International Journal of Biomathematics
Vol. 6, No. 2 (2013) 1350003 (14 pages)
© World Scientific Publishing Company
DOI: 10.1142/S1793524513500034

PREDICTING SUBCHLOROPLAST LOCATIONS OF PROTEINS BASED ON THE GENERAL FORM OF CHOU'S PSEUDO AMINO ACID COMPOSITION: APPROACHED FROM OPTIMAL TRIPEPTIDE COMPOSITION

HAO LIN[*,§], CHEN DING[*], LU-FENG YUAN[*], WEI CHEN[†,¶],
HUI DING[*], ZI-QIANG LI[‡], FENG-BIAO GUO[*],
JIAN HUANG[*] and NI-NI RAO[*]

[*]*Key Laboratory for NeuroInformation of Ministry of Education*
Center of Bioinformatics, School of Life Science and Technology
University of Electronic Science and Technology of China
Chengdu 610054, P. R. China

[†]*Center for Genomics and Computational Biology*
Department of Physics, College of Sciences
Hebei United University
Tangshan 063000, P. R. China

[‡]*School of Information and Engineering*
Sichuan Agricultural University
Yaan 625014, P. R. China
[§]*hlin@uestc.edu.cn*
[¶]*chenwei_imu@yahoo.com.cn*

Received 10 March 2012
Accepted 20 December 2012
Published 28 February 2013

Chloroplasts are organelles found in plant cells that conduct photosynthesis. The subchloroplast locations of proteins are correlated with their functions. With the availability of a great number of protein data, it is highly desired to develop a computational method to predict the subchloroplast locations of chloroplast proteins. In this study, we proposed a novel method to predict subchloroplast locations of proteins using tripeptide compositions. It first used the binomial distribution to optimize the feature sets. Then the support vector machine was selected to perform the prediction of subchloroplast locations of proteins. The proposed method was tested on a reliable and rigorous dataset including 259 chloroplast proteins with sequence identity $\leq 25\%$. In the jack-knife cross-validation, 92.21% envelope proteins, 93.20% thylakoid membrane, 52.63% thylakoid lumen and 85.00% stroma can be correctly identified. The overall accuracy achieves 88.03% which is higher than that of other models. Based on this method, a predictor called ChloPred has been built and can be freely available

[§,¶]Corresponding authors.

from http://cobi.uestc.edu.cn/people/hlin/tools/ChloPred/. The predictor will provide important information for theoretical and experimental research of chloroplast proteins.

Keywords: Subchloroplast localization; tripeptide; binomial distribution; support vector machine.

Mathematics Subject Classification 2010: 92D20

1. Introduction

The chloroplast is one of key organelles in green plant cells. It houses the machinery necessary for photosynthesis, amino acid biosynthesis, pigment biosynthesis and so on [2]. The chloroplast is divided into four parts: stroma, thylakoid lumen, thylakoid membrane and envelope according to their structures and functions [2]. The proteins located in these four subchloroplast locations play different biological roles. The stroma is an internal space enclosed by the chloroplast double membrane but excluding the thylakoid. It contains one or more small circular DNA, some ribosomes and some temporary products of photosynthesis. The thylakoid membrane, an internal system of interconnected membranes, carries out the light reactions of photosynthesis. The thylakoid lumen is the chloroplast compartment bounded by the thylakoid membranes. The chloroplast envelope comprises the inner and outer chloroplast membrane.

For timely understanding protein functions and realizing the process of photosynthesis, it needs to accurately identify the subchloroplast location of chloroplast proteins. Unfortunately, it is both time-consuming and costly for experimental approach to confirm proteins location in chloroplast. Phylogenetic tree is a traditional method for most experimental scholars to predict the sub-subcellular locations of proteins. Although this method is not particularly expensive, it is more time consuming than machine learning approaches. Furthermore, for the sequences which do not have homologue sequences in benchmark data, phylogenetic tree will produce ineffective, inexact and even wrong information. In the past several years, lots of works have been proposed for protein subcellular localization prediction [4, 5, 10, 12–15, 32, 34–36, 38, 41, 43, 45, 57, 59, 64, 65, 69, 68, 70, 74, 79, 75, 81]. In parallel with these theoretical methods, large numbers of proteins have been sequenced and annotated which promote the developments of machine learning approaches to predict and annotate chloroplast proteins [20, 29, 58, 61]. For example, Emanuelsson *et al.* [21] have developed a predictor called ChloroP to predict chloroplast transit peptides and their cleavage sites. Tung *et al.* [61] proposed a Random Forest model to predict of protein subchloroplast locations. Recently, Du *et al.* [20] have used the pseudo-amino acid composition (PseAAC) to predict subchloroplast locations of proteins and developed a server, called SubChlo. Overall accuracy (OA) of jackknife test is 67.18% for the dataset with the sequence identity of 60%. Based on the same benchmark data, Shi *et al.* [58] have improved the accuracy to 89.31% by using discrete wavelet transform to exact feature. However, many proteins with just about 40% sequence identity might be homologous to each other. It has been proved that there is a close relationship between predictive accuracy and sequence identity [50, 72]. High similarity data can surely lead to overestimation of the performance of the methods considered.

