PLOS ONE

RESEARCH ARTICLE

iDPF-PseRAAAC: A Web-Server for Identifying the Defensin Peptide Family and Subfamily Using Pseudo Reduced Amino Acid Alphabet Composition

Yongchun Zuo[1]*, Yang Lv[1], Zhuying Wei[1], Lei Yang[2], Guangpeng Li[1], Guoliang Fan[3]*

1 The Key Laboratory of Mammalian Reproductive Biology and Biotechnology of the Ministry of Education, College of life sciences, Inner Mongolia University, Hohhot, China, **2** College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, **3** Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot, China

* yczuo@imu.edu.cn (Y-CZ); eeguoliangfan@sina.com (G-LF)

Citation: Zuo Y, Lv Y, Wei Z, Yang L, Li G, Fan G (2015) iDPF-PseRAAAC: A Web-Server for Identifying the Defensin Peptide Family and Subfamily Using Pseudo Reduced Amino Acid Alphabet Composition. PLoS ONE 10(12): e0145541. doi:10.1371/journal.pone.0145541

Editor: Junwen Wang, The University of Hong Kong, HONG KONG

Received: September 17, 2015

Accepted: December 4, 2015

Published: December 29, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by The National Nature Scientific Foundation of China (No:11447201, 61561036), the Specialized Research Fund for the Doctoral Program of Higher Education (20131501120009) and the Natural Science Foundation of Inner Mongolia Autonomous Region (No:2013MS0503).

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Defensins as one of the most abundant classes of antimicrobial peptides are an essential part of the innate immunity that has evolved in most living organisms from lower organisms to humans. To identify specific defensins as interesting antifungal leads, in this study, we constructed a more rigorous benchmark dataset and the **iDPF-PseRAAAC** server was developed to predict the defensin family and subfamily. Using reduced dipeptide compositions were used, the overall accuracy of proposed method increased to 95.10% for the defensin family, and 98.39% for the vertebrate subfamily, which is higher than the accuracy from other methods. The jackknife test shows that more than 4% improvement was obtained comparing with the previous method. A free online server was further established for the convenience of most experimental scientists at http://wlxy.imu.edu.cn/college/biostation/fuwu/iDPF-PseRAAAC/index.asp. A friendly guide is provided to describe how to use the web server. We anticipate that **iDPF-PseRAAAC** may become a useful high-throughput tool for both basic research and drug design.

Introduction

Defensins are important small, basic, cysteine-rich, antimicrobial, and cationic peptides that are abundant and widely distributed [1, 2]. Defensins are widely distributed in multiple tissues in the body, most notably leukocytes and epithelial surfaces. They are often present at high concentrations[3, 4] and play an essential role in the innate immunity of their hosts from insects and plants to amphibians and mammals[5]. Membrane permeabilization is the crucial step in defensin-mediated antimicrobial activity and cytotoxicity[3, 6]. Defensins from different origins exhibit structural and functional similarities with phylogenetic relationships between different types of defensins. Mature defensins amino acids sequences are highly

variable in each defensin family and subfamily [7–9]. Accurately identifying the types of defensins will be helpful in analyzing their specificitis for various microbial targets, provide novel insights for understanding their function, and facilitate antimicrobial drugs targets discovery.

Biochemical experimental methods are highly reliable for elucidating types of defensins, such as nuclear magnetic resonance (NMR) spectroscopy [10]. However, such the experimental techniques are time-consuming and expensive. Bioinformatics methods can timely provide useful information and insights for both basic research and antibiotics design[11, 12]. Thus, better understanding the distinct functions of defensin proteins requires an automated method for timely and reliably annotating the families of many defensin proteins. In our previous study, four defensin families (vertebrate, plant, insect and other defensins) were successfully classified using the increment of diversity (ID) method [13]. In another work the authors developed the DEFENSINPRED classifier to predict human defensin proteins and their types based on pseudo amino acid compositions[14].

However, further work is necessary because the datasets constructed in those methods were too small to reflect a statistical profile and did not impose a rigorous cutoff threshold[15, 16] to exclude the redundant samples in the existing defensin datasets. Moreover, a better web-server for defensins is also needed. In the present work, we constructed a more rigorous benchmark dataset to train the program, and a support vector machine (SVM) classifier was further proposed to classify these five defensin families. An 4% improvement was obtained compared with the previous method. For the convenience of experimental scientists, a free online server **iDPF-PseRAAAC** was first established. A friendly guide was further provided to describe how to use the web server.

Materials and Methods

Dataset

With rapidly increasing interest in defensins, the Defensins Knowledgebase is available, which is a manually curated database and information source devoted to the defensin family of antimicrobial peptides [17]. The benchmark data set $\mathbb{S}$ for the defensin proteins in this study was taken from the Defensins Knowledgebase, which currently contains more than 500 defensin sequences ranging from prokaryotes to eukaryotes. To prepare a high-quality dataset, the program CD-HIT[18] was used to remove the defensin proteins with $\geq$ 80% pairwise sequence identity to any other protein. Highly similar data will surely lead to overestimation of the performance of the proposed methods. If the sequence identity cutoff is set to a lower percentage (such as 25%), the results will be more objective and reliable. However, in this study we did not use such a stringent criterion because the currently available data do not allow us to do so. The proposed method is a sequence-dependent predictor, the input feature vectors are only derived from the primary amino acids sequence. So there is needed enough amino acids and dipeptide compositions to train the multi-classifier module. For the defensin peptides are polypeptides of fewer than 100 amino acids (See Fig 1) [3]. Besides, for ensure the data reliable, this dataset is a manually curated database. All of the families annotation are gathered from bibliographic databases and sequence databases literature sources. If the sequence identity cutoff is set to a lower percentage (such as 25%), the numbers of proteins for family subsets would have been too few to have statistical significance. And the imbalanced data cause classifiers to tend to overfit and to perform poorly in particular on the minority class[19, 20]. Finally, we obtained a dataset $\mathbb{S}$ composed of 333 defensin proteins classified into five families, as formulated by the following equation:

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_3 \cup \mathbb{S}_4 \cup \mathbb{S}_5 \quad (1)$$

