Send Orders of Reprints at bspsaif@emirates.net.ae

Predicting Antibacterial Peptides by the Concept of Chou's Pseudo-amino Acid Composition and Machine Learning Methods

Maede Khosravian[a], Fateme Kazemi Faramarzi[b], Majid Mohammad Beigi[b], Mandana Behbahani[a] and Hassan Mohabatkar [a,*]

[a]*Department of Biotechnology, Faculty of Advanced Sciences and Technologies, University of Isfahan, Isfahan, Iran;* [b]*Department of Biomedical Engineering, Faculty of Engineering, University of Isfahan, Isfahan, Iran*

Abstract: Microbial resistance to antibiotics is a rising concern among health care professionals, driving them to search for alternative therapies. In the past few years, antimicrobial peptides (AMPs) have attracted a lot of attention as a substitute for conventional antibiotics. Antimicrobial peptides have a broad spectrum of activity and can act as antibacterial, antifungal, antiviral and sometimes even as anticancer drugs. The antibacterial peptides have little sequence homology, despite common properties. Since there is a need to develop a computational method for predicting the antibacterial peptides, in the present study, we have applied the concept of Chou's pseudo-amino acid composition (PseAAC) and machine learning methods for their classification. Our results demonstrate that using the concept of PseAAC and applying Support Vector Machine (SVM) can provide useful information to predict antibacterial peptides.

Keywords: Antibacterial peptides, bioinformatics, Chou's pseudo amino acid composition, machine learning methods, clustering, fivefold cross-validation.

1. INTRODUCTION

AMPs are naturally present in all organisms and play a vital role in innate immunity [1]. Over the past decades numerous membrane-lytic peptides have been isolated from insects, amphibians, and mammals. Among them are melittin from bee venom, mastoparans from wasp venom, cecropins from insects, defensins from mammalian neutrophils, and magainins from frog skin [2]. These peptides are very diverse with respect to amino acid sequence and secondary structure but share certain properties, such as affinity for the negatively charged phospholipids that are present on the outer surfaces of the cytoplasmic membrane of many microbial species [3]. These peptides cause cell death either by disrupting the microbial cell membrane; inhibiting extracellular polymer synthesis or intracellular functions [4,5,6]. Microbial resistance to antibiotics is a rising concern among health care professionals, driving them to search for alternative therapies. In the past few years, AMPs have attracted lot of attention as a substitute for conventional antibiotics [7]. Their short length and fast and efficient action against microbes has made them potential candidates as peptide drugs [8,9].

Antimicrobial peptides have a broad spectrum of activity and can act as antibacterial, antifungal, antiviral and sometimes even as anticancer peptide [10]. Some AMPs also possess antitumor activity and can act as mitogens and signaling molecules [11]. Generally they contain 15–45 amino acid residues and the net charge is positive [12]. In 1987, Zasloff discovered that a cationic peptide in the skin of the African clawed frog Xenopus laevis had broad-spectrum antibacterial activity based on a "pore-formation" mechanism [13]. He called it magainin. Pexiganan, a synthetic 22-amino-acid analogue of magainin 2, demonstrated excellent *in vitro* broad-spectrum activity against several bacterial clinical isolates [14]. Iseganan (IB-367) is a synthetic protegrin 1 derived from the naturally occurring protegrins in pig leucocytes [15]. As a cationic antimicrobial peptide, it has broad-spectrum *in vitro* antibacterial and antifungal inhibitory activity. Several peptides have shown promise for possible drug development in preclinical studies [14].

The antibacterial peptides have little sequence homology, despite common properties. Thus it is difficult to develop a method for predicting them based on sequence similarity. Moreover, experimental methods for identification and designing of antibacterial peptides are costly, time consuming and resource intensive. Thus there is a need to develop computational tools for predicting antibacterial peptides, which could be used to design potent peptides against bacterial pathogens [3].