The present study was dedicated to develop a new and more powerful predictor, called ChloPred, for predicting subchloroplast localization of proteins. According to a recent comprehensive review [9], 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 samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute 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 one-by-one.

2. Materials and Methods

2.1. *Dataset*

Both amino acid sequences and annotation information of chloroplast proteins were extracted from universal protein resource (Uniprot) [60]. To construct a reliable benchmark dataset, the following steps were used to prepare high quality datasets: (1) Although proteins with multiple subchloroplast locations have some special biological functions, we collected the proteins with only one subchloroplast location because the number of proteins with multiple subchloroplast locations is too small to have statistical significance. (2) Proteins with ambiguous protein existence annotations, such as “uncertain”, “predicted” and “inferred from homology” were excluded because they lack confidence. (3) Only those proteins with experimental confirmed subchloroplast location were included because they can provide correct and validated information. (4) The sequences which are fragment of other proteins were excluded because their information is redundant and not integral. (5) Sequences containing nonstandard letters, such as “B”, “X” or “Z”, were excluded because their meanings are ambiguous. (6) To avoid any homology bias, the proteins with $>25\%$ sequence identity to any other in the same subset were excluded using PISCES [63]. After strictly following the above procedures, we finally obtained 259 proteins including 60 stroma proteins, 19 thylakoid lumen proteins, 103 thylakoid membrane proteins and 77 envelope proteins.

2.2. *Tripeptide compositions*

It is one of the most important parts for pattern recognition to generate a set of informative parameters. To avoid losing many important information hidden in protein sequences, the PseAAC was proposed to replace the simple amino acid composition (AAC) for representing the sample of a protein [6, 7].

For a brief introduction about Chou's PseAAC, visit the Wikipedia web-page at http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition. For a summary about its recent development and applications, see a comprehensive review [8]. Ever since the concept of PseAAC was proposed by Chou [6] in 2001, it has rapidly penetrated into almost all the fields of protein attribute prediction, such as predicting protein structural classes [37, 56], predicting protein quaternary structure [76], identifying bacterial virulent proteins [52], identifying cell wall lytic enzymes [18], identifying risk type of human papillomaviruses [22], identifying DNA-binding proteins [24], predicting homo-oligomeric proteins [55], predicting protein secondary structure content [3], predicting supersecondary structure [83], predicting enzyme family and sub-family classes [54, 66, 82], predicting protein subcellular location [35, 36, 80], predicting subcellular localization of apoptosis proteins [32, 35, 44, 19], predicting protein subnuclear location [33], predicting protein submitochondria locations [75, 51], predicting G-Protein-Coupled Receptor Classes [27, 53], predicting protein folding rates [28], predicting outer membrane proteins [39], predicting cyclin proteins [48], predicting GABA(A) receptor proteins [49], identifying bacterial secreted proteins [73], identifying the cofactors of oxidoreductases [77], identifying lipase types [78], identifying protease family [30], predicting Golgi protein types [17], classifying amino acids [26], among many others.

Recently, Anishetty *et al.* [1] demonstrated that the tripeptide may be used to predict plausible structures for oligopeptides and denovo protein design. Tripeptide motifs represent potentially important starting points for design of small molecule biological modulators [62]. Thus, tripeptide composition was employed to encode chloroplast protein sequences in this study. Actually, like dipeptide composition [40, 42], tripeptide composition, tetrapeptide composition, pentapeptide composition *et al.* are just different modes of Chou's PseAAC. According to the general form of Chou's (PseAAC) (see [8, Eq. 6]), the general form of Chou's PseAAC can be formulated as