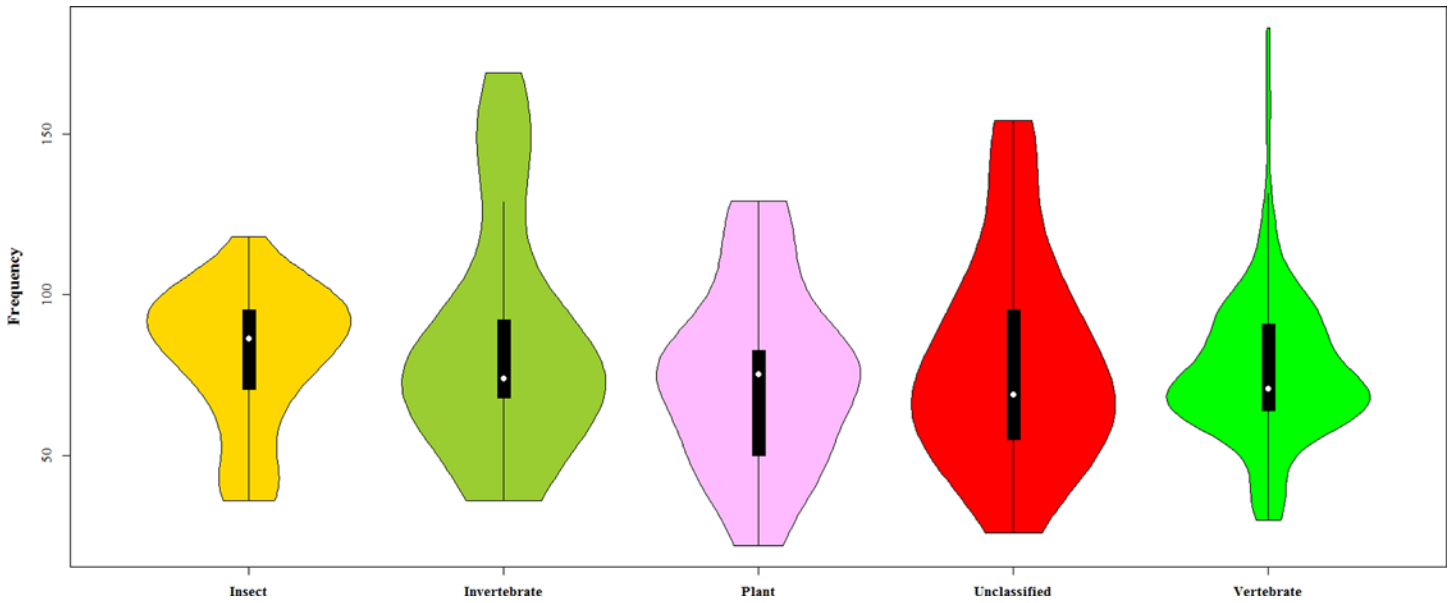

Fig 1. Violin plots show the length distribution of five defensin families.

doi:10.1371/journal.pone.0145541.g001

where the subset $\mathbb{S}_1$ contains 60 insect defensins, $\mathbb{S}_2$ contains 34 invertebrate defensins, $\mathbb{S}_3$ contains 42 plant defensins, $\mathbb{S}_4$ contains 40 unclassified defensins and $\mathbb{S}_5$ contains 157 vertebrate defensins, and $\cup$ represents the symbol for "union" in set theory. The length distribution of the five families is depicted in Fig 1. For the readers' convenience, the 333 defensin proteins sequences and codes are in S1 File.

Reduced Amino Acid Alphabet

In this study, the reduced amino acid alphabet composition(RAAAC) clustered by Protein Blocks(PBs) are used to predict defensins family and subfamily[21, 22], which is composed of 16 average protein fragments of 5 residues in length. The Protein Blocks have proven their efficiency both in description and prediction of longer fragments[23–25]. Once the databank was encoded in terms of PBs, sequence specificity was computed. Each PB was so associated with a set of enlarged sequence windows [–w; +w] of length l. An amino acid occurrence matrix of dimension 20 × l was computed for each PB. Then, each matrix was transformed into propensities matrix. Finally, all the matrices were compiled to create a matrix F of size 20 ×m with m, a vector of length 16l. The distance between two kinds of amino acids *i* and *j* was computed by using the D(aa*i*, aa*j*). Then a hierarchical clustering using all the amino acid occurrence matrices of the 16 PBs was performed, each resulting amino acid cluster represents amino acids that showed the same over- and under-representations upon all the PBs.

Different defensin peptides usually have specific functional regions, such as β-sheet-rich fold and framework of six disulphide-linked cysteines. Based on the similarity of their functional and physicochemical features in proteins, the 20 amino acids can be clustered into some smaller groups [26, 27]. The reduced amino acids not only can simplify the complexity of the protein system, but also improve the ability in finding structurally conserved regions and the structural similarity of entire proteins[13, 28]. The reduced amino acid alphabet derived from Protein Blocks method has the ability for abstracting useful functional and conservative feature. And it also is helpful for simplifying the amino acids composition of defensin peptide and

improving the ability in finding structurally conserved regions and the structural similarity of entire proteins.

Up to now, the Protein Blocks method has successful been used to analyze long protein fragments and to predict functional regions[21, 22], and the results have proven their efficiency both in description and prediction of longer fragments, such as protein structure mining[23, 24], outer membrane proteins analysis[23] and backbone structure prediction of proteins[25]. Our previous researches have also demonstrated that this feature selection method will be useful for analyzing the conservative domain and understanding function evolution of defensin protein [13, 29].

Support Vector Machine (SVM)

SVM is a powerful and popular method for pattern recognition that has been widely used in biology classification based on statistical learning theory[30–32]. In training an SVM classification system, proteins are represented by sequence-derived properties and are projected onto a hyperspace where the proteins in a family are separated from proteins outside the family by a hyperplane. By projecting a new sequence onto this hyperspace, the SVM system can determine whether or not the corresponding protein belongs to the family based on its location with respect to the hyperplane [33].

In the current study, the LIBSVM 3.0 package was used to implement of SVM[34]; it can be downloaded for free from the website (http://www.csie.ntu.edu.tw/_cjlin/libsvm). Four types of kernel functions, a linear function, polynomial function, sigmoid function and radial basis function (RBF), can be used for predictions in this software. Empirical studies show demonstrated that the RBF outperforms the remaining three types of kernel functions in nonlinear classification[31]. Thus, the RBF kernel function was used in the current work. All the computations were performed using LIBSVM-3.0 standard package (Chang and Lin, 2001). The various user-defined parameters, e.g., kernel parameter γ and regularization parameter C were optimized on the training dataset. The predictor obtained via the aforementioned procedure is called **iDPF-PseRAAAC**, where "i" stands for "identify", "DPF" for "defensin peptide family", "Pse" for "pseudo", "R" for ""reduced", "AAA" for "amino acid alphabet", and "C" for "composition".

