Send Orders for Reprints to reprints@benthamscience.ae

The Open Bioinformatics Journal, 2015, *9*, 13-19

Open Access

Using Chou's Pseudo Amino Acid Composition and Machine Learning Method to Predict the Antiviral Peptides

Maryam Zare[1], Hassan Mohabatkar[1,*], Fateme Kazemi Faramarzi[2], Majid Mohammad Beigi[2] and Mandana Behbahani[1]

[1]*Department of Biotechnology, Faculty of Advanced Sciences and Technologies, University of Isfahan, Isfahan, Iran;* [2]*Department of Biomedical Engineering, Faculty of Engineering, University of Isfahan, Isfahan, Iran*

Abstract: Traditional antiviral therapies are expensive, limitedly available, and cause several side effects. Currently, designing antiviral peptides is very important, because these peptides interfere with the key stage of virus life cycle. Most of the antiviral peptides are derived from viral proteins for example peptide derived from HIV-1 capsid protein. Because of the importance of these peptides, in this study the concept of pseudo-amino acid composition (PseAAC) and machine learning methods are used to classify or identify antiviral peptides.

Keywords: Antiviral peptides, machine learning approach, pseudo-amino acid composition.

INTRODUCTION

Antimicrobial peptides exist naturally in all organisms. They have an important role in immune system. These peptides have a broad spectrum of antimicrobial activity [1, 2]. They act as antibacterials; antifungals; antivirals and sometimes anticancers molecules [3-5]. Some peptides with antiviral activity such as C34 from immunodeficiency virus [5], NS5A from hepatites C virus [6] and 7524BVS7 from hepatitis B virus have been previously studied [7]. In the recent decades, many anti-HIV peptides are isolated form HIV-1 proteins and most of them are fusion inhibitor [8]. Some of these peptides have clinical use, for example Enfuvirtide (T20). T20 was isolated from heptads repeat region of gp41 protein [9]. Capsid (p24) is another protein of HIV-1which plays an important role in maturation and viral assembly. It has been demonstrated that corresponding peptides to HIV-1 capsid protein can prevent the spread of viral infection [10].

Antiviral peptides are sometimes better than other antiviral agents, because they have low molecular weight, low toxicity, rapid elimination from the host cells and low side effects [11, 12]. Currently, antiviral peptides have been developed to block virus attachment or entry in to host cells or inhibiting viral replication. Despite the same mechanism of action, antiviral peptides have low sequence homology. Thus, it is difficult to predict the antiviral peptides based on the sequence homology. Due to the cost of computational methods prediction of antiviral peptides is valuable previous to carry out the bench work [13]. Lots of different characters of these molecules can be predicted by variety of computational methods. In fact these methods employ several of protein features for instance amino acid sequence [14, 15], template [16-18] and amino acid composition (AAC) [19, 20]. On the other hand one of the most important and also most difficult problems in computational biology is how to formulate biological sequences with vectors but still considerably keep their sequence-order information. To address this challenging problem, the pseudo amino acid composition (PseAAC) was proposed by Chou. Since the concept of PseAAC or Chou's PseAAC [21] was proposed in 2001, it has rapidly penetrated into almost all the areas of computational proteomics, such as cyclin [22], metalloproteinase family [23], risk type of human papillomaviruses [24], protein quaternary structure [25], Discriminating protein structure classes [26], Predicting anticancer peptides [27], Prediction of bacterial protein subcellular localization [28], predict membrane protein types [29], Predict cysteine S-nitrosylation sites in proteins [30], Identifying the heat shock protein families [31] and Predicting hydroxyproline and hydroxylysine in proteins [32], more example and the like [33-47]. Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA and nucleotides [41, 48-53], as well as other biological samples (see, e.g., [54]). Because it has been widely and increasingly used, in addition to the web-server 'PseAAC' built in 2008, recently three powerful open access soft-wares, called 'PseAAC-Builder' [55], 'propy' [56], and 'PseAAC-General' [39], were established: the former two are for generating various modes of Chou's special PseAAC; while the 3rd one for those of Chou's general PseAAC. Also, similar to PseAAC for protein/peptide sequences, two powerful web-server predictors have also been established to generate pseudo K-tuble nucleotide composition or PseKNC for DNA/RNA sequences. In the present study, concepts of PseAAc and machine learning methods were used for classification and prediction of antiviral peptides.

METHODS

Dataset

In this work two datasets were used. Positive set includes 614 sequences of antiviral peptides and the negative one includes 452 non-antiviral peptides. These sequences were

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

1875-0362/15

chosen from Antiviral Peptides prediction Database (http://crdd.osdd.net/server/avppred). All peptides contain 10 to 100 amino acids. Cd-HIT program was applied to eliminate peptides with 90% similarity [57]. Using Cd-HIT, the number of antiviral peptides was reduced to 342 and the number of negative ones was reduced to 312.