$$P = [\psi_1, \psi_2, \ldots, \psi_i, \ldots, \psi_\Omega]^T, \tag{2.1}$$

where T is a transpose operator, while the subscript Ω is an integer and its value as well as the components $\psi_1, \psi_2, \ldots$ will depend on how to extract the desired information from the amino acid sequence of P. Based on the above general equation, for the general tripeptide composition, a chloroplast protein with length of L can be characterized as an $\Omega = 20 \times 20 \times 20 = 8000$ dimension feature vector and described as follows:

$$F_{8000} = [f_1, f_2, \ldots, f_i, \ldots, f_{8000}]^T, \tag{2.2}$$

here symbol T denotes the transposition of vector. f_i is the frequency of the ith-tripeptide and expressed as:

$$f_i = n_i \Bigg/ \sum_{i=1}^{8000} n_i = n_i/(L-2), \tag{2.3}$$

here n_i and L denote the number of the ith-tripeptide and length of the protein, respectively.

2.3. *Feature selection*

In machine learning problems, to avoid the high-dimensional problems such as "dimension disaster", overfitting or redundancy [67], dimensionality reduction is an important technique for removing irrelevant features (or redundant features) and building robust models. Some algorithms such as principal component analysis [46], minimal-redundancy-maximal-relevance (mRMR) [31], diffusion Maps [71] and the analysis of variance (ANOVA) [40] have been proposed for reducing the dimensionality. This study will introduce a new algorithm based on binomial distribution to optimize the feature sets [25]. Eight thousands kinds of tripeptides may occur in four classes of chloroplast protein dataset. Each kind of tripeptide occurring in one type may be a stochastic event. Then, the probability of the ith-tripeptide occurring in the jth-class ($j =$ stroma, thylakoid lumen, thylakoid membrane and envelope) can be defined by:

$$\mathrm{CL}_{ij} = 1 - \sum_{n=n_{ij}}^{N_i} \frac{N_i!}{n!(N_i - n)!} p_j^n (1 - p_j) N_i - n, \tag{2.4}$$

here probability CL_{ij} is also called the confidence level (CL) of ith-tripeptide in jth-class. N_i denotes the total number of ith-tripeptide in the dataset. n_{ij} denotes the occurrence number of ith-tripeptide in jth-class. The sum is taken from n_{ij} to N_i. The probability p_j is the relative frequency of class j in the dataset and defined as:

$$p_j = \sum_{i=1}^{8000} n_{ij} \Bigg/ \sum_{i=1}^{8000} N_i, \tag{2.5}$$

here $\sum_{i=1}^{8000} N_i$ and $\sum_{i=1}^{8000} n_{ij}$ are the total occurrence number of all tripeptides in the dataset and in jth-class proteins, respectively.

If there are Ω tripeptides whose CL_{ij} is larger than a given cutoff CL_o, the frequencies of these tripeptides are selected as optimized features expressed as:

$$F_\Omega = [f_1, f_2, \ldots, f_i, \ldots, f_\Omega]^T. \tag{2.6}$$

If CL_o is set to zero, 8000 tripeptides are all selected. If $\mathrm{CL}_o > 1$, no tripeptides are selected. For different cutoff threshold of CL_o, the value of Ω will be different. Based on CL (Eq. (2.3)), high-dimensional data can be projected into low-dimensional space. The final Ω will be determined by cross-validation.

2.4. *Support vector machine*

Support vector machine (SVM) is a wonderful and popular machine learning method based on statistical learning theory. Because of its easy-to-use and good performance, SVM has been widely applied in protein bioinformatics. For multiclass problems, several strategies such as one-versus-rest (OVR) and one-versus-one (OVO) can be used to extend the traditional SVM. This paper adopts OVO strategy for multi-class classification. The software toolbox used to implement SVM is Libsvm written by Lin's lab and can be freely downloaded from: http://www.csie.ntu.edu.tw/∼cjlin/libsvm [23]. Usually, four kinds of kernel functions, i.e. linear function, polynomial function, sigmoid function and radial basis function (RBF), are applied to perform predictions. Empirical studies have demonstrated that the RBF outperforms the other three kinds of kernel functions. Hence, we used the RBF to perform the prediction. The grid search program was applied to optimize the regularization parameter C and kernel parameter γ using five-fold cross-validation.