Multi-class SVM

Prediction of defensin family classes is a multi-classification problem. SVM is regarded as a typical binary classifier[35, 36]. The methods of applying SVM to solve multi-class classification problems have one-against-one(OAO), one-against-all(OAA) and directed acyclic graph SVM (DAGSVM)[37]. In the present study, we adopt the "One-against-one" approach to transfer it into a two-class problem. This method involves construction of individual binary SVM classifier corresponding to each pair of the classes. Hence, if there are K classes, OAO will construct a total of K(K-1)/2 classifiers. Each classifier plays a role in classifying of one class and another class. Classifier *i j*, named f_{ij}, is trained using all the patterns from class *i* as positive instances, all the patterns from class *j* as negative instances, and disregarding the rest. The classifiers then are combined using majority voting scheme. Predictions are made with each binary classifiers and label is assigned to a class with maximum number of votes. In case when tie arise, i.e. two classes have identical votes, label assignment to the class is made on the basis of smallest index. More details for one-against-one(OAO) of SVM classification can be found in can be found in[37].

Performance Evaluation

This method's performance was measured based on sensitivity (Sn), specificity (Sp), Matthew's correlation coefficient (MCC) and overall accuracy (OA), which were defined as follows:

$$\begin{cases} \mathrm{Sn}(i) = \dfrac{\mathrm{TP}(i)}{\mathrm{TP}(i) + \mathrm{FN}(i)} \\ \mathrm{Sp}(i) = \dfrac{\mathrm{TN}(i)}{\mathrm{TN}(i) + \mathrm{FP}(i)} \\ \mathrm{MCC}(i) = \dfrac{\mathrm{TP}(i) \times \mathrm{TN}(i) - \mathrm{FP}(i) \times \mathrm{FN}(i)}{\sqrt{[\mathrm{TP}(i) + \mathrm{FP}(i)][\mathrm{TP}(i) + \mathrm{FN}(i)][\mathrm{TN}(i) + \mathrm{FP}(i)][\mathrm{TN}(i) + \mathrm{FN}(i)]}} \\ \mathrm{OA} = \dfrac{1}{N}\sum_{i=1}^{M}\mathrm{TP}(i) \end{cases} \quad (2)$$

where TP(i), TN(i), FP(i), and FN(i) represent true positive, true negative, false positive and false negative of family i; $M = 5$ is the number of subsets while N the number of the total samples in $\mathbb{S}$.

Results and Discussion

Cross Validation

Three cross-validation methods, namely the sub-sampling (or K-fold cross-validation) test, independent dataset test and jackknife test, are often used to evaluate the quality of a predictor [38]. Among the three methods, the jackknife test is the least arbitrary and most objective as demonstrated in[39] and can always yield a unique result for a given benchmark dataset hence. The jackknife test has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors[40–44]. Accordingly, the jackknife test was used to examine the performance of the model proposed in the current study.

Defensin Family Prediction

The jackknife results obtained using the **iDPF-PseRAAAC** and the benchmark dataset $\mathbb{S}$ based on different sizes(S) and N-peptide compositions(N) are depicted in Fig 2. Fig 2 shows the prediction results for the overall accuracy of the defensin families based on N-peptide composition with S size alphabet (N, S). As the dimensions increases, N-peptides provide progressively more detailed sequential information. However, the predictive ability did not increased linearly with dimension increase; for example, when the tripeptides composition (3, 20), 8000 dimensions, was selected as the input parameter, the overall accuracy for predicting five defensins families was only 79.28%. The results reflect the notion that a larger dimension does not necessarily result in better performance, and the prediction ability is not always better when the feature dimensions increase. Excessively large dimensions typically lead to information redundancy or noise, which results in bad prediction accuracy.

The Fig 3 heatmap shows the adjacent correlation of 13 reduced amino acids for five different defensin families. From the prediction performance based on different vector dimensions depicted in Fig 2, we observed that the overall accuracy reached a maximum 85.59% based on 2-peptide composition of 13 reduced amino acids (N = 2, S = 13). Table 1 shows the Jackknife results obtained using **iDPF-PseRAAAC** to identify defensin family with dipeptide (N = 2) composition based on different reduced amino acid alphabet approaches. As shown in Table 2, 2-peptide compositions with alphabet of 13 (N, S) outperformed the other reduced amino acid alphabet sizes. The largest defensin family, vertebrate defensins, yielded the best success rate at