There are various approaches to predict different aspects of proteins. Some of these approaches are based on amino acid sequence [16,17], template [18-20] and amino acid composition [21,22]. One of the classification methods is PseAAC, originally developed by Chou for prediction of protein sub-cellular localization and membrane protein type [23,24]. PseAAC concept has been widely used to predict many aspects of proteins, including prediction of secondary structure [19,20], super- secondary structure [16], protein quaternary structure [25] and functional classification of enzyme family classes [26], cyclins [27], risk type of human papilloma viruses [28], GABA A receptors [29] and metallo-

*Address correspondence to this author at the Department of Biotechnology, Faculty of Advanced Sciences and Technologies, University of Isfahan, Isfahan, Iran; Tel: +98311-7934391; Mob: 09134019436;
Fax: +98311-7932342; E-mail: h.mohabatkar@ast.ui.ac.ir

1875-5305/13 $58.00+.00

proteinase family [30]. Recently Du *et al* [31] have proposed a new cross-platform stand-alone software program, called PseAAC-Builder (http://www.pseb.sf.net), which can be used to generate various modes of Chou's pseudo-amino acid composition in an efficient and flexible way. In the present study, we have applied the concept of Chou's pseudo-amino acid composition (PseAAC) and machine learning methods including Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) for the classification of antibacterial peptides.

2. METHODS

As summarized in a recent comprehensive review [32], to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the protein or peptide samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us describe how to deal with these steps.

2.1. Dataset

Amino acid sequences of non-redundant 1086 antibacterial peptides as positive set and 8860 non-antibacterial peptides as negative set were analyzed. Positive and negative data were fetched from Antimicrobial Peptides Database (http://aps.unmc.edu/AP/database/antiB.php) and AMP database (http://amp.biosino.org/) respectively. To avoid bias in classifier, the homologous proteins (more than 90% similarity) were removed from our dataset by cd-hit program (It is available in: http://cd-hit.org/) [33,34].

2.2. Generating Chou's PseAAC

Two kinds of models are usually used to represent protein samples. One is the sequential model, and the other one the discrete model. The most straight forward sequential model for a protein sample is its entire amino acid sequence, as expressed by

P =R1R2R3R4… RL

Where R1 represents the 1st residue of the protein P, R2 the 2nd residue …, RL the L-th residue, and they each belongs to one of the 20 native amino acid types.

Various non-sequential models, or discrete models, were proposed to mathematically represent the protein. In this study, to avoid losing much important information hidden in protein sequences, we used the concept of Chou's PseAAC [35,36] to formulate the feature vector for protein or peptide samples.

According to Eq.**6** of a recent comprehensive review [32], the form of PseAAC can be generally formulated as:

$$\mathbf{P}= [\psi_1 \psi_2 \ \ldots \ \psi_u \ldots \ \psi_\Omega]^{\mathrm{T}} \quad (1)$$

Where $\mathbf{T}$ is a transpose operator. The subscript Ω is an integer and its value as well as the components $\psi_1, \psi_2, \ldots, \psi_\Omega$ will depend on how to extract the desired information from the amino acid sequence of $\mathbf{P}$. The following relation was used for extraction:

$$\psi_u = \begin{cases} f_u \Big/ \left(\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\theta_j\right) & (1 \le u \le 20) \\ w\theta_{u-20} \Big/ \left(\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\theta_j\right) & (20+1 \le u \le 20+\lambda = \Omega;\ \ \lambda < \mathrm{L}) \end{cases} \quad (2)$$

Where f_i is the normalized occurrence frequency of the 20 amino acid in the protein, w is weight factor, designed for the user to put weight on the additional PseAAC components, λ is the counted rank (or tier) of the correlation along a protein sequence and θj is the j-tier sequence correlation factor computed according:

$$\theta_j = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}\varphi(R_i, R_{i+\lambda}) \quad (3)$$

In above equation, correlation function (φ) is given by:

$$\varphi(R_i, R_j) = \frac{1}{n}\sum_{k}^{n}\left(H_k(R_i) - H_k(R_j)\right)^2 \quad (4)$$