Producing Chou's PseAAC

The concept of Chou's PseAAC was presented in 2001 and then it quickly pierced into many areas of computational proteomics [23, 24, 58-61]. A flexible web server creates a variety of protein PseAACs (http://chou.med.harvard.edu/bioinf/PseAAC). PseAAC of a protein or peptide is shown by more than 20 different factors. The first 20 factors are related to components of their conservative amino acid composition, whereas the extra factors incorporate their sequence order information through a variety of methods [62-64]. Three factors are often used to produce various types of PseAAc: quantitative parameter of amino acid composition, weight factor and grad of correlation. PseAAc back up six amino acid features are (1) hydrophobicity, (2) hydrophilicity, (3) side chain mass, (4) pK1 (alpha-COOH), (5) pK2 (NH3) and (6) pI (Isoelectric point). These characters are applied to value the effect of different locations of amino acid along the peptide sequence.

In this study, type 1of PseAAC or parallel type, λ =1 and weight factor = 0.05 were selected as in Chou`s original paper and similar papers were selected [27, 65]. Applying six characters and their combination produced 126 features of per peptide that were used to classify dataset [66].

Adaboost

Freund and Schapire presented the Adaboost (Adaptive boosting) as a Meta algorithm in 1995 [67]. This algorithm uses combination of simple weak (base) classifiers to construct a strong classifier. In contrast with bagging, this algorithm uses training set re-weighting, instead of re-sampling. Initially, equal weights are assigned to all training samples. In other words, if the training set consist of: $(x_1,y_1),\ldots,(x_m,y_m)$, the initialize weights are $D_1(i)=1/m$, $i=1,\ldots m$. With these weights, a weak classifier is trained on the dataset. Then, depending upon how the classifier is learned the training dataset; the values of weights are updated and this process done frequently in a series of T round.

The weights of incorrectly classified examples are increased in each round. Thereupon the emphasis of new classifier is on the hard examples in the training set. After predefined cycles (T), the prediction class for each sample is obtained by taking a weighted vote of the predictions of each classifier that these weights being commensurate to accuracy of the classifier on its training set.

Selecting the base classifier is one of the factors that have much effect on the performance of Adaboost algorithm. In the present study, three decision trees were used as base classifier; consist of reduced-error pruning tree (REPTree), J48 and Decision Stump. REP Tree uses information gain/variance reduction and prunes to build a decision/regression tree. By using this fast algorithm, the classification process would be less complex. In addition, to find the best sub-tree, the initially grown tree is pruned.

The complexity in the classification process is reduced by using this fast algorithm. In addition, pruning is used to find the best sub-tree of the initially grown tree with the minimum error for the test set [68].

J48 is the C4.5 algorithm that is implemented in the WEKA data mining tool and produces decision trees. This is a standard algorithm that commonly used for practical machine learning [68]. Decision stumps are basically one level decision trees that use a single feature value to prediction. This type of decision trees are not useful in prediction on their own and often used as weak (base) learner in ensemble technique such as bagging and boosting [69].

Moreover, Radial Basis Function (RBF) and Naïve Bayes were used as base classifier. RBFis a feed forward multilayer neural network to classify the non-linearly separable data. This algorithm uses non-linear functions to transform the input feature space into a working feature space. Then, it applies a linear function on the working feature space to produce the output space. This method can be used for classification and/or function approximation problems. The Naïve Bayes algorithms a supervised learning method that exploits the Bayes rule and assumes that attributes of the training set are conditionally independent. This method calculates the maximum posterior probability for each class [70].

RESULTS

To assess the performance of predictors in statistical prediction, three cross-validation methods are used: independent dataset test, sub sampling test, and jackknife test [62]. In this study, we used the five-fold cross-validation for assessing the validity of the proposed predictors. In the five-fold cross-validation, the dataset is divided into two subsets consist of training and testing data in 5 different ways. With these subsets, each time

Four subsets are used for training and one subset is used for testing. The same process is repeated 5 times and the average performance is calculated. In this article, to estimate the performance of the predictor different measures such as overall accuracy (ACC), sensitivity (SEN), specificity (SPEC), Matthew's Correlation Coefficient (MCC), and area under curve (AUC) were used. Overall accuracy is the total accuracy rate of predictor. Sensitivity shows the ability of predictor to correctly classifying the AVPs. Specificity expresses the correct prediction of non-AVPs. Matthew's correlation coefficient is a measure of the quality of binary classifications.

In AUC the area under Receiver Operating Characteristic (ROC) curve is calculated to measure the quality of the prediction [71]. If AUC is equal one, the predictor is perfect [72-76].