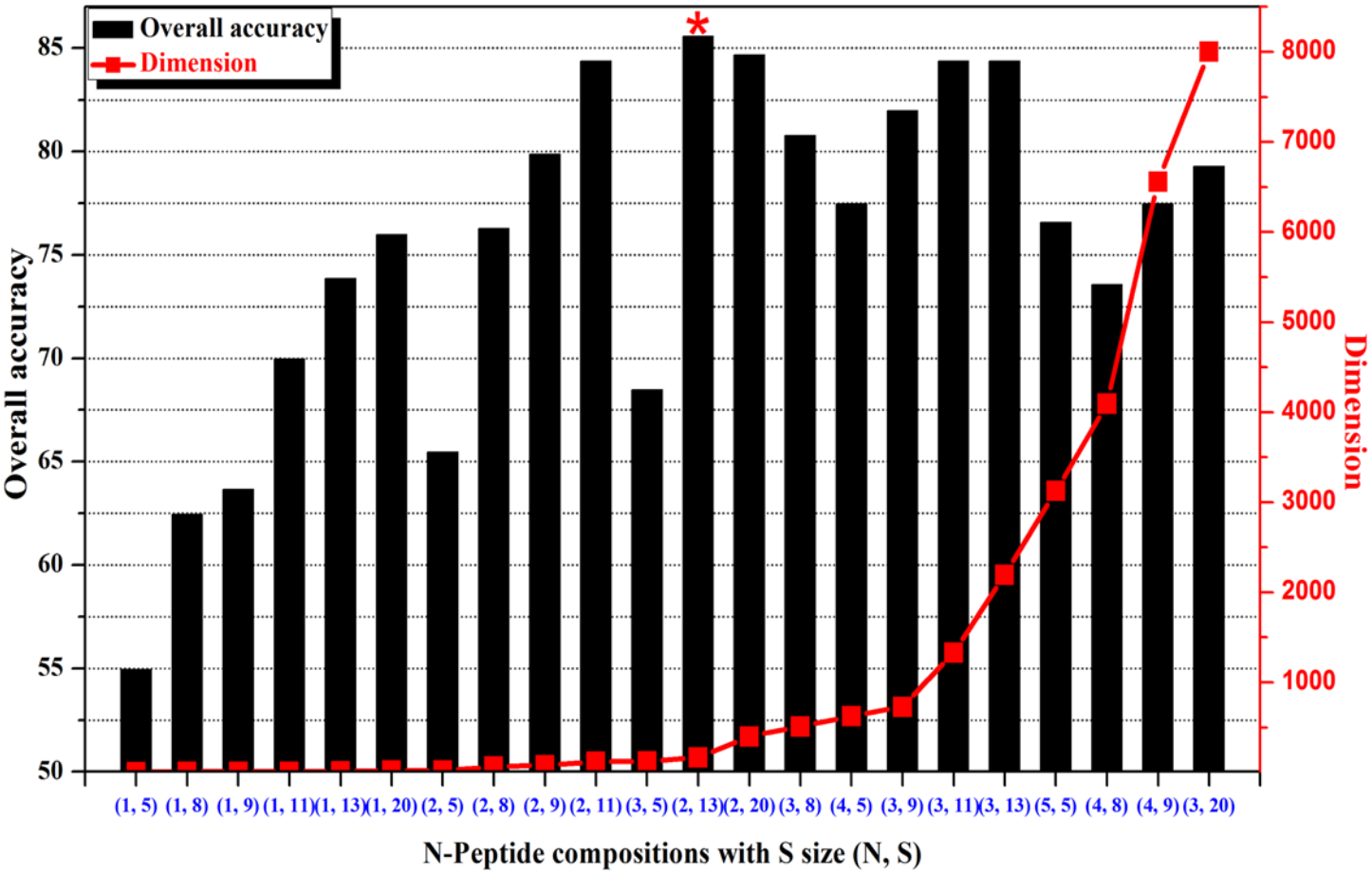

Fig 2. The predictive overall accuracy of defensins families based on different *N*-peptide composition with S size alphabet (N, S).

doi:10.1371/journal.pone.0145541.g002

99.36%. The 10-fold cross-validation has been performed to examine the comparability of our method. The prediction results are similar to the jackknife test (Total accuracy: 83.78% Vs 85.59%, S1 Table).

For further comparison, the amino acid (i.e., N = 1) and tripeptide (i.e., N = 3) results were also calculated and were in S2 Table, which shows that none of the results exhibit a higher success rate than N = 2. The data indicat that the reduced amino acid composition provides a greater weight of compositional bias to proteins with a signal at different sequence regions. Subsequently, the adjacent correlation for 13 reduced amino acids was analyzed and depicted using a heatmap plot.

Vertebrate Defensin Subfamily Prediction

Vertebrates include three distinct defensin subfamilies, Alpha-, Beta-, and Theta-defensins, which exhibit a broad spectrum of antimicrobial activities against bacteria, fungi, and viruses. Subsequently, the proposed method was used to predict the vertebrate defensins subfamily. The prediction results in Table 2 show that the best overall accuracy was 98.39%, and the Mathew's correlation coefficients (MCC) for the Alpha-type, Beta-type and Theta-type are 0.97, 0.96 and 0.89, respectively. Such high accuracies demonstrate that the proposed method is an effective and powerful approach for predicting defensin subfamilies.

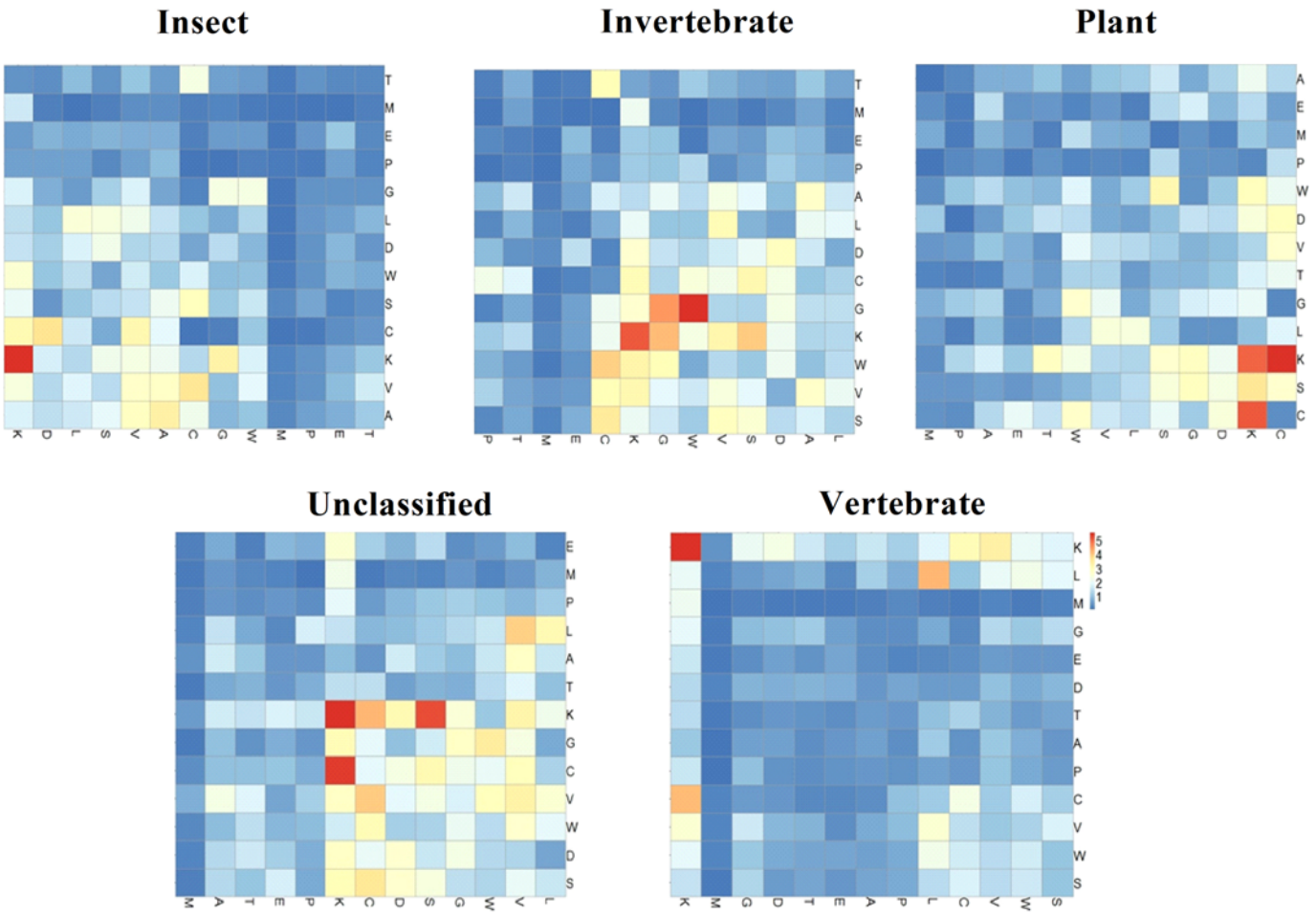

Fig 3. The heatmap shows the adjacent correlation of 13 reduced amino acids for five different defensin families.