PseAAC is a flexible web server for generating various kinds of protein pseudo-amino acid composition, which is available at http://chou.med.harvard.edu/bioinf/PseAAC. The essence of PseAAC is based on the protein representation using a set of more than 20 discrete factors, where the first 20 factors represent the components of its conventional amino acid composition, whereas the additional factors incorporate some of its sequence order information via various modes. These additional factors are a series of rank-different correlation factors along a protein chain, but they can also be of any combination of other factors as long as they can reflect some sort of sequence order effects in one way or the other [35-37]. Three different types of parameters are often used to generate various kinds of PseAAC: quantitative characters of amino acids, weight factor and rank of correlation. The following six amino acid characters are supported by PseAAC server to calculate the correlations between amino acids at different positions along the protein chain: (1) hydrophobicity, (2) hydrophilicity, (3) side chain mass, (4) pK1 (alpha-COOH), (5) pK2 (NH3) and (6) pI. Selection of any character or combination of characters as part of the input is user-dependent. The weight factor is designed for the user to put weight on the additional PseAA components with respect to the conventional AA components. The user can select any value within the region from 0.05 to 0.70 for the weight factor. The counted rank (or tier) of the correlation along a protein sequence is represented by λ. For the present study, type 1 PseAAC, which is also called the parallel correlation type, λ=2, and weight factor=0.05 were applied.

In this study, calculations by PseAAC server for all six characters and their all combinations (57 possible combinations) (Table **1**) have been considered. Thus:

$$\Omega = C(6,1)+C(6,2)+C(6,3)+C(6,4)+C(6,5)+C(6,6)=63 \quad (5)$$

Table 1. Different Combination of Six Characters of Antibacterial Peptides and Negative Instances

No.	Character (s)	No.	Character (s)
1	Hydrophobicity	**33**	Hydrophilicity, mass and pk2
2	Hydrophilicity	**34**	Hydrophilicity, mass and pI
3	Mass	**35**	Hydrophilicity, pk1 and pk2
4	pk1	**36**	Hydrophilicity, pk1 and pI
5	pk2	**37**	Hydrophilicity, pk2 and pI
6	pI	**38**	Mass, pk1 and pk2
7	Hydrophobicity and hydrophilicity	**39**	Mass, pk1 and pI
8	Hydrophobicity and mass	**40**	Mass, pk2 and pI
9	Hydrophobicity and pk1	**41**	pk1 and pk2 and pI
10	Hydrophobicity and pk2	**42**	Hydrophobicity, hydrophilicity and mass and pk1
11	Hydrophobicity and pI	**43**	Hydrophobicity, hydrophilicity and mass and pk2
12	Hydrophilicity and mass	**44**	Hydrophobicity, hydrophilicity and mass and pI
13	Hydrophilicity and pk1	**45**	Hydrophobicity, hydrophilicity and pk1 and pk2
14	Hydrophilicity and pk2	**46**	Hydrophobicity, hydrophilicity and pk1 and pI
15	Hydrophilicity and pI	**47**	Hydrophobicity, hydrophilicity and pk2 and pI
16	Mass and pk1	**48**	Hydrophobicity, mass, pk1 and pk2
17	Mass and pk2	**49**	Hydrophobicity, mass, pk1 and pI
18	Mass and pI	**50**	Hydrophobicity, mass, pk2 and pI
19	pk1 and pk2	**51**	Hydrophobicity and pk1 and pk2 and pI
20	pk1 and pI	**52**	Hydrophilicity, mass, pk1 and pk2
21	pk2 and pI	**53**	Hydrophilicity, mass, pk1 and pI
22	Hydrophobicity, hydrophilicity and mass	**54**	Hydrophilicity, mass, pk2 and pI
23	Hydrophobicity, hydrophilicity and pk1	**55**	Hydrophilicity, pk1, pk2 and pI
24	Hydrophobicity, hydrophilicity and pk2	**56**	Mass, pk1, pk2 and pI
25	Hydrophobicity, hydrophilicity and pI	**57**	Hydrophobicity, hydrophilicity and mass and pk1 and pk2
26	Hydrophobicity, mass and pk1	**58**	Hydrophobicity, hydrophilicity and mass and pk1 and pI
27	Hydrophobicity, mass and pk2	**59**	Hydrophobicity, hydrophilicity and mass and pk2 and pI
28	Hydrophobicity, mass and pI	**60**	Hydrophobicity, hydrophilicity and pk1 and pk2 and pI
29	Hydrophobicity and pk1 and pk2	**61**	Hydrophobicity, mass, pk1, pk2 and pI
30	Hydrophobicity and pk1 and pI	**62**	Hydrophilicity, mass, pk1, pk2 and pI
31	Hydrophobicity and pk2 and pI	**63**	Hydrophobicity, hydrophilicity and mass and pk1 and pk2 and pI
32	Hydrophilicity, mass and pk1		

Since lambda parameter was chosen as two. Therefore, for any protein (besides the first 20 numbers, which represent the classic amino acid composition) there were two tier values. This means that, 126 features for each protein were used to classify the dataset.