These parameters are also given by following equations (1-4) [77]:

$$ACC = (TP + TN) / (TP + TN + FP + FN) \tag{1}$$

$$SEN = TP / (TP + FN) \tag{2}$$

$$SPEC = TN / (TN + FP) \tag{3}$$

$$MCC = (TP.TN - FP.FN) \Big/ \sqrt{(TP+FN).(TP+FP).(TN+FN).(TN+FP)} \quad (4)$$

Where, TP, TN, FP and FN refer to the numbers of true positive (AVPs predicted as AVPs), true negative (non-AVPs predicted as non-AVPs), false positive (non-AVPs predicted as AVPs) and false negative (AVPs predicted as non-AVPs), respectively.

Also, AUC is a measure that determines the quality of the prediction by calculating the area under ROC curve [78]. ROC curve is a graphical plot of the true-positive rate vs. false-positive rate. For the perfect predictor, the AUC is equal one.

To most biologists, unfortunately, the four metrics as formulated in **Eqs. 1-4** are not quite intuitive and easy to understand, particularly the equation for MCC. Here we adopt the formulation proposed recently in [30, 41, 79] based on the symbols introduced by Chou [80, 81] in predicting signal peptides. According to the Chou's formulation, the same four metrics can be expressed as

$$\begin{cases} \mathrm{S_n} = 1 - \dfrac{\mathrm{N}_{-}^{+}}{\mathrm{N}^{-},} \; 0 \le \mathrm{S_n} \le 1 \\ \mathrm{S_p} = 1 - \dfrac{\mathrm{N}_{+}^{-}}{\mathrm{N}^{-},} \; 0 \le \mathrm{S_p} \le 1 \\ \mathrm{A_{cc}} = 1 - \dfrac{\mathrm{N}_{-}^{+} + \mathrm{N}_{+}^{-}}{\mathrm{N}^{-} + \mathrm{N}^{-},} \; 0 \le \mathrm{A_{cc}} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left[\dfrac{\mathrm{N}_{-}^{+}}{\mathrm{N}^{+}} + \dfrac{\mathrm{N}_{+}^{-}}{\mathrm{N}^{+}}\right]}{\sqrt{\left[1 + \dfrac{\mathrm{N}_{-}^{-} - \mathrm{N}_{-}^{+}}{\mathrm{N}^{+}}\right]\left[1 + \dfrac{\mathrm{N}_{-}^{+} - \mathrm{N}_{+}^{-}}{\mathrm{N}^{-} + \mathrm{N}^{-},}\right]}} \; -1 \le \mathrm{MCC} \le 1 \end{cases} \quad (5)$$

Where N^{-} is the total number of the AVPs investigated white N_{-}^{+} ubiquitination peptides incorrectly predicted as the non-AVPs; N^{-} the total number of the non-AVPs investigated while N_{-}^{+} the number of the non-AVPs incorrectly predicted as the AVPs [82]. Now, it is crystal clear from **Eq. 5** that when $\mathrm{N}_{-}^{+} = 0$ meaning none of the AVPs was incorrectly predicted to be a non-AVPs, we have the sensitivity $\mathrm{S_n} = 1$ when $\mathrm{N}_{-}^{+} = \mathrm{N}^{-}$ meaning that all the AVPs were incorrectly predicted as the non- AVPs, we have the ty$\mathrm{S_n} = 0$. Likewise, when $\mathrm{N}_{-}^{+} = 0$ meaning none of the non-AVPs was incorrectly predicted to be the AVPs, we have the specificity$\mathrm{S_p} = 1$; whereas$\mathrm{N}_{-}^{+} = \mathrm{N}^{-}$meaning all the non-AVPs were incorrectly predicted as the AVPs, we have the specificity $\mathrm{S_p} = 0$ when $\mathrm{N}_{-}^{+} = \mathrm{N}_{-}^{+} = 0$ meaning that none of AVPs in the positive dataset S^{+} and none of the non- AVPs in the negative dataset S^{-} was incorrectly predicted, we have the overall accuracy $\mathrm{A_{cc}} = 1$ and $\mathrm{MCC} = 1$; when $\mathrm{N}_{-}^{+} = \mathrm{N}^{+}$and $\mathrm{N}_{+}^{-} = \mathrm{N}^{-}$meaning that all the AVPs in the positive dataset S^{+} and all the non- AVPs in the negative taset S^{-} were incorrectly predicted, we have the overall accuracy $\mathrm{A_{cc}} = 0$ and $\mathrm{MCC} = -1$; whereas when $\mathrm{N}_{-}^{+} = \mathrm{N}^{+}/2$ and $\mathrm{N}_{+}^{-} = \mathrm{N}^{-}/2$ we have $\mathrm{A_{cc}} = 0.5$ and $\mathrm{MCC} = 0$ meaning no better than random prediction. As we can see from the above discussion based on **Eq. 5**, the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient have become much more intuitive and easier-to-understand. It is instructive to point out, however, the set of metrics in **Eqs. 1-4** as well as **Eq. 5** are valid only for the single-label systems. For the multi-label systems, such as those for the subcellular localization of multiplex proteins (see, e.g., [83-95]) where a protein may have two or more locations, and those for the functional types of antimicrobial peptides (see, e.g., [96] where a peptide may possess two or more functional types, a completely different set of metrics is needed as elaborated in [97].