2.5. *Performance evaluation*

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, sub-sampling test and jack-knife test [16]. However, of the three test methods, the jack-knife test is deemed the least arbitrary and objective as elucidated in [9] and demonstrated by Eqs. (28)–(31) therein. Accordingly, the jack-knife test has been widely and increasingly used to examine the power of various statistical predictors [75, 41, 43, 47, 35, 18, 22, 3, 54, 82, 27, 48]. Thus the jack-knife cross-validation was used to evaluate the performance of the proposed model. Two important evaluating parameters: sensitivity (Sn) and OA were calculated as the following formulas:

$$\mathrm{Sn}_i = TP_i/N_i, \tag{2.7}$$

$$\mathrm{OA} = \sum_i TP_i/N, \tag{2.8}$$

here TP_i and N_i are the numbers of correctly predicted proteins and total number of the ith-class, respectively. N is the total number of four classes of proteins in the dataset.

3. Results and Discussion

3.1. *Prediction accuracy*

The specific tripeptides can be selected by using Eq. (2.4). In our statistics, only tripeptides with $N_i \geq 3$ are considered, because occurrence of a tripeptide with $N_i < 3$ in chloroplast proteins is an event of small probability ($p < 0.0001$). Therefore, we selected the tripeptides with different confidence levels under the constraint $N_i \geq 3$. There are 6723 tripeptides with $N_i \geq 3$ in the benchmark dataset.

In general, the tripeptide sets with high CL give more reliable information for classification. However, the number of these words is too small to afford enough information, which deduces the poor predictive accuracy. For example, using $> 99.9\%$ as CL, we can achieve 28 tripeptides. But the OA is only 52.51% in five-fold cross-validation. In contrast, the tripeptide sets with low confidence contains too many components. But it would reduce the cluster-tolerant capacity so as to lower down the cross-validation accuracy. For instance, 6510 tripeptides with $> 50\%$ of CL can only produce the OA of 53.28% in five cross-validation. Therefore, using appropriate tripeptide sets would yield a prediction with higher accuracy. By changing the cutoff of CL, we can obtain a series of tripeptide sets. For economizing time and improving efficiency, we first used five-fold cross-validation to optimize the regularization parameter C and kernel parameter γ. The three dimension graph for feature dimension, CL and OA is shown in Fig. 1. It exhibits that the five-cross-validated accuracy increases to 87.26% when using $> 97.04\%$ as CL. The optimized tripeptide set contains 571 dimension feature vector. The regularization parameter C and kernel parameter γ are 512 and 0.0078125, respectively. The numbers of tripeptide with this CL are 122, 167, 105 and 178, respectively for envelope, thylakoid membrane, thylakoid lumen and stroma.

Furthermore, we examined the jack-knife-cross-validated accuracy using 571 dimension features. Results are recorded in Table 1. As it can be seen from Table 1, 92% (71/77) envelope, 93% (96/103) thylakoid membrane, 53% (10/19) thylakoid lumen and 85% (51/60) stroma proteins can be correctly predicted. OA achieves 88.03%. It should be noticed that the accuracy of thylakoid lumen is dramatically lower than that of another three classes. The reason is that the benchmark data is unbalance and fewer features (105) are selected from thylakoid lumen proteins.

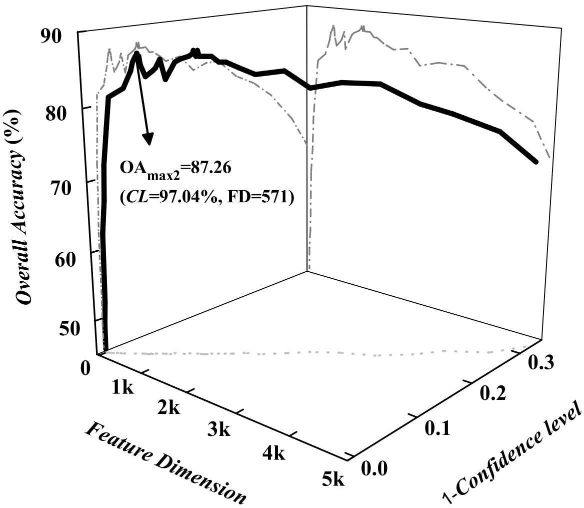

Fig. 1. (Color online) The graph for predicting subchloroplast locations of proteins. Dark line denotes 3D curve. Three gray lines are projections on three planes (OA/feature dimension plane, OA/confidence level plane, confidence level/feature dimension plane).

Table 1. The comparison of performance of proposed model and other models.