2.3. Support Vector Machine

SVMs, an algorithm for the classification of both linear and nonlinear data, map the original data into a higher dimension, where we can find a hyper plane as a discriminant function for the separation of data using some instances

called support vector [38-40]. This discriminant function is represented as a linear function in feature space in the form of $f(x)=w^T\varphi(x)$ for some weight vector $w\in F$. Given a training set of instance-label pairs (x_i , y_i), i=1,2,3,…,l where $x_i\in R^n$ and $y_i\in\{1,-1\}$, to map the input data samples x_i into a higher dimensional feature space $\varphi(x_i)$, a set of nonlinearly separable problem is solved. The classical maximum margin SVM classifier aims to find a hyper plane of the form $w^t\varphi(x)+b=0$, which separates the patterns of the two classes.

In the case of noisy data, to avoid poor generalization for unseen data, a vector of slack variables $X=(\xi_1,\xi_2,\ldots,\xi_l)^T$ should be taken into account. The problem can then be written as:

$$\text{Minimize } \frac{1}{2}w^T w+C\sum_{i=1}^{l}\xi_i$$

subject to

$$y_i(w^T\phi(x_i)+b)\geq 1-\xi_i,\ \xi_i\geq 0 \text{ where } i=1,\ 2,\ 3,\ldots l \tag{6}$$

The solution then yields the soft margin classifier. By introducing a set of Lagrange multipliers α_i and setting the derivation of Lagrangian function equal to zero we obtain:

$$\underset{\alpha}{\text{Minimize}}\ \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_i y_j\alpha_i\alpha_j k(x_i,x_j)-\sum_{i=1}^{l}\alpha_i$$

subject to

$$\sum_{i=1}^{l}\alpha_i y_i=0,\ \alpha_i\geq 0\ \ i=1,\ 2,\ 3,\ldots n \tag{7}$$

Where $K(x_i , x_j)= \varphi(x_i)\ \varphi(x_j)$, termed as kernel matrix, is an implicit mapping of the input data into the high dimensional feature space by a kernel function. In this paper we focus on the RBF kernels:

$$K(x_i,x_j)=\phi(x_i).\phi(x_j)=\exp\left(-\left\|x_i-x_j\right\|^2\Big/2\sigma^2\right) \tag{8}$$

For this study the publicly available LIBSVM software [41] with the radial basis function as a kernel is used.

2.4. Multi-Layer Perceptron

MLP is a feed-forward artificial neural network. This network consists of multiple layers of nodes so that two nodes in successive layers are connected with a certain weight (w). The architecture of MLP consists of three layers, input, output, and hidden layer. The number of hidden layers is determined by considering the data and can be more than one. MLP use a supervised learning method to train the network, called back propagation. In supervised learning method, we need to train data so that the desired outputs (d) for them are given. The error in output node j in nth data point is given by following formula:

$$e_j(n)=d_j(n)-y_j(n) \tag{9}$$

Where y is the output that calculated by perceptron. The error in the entire output is given by:

$$\varepsilon(n)=\frac{1}{2}\sum_j e_j^{\ 2}(n) \tag{10}$$

Minimizing this error is goal of this method. For this reason, after processing each data, connection weights are adjusted. For this adjustment, gradient descent is used. The new weight is obtained by following relation:

$$w_{new}(n)=w_{old}(n)-\eta\frac{\partial\varepsilon(n)}{\partial w} \tag{11}$$

That η is learning rate. This parameter determines the step size.