In this work, the concept of PseAAC was applied. Then the Adaboost algorithm with five different base classifiers consists of; RBF, Naïve Bayes, J48, REPTree, and Decision Stump, was used as classier. The results of applying Adaboost with different base classifiers using five-fold cross-validation are shown in Table **1**. According to the results, when Adaboost was applied with J48 as base classifier, the maximum values of evaluation parameters were obtained. In this condition, the accuracy, Matthew's correlation coefficient and area under curve are 93.26%, 0.86 and 0.982, respectively.

The Adaboost algorithm is known as a successful meta-technique to improve the predictive power of classifier. This algorithm constructs a strong classifier as linear combination of base classifiers. Indeed, if individual classifiers make errors on different instances, a strategic combination of these classifiers can reduce the total error. In order to examine this issue, the results of applying these five classifiers to classify the data are shown in Table **2**.

According to the results in Tables **1** and **2**, using Adaboost to combine a set of classifiers by voting, the evaluation parameters are improved in comparison with a single classifier.

In this work, the concept of Pse-AAC was applied. Then the Adaboost algorithm with five different base classifiers consists of; RBF, Naïve Bayes, J48, REPTree, and Decision Stump, was used as classier. Since the number of cycle (T) is an important value for Adaboost method, we used three different T value to assess the effect of this parameter on the accuracy. The results of applying Adaboost with different base classifiers using five-fold cross-validation for different T values are shown in Table **1**. The results showed that the larger T value yields a better classification performance, but the running time and computational complexity are increases. For example, the running time for T=20 is twice the time for T=10. When the number of data is enormous, running time would be problematic, therefore we select T=10 as optimum value in this study. According to the results, when Adaboost was applied with J48 as base classifier, the maximum values of evaluation parameters were obtained. In this condition with T=10, the accuracy, Matthew's correlation coefficient and area under curve are 93.26%, 0.86 and 0.982, respectively.

DISCUSSION

There are a few antiviral drugs to battle against viral infections. Also emergence of drug-resistant strains is not a rare occurrence. Therefore, there are considerable focuses on the new antiviral agents including antiviral peptides. These peptides are preferable due to low toxicity, low molecular weight, rapid elimination from the host cell and selectivity function. Since prediction of antiviral activity of peptides is faster and cheaper than experimental methods, in this study an antiviral peptide prediction method has been designed. Antiviral peptides have very low sequence homology, so concept of PseAAC and Machine Learning Methods have been applied for their classification and prediction.

Table 1. Performance of the Adaboost classifier with different base classifiers for a five-fold cross-validation that measured in terms of Accuracy (ACC), Sensitivity (SEN), Specificity (SPEC), Area under Curve (AUC), and Matthew's Correlation Coefficient (MCC).

	T	ACC	SEN	SPEC	AUC	MCC
	5	71.82	0.682	0.757	0.791	0.44
Adaboost (RBF)	10	74.27	0.732	0.754	0.835	0.49
	20	75.80	0.774	0.741	0.853	0.51
	5	75.04	0.750	0.751	0.813	0.50
Adaboost (Naive Bayes)	10	78.87	0.768	0.812	0.872	0.58
	20	81.01	0.818	0.820	0.894	0.62
	5	91.73	0.915	0.920	0.962	0.83
Adaboost (J48)	**10**	**93.26**	**0.926**	**0.939**	**0.982**	**0.86**
	20	93.26	0.929	0.936	0.984	0.86
	5	69.06	0.653	0.732	0.789	0.38
Adaboost (Decision Stump)	10	72.43	0.765	0.681	0.820	0.45
	20	75.65	0.774	0.738	0.854	0.51
	5	86.98	0.856	0.885	0.938	0.74
Adaboost (REPTree)	10	87.59	0.874	0.879	0.948	0.75
	20	88.51	0.888	0.882	0.957	0.77

Table 2. Performance of five different classifiers for a five-fold cross-validation that measured in terms of accuracy (ACC), Sensitivity (SEN), Specificity (SPEC), Area under Curve (AUC), and Matthew's Correlation Coefficient (MCC).

Classifire	ACC	SEN	SPEC	AUC	MCC
RBF	62.63	0.582	0.674	0.646	0.26
Naive Bayes	61.56	0.482	0.76	0.667	0.25
J48	87.59	0.862	0.891	0.894	0.75
Decision Stump	69.98	0.797	0.594	0.702	0.40
REPTree	83.15	0.821	0.843	0.885	0.66

In the present study, PseAACs was extracted from sequences and then Adaboost algorithm with five different base classifiers, was used for the classification task. Adaboost algorithm adjusts the weights of training samples and produces a classifier as linear combination of weak classifiers. This algorithm is reported as a successful method to improve the accuracy of classifier learning system. The results show that employing PseAAc and Adaboost with J48 as base classifier can be useful in predicting antiviral peptides. Due to the advantages of a web server [98], we try to develop a server based on the method presented in the present paper.