	Accuracy (%)				
	Envelope	Thylakoid membrane	Thylakoid lumen	Stroma	Overall
SVM (571 tripeptides)	92.21	93.20	52.63	85.00	88.03
SVM (400 dipeptides)	55.84	78.64	0	35.00	55.98
SVM (20 AAs)	45.45	88.35	0	25.00	53.28
SVM (20 AAs + 400 dipeptides)	57.14	70.87	0	48.33	56.37
SVM (PseAAC)	66.23	75.73	10.53	43.33	60.62
Naïve Bayes (576 tripeptides)	85.71	70.87	0	55.00	66.41
Naïve Bayes (PseAAC)	38.96	49.51	26.32	60.00	47.10
RBF Network (560 tripeptides)	80.52	54.37	52.63	81.67	68.34
RBF Network (PseAAC)	59.74	67.96	10.53	36.67	54.05

With the rapid expansion of the chloroplast protein dataset, more tripeptides with a higher CL will be obtainable, making the prediction more accurate.

3.2. *Comparison accuracies*

It is necessary to investigate whether the proposed method has a better performance than other existing approaches. Du *et al.* [20] have constructed a dataset ($S60$) containing 262 proteins with identity of 60% and predicted them using PseAAC. The accuracy is only 67%. Shi *et al.* [58] achieved an accuracy of 86% using the same benchmark dataset. Nevertheless, we cannot provide direct comparison with these works because the location annotation of some proteins in the dataset $S60$ have been changed with the update of Uniprot. We are only able to give a rough comparison between our method and the two methods. The benchmark dataset in our study has the same scale as $S60$, but the sequence identity of our study is much lower than that of $S60$. That is to say our dataset is more rigorous and objective. Moreover, on this dataset, we achieved 88% accuracy which is better than that of other methods in the literatures [20, 58].

Furthermore, we compared the accuracy of the proposed method with that of other methods using our dataset. First, we compared the performance of optimized tripeptides with other parameters, such as: dipeptides, amino acid and PseAAC. As it can be seen from Table 1, optimized tripeptides achieve the best results among all parameters. Second, we compared the performance of SVM algorithm with Naïve Bayes and RBF Network using tripeptides. We repeated the process of feature selection for finding highest accuracies of Naïve Bayes and RBF Network. Results in Table 1 show that the highest accuracies are 66.41 and 68.34% for Naïve Bayes and RBF Network, respectively. The optimized feature sets for the two methods contain 576 and 560 vectors, respectively. Table 1 also records the results of Naïve Bayes and RBF Network using PseAAC as parameters. It is obvious that the OA of our method is the best one among all listed methods. This result indicates that our method can be used for the prediction of subchloroplast protein location.

4. Conclusion

In this study, we developed a SVM-based method to predict the subchloroplast locations of chloroplast proteins using primary sequence information. A novel feature selection technique based on binomial distribution is proposed to optimize the feature set. Results in Table 1 show that the proposed method achieves an OA of 88.03% in the jack-knife test on a very rigorous and objective dataset, which demonstrates the capability of binomial distribution technique in the process of feature selection. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors [11], a web-server for the method presented in this paper is constructed and can be freely available from http://cobi.uestc.edu.cn/people/hlin/tools/ChloPred/.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (ZYGX2012J113), the Project of Education Department in Sichuan (12ZA112) and the National Natural Science Foundation of China (61202256, 61100092).

References

[1] S. Anishetty, G. Pennathur and R. Anishetty, Tripeptide analysis of protein structures, *BMC Struct. Biol.* **2** (2002) 9.

[2] N. A. Campbell, B. Williamson and R. J. Heyden, *Biology: Exploring Life* (Pearson Prentice Hall, Boston, MA, 2006).

[3] 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 Peptide Lett.* **16**(1) (2009) 27–31.

[4] Y. L. Chen and Q. Z. Li, Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition, *J. Theor. Biol.* **248**(2) (2007) 377–381.

[5] Y. L. Chen and Q. Z. Li, Prediction of the subcellular location of apoptosis proteins, *J. Theor. Biol.* **245**(4) (2007) 775–783.

[6] K. C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, *Proteins* **43**(3) (2001) 246–255.

[7] K. C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, *Bioinformatics* **21**(1) (2005) 10–19.

[8] K. C. Chou, Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology *Curr. Proteomics* **6**(4) (2009) 262–274.

[9] K. C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review), *J. Theor. Biol.* **273** (2011) 236–247.

[10] 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.* **2**(10) (2010) 1090–1103), *Nat. Protoc.* **3**(1) (2008) 153–162.