3. RESULTS

In statistical prediction, three cross-validation methods are often used to examine the efficiency of the predictor in practical application: independent dataset test, subsampling test, and jackknife test [42]. Among these methods, the jackknife test can always yield a unique result for a given benchmark dataset and is considered as least arbitrary method. For the independent dataset test, all the samples used to test the predictor are outside the training dataset, used to train it so as to exclude the "memory" effect or bias; however the way of how to select the independent samples to test the predictor could be quite arbitrary unless the number of independent samples is sufficiently large. This kind of arbitrariness might lead to completely different conclusions. For instance, a predictor might achieve a higher success rate than the other predictor for a given independent testing dataset but fail to keep so when tested by another independent testing dataset [42]. For the subsampling test, the concrete procedure usually used in literatures is the 5-fold, 7-fold or 10-fold cross-validation. The problem with this kind of subsampling test is that the number of possible selections in dividing a benchmark dataset is an astronomical figure even for a very simple dataset, as demonstrated by Eqs.28-30 in [32]. Therefore, in any actual subsampling cross-validation tests, only an extremely small fraction of the possible selections are taken into account. Since different selections will always lead to different results even for a same benchmark dataset and a same predictor, the subsampling test cannot avoid the arbitrariness either. Obviously, a test method that is unable to yield a unique outcome cannot be deemed as a good one. In the jackknife test, all the samples in the benchmark dataset will be singled out one-by-one and tested by the predictor trained by the remaining samples. During the process of jackknifing, both the training dataset and testing dataset are actually open, and each sample will be in turn moved between the two. The jackknife test can exclude the "memory" effect. Also, the outcome obtained by the jackknife cross-validation is always unique for a given benchmark dataset, thus the arbitrariness problem as mentioned above for the independent dataset test and subsampling test can be avoided with this method. Accordingly, the jackknife test has been increasingly and widely used by those investigators with strong math background to examine the quality of various predictors (see, e.g., [43-47]). However, to reduce the computational time, we adopted the independent testing dataset fivefold cross-validation in this study was

done by many investigators with SVM as the prediction engine.

In the fivefold cross-validation, the dataset is randomly divided into five subsets with equal samples. With these subsets, each time four subsets are used for training and one subset is used for testing. Therefore the training and testing are performed five times. Finally the average performance is calculated using the definition of accuracy:

$$ACC = (TP + TN) / (TP + TN + FP + FN) \quad (12)$$

In our classification task, the minority class is labeled as positive, and the majority class is labeled as negative. TP, TN, FP and FN are the numbers of true positive, true negative, false positive and false negative, respectively.

Besides accuracy, other parameters such as sensitivity (SEN), specificity (SPEC), Matthew's Correlation Coefficient (MCC) and Area under Curve (AUC) are used to evaluate the performance of the predictor. AUC is a measure that determines the quality of the prediction by calculating the area under Receiver Operating Characteristic (ROC) curve [48]. ROC curve is a graphical plot of the true-positive rate vs. false-positive rate. For the perfect predictor, the AUC is equal to one. Sensitivity, Specificity and Matthew's correlation coefficient are also given by following equations (13-15):

$$SEN = TP / (TP + FN) \quad (13)$$

$$SPEC = TN / (TN + FP) \quad (14)$$

$$MCC = (TP.TN - FP.FN) \Big/ \sqrt{(TP + FN).(TP + FP).(TN + FN).(TN + FP)} \quad (15)$$

Matthew's correlation coefficient is a measure of the quality of binary classifications. This measure has a value between -1 and +1. It is equal to +1 for perfect predictors.

In this study, the concept of PseAAC is used. Since, the dataset, used in this work, are imbalanced, a pre-processing step is applied on the data to balance the class distribution and then, two machine learning methods are used to classify the data.

The results of applying MLP with different learning rate (η) using fivefold cross-validation are shown in Table **2**. As is shown in equation (**11**), η is used in weight updating. According to the results, when MLP is applied with η=0.2, the maximum values of evaluation parameters are obtained. In this condition, the accuracy, Matthew's correlation coefficient and area under curve are 88.16%, 0.763 and 0.941, respectively.

The results of applying SVM with different values of gamma (γ) and cost (C) using fivefold cross-validation are listed in Table **3**. C and γ have much effects on the performance of SVMs classifier. The former parameter is used to control over-fitting and the latter is RBF kernel-specific parameter. Optimum values of these parameters are explored using a grid search algorithm.

By comparing the results, shown in Table **3**, the best performance of SVM classifier is obtained with C=5 and γ=0.02. Using SVM with these optimum parameters, the accuracy, Matthew's correlation coefficient and area under curve are 95.51%, 0.91 and 0.955, respectively.