CONFLICT OF INTEREST

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

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of this study by the University of Isfahan.

PATIENT'S CONSENT

Declared none.

REFERENCES

[1] S. Thomas, S. Karnik, R. S. Barai, V.k. Jayaraman, and S. I. Thomas "CAMP: a useful resource for research on antimicrobial peptides", *Nucleic. Acids Res*., vol. 38, pp.774-780, October 2010.

[2] S. Lata, N. K. Mishra, and G. P. Raghava, "AntiBP2: improved version of antibacterial peptide prediction", *BMC Bioinformatics*, vol. 11, pp. s1-s19, January 2010.

[3] B. P. Pelegrini, R. P. Del Sarto, O. N. Silva, R. Pogue, and O. L. Franco, "Antimicrobial peptide from plant: what they are and how they work?", *Biochem. Res. Int.*, vol. 2011, pp. 23-2, January 2011.

[4] A, Tossi ,"Host defense peptide: role and application", *Curr. Protein Pept. Sci.*, vol. 6, pp. 1-3, 2005.

[5] B. Soonthornsata, Y. S. Tian, P. Utachee, S. Supsutthipas, P. Isarangkura-na-Ayathaya, W. Auwanit, and T. Takaji, "Design and evaluation of antiretroviral peptide corresponding to the C-terminal heptad repeat region (C-HR) of human immunodeficiency virus type 1 envelop glycoprotein gp41", *Virology*, vol. 405, pp. 157-164, September 2010.

[6] M. D. Bobardt, G. Chang, L. Witte, S. Selvarajah, U. Chatterji, and F. V. Chisari, "Hepatities C NS5A anchor peptide disrupts human immunodeficiency virus", *Proc. Natl. Acad. Sci.*, vol. 105, pp. 5525-5530, February 2008.

[7] D. H. Kim, Y. Ni, S. H. Lee, S. Urban, and K. H. Han, "An antiviral peptide derived from the preS1 surface protein of hepatitis B virus", *BMB reports,* vol. 41, pp. 640-644, April 2008.

[8] W. Pang, S. C. Tam and Y. T. Zheng, "Current peptide HIV type-1 fusion inhibitor", *Antiviral Chem. Chemother.*, vol. 20, pp. 1-18, April 2010.

[9] T. Jesus, L. Rogelio, C. Abraham, L. Uriel, G. J- Daniel, M. T. Alfonso, and B. B. Lilia, " Prediction of antiviral peptides derived from viral fusion proteins potentially active against herpes simplex and influenza A viruses" *Bioinformation*, vol. 8, pp. 870-874, September 2012.

[10] H. Zhang, F. Curreli, X. Zhang, S. A. Bhattacharya, A. Waheed, A. Cooper, D. O. Cowburn, E. K. Freed, and A. Debnath, " Antiviral activity of a-helical stapled peptides designed from the HIV-1 capsid dimerization domain" *Retrovirology*, vol. 8, pp. 28-46, 2010.

[11] N. Thakur, A. Qureshi, and K. Manoj, "AVP: Collection and Prediction of highly effective antiviral peptides", *Nucleic. Acids Res.*, vol. 40, pp. 199-204, April 2012.

[12] Z. Wang, and G. Wang, "APD: the antimicrobial peptide database", *Nucleic. Acids Res.*, vol. 32, pp. 590-592, September 2003.

[13] S. Lata, B. Sharma, and G. P. Raghava, "Analysis and prediction of antibacterial peptides", *BMC Bioinformatics*, vol. 8, pp. 263-273, July 2007.

[14] Y. C. Liu, M. H. Yang, W.L. Lin, C. K. Huang, and Y.J. Oyang, "A sequence-based hybrid predictor for identifying conformationally ambivalent regions in proteins", *BMC Genomics*, vol. 10, pp. s1-s22, December 2009.

[15] D, Zou, Z. He, J. He, and Y. Xia, "Super secondary structure prediction using Chou's pseudo amino acid composition", *J. Comput. Chem.*, vol. 32, pp. 271-278, January 2011.

[16] H. Chen, and D. Kihara, "Effect of using suboptimal alignments in template-based protein structure prediction", *Proteins,* vol. 79, pp. 315-334, January 2011.

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

[18] C. C. Chen, J. K. Hwang, and J. M. Yang, "(PS) 2-v2: template-based protein structure prediction server", *BMC Bioinformatics*, vol. 1, pp. 366, October 2009.