[11] K. C. Chou and H. B. Shen, Review: Recent advances in developing web-servers for predicting protein attributes, *Nat. Sci.* **2** (2009) 63–92.

[12] 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* **5**(4) (2010) e9931.

[13] 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* **5**(6) (2010) e11335.

[14] K. C. Chou, Z. C. Wu and X. Xiao, iLoc-Euk: A multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins, *PLoS ONE* **6**(3) (2011) e18258.

[15] 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, *Molec. Biosyst.* **8**(2) (2012) 629–641.

[16] K. C. Chou and C. T. Zhang, Review: Prediction of protein structural classes, *Crit. Rev. Biochem. Molec. Biol.* **30**(4) (1995) 275–349.

[17] H. Ding, L. Liu, F. B. Guo, J. Huang and H. Lin, Identify Golgi protein types with modified Mahalanobis discriminant algorithm and pseudo amino acid composition, *Protein Peptide Lett.* **18**(1) (2011) 58–63.

[18] H. Ding, L. F. Luo and H. Lin, Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition, *Protein Peptide Lett.* **16**(4) (2009) 351–355.

[19] Y. S. Ding and T. L. Zhang, Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: An approach with immune genetic algorithm-based ensemble classifier, *Pattern Recogn. Lett.* **29**(13) (2008) 1887–1892.

[20] P. Du, S. Cao and Y. Li, SubChlo: Predicting protein subchloroplast locations with pseudo-amino acid composition and the evidence-theoretic K-nearest neighbor (ET-KNN) algorithm, *J. Theor. Biol.* **261**(2) (2009) 330–335.

[21] O. Emanuelsson, H. Nielsen and G. von Heijne, ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites, *Protein Sci.* **8** (1999) 978–984.

[22] 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.* **263**(2) (2010) 203–209.

[23] R. E. Fan, P. H. Chen and C. J. Lin, Working set selection using the second order information for training SVM, *J. Mach. Learn. Res.* **6** (2005) 1889–1918.

[24] 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* **34**(1) (2008) 103–109.

[25] Y. Feng and L. Luo, Use of tetrapeptide signals for protein secondary-structure prediction, *Amino Acids* **35**(3) (2008) 607–614.

[26] 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.* **257**(1) (2009) 17–26.

[27] Q. Gu, Y. S. Ding and T. L. Zhang, Prediction of G-protein-coupled receptor classes in low homology using Chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns, *Protein Peptide Lett.* **17**(5) (2010) 559–567.

[28] J. Guo, N. Rao, G. Liu, Y. Yang and G. Wang, Predicting protein folding rates using the concept of Chou's pseudo amino acid composition, *J. Comput. Chem.* **32**(8) (2011) 1612–1617.

[29] J. Hu and X. Yan, BS-KNN: An effective algorithm for predicting protein subchloroplast localization, *Evol. Bioinform. Online* **8** (2012) 79–87.

[30] L. Hu, L. Zheng, Z. Wang, B. Li and L. Liu, Using pseudo amino acid composition to predict protease families by incorporating a series of protein biological features, *Protein Peptide Lett.* **18**(6) (2011) 552–558.

[31] T. Huang, L. Chen, Y. D. Cai and K. C. Chou, Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property, *PLoS ONE* **6**(9) (2011) e25297.

[32] X. Jiang, R. Wei, T. L. Zhang and Q. Gu, Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: An approach by approximate entropy, *Protein Peptide Lett.* **15**(4) (2008) 392–396.

[33] X. Jiang, R. Wei, Y. Zhao and T. Zhang, Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location, *Amino Acids* **34**(4) (2008) 669–675.

[34] Y. Jin, B. Niu, K. Y. Feng, W. C. Lu, Y. D. Cai and G. Z. Li, Predicting subcellular localization with AdaBoost learner, *Protein Peptide Lett.* **15**(3) (2008) 286–289.

[35] K. K. Kandaswamy, G. Pugalenthi, S. Moller, E. Hartmann, K. U. Kalies, P. N. Suganthan and T. Martine, Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition, *Protein Peptide Lett.* **17**(12) (2010) 1473–1479.

[36] F. M. Li and Q. Z. Li, Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach, *Protein Peptide Lett.* **15**(6) (2008) 612–616.

[37] Z. C. Li, X. B. Zhou, Z. Dai and X. Y. Zou, Prediction of protein structural classes by Chou's pseudo amino acid composition: Approached using continuous wavelet transform and principal component analysis, *Amino Acids* **37**(2) (2009) 415–425.

