. Position-specific scoring matrix and hidden markov model complement each other for the prediction of conopeptide superfamilies. *Biochim. Biophys. Acta* **2013**, *1834*, 717–724. [CrossRef] [PubMed]
40. Gowd, K.H.; Dewan, K.K.; Iengar, P.; Krishnan, K.S.; Balaram, P. Probing peptide libraries from *Conus* achatinus using mass spectrometry and cdna sequencing: Identification of delta and omega-conotoxins. *J. Mass Spectrom. JMS* **2008**, *43*, 791–805. [CrossRef] [PubMed]
41. Yuan, L.F.; Ding, C.; Guo, S.H.; Ding, H.; Chen, W.; Lin, H. Prediction of the types of ion channel-targeted conotoxins based on radial basis function network. *Toxicol. Int. J. Publ. Assoc. BIBRA.* **2013**, *27*, 852–856. [CrossRef] [PubMed]
42. Ding, H.; Deng, E.Z.; Yuan, L.F.; Liu, L.; Lin, H.; Chen, W.; Chou, K.C. Ictx-type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *BioMed Res. Int.* **2014**, *2014*, 286419. [CrossRef] [PubMed]
43. Zhang, L.; Zhang, C.; Gao, R.; Yang, R.; Song, Q. Using the smote technique and hybrid features to predict the types of ion channel-targeted conotoxins. *J. Theor. Biol.* **2016**, *403*, 75–84. [CrossRef] [PubMed]
44. Wu, Y.; Zheng, Y.; Tang, H. Identifying the types of ion channel-targeted conotoxins by incorporating new properties of residues into pseudo amino acid composition. *BioMed Res. Int.* **2016**, *2016*, 3981478. [CrossRef] [PubMed]
45. Wang, X.; Wang, J.; Wang, X.; Zhang, Y. Predicting the types of ion channel-targeted conotoxins based on avc-svm model. *BioMed Res. Int.* **2017**, *2017*, 2929807.
46. He, B.; Chai, G.; Duan, Y.; Yan, Z.; Qiu, L.; Zhang, H.; Liu, Z.; He, Q.; Han, K.; Ru, B.; et al. Biopanning data bank. *Nucleic Acids Res.* **2016**, *44*, D1127–D1132. [CrossRef] [PubMed]
47. Ru, B.; Huang, J.; Dai, P.; Li, S.; Xia, Z.; Ding, H.; Lin, H.; Guo, F.; Wang, X. Mimodb: A new repository for mimotope data derived from phage display technology. *Molecules* **2010**, *15*, 8279–8288. [CrossRef] [PubMed]
48. Huang, J.; Ru, B.; Zhu, P.; Nie, F.; Yang, J.; Wang, X.; Dai, P.; Lin, H.; Guo, F.B.; Rao, N. Mimodb 2.0: A mimotope database and beyond. *Nucleic Acids Res.* **2012**, *40*, D271–D277. [CrossRef] [PubMed]
49. Liang, Z.Y.; Lai, H.Y.; Yang, H.; Zhang, C.J.; Yang, H.; Wei, H.H.; Chen, X.X.; Zhao, Y.W.; Su, Z.D.; Li, W.C.; et al. Pro54db: A database for experimentally verified sigma-54 promoters. *Bioinformatics* **2017**, *33*, 467–469. [CrossRef] [PubMed]
50. The UniProt, Consortium. Uniprot: The universal protein knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169.
51. Rose, P.W.; Prlic, A.; Altunkaya, A.; Bi, C.; Bradley, A.R.; Christie, C.H.; Costanzo, L.D.; Duarte, J.M.; Dutta, S.; Feng, Z.; et al. The rcsb protein data bank: Integrative view of protein, gene and 3d structural information. *Nucleic Acids Res.* **2017**, *45*, D271–D281. [PubMed]