[19] K. Coeytaux, and A. Poupon," Prediction of unfolded segments in a protein sequence based on amino acid composition", *Bioinformatics*, vol. 21, pp. 1891-1900, January 2005.

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

[21] S. X. Lin, and J. Lapointe, "Theoretical and experimental biology in one —A symposium in honour of Professor Kuo-Chen Chou’s 50th anniversary and Professor Richard Giegé’s 40th anniversary of their scientific careers*", J. Biomed. Sci. Eng.*, vol. 6, pp. 435-442, April 2013.

[22] H. Mohabatkar, "Prediction of Cyclin Proteins Using Chou’s Pseudo Amino Acid Composition", *Protein Pept. Lett.*, vol. 17, pp. 1207-1214, October 2010.

[23] M. Mohammad Beigi, M. Behjati, and H. Mohabatkar, " Prediction of metalloproteinase family based on the concept of Chou's pseudoamino acid composition using a machine learning approach", *Struct. Funct. Genomics*, vol. 12. pp. 191-197, November 2011.

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

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

[26] M. Hayat, and N. Iqbal, "Discriminating protein structure classes by incorporating Pseudo Average Chemical Shift to Chou's general PseAAC and Support Vector Machine", *Comput. Methods Programs Biomed.*, vol. 116, pp. 184-192, October 2014.

[27] Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani, and H. Mohabatkar, "Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test", *J. Theor. Biol.*, vol. 341, pp. 34-40, January 2014.

[28] L. Li, S. Yu, W. Xiao, Y. Li, M. Li, L. Huang, X. Zheng, S. Zhou, and H. Yang," Prediction of bacterial protein subcellular localization by incorporating various features into Chou's PseAAC and a backward feature selection approach", *Biochimie.*, vol. 104, pp. 100-107, September 2014.

[29] G. Han, S. Z. G. Yu, and V. Anh, "A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of Chou's PseAAC", *J. Theor. Biol.*, vol. 344, pp. 31-39, March 2014.

[30] Y. Xu, J. Ding, and L. Y. Wu," iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition", *PLoS ONE*, vol. 8, pp. e55844, February 2013.

[31] P. M. Feng, W. Chen, and H. Lin, "iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition", *Anal. Biochem.*, vol. 442, pp. 118-125, November 2013.

[32] Y. Xu, X. Wen, X. J. Shao, and N. Y. Deng, "iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition", *Int. J. Mol. Sci.*, vol. 15, pp. 7594-7610, 2014

[33] L. Kong, L. Zhang, and J. Lv, "Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition", *J. Theor. Biol.*, vol. 344, pp. 12-18, March 2014.

[34] S. Mondal, and P. P Pai, "Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction", *J. Theor. Biol.*, vol. 356, pp. 30-35, September 2014.

[35] L. Zhang, X. Zhao, and L. Kong, "Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition" *J. Theor. Biol.*, vol. 355, pp. 105-110, March 2014.

[36] Y. C. Zuo, Y. Peng, L. Liu, W. Chen, L.Yang, and G. L. Fan, "Predicting peroxidase subcellular location by hybridizing different descriptors of Chou's pseudo amino acid patterns", *Anal. Biochem.*, vol. 458, pp. 14-19, August 2014.

[37] C. Jia, X. Lin, and Z. Wang, "Prediction of Protein S-Nitrosylation Sites Based on Adapted Normal Distribution Bi-Profile Bayes and Chou's Pseudo Amino Acid Composition", *Int. J. Mol. Sci.*, vol. 15, pp. 10410-10423, May 2014.

[38] L. Nanni, S. Brahnam, and A. Lumini, "Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition", *J. Theor. Biol.*, vol. 360C, pp. 109-116, November 2014.

[39] P. Du, S. Gu, and Y. Jiao, "PseAAC-General: Fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets", *Int. J. Mol. Sci.*, vol. 15, pp. 3495-3506, February 2014.

[40] J. Zhang, X. Zhao, P. Sun, and Z. Ma, "PSNO: Predicting Cysteine S-Nitrosylation Sites by Incorporating Various Sequence-Derived Features into the General Form of Chou's PseAAC", *Int. J. Mol. Sci.*, vol. 15, pp. 11204-11219, May 2014.

[41] W. Chen, P. M. Feng, H. Lin, and K. C. Chou, " iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition", *Nucleic. Acids Res.*, vol. 41, pp. e68, December 2013.

[42] D. N. Georgiou, T. E. Karakasidis, and A. C. Megaritis, "A short survey on genetic sequences, Chou's pseudo amino acid composition and its combination with fuzzy set theory" *Open Bioinform J.*, vol. 7, pp. 41-48; 2013.

[43] H. Ding, E. Z. Deng, L. F. Yuan, L. Liu, and H. Lin, "iCTX-Type: A sequence-based predictor for identifying the types of conotoxins

in targeting ion channels" *Biomed. Res. Int.*, vol. 2014, pp. 286419, May 2014.