doi:10.1371/journal.pone.0145541.g003

Comparison with Previous Methods

To further demonstrate the performance of the proposed method, it is necessary to be compared with other existing methods. However, directly comparing the results is not objective and strict examination due to the different benchmark datasets used. Therefore, we repeated the feature selection and prediction process on the previous dataset. The jackknife cross-validated accuracies are depicted in Fig 4. Obviously, our proposed method yields the highest predictive success rate.

According to Fig 4, when the 169 reduced dipeptides are used, our method can achieve a maximum overall accuracy (OA) of 95.10% for the defensin family, and 98.39% for the vertebrate subfamily, which is higher than the maximum accuracy obtained using other methods. Although the success rate for others family and subfamily obtained using our method is not better, the accuracy of the other families and subfamilies are dramatically better than using other methods, and Fig 4 clearly indicate that the proposed method is more powerful than our previous method.

Subsequently, it is instructive to compare the overall success rate from **iDPF-PseRAAAC** with the success rate for weighted random guess (WRG)[45]. The overall success rate based on

Table 1. Results obtained by iDPF-PseRAAAC in identifying defensin peptide families with dipeptide case(N = 2).

			N-Peptide compositions of RAAA with S size (N, S)					
Family	**Subset**	**Metrics**	**(2, 20)**	**(2, 13)**	**(2, 11)**	**(2, 9)**	**(2, 8)**	**(2, 5)**
			400	**169**	121	81	64	25
Insect	$\mathbb{S}_1$	Sn(%)	88.33	**90.00**	83.33	76.67	81.67	58.33
		Sp(%)	98.53	**97.07**	96.34	95.60	96.70	91.94
		MCC	0.89	**0.86**	0.80	0.73	0.79	0.51
Invertebrate	$\mathbb{S}_2$	Sn(%)	64.71	**61.76**	61.76	64.71	55.88	38.24
		Sp(%)	96.32	**97.32**	96.99	95.32	97.66	95.99
		MCC	0.62	**0.64**	0.62	0.59	0.60	0.39
Plant	$\mathbb{S}_3$	Sn(%)	83.33	**90.48**	90.48	80.95	64.29	57.14
		Sp(%)	99.66	**98.97**	97.94	98.63	98.63	97.25
		MCC	0.89	**0.90**	0.87	0.83	0.72	0.61
Unclassified	$\mathbb{S}_4$	Sn(%)	45.00	**40.00**	47.50	42.50	17.50	17.50
		Sp(%)	96.59	**96.93**	95.90	95.90	96.93	95.90
		MCC	0.49	**0.46**	0.49	0.44	0.22	0.19
Vertebrate	$\mathbb{S}_5$	Sn(%)	98.09	**99.36**	97.45	93.63	96.82	88.54
		Sp(%)	85.80	**88.64**	91.48	85.80	71.59	65.34
		MCC	0.84	**0.88**	0.89	0.79	0.70	0.55
	OA(%)		84.68	**85.59**	84.38	79.88	76.28	65.47

The bold values show the best results.

doi:10.1371/journal.pone.0145541.t001

the OA to identify the defensin proteins among their four subfamilies WRG is given by [45].

$$\mathrm{OA(WRG)} = \frac{(N_1)^2 + (N_2)^2 + (N_3)^2 + (N_4)^2 + (N_5)^2}{N^2} \tag{3}$$

where N is the number of defensin proteins in the benchmark dataset $\mathbb{S}$, N_1 the number of defensin proteins in the subset $\mathbb{S}_1$, N_2 the number of defensin proteins in the subset $\mathbb{S}_2$, and so forth (see **Eq 1**). Substituting these data into **Eq 3**, we obtain the following.

$$\mathrm{OA(WRG)} = \frac{(60)^2 + (34)^2 + (42)^2 + (40)^2 + (157)^2}{333^2} = 29.55\% \tag{4}$$

In contrast, the best overall success rate for **iDPF-PseRAAAC** was 85.59%. Compared with the results in Eq 4, the overall success rate for **iDPF-PseRAAAC** is approximately 56% higher than using a weighted random guess, which indicate that **iDPF-PseRAAAC** may be an easy and useful tool for timely identifying defensin proteins families.

Table 2. The prediction results for vertebrate subfamilies based on 2-peptide composition of 13 reduced amino acids (N = 2, S = 13).

Subfamily	Alpha-type	Beta-type	Theta-type	Sn(%)	Sp(%)	MCC
Alpha-type	69	3	0	95.83	100	0.97
Beta-type	0	172	0	100	94.81	0.96
Theta-type	0	1	4	80	100	0.89
Overall accuracy(%)			98.39			