As can be seen, using the proposed method, a good performance is obtained. By comparing the results, it is clear that, performance of LIBSVM classifier is better than MLP.

Table 2. Performance of MLP Method with Different Learning Rates

Method	ACC	SEN	SPEC	MCC	AUC
MLP (η=0.1)	87.41	86.4	89.5	0.759	0.954
MLP (η=0.2)	88.16	86.5	89.8	0.763	0.941
MLP (η=0.3)	87.76	86.5	89.1	0.755	0.939
MLP (η=0.4)	87.24	85.8	88.6	0.744	0.934

Table 3. Performance of SVM Method with Different Parameters

Method	ACC	SEN	SPEC	MCC	AUC
SVM (γ=0.01, C=5)	94.57	96.2	93.0	0.892	0.946
SVM (γ=0.02, C=5)	95.51	96.4	94.7	0.910	0.955
SVM (γ=0.03, C=5)	95.04	94.4	95.7	0.901	0.950
SVM (γ=0.02, C=1)	94.50	95.1	93.9	0.890	0.945
SVM (γ=0.02, C=10)	95.41	96.1	94.7	0.908	0.954
SVM (γ=0.02, C=15)	95.40	96.1	94.7	0.907	0.954

4. DISCUSSION

Microbial resistance to common antibiotics is an important and increasing concern in human communities [49]. AMPs are gaining popularity as better substitute to antibiotics [50]. Because these peptides generally have properties such as short length, positive charge, fast and effective effect, they become appropriate candidate as new drugs. These peptides are shown to be active against several bacteria, fungi, viruses, protozoa and even cancerous cells [1]. Antibacterial peptides can be effective anti gram negative and positive bacteria. The driving physical forces behind antibacterial activity include net positive charge (enhancing interaction with anionic lipids and other bacterial targets), hydrophobicity (required for membrane insertion and often driven by this process), and flexibility (permitting the peptide to transition from its solution conformation to its membrane-interacting conformation).

Since antibacterial peptides have little similarity in primary and secondary structures, their prediction is difficult. We performed a classification, based on Chou's PseAAC concept for prediction of antibacterial peptides. Recently, this concept has been used by many scientists for prediction of different features of proteins [51-53]. In this study, PseAACs are extracted from sequences and then a preprocessing step is applied to balance the class distribution. Finally, MLP and SVM are used for the classification task. Fivefold cross-validation is used to examine the efficiency of the predictor.

Using MLP with learning rate equal to 0.2, the maximum values of evaluation parameters are obtained. In this condition, the values of accuracy and Matthew's correlation coefficient and area under curve are 88.16%, 0.763 and 0.941, respectively. Also, when SVM classifier with optimum values of gamma (γ=0.02) and cost (C=5) is applied on the data, the values of accuracy and Matthew's correlation coefficient and area under curve are 95.51%, 0.91 and 0.955, respectively.

As can be seen, both of approaches are impressive but the SVM classifier is efficient than MLP to predict the antibacterial peptides. In other word, the results demonstrate that, using the concept of PseAAC and applying SVM, is a successful method to predict the antibacterial peptides. Information derived from PseAAC might be helpful in predicting antibacterial peptides and designing novel therapeutic agents against bacteria.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors [54], we shall make efforts in our future work to provide a web-server for the method presented in this paper.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

Support of this study by University of Isfahan is acknowledged.

REFERENCES

[1] Thomas, S.; Karnik, S.; Barai, R.S.; Jayaraman, V.K.; Idicula-Thomas, S. CAMP: a useful resource for research on antimicrobial peptides. *Nucleic Acids Res.*, **2010**, *38*, D774-780.

[2] Wieprecht, T.; Dathe, M.; Epand, R.M.; Beyermann, M.; Krause, E.; Maloy, W.L.; MacDonald, D.L.; Bienert, M. Influence of the angle subtended by the positively charged helix face on the membrane activity of amphipathic, antibacterial peptides. *Biochemistry*, **1997**, *36*, 12869-12880.

[3] Lata, S.; Sharma, B.K.; Raghava, G.P. Analysis and prediction of antibacterial peptides. *BMC Bioinformatics*, **2007**, *8*, 263.

[4] 4. Brogden, K.A. Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria?. *Nat. Rev. Microbiol.*, **2005**, *3*, 238-250.