[44] B. Liu, J. Xu, X. Lan, R. Xu, and J. Zhou, "iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition" *PLoS ONE*, vol. 9, pp. e106691, September 2014.

[45] W. R Qiu, X. Xiao, and W. Z. Lin, "iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach", *Biomed. Res. Int.*, vol. 2014, pp. 947416, April 2014.

[46] Y. Xu, X. Wen, L. S. Wen, and L. Y. Wu, "iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition", *PLoS ONE*, vol. 9, pp. e105018, Agust 2014.

[47] Y. N. Fan, X. Xiao, and J. L. Min, "iNR-Drug: Predicting the interaction of drugs with nuclear receptors in cellular networking", *Int. J. Mol. Sci.*, vol. 15, pp. 4915-4937, February 2014.

[48] W. R. Qiu, X. Xiao, and K. C. Chou, "iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components" *Int. J. Mol. Sci.*, vol. 15, pp. 1746-1766, January 2014.

[49] S. H. Guo, E. Z. Deng, L. Q. Xu, H. Ding, and H. Lin, "iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition", *Bioinformatics*, vol. 30, pp. 1522-1529, February 2014.

[50] W. Chen, P. M. Feng, and H. Lin, "iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition" *Biomed. Res. Int.*, vol. 2014, pp. 623149, April 2014.

[51] W. Chen, P. M. Feng, and E. Z. Deng, "iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition", *Anal. Biochem.*, vol. 462, pp. 76-83, October 2014.

[52] W. Chen, X. Zhang, J. Brooker, and H. Lin, "PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions", *Bioinformatics*, vol. 2014, pp. 9-25, August 2014.

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

[54] Y. Jiang, T. Huang, L. Chen, and Y. F. Gao, "Signal propagation in protein interaction network during colorectal cancer progression", *Biomed. Res. Int.*, vol. 2013, pp. 287019, February 2013.

[55] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou'spseudo-amino acid compositions", *Anal. Biochem.*, vol. 425, pp. 117-119, June 2012.

[56] D. S. Cao, Q. S. Xu, and Y. Z. Liang, "propy: a tool to generate various modes of Chou's PseAAC", *Bioinformatics*, vol. 29, pp. 960-962, February 2013.

[57] W. Li, and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences", *Bioinformatics*, vol. 22, pp. 1658-1659, April 2006.

[58] D. N. Georgiou, T. E. Karakasidis, J. J. Nieto, and A. Torres, "Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition", *J. Theor. Biol.*. Vol. 257, pp. 17-26, March 2009.

[59] M. K. Gupta, R. Niyogi, and M. Misra, "An alignment-free method to find similarity among protein sequences *via* the general form of Chou's pseudo amino acid composition", *SAR QSAR Environ. Res.*, vol. 24, pp. 597-609, January 2013.

[60] H. Mohabatkar, M. M. Beigi, K. Abdolahi, and S. Mohsenzadeh, "Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach", *Med. Chem*., vol. 9, pp. 133-137, February 2013.

[61] H. Mohabatkar, M. Mohammad Beigi, and A. Esmaeili, "Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine", *J. Theor. Biol.*, vol. 281, pp. 18-23, July 2011.

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

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

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

[65] K.C. Chou, and H.B. Shen, “Recent progress in protein subcellular location prediction”, *Anal. Biochem.*, vol. 370, pp. 1-16, 2007.

[66] K. Khosravian, F. Kazemi Faramarzi, M. Mohammad Beigi, M. Behbahani, and H. Mohabatkar. "Predicting antibacterial peptides by the concept of Chou’s pseudo-amino acid composition and machine learning methods", *Protein Pept. Lett.,* vol. 20, pp.1-7, February 2013.

[67] Y. Freund, and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", *J. Comput. Syst. Sci.*, vol. 55, pp.119-139, August 1997.

[68] V. Ramesh, P. Parkavi, and P. Yasodha, "Performance analysis of data mining techniques for placement chance prediction", *Int J. Sci. Eng. Res.,* vol. 2, August 2011.

[69] W. Iba, and P. Langley, "Induction of One-Level Decision Trees", In: *Proc of the 9th International Workshop on Machine Learning*. 1992, pp. 233-240.

[70] S. Ali, and K. A. Smith, "On learning algorithm selection for classification", *Appl. Soft. Comput.*, vol. 6, pp. 119-138, January 2006.

[71] K. C. Chou, and C. T. Zhang, "Review: Prediction of protein structural classes", *Crit. Rev. Biochem. Mol. Biol.*, vol. 3, pp. 275-349,1995.

[72] K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review)", *J. Theor. Biol.* , vol. 273, pp. 236-247, March 2011.