doi:10.1371/journal.pone.0145541.t002

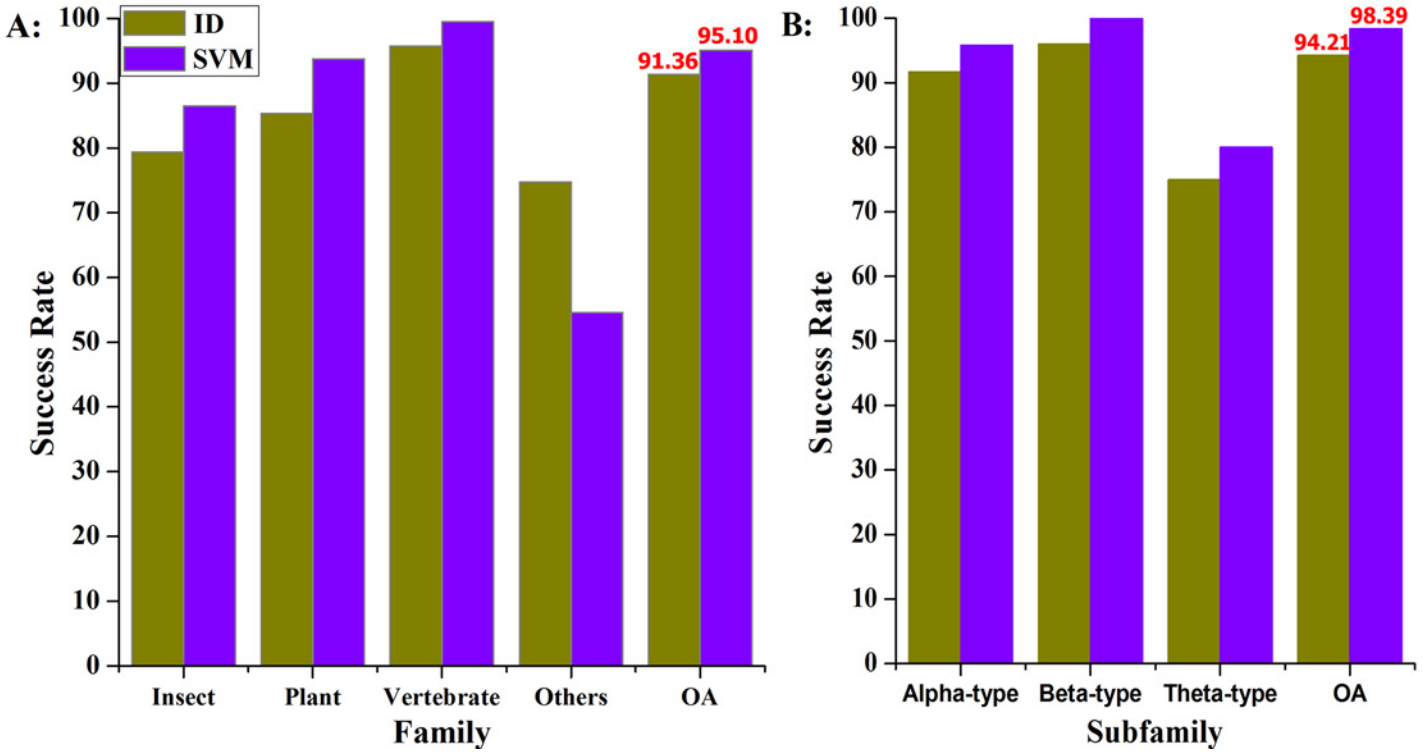

Fig 4. Comparing the performance of the proposed method with our previous methods. A: indicated the prediction results of defensin family; **B:** indicated the prediction results of vertebrate defensin subfamily.

doi:10.1371/journal.pone.0145541.g004

Web-Server Guide

For the convenience of most experimental scientists, below, we provide a step-by-step guide on how to use the **iDPF-PseRAAAC** web-server to achieve the desired results.

Step 1. Open the web server at http://wlxy.imu.edu.cn/college/biostation/fuwu/iDPF-PseRAAAC/index.asp and you will see the top page for **iDPF-PseRAAAC** on your computer screen, as shown in Fig 5. Click on the 'Read Me' button to see a brief introduction about the predictor and the caveat for using it.

Step 2. Either type or copy/paste the query defensin peptide sequence into the input box at the center of Fig 5. The input sequence should be in the FASTA format. A sequence in the 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 immediately following the "$>$" symbol in the single initial line are optional and only used for the identification and description. The sequence ends if another line starting with a "$>$" appears; this indicates the start of another sequence. Example sequences in FASTA format can be viewed by clicking on the 'Example' button right above the input box.

Step 3. Click on the 'Submit' button to see the predicted result. For example, if you use the query defensin protein sequences in the 'Example' window as the input, after clicking the 'Submit' button, you will see the "Result page" shown on the screen of your computer. All these results are fully consistent with the experimental observations. It takes approximately a few seconds for the above computation before the predicted result appears on your computer screen; for query sequences and longer each sequences, more time is typically required.

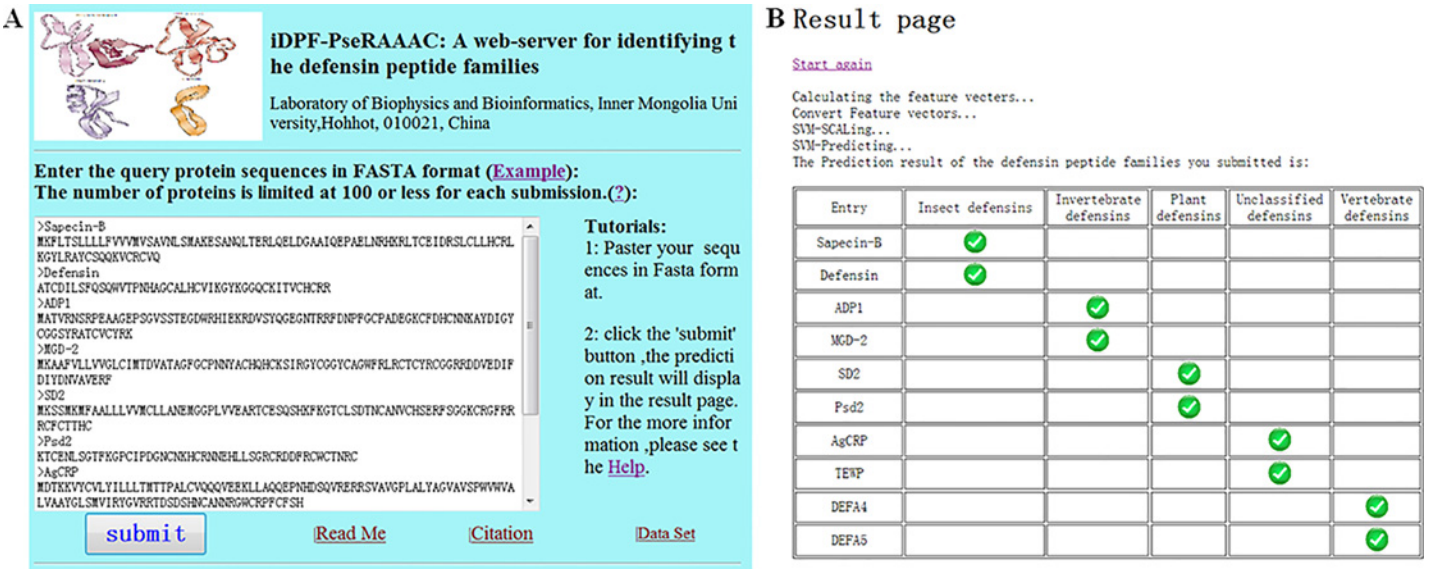

Fig 5. A semi-screenshot to show the top page of the iDPF-PseRAAAC web-server.

doi:10.1371/journal.pone.0145541.g005

Step 4. Click on the 'Citation' button to find the relevant papers that document the detailed development and algorithm of **iDPF-PseRAAAC**.