[5] Yeaman, M.R.; Yount, N.Y. Mechanisms of antimicrobial peptide action and resistance. *Pharmacol. Rev.*, **2003**, *55*, 27-55.

[6] Ong, P.Y.; Ohtake, T.; Brandt, C.; Strickland, I.; Boguniewicz, M.; Ganz, T.; Gallo, R.L.; Leung, D.Y. Endogenous antimicrobial peptides and skin infections in atopic dermatitis. *N. Engl. J. Med.*, **2002**, *347*, 1151-1160.

[7] Jenssen, H.; Hamill, P.; Hancock, R.E. Peptide antimicrobial agents. *Clin. Microbiol. Rev.*, **2006**, *19*, 491-511.

[8] Loffet, A. Peptides as drugs: is there a market? *J. Pept. Sci.*, **2002**, *8*, 1-7.

[9] van 't Hof, W.; Veerman, E.C.; Helmerhorst, E.J.; Amerongen, A.V. Antimicrobial peptides: properties and applicability. *Biol. Chem.*, **2001**, *382*, 597-619.

[10] Lata, S.; Mishra, N.K.; Raghava, G.P. AntiBP2: improved version of antibacterial peptide prediction. *BMC Bioinformatics*, **2010**, *11 Suppl 1*, S19.

[11] Kamysz, W.; Okroj, M.; Lukasiak, J. Novel properties of antimicrobial peptides. *Acta Biochim. Pol.*, **2003**, *50*, 461-469.

[12] Boman, H.G. Antibacterial peptides: basic facts and emerging concepts. *J. Intern. Med.*, **2003**, *254*, 197-215.

[13] Zasloff, M. Magainins, a class of antimicrobial peptides from Xenopus skin: isolation, characterization of two active forms, and partial cDNA sequence of a precursor. *Proc. Natl. Acad. Sci. USA*, **1987**, *84*, 5449-5453.

[14] Gordon, Y.J.; Romanowski, E.G.; McDermott, A.M. A review of antimicrobial peptides and their therapeutic potential as anti-infective drugs. *Curr. Eye. Res.*, **2005**, *30*, 505-515.

[15] Panyutich, A.; Shi, J.; Boutz, P.L.; Zhao, C.; Ganz, T. Porcine polymorphonuclear leukocytes generate extracellular microbicidal activity by elastase-mediated activation of secreted proprotegrins. *Infect. Immun.*, **1997**, *65*, 978-985.

[16] Zou, D.; He, Z.; He, J.; Xia, Y. Supersecondary structure prediction using Chou's pseudo amino acid composition. *J. Comput. Chem.*, **2011**, *32*, 271-278.

[17] Liu, Y.C.; Yang, M.H.; Lin, W.L.; Huang, C.K.; Oyang, Y.J. A sequence-based hybrid predictor for identifying conformationally ambivalent regions in proteins. *BMC Genomics*, **2009**, *10 Suppl 3*, S22.

[18] Chen, H.; Kihara, D. Effect of using suboptimal alignments in template-based protein structure prediction. *Proteins*, **2011**, *79*, 315-334.

[19] Chen, C.C.; Hwang, J.K.; Yang, J.M. (PS)2-v2: template-based protein structure prediction server. *BMC Bioinformatics*, **2009**, *10*, 366.

[20] Chen, C.; Chen, L.; Zou, X.; Cai, P. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. *Protein Pept. Lett.*, **2009**, *16*, 27-31.

[21] Lee, S.; Lee, B.C.; Kim, D. Prediction of protein secondary structure content using amino acid composition and evolutionary information. *Proteins*, **2006**, *62*, 1107-1114.

[22] Coeytaux, K.; Poupon, A., Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics*, **2005**, *21*, 1891-1900.

[23] Chou, K.C.; Shen, H.B. Recent progress in protein subcellular location prediction. *Anal. Biochem.*, **2007**, *370*, 1-16.

[24] Shen, H.B.; Chou, K.C. Using ensemble classifier to identify membrane protein types. *Amino Acids*, **2007**, *32*, 483-488.

[25] Zhang, S.W.; Chen, W.; Yang, F.; Pan, Q. Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach. *Amino Acids*, **2008**, *35*, 591-598.