52. Coordinators, N.R. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **2017**, *45*, D12–D17.
53. Kaas, Q.; Yu, R.; Jin, A.H.; Dutertre, S.; Craik, D.J. Conoserver: Updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Res.* **2012**, *40*, D325–D330. [CrossRef] [PubMed]
54. Kaas, Q.; Westermann, J.C.; Halai, R.; Wang, C.K.; Craik, D.J. Conoserver, a database for conopeptide sequences and structures. *Bioinformatics* **2008**, *24*, 445–446. [CrossRef] [PubMed]
55. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef] [PubMed]
56. Yan, K.; Xu, Y.; Fang, X.; Zheng, C.; Liu, B. Protein fold recognition based on sparse representation based classification. *Artif. Intell. Med.* **2017**. [CrossRef] [PubMed]
57. Tang, H.; Chen, W.; Lin, H. Identification of immunoglobulins using chou's pseudo amino acid composition with feature selection technique. *Mol. Biosyst.* **2016**, *12*, 1269–1275. [CrossRef] [PubMed]

58. Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L.; Chou, K.C. Pse-in-one: A web server for generating various modes of pseudo components of DNA, rna, and protein sequences. *Nucleic Acids Res.* **2015**, *43*, W65–W71. [CrossRef] [PubMed]
59. Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* **2001**, *43*, 246–255. [CrossRef] [PubMed]
60. Mathura, V.S.; Kolippakkam, D. Apdbase: Amino acid physico-chemical properties database. *Bioinformation* **2005**, *1*, 2–4. [CrossRef] [PubMed]
61. Leise, T.L. Wavelet-based analysis of circadian behavioral rhythms. *Methods Enzymol.* **2015**, *551*, 95–119. [PubMed]
62. Ding, C.; Yuan, L.F.; Guo, S.H.; Lin, H.; Chen, W. Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions. *J. Proteom.* **2012**, *77*, 321–328. [CrossRef] [PubMed]
63. Yong, S.K.; Street, W.N.; Menczer, F. Feature selection in data mining. *Data Min Oppor. Chall.* **2003**, *9*, 80–105.
64. Rocchi, L.; Chiari, L.; Cappello, A. Feature selection of stabilometric parameters based on principal component analysis. *Med. Biol. Eng. Comput.* **2004**, *42*, 71–79. [CrossRef] [PubMed]
65. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
66. Zou, Q.; Zeng, J.; Cao, L.; Ji, R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* **2016**, *173*, 346–354. [CrossRef]
67. Lin, H.; Ding, H. Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. *J. Theor. Biol.* **2011**, *269*, 64–69. [CrossRef] [PubMed]
68. Kira, K.; Rendell, L.A. He feature selection problem: Traditional methods and a new algorithm. In Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA, USA, 12–16 July 1992; pp. 129–134.
69. Sun, Y. Iterative relief for feature weighting: Algorithms, theories, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1035–1051. [CrossRef] [PubMed]
70. Lafon, S.; Lee, A.B. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1393–1403. [CrossRef] [PubMed]
71. Zou, Q.; Xie, S.; Lin, Z.; Wu, M.; Ju, Y. Finding the best classification threshold in imbalanced classification. *Big Data Res.* **2016**, *5*, 2–8. [CrossRef]
72. Ding, H.; Li, D. Identification of mitochondrial proteins of malaria parasite using analysis of variance. *Amino Acids* **2015**, *47*, 329–333. [CrossRef] [PubMed]
73. Tang, H.; Zou, P.; Zhang, C.; Chen, R.; Chen, W.; Lin, H. Identification of apolipoprotein using feature selection technique. *Sci. Rep.* **2016**, *6*, 30441. [CrossRef] [PubMed]
74. Chen, X.X.; Tang, H.; Li, W.C.; Wu, H.; Chen, W.; Ding, H.; Lin, H. Identification of bacterial cell wall lyases via pseudo amino acid composition. *BioMed Res. Int.* **2016**, *2016*, 1654623. [CrossRef] [PubMed]
75. Yang, H.; Tang, H.; Chen, X.X.; Zhang, C.J.; Zhu, P.P.; Ding, H.; Chen, W.; Lin, H. Identification of secretory proteins in mycobacterium tuberculosis using pseudo amino acid composition. *BioMed Res. Int.* **2016**, *2016*, 5413903. [CrossRef] [PubMed]
76. Wu, Y.; Tang, H.; Chen, W.; Lin, H. Predicting human enzyme family classes by using pseudo amino acid composition. *Curr. Proteom.* **2016**, *13*, 99–104. [CrossRef]
77. Zhao, Y.W.; Lai, H.Y.; Tang, H.; Chen, W.; Lin, H. Prediction of phosphothreonine sites in human proteins by fusing different features. *Sci. Rep.* **2016**, *6*, 34817. [CrossRef] [PubMed]