Step 5. Click on the 'Data' button to download the benchmark datasets used to train and test the **iDPF-PseRAAAC** predictor.

Conclusions

Defensins also play important regulatory roles in the immune systems of animals and plants, acting as a bridge between innate and adaptive immunity in vertebrates. In this study, a promising method, **iDPF-PseRAAAC,** was developed to improve prediction performance for defensin proteins. The use of reduced amino acid alphabets not only provides an efficient and accurate means of protein vectorization for sequence-based protein classification systems but also remarkably improves computational efficiency. High predictive accuracies demonstrate that our proposed method is a potentially useful tool for classifying defensin family.

Supporting Information

S1 File. The 333 defensin proteins sequences and codes.docx.
(DOCX)

S1 Table. The prediction results of 10-fold cross-validation for the benchmark dataset.
(DOCX)

S2 Table. The results obtained by iDPF-PseRAAAC in identifying defensin peptide families with (A) single amino acid case (N = 1) and (B) tripeptide case (N = 3).
(DOCX)

Acknowledgments

We wish to express our gratitude to the editor and three anonymous reviewers whose constructive comments were very helpful in strengthening the presentation of this paper.

Author Contributions

Conceived and designed the experiments: YCZ GLF. Performed the experiments: YCZ YL. Analyzed the data: LY GLF. Contributed reagents/materials/analysis tools: ZYW GPL. Wrote the paper: YCZ YL GLF.

References

1. Aerts AM, Francois IE, Cammue BP, Thevissen K. The mode of antifungal action of plant, insect and human defensins. Cellular and molecular life sciences: CMLS. 2008; 65(13):2069–79. Epub 2008/03/25. doi: 10.1007/s00018-008-8035-0 PMID: 18360739.
2. Lay FT, Anderson MA. Defensins—components of the innate immune system in plants. Current protein & peptide science. 2005; 6(1):85–101. Epub 2005/01/11. PMID: 15638771.
3. Ganz T. Defensins: antimicrobial peptides of innate immunity. Nature reviews Immunology. 2003; 3 (9):710–20. Epub 2003/09/02. doi: 10.1038/nri1180 PMID: 12949495.
4. Oppenheim JJ, Biragyn A, Kwak LW, Yang D. Roles of antimicrobial peptides such as defensins in innate and adaptive immunity. Annals of the rheumatic diseases. 2003; 62 Suppl 2:ii17–21. Epub 2003/10/09. PMID: 14532141; PubMed Central PMCID: PMC1766745.
5. Menendez A, Brett Finlay B. Defensins in the immunology of bacterial infections. Current opinion in immunology. 2007; 19(4):385–91. Epub 2007/08/19. doi: 10.1016/j.coi.2007.06.008 PMID: 17702560.
6. Thomma BP, Cammue BP, Thevissen K. Plant defensins. Planta. 2002; 216(2):193–202. Epub 2002/11/26. doi: 10.1007/s00425-002-0902-6 PMID: 12447532.
7. Palma MS. Peptides as toxins/defensins. Amino acids. 2011; 40(1):1–4. Epub 2010/08/25. doi: 10.1007/s00726-010-0726-9 PMID: 20734212.
8. Li D, Zhang L, Yin H, Xu H, Satkoski Trask J, Smith DG, et al. Evolution of primate alpha and theta defensins revealed by analysis of genomes. Molecular biology reports. 2014; 41(6):3859–66. Epub 2014/02/22. doi: 10.1007/s11033-014-3253-z PMID: 24557891.
9. Jarczak J, Kosciuczuk EM, Lisowski P, Strzalkowska N, Jozwik A, Horbanczuk J, et al. Defensins: natural component of human innate immunity. Human immunology. 2013; 74(9):1069–79. Epub 2013/06/13. doi: 10.1016/j.humimm.2013.05.008 PMID: 23756165.
10. de Medeiros LN, Angeli R, Sarzedas CG, Barreto-Bergter E, Valente AP, Kurtenbach E, et al. Backbone dynamics of the antifungal Psd1 pea defensin and its correlation with membrane interaction by NMR spectroscopy. Biochimica et biophysica acta. 2010; 1798(2):105–13. Epub 2009/07/28. doi: 10.1016/j.bbamem.2009.07.013 PMID: 19632194.
11. Xiao X, Wang P, Lin WZ, Jia JH, Chou KC. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Analytical biochemistry. 2013; 436(2):168–77. Epub 2013/02/12. doi: 10.1016/j.ab.2013.01.019 PMID: 23395824.
12. Chen W, Luo L. Classification of antimicrobial peptide using diversity measure with quadratic discriminant analysis. Journal of microbiological methods. 2009; 78(1):94–6. Epub 2009/04/08. doi: 10.1016/j.mimet.2009.03.013 PMID: 19348863.
13. Zuo YC, Li QZ. Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet. Peptides. 2009; 30(10):1788–93. Epub 2009/07/14. doi: 10.1016/j.peptides.2009.06.032 PMID: 19591890.
14. Kumari SR, Badwaik R, Sundararajan V, Jayaraman VK. Defensinpred: defensin and defensin types prediction server. Protein and peptide letters. 2012; 19(12):1318–23. Epub 2012/06/08. PMID: 22670676.
15. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of theoretical biology. 2011; 273(1):236–47. Epub 2010/12/21. doi: 10.1016/j.jtbi.2010.12.024 PMID: 21168420.
16. Xiao X, Wang P, Chou KC. Cellular automata and its applications in protein bioinformatics. Current protein & peptide science. 2011; 12(6):508–19. Epub 2011/07/27. PMID: 21787298.
17. Seebah S, Suresh A, Zhuo S, Choong YH, Chua H, Chuon D, et al. Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides. Nucleic acids research. 2007; 35(Database issue):D265–8. Epub 2006/11/09. doi: 10.1093/nar/gkl866 PMID: 17090586; PubMed Central PMCID: PMC1669742.
18. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics (Oxford, England). 2006; 22(13):1658–9. Epub 2006/05/30. doi: 10.1093/bioinformatics/btl158 PMID: 16731699.

19. Zhao XM, Wang Y, Chen L, Aihara K. Gene function prediction using labeled and unlabeled data. Bmc Bioinformatics. 2008; 9. doi: 10.1186/1471-2105-9-57 PMID: WOS:000254435900002.