[26] Qiu, J.D.; Huang, J.H.; Shi, S.P.; Liang, R.P. Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform. *Protein Pept. Lett.*, **2010**, *17*, 715-722.

[27] Mohabatkar, H. Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein Pept. Lett.*, **2010**, *17*, 1207-1214.

[28] Esmaeili, M.; Mohabatkar, H.; Mohsenzadeh, S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol.*, **2010**, *263*, 203-209.

[29] Mohabatkar, H.; Mohammad Beigi, M.; Esmaeili, A. Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.*, **2011**, *281*, 18-23.

[30] Mohammad Beigi, M.; Behjati, M.; Mohabatkar, H. Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. *J. Struct. Funct. Genomics*, **2011**, *12*, 191-197.

[31] Du, P.; Wang, X.; Xu, C.; Gao, Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.*, **2012**, *425*, 117-119.

[32] Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.*, **2011**, *273*, 236-247.

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

[34] Li, W.; Jaroszewski, L.; Godzik, A. Sequence clustering strategies improve remote homology recognitions while reducing search times. *Protein Eng.*, **2002**, *15*, 643-649.

[35] Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **2001**, *43*, 246-255.

[36] Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **2005**, *21*, 10-19.

[37] Shen, H.B.; Chou, K.C. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.*, **2008**, *373*, 386-388.

[38] Schölkopf, B.; Burges, C.J.C.; Smola, A.J. *Advances in kernel methods support vector learning.* MIT Press: Cambridge, Mass., 1998.

[39] Scholkopf, B.; Smola, A. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press: 2001.

[40] Vapnik, V.N. *The nature of statistical learning theory*. Springer: New York, 1995.

[41] Chang, C.; Lin, C. {LIBSVM}: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

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

[43] Wu, Z.C.; Xiao, X.; Chou, K.C. iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex Gram-positive bacterial proteins. *Protein Pept. Lett.*, **2012**, *19*, 4-14.

[44] Hayat, M.; Khan, A. Discriminating outer membrane proteins with Fuzzy K-nearest Neighbor algorithms based on the general form of Chou's PseAAC. *Protein Pept. Lett.*, **2012**, *19*, 411-421.

[45] Chou, K.C.; Wu, Z.C.; Xiao, X. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.*, **2012**, *8*, 629-641.

[46] Liu, L.; Hu, X.Z.; Liu, X.X.; Wang, Y.; Li, S.B. Predicting protein fold types by the general form of Chou's pseudo amino acid composition: approached from optimal feature extractions. *Protein Pept. Lett.*, **2012**, *19*, 439-449.

[47] Chou, K.C.; Wu, Z.C.; Xiao, X. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PLoS ONE*, **2011**, *6*, e18258.

[48] Bradley, A.P. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognit.*, **1997**, *30*, 1145-1159.

[49] Wang, J.F.; Chou, K.C. Insights from modeling the 3D structure of New Delhi metallo-beta-lactamse and its binding interactions with antibiotic drugs. *PLoS ONE*, **2011**, *6*, e18414.

[50] Wang, P.; Hu, L.; Liu, G.; Jiang, N.; Chen, X.; Xu, J.; Zheng, W.; Li, L.; Tan, M.; Chen, Z.; Song, H.; Cai, Y.D.; Chou, K.C. Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PLoS ONE*, **2011**, *6*, e18476.

[51] Rehman ZU, K.A. Identify GPCRs and their types with Chou's pseudo amino acid composition: an approach from multi-scale energy representation and position specific scoring matrix. *Protein Pept. Lett.*, **2012**, [Epub ahead of print].

[52] Zhao, X.W.; Li, X.T.; Ma, Z.Q.; Yin, M.H. Identify DNA-binding proteins with optimal Chou's amino acid composition. *Protein Pept. Lett.*, **2012**, *19*, 398-405.

[53] Gao, Q.B.; Zhao, H.; Ye, X.; He, J. Prediction of pattern recognition receptor family using pseudo-amino acid composition. *Biochem. Biophys. Res. Commun.*, **2012**, *417*, 73-77.

[54] Chou, K.C.; Shen, H.B. Review: recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.*, **2009**, *1*, 63-92.

Received: March 14, 2012 **Revised: May 16, 2012** **Accepted: May 19, 2012**