[38] B. Liao, J. B. Jiang, Q. G. Zeng and W. Zhu, Predicting apoptosis protein subcellular location with PseAAC by incorporating tripeptide composition, *Protein Peptide Lett.* **18**(11) (2011) 1086–1092.

[39] H. Lin, The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition, *J. Theor. Biol.* **252**(2) (2008) 350–356.

[40] H. Lin and H. Ding, Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition, *J. Theor. Biol.* **269**(1) (2011) 64–69.

[41] H. Lin, H. Ding, F. B. Guo, A. Y. Zhang and J. Huang, Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition, *Protein Peptide Lett.* **15**(7) (2008) 739–744.

[42] H. Lin and Q. Z. Li, Using pseudo amino acid composition to predict protein structural class: Approached by incorporating 400 dipeptide components, *J. Comput Chem.* **28**(9) (2007) 1463–1466.

[43] J. Lin and Y. Wang, Using a novel AdaBoost algorithm and Chou's pseudo amino acid composition for predicting protein subcellular localization, *Protein Peptide Lett.* **18**(12) (2011) 1219–1225.

[44] H. Lin, H. Wang, H. Ding, Y. L. Chen and Q. Z. Li, Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition, *Acta Biotheor.* **57**(3) (2009) 321–330.

[45] T. Liu, X. Zheng, C. Wang and J. Wang, Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: An approach from autocovariance transformation, *Protein Peptide Lett.* **17**(10) (2010) 1263–1269.

[46] J. Ma and H. Gu, A novel method for predicting protein subcellular localization based on pseudo amino acid composition, *BMB Rep.* **43**(10) (2010) 670–676.

[47] S. Mei, Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization, *J. Theor. Biol.* **293** (2012) 121–130.

[48] H. Mohabatkar, Prediction of cyclin proteins using Chou's pseudo amino acid composition, *Protein Peptide Lett.* **17**(10) (2010) 1207–1214.

[49] H. Mohabatkar, B. M. Mohammad 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.* **281**(1) (2011) 18–23.

[50] R. Nair and B. Rost, Sequence conserved for subcellular localization, *Protein Sci.* **11**(12) (2002) 2836–2847.

[51] L. Nanni and A. Lumini, Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization, *Amino Acids* **34**(4) (2008) 653–660.

[52] L. Nanni, A. Lumini, D. Gupta and A. Garg, Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information, *IEEE/ACM Trans. Comput. Biol. Bioinform.* **9** (2012) 467–475.

[53] J. D. Qiu, J. H. Huang, R. P. Liang and X. Q. Lu, Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: An approach from discrete wavelet transform, *Anal. Biochem.* **390**(1) (2009) 68–73.

[54] J. D. Qiu, J. H. Huang, S. P. Shi and R. P. Liang, 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 Peptide Lett.* **17**(6) (2010) 715–722.

[55] J. D. Qiu, S. B. Suo, X. Y. Sun, S. P. Shi and R. P. Liang, OligoPred: A web-server for predicting homo-oligomeric proteins by incorporating discrete wavelet transform into Chou's pseudo amino acid composition, *J. Molec. Graph Model.* **30** (2011) 129–134.

[56] S. S. Sahu and G. Panda, A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction, *Comput. Biol. Chem.* **34**(5–6) (2010) 320–327.

[57] H. B. Shen and K. C. Chou, Gpos-mPLoc: A top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins, *Protein Peptide Lett.* **16**(12) (2009) 1478–1484.

[58] S. P. Shi, J. D. Qiu, X. Y. Sun, J. H. Huang, S. Y. Huang, S. B. Suo, R. P, Liang and L. Zhang, Identify submitochondria and subchloroplast locations with pseudo amino acid composition: Approach from the strategy of discrete wavelet transform feature extraction, *Biochim. Biophys. Acta* **1813**(3) (2011) 424–430.

[59] R. Shi and C. Xu, Prediction of rat protein subcellular localization with pseudo amino acid composition based on multiple sequential features, *Protein Peptide Lett.* **18**(6) (2011) 625–633.

[60] The UniProt Consortium, Ongoing and future developments at the Universal Protein Resource, *Nucleic Acids Res.* **39** (2011) D214–D219.

[61] C. W. Tung, C. Liaw, S. J. Ho and S. Y. Ho, Prediction of protein subchloroplast locations using random forests, *World Acad. Sci. Engrg. Tech.* **65** (2010) 903–907.