20. Zhao XM, Li X, Chen L, Aihara K. Protein classification with imbalanced data. Proteins-Structure Function and Bioinformatics. 2008; 70(4):1125–32. doi: 10.1002/prot.21870 PMID: WOS:000253567400001.

21. Etchebest C, Benros C, Bornot A, Camproux AC, de Brevern AG. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. European biophysics journal: EBJ. 2007; 36(8):1059–69. Epub 2007/06/15. doi: 10.1007/s00249-007-0188-5 PMID: 17565494.

22. de Brevern AG. New assessment of a structural alphabet. In silico biology. 2005; 5(3):283–9. Epub 2005/07/06. PMID: 15996119; PubMed Central PMCID: PMC2001288.

23. Martin J, de Brevern AG, Camproux AC. In silico local structure approach: a case study on outer membrane proteins. Proteins. 2008; 71(1):92–109. Epub 2007/10/13. doi: 10.1002/prot.21659 PMID: 17932925.

24. Tyagi M, de Brevern AG, Srinivasan N, Offmann B. Protein structure mining using a structural alphabet. Proteins. 2008; 71(2):920–37. Epub 2007/11/16. doi: 10.1002/prot.21776 PMID: 18004784.

25. Tyagi M, Sharma P, Swamy CS, Cadet F, Srinivasan N, de Brevern AG, et al. Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. Nucleic acids research. 2006; 34(Web Server issue):W119–23. Epub 2006/07/18. doi: 10.1093/nar/gkl199 PMID: 16844973; PubMed Central PMCID: PMC1538797.

26. Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. Proceedings of the National Academy of Sciences of the United States of America. 1996; 93 (21):11628–33. Epub 1996/10/15. PMID: 8876187; PubMed Central PMCID: PMC38109.

27. Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. Journal of molecular biology. 1999; 291(1):177–96. Epub 1999/08/10. doi: 10.1006/jmbi.1999.2911 PMID: 10438614.

28. Zuo YC, Chen W, Fan GL, Li QZ. A similarity distance of diversity measure for discriminating mesophilic and thermophilic proteins. Amino acids. 2013; 44(2):573–80. Epub 2012/08/02. doi: 10.1007/s00726-012-1374-z PMID: 22851052.

29. Zuo YC, Li QZ. Using K-minimum increment of diversity to predict secretory proteins of malaria parasite based on groupings of amino acids. Amino acids. 2010; 38(3):859–67. Epub 2009/04/24. doi: 10.1007/s00726-009-0292-1 PMID: 19387791.

30. Zuo YC, Peng Y, Liu L, Chen W, Yang L, Fan GL. Predicting peroxidase subcellular location by hybridizing different descriptors of Chou' pseudo amino acid patterns. Analytical biochemistry. 2014; 458:14–9. Epub 2014/05/08. doi: 10.1016/j.ab.2014.04.032 PMID: 24802134.

31. Ding C, Yuan LF, Guo SH, Lin H, Chen W. Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions. Journal of proteomics. 2012; 77:321–8. Epub 2012/09/25. doi: 10.1016/j.jprot.2012.09.006 PMID: 23000219.

32. Lin H, Ding H. Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. Journal of theoretical biology. 2011; 269(1):64–9. Epub 2010/10/26. doi: 10.1016/j.jtbi.2010.10.019 PMID: 20969879.

33. Lin HH, Han LY, Cai CZ, Ji ZL, Chen YZ. Prediction of transporter family from protein sequence by support vector machine approach. Proteins. 2006; 62(1):218–31. Epub 2005/11/16. doi: 10.1002/prot.20605 PMID: 16287089.

34. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011;2(3):1–27. (software available at http://www.csie.ntu.edu.tw/_cjlin/libsvm).

35. Zhao XM, Cheung YM, Huang DS. A novel approach to extracting features from motif content and protein composition for protein sequence classification. Neural networks: the official journal of the International Neural Network Society. 2005; 18(8):1019–28. Epub 2005/09/13. doi: 10.1016/j.neunet.2005.07.002 PMID: 16153801.

36. Zhao XM, Du JX, Wang HQ, Zhu YP, Li YX. A new technique for selecting features from protein sequences. International Journal of Pattern Recognition and Artificial Intelligence. 2006; 20(2):271–83. doi: 10.1142/s021800140600465x PMID: WOS:000237140000012.

37. Hsu CW, Lin CJ. A comparison of methods for multiclass support vector machines. IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council. 2002; 13(2):415–25. Epub 2008/02/05. doi: 10.1109/72.991427 PMID: 18244442.

38. Chou KC, Zhang CT. Prediction of protein structural classes. Critical reviews in biochemistry and molecular biology. 1995; 30(4):275–349. Epub 1995/01/01. doi: 10.3109/10409239509083488 PMID: 7587280.

39. Chou KC, Shen HB. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nature protocols. 2008; 3(2):153–62. Epub 2008/02/16. doi: 10.1038/nprot.2007.494 PMID: 18274516.

40. Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q, et al. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics (Oxford, England). 2014; 30(4):472–9. Epub 2013/12/10. doi: 10.1093/bioinformatics/btt709 PMID: 24318998.

41. Liu B, Chen J, Wang X. Protein remote homology detection by combining Chou's distance-pair pseudo amino acid composition and principal component analysis. Molecular genetics and genomics: MGG. 2015. Epub 2015/04/22. doi: 10.1007/s00438-015-1044-4 PMID: 25896721.

42. Du P, Tian Y, Yan Y. Subcellular localization prediction for human internal and organelle membrane proteins with projected gene ontology scores. Journal of theoretical biology. 2012; 313:61–7. Epub 2012/09/11. doi: 10.1016/j.jtbi.2012.08.016 PMID: 22960368.

43. Du P, Yu Y. SubMito-PSPCP: predicting protein submitochondrial locations by hybridizing positional specific physicochemical properties with pseudoamino acid compositions. BioMed research international. 2013; 2013:263829. Epub 2013/09/13. doi: 10.1155/2013/263829 PMID: 24027753; PubMed Central PMCID: PMC3763570.

44. Du P, Wang L. Predicting human protein subcellular locations by the ensemble of multiple predictors via protein-protein interaction network with edge clustering coefficients. PloS one. 2014; 9(1):e86879. Epub 2014/01/28. doi: 10.1371/journal.pone.0086879 PMID: 24466278; PubMed Central PMCID: PMC3900678.

45. Chou KC. Some remarks on predicting multi-label attributes in molecular biosystems. Molecular bioSystems. 2013; 9(6):1092–100. Epub 2013/03/29. doi: 10.1039/c3mb25555g PMID: 23536215.