[73] Z. C. Wu, X. Xuan, and K. C. Chou, "iLoc-Gpos: A multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins", *Protein Pept. Lett.*, vol. 19, pp. 4-14, January 2012.

[74] M. Hayat, and A. Khan, "Discriminating outer membrane proteins with fuzzy K-Nearest neighbor algorithms based on the general form of Chou's PseAAC", *Protein Pept. Lett.*, vol. 19, pp. 411-421, February 2012.

[75] K. C. Chou, Z. C. Wu, and X. Xiao, "iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites", *Mol. Biosyst.*, vol. 8, pp. 629-641, Novamber 2012.

[76] L. Liu, H. X. Zhen, L. Xing-Xing, W. Ying, and L. Shao-Bo "Predicting protein fold types by the general form of Chou's pseudo amino acid composition: approached from optimal feature extractions", *Protein Pept. Lett*, vol. 19, pp. 439-449, February 2012.

[77] Y. Fang, Y. Guo, Y. Feng, and M. Li, "Predicting DNA-binding proteins: approached from Chou’s pseudo amino acid composition and other specific sequence features", *Amino Acids*, vol. 34, pp. 103–109, January 2008.

[78] T. Fawcett, "An introduction to ROC analysis", *Pattern Recognit. Lett.*, vol. 27, pp. 861-874, June 2006.

[79] Y. Xu, X. J. Shao, L. Y. Wu, and N. Y. Deng, "iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins" *Peer J.*, vol. 1, pp. e171, October 2013.

[80] K. C. Chou, "Using subsite coupling to predict signal peptides", *Protein Eng. Des. Sel.*, vol. 14, pp. 75-79, December 2000.

[81] K. C. Chou, "Prediction of signal peptides using scaled window", *Peptides*, vol. 22, pp. 1973-1979, December 2001.

[82] K. C. Chou, "Prediction of protein signal sequences and their cleavage sites", *Proteins: Struct., Funct., Bioinf.*, vol. 42, pp. 136-139, Jaunary 2001.

[83] K. C. Chou, and H. B. Shen, "Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms (updated version: Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms)", *Nat. Sci.*, vol. 2, pp. 1090-1103, September 2010.

[84] K. C. Chou, and H. B. Shen, "Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites", *J. Proteome Res.*, vol. 6, pp. 1728-1734, March 2007.

[85] H. B. Shen, "Gneg-mPLoc: A top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins", *J. Theor. Biol.* , vol. 264, pp. 326-333, May 2010.

[86] H. B. Shen, "Gpos-mPLoc: A top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins", *Protein Pept. Lett.*, vol. 16, pp. 1478-1484, December 2009.

[87] H. B. Shen, "Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites", *Biochem. Biophys. Res. Commun.*, vol. 355, pp. 1006-1011, April 2007.

[88] H. B. Shen, "A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0", *Anal. Biochem.*, vol. 394, pp. 269-274, November 2009.

[89] H. B. Shen, "Virus-mPLoc: A Fusion Classifier for Viral Protein Subcellular Location Prediction by Incorporating Multiple Sites", *J. Biomol. Struct. Dyn.* (JBSD), vol. 28, pp. 175-186, February 2010.

[90] K. C. Chou, and H. B. Shen, "A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0" *PLoS ONE*, vol. 5, pp. e9931, April 2010.

[91] K. C. Chou, and H. B. Shen, "Plant-mPLoc: A top-down strategy to augment the power for predicting plant protein subcellular localization" *PLoS ONE*, vol. 5, pp. e11335, 2010.

[92] X. Xiao, and Z. C. Wu, "iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites" *J. Theor. Biol.* , vol. 284, pp. 42-51, September 2011.

[93] Z. C. Wu, and X. Xiao, "iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites", *Mol. Biosyst.*, vol. 7, pp. 3287-3297, October 2011.

[94] W. Z. Lin, J. A. Fang, and X. Xiao, "iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins", *Mol. Biosyst.*, vol. 9, pp. 634-644, Jaunary 2013.

[95] K. C. Chou, Z. C. Wu, and X. Xiao, "iLoc-Euk: A multi-labelclassifier for predicting the subcellular localization of singleplex andmultiplex eukaryotic proteins", *PLoS ONE*, vol. 6, pp. e18258, March 2011.

[96] X. Xiao, P. Wang, W. Z. Lin, and J. H. Jia, "iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types", *Anal. Biochem.,* vol. 436, pp. 168-177, May 2013.

[97] K. C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystem", *Mol. Biosyst.*, vol. 9, pp. 1092-1100, February 2013.

[98] K. C. Chou, and H. B. Shen, "Review: recent advances in developing web-servers for predicting protein attributes", *Nat. Sci.*, vol. 2, pp. 63-92, August 2009.

Received: September 18, 2014 Revised: December 05, 2014 Accepted: December 23, 2014