78. Vapnik, V.N.; Vapnik, V. *Statistical Learning Theory*; John Wiley and Sons Inc.: New York, NY, USA, 1998.
79. Liu, B.; Zhang, D.; Xu, R.; Xu, J.; Wang, X.; Chen, Q.; Dong, Q.; Chou, K.C. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* **2014**, *30*, 472–479. [CrossRef] [PubMed]
80. Lin, H.; Liang, Z.Y.; Tang, H.; Chen, W. Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**. [CrossRef] [PubMed]
81. Chen, W.; Tang, H.; Ye, J.; Lin, H.; Chou, K.C. IRNA-pseu: Identifying rna pseudouridine sites. *Mol. Ther. Nucleic Acids* **2016**, *5*, e332. [PubMed]

82. Lai, H.Y.; Chen, X.X.; Chen, W.; Tang, H.; Lin, H. Sequence-based predictive modeling to identify cancerlectins. *Oncotarget* **2017**, *8*, 28169–28175. [CrossRef] [PubMed]
83. Chen, W.; Tang, H.; Lin, H. Methyrna: A web server for identification of n6-methyladenosine sites. *J. Biomol. Struct. Dyn.* **2017**, *35*, 683–687. [CrossRef] [PubMed]
84. He, B.; Kang, J.; Ru, B.; Ding, H.; Zhou, P.; Huang, J. Sabinder: A web service for predicting streptavidin-binding peptides. *BioMed Res. Int.* **2016**, *2016*, 9175143. [CrossRef] [PubMed]
85. Tang, Q.; Nie, F.; Kang, J.; Ding, H.; Zhou, P.; Huang, J. Nieluter: Predicting peptides eluted from hla class i molecules. *J. Immunol. Methods* **2015**, *422*, 22–27. [CrossRef] [PubMed]
86. Ru, B.; Pa, T.H.; Nie, F.; Lin, H.; Guo, F.B.; Huang, J. Phd7faster: Predicting clones propagating faster from the ph.D.-7 phage display peptide library. *J. Bioinform. Comput. Biol.* **2014**, *12*, 1450005. [CrossRef] [PubMed]
87. Liu, B.; Fang, L.; Liu, F.; Wang, X.; Chen, J.; Chou, K.C. Identification of real microrna precursors with a pseudo structure status composition approach. *PLoS ONE* **2015**, *10*, e0121501. [CrossRef] [PubMed]
88. Li, D.; Ju, Y.; Zou, Q. Protein folds prediction with hierarchical structured svm. *Curr. Proteom.* **2016**, *13*, 79–85. [CrossRef]
89. Chang, C.C.; Hsu, C.W.; Lin, C.J. The analysis of decomposition methods for support vector machines. *IEEE Trans. Neural Netw.* **2000**, *11*, 1003–1008. [CrossRef] [PubMed]
90. Pedrycz, W. Advances in kernel methods: Support vector learning. *Neurocomputing* **2002**, *47*, 303–304. [CrossRef]
91. Eddy, S.R. Profile hidden markov models. *Bioinformatics* **1998**, *14*, 755–763. [CrossRef] [PubMed]
92. Eddy, S.R. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput. Biol.* **2008**, *4*, e1000069. [CrossRef] [PubMed]
93. Wheeler, T.J.; Eddy, S.R. Nhmmer: DNA homology search with profile hmms. *Bioinformatics* **2013**, *29*, 2487–2489. [CrossRef] [PubMed]
94. Chai, G.; Yu, M.; Jiang, L.; Duan, Y.; Huang, J. Hmmcas: A web tool for the identification and domain annotations of cas proteins. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**. [CrossRef] [PubMed]
95. Boudaren, E.M.Y.; Pieczynski, W. Dempster-shafer fusion of multisensor signals in nonstationary markovian context. *EURASIP J. Adv. Signal Process.* **2012**, *2012*, 134. [CrossRef]
96. Boudaren, M.E.; Monfrini, E.; Pieczynski, W. Unsupervised segmentation of random discrete data hidden with switching noise distributions. *IEEE Signal Process. Lett.* **2012**, *19*, 619–622. [CrossRef]
97. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
98. Vincent, P.; Bengio, Y. K-local hyperplane and convex distance nearest neighbor algorithms. *Adv. Neural Inf. Process. Syst.* **2002**, *14*, 985–992.