[62] P. Ung and D. A. Winkler, Tripeptide motifs in biology: Targets for peptidomimetic design, *J. Med. Chem.* **54**(5) (2011) 111–1125.

[63] G. Wang and R. L. Dunbrack, Jr., PISCES: Recent improvements to a PDB sequence culling server, *Nucleic Acids Res.* **33** (2005) W94–W98.

[64] W. Wang, X. B. Geng, Y. Dou, T. Liu and X. Zheng, Predicting protein subcellular localization by pseudo amino acid composition with a segment-weighted and features-combined approach, *Protein Peptide Lett.* **18**(5) (2011) 480–487.

[65] K. Wang, L. L. Hu, X. H. Shi, Y. S. Dong, H. P. Li and T. Q. Wen, PSCL: Predicting protein subcellular localization based on optimal functional domains, *Protein Peptide Lett.* **19**(1) (2012) 15–22.

[66] Y. C. Wang, X. B. Wang, Z. X. Yang and N. Y. Deng, Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature, *Protein Peptide Lett.* **17**(11) (2010) 1441–1449.

[67] T. Wang, J. Yang, H. B. Shen and K. C. Chou, Predicting membrane protein types by the LLDA algorithm, *Protein Peptide Lett.* **15**(9) (2008) 915–921.

[68] Z. C. Wu, X. Xiao and K. C. Chou, iLoc-Plant: A multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites, *Molec. BioSyst.* **7**(12) (2011) 3287–3297.

[69] Z. C. Wu, X. Xiao and K. C. Chou, iLoc-Gpos: A multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins, *Protein Peptide Lett.* **19**(1) (2012) 4–14.

[70] X. Xiao, Z. C. Wu and K. C. Chou, iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites, *J. Theor. Biol.* **284**(1) (2011) 42–51.

[71] J. B. Yin, Y. X. Fan and H. B. Shen, Conotoxin superfamily prediction using diffusion maps dimensionality reduction and subspace classifier, *Curr. Protein Peptide Sci.* **12**(6) (2011) 580–588.

[72] C. S. Yu, Y. C. Chen, C. H. Lu and J. K. Hwang, Prediction of protein subcellular localization, *Proteins* **64**(3) (2006) 643–651.

[73] L. Yu, Y. Guo, Y. Li, G. Li, M. Li, J. Luo, W. Xiong and W. Qin, SecretP: Identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition, *J. Theor. Biol.* **267**(1) (2010) 1–6.

[74] P. Zakeri, B. Moshiri and M. Sadeghi, Prediction of protein submitochondria locations based on data fusion of various features of sequences, *J. Theor. Biol.* **269**(1) (2011) 208–216.

[75] Y. H. Zeng, Y. Z. Guo, R. Q. Xiao, L. Yang, L. Z. Yu and M. L. Li, Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on autocovariance approach, *J. Theor. Biol.* **259**(2) (2009) 366–372.

[76] 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* **35**(3) (2008) 591–598.

[77] G. Y. Zhang and B. S. Fang, Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition, *J. Theor. Biol.* **253**(2) (2008) 310–315.

[78] G. Y. Zhang, H. C. Li, J. Q. Gao and B. S. Fang, Predicting lipase types by improved Chou's pseudo amino acid composition, *Protein Peptide Lett.* **15**(10) (2008) 1132–1137.

[79] L. Zhang, B. Liao, D. Li and W. Zhu, A novel representation for apoptosis protein subcellular localization prediction using support vector machine, *J. Theor. Biol.* **259**(2) (2009) 361–365.

[80] S. W. Zhang, Y. L. Zhang, H. F. Yang, C. H. Zhao and Q. Pan, Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: An approach by incorporating evolutionary information and von Neumann entropies, *Amino Acids* **34**(4) (2008) 565–572.

[81] Z. L. Zheng, L. Guo, J. Jia, C. M. Xie, W. C. Zeng and J. Yang, Compressed learning and its applications to subcellular localization, *Protein Peptide Lett.* **18**(9) (2011) 925–934.

[82] X. B. Zhou, C. Chen, Z. C. Li and X. Y. Zou, Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes, *J. Theor. Biol.* **248**(3) (2007) 546–551.

[83] D. Zou, Z. He, J. He and Y. Xia, Supersecondary structure prediction using Chou's pseudo amino acid composition, *J. Comput. Chem.* **32**(2) (2011) 271–278.