99. Mahalanobis, P.C. On the generalised distance in statistics. *Proc. Natl. Inst. Sci. India* **1936**, *2*, 49–55.
100. Lin, H. The modified mahalanobis discriminant for predicting outer membrane proteins by using chou's pseudo amino acid composition. *J. Theor. Biol.* **2008**, *252*, 350–356. [CrossRef] [PubMed]
101. Feng, Y.; Luo, L. Use of tetrapeptide signals for protein secondary-structure prediction. *Amino Acids* **2008**, *35*, 607–614. [CrossRef] [PubMed]
102. Chen, S.A.; Ou, Y.Y.; Lee, T.Y.; Gromiha, M.M. Prediction of transporter targets using efficient rbf networks with pssm profiles and biochemical properties. *Bioinformatics* **2011**, *27*, 2062–2067. [CrossRef] [PubMed]
103. Jiang, L.; Zhang, J.; Xuan, P.; Zou, Q. Bp neural network could help improve pre-mirna identification in various species. *BioMed Res. Int.* **2016**, *2016*, 9565689. [CrossRef] [PubMed]
104. Witten, I.H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*; MorganKaufmann: SanFrancisco, CA, USA, 2005.
105. Zhang, C.J.; Tang, H.; Li, W.C.; Lin, H.; Chen, W.; Chou, K.C. Iori-human: Identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget* **2016**, *7*, 69783–69793. [CrossRef] [PubMed]
106. Liao, Z.; Ju, Y.; Zou, Q. Prediction of g protein-coupled receptors with svm-prot features and random forest. *Scientifica* **2016**, *2016*, 8309253. [CrossRef] [PubMed]
107. Zhao, X.; Zou, Q.; Liu, B.; Liu, X. Exploratory predicting protein folding model with random forest and hybrid features. *Curr. Proteom.* **2014**, *11*, 289–299. [CrossRef]

108. Liu, B.; Long, R.; Chou, K.C. Idhs-el: Identifying dnase i hypersensitive-sites by fusing three different modes of pseu-do nucleotide composition into an ensemble learning framework. *Bioinformatics* **2016**, *32*, 2411–2418. [CrossRef] [PubMed]
109. Chou, K.C.; Zhang, C.T. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275–349. [CrossRef] [PubMed]
110. Liu, B.; Fang, L.; Liu, F.; Wang, X.; Chou, K.C. Imirna-psedpc: Microrna precursor identification with a pseudo distance-pair composition approach. *J. Biomol. Struct. Dyn.* **2016**, *34*, 223–235. [CrossRef] [PubMed]
111. Metz, C.E. Some practical issues of experimental design and data analysis in radiological roc studies. *Investig. Radiol.* **1989**, *24*, 234–245. [CrossRef]
112. Johnson, L.S.; Eddy, S.R.; Portugaly, E. Hidden markov model speed heuristic and iterative hmm search procedure. *BMC Bioinform.* **2010**, *11*, 431. [CrossRef] [PubMed]