Analytical Biochemistry 458 (2014) 14–19

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

Predicting peroxidase subcellular location by hybridizing different descriptors of Chou' pseudo amino acid patterns

Yong-Chun Zuo [a,*], Yong Peng [b], Li Liu [b], Wei Chen [c], Lei Yang [d,*], Guo-Liang Fan [b,*]

[a] The Key Laboratory of Mammalian Reproductive Biology and Biotechnology of the Ministry of Education, Inner Mongolia University, Hohhot 010021, China
[b] Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China
[c] Center of Genomics and Computational Biology, College of Sciences, Hebei United University, Tangshan 063000, China
[d] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

ARTICLE INFO

Article history:
Received 23 February 2014
Received in revised form 22 April 2014
Accepted 25 April 2014
Available online 4 May 2014

Keywords:
Peroxidase proteins
Chou' pseudo amino acid patterns
GO-homology annotation
Prediction performance

ABSTRACT

Peroxidases as universal enzymes are essential for the regulation of reactive oxygen species levels and play major roles in both disease prevention and human pathologies. Automated prediction of functional protein localization is rarely reported and also is important for designing new drugs and drug targets. In this study, we first propose a support vector machine (SVM)-based method to predict peroxidase subcellular localization. Various Chou' pseudo amino acid descriptors and gene ontology (GO)-homology patterns were selected as input features to multiclass SVM. Prediction results showed that the smoothed PSSM encoding pattern performed better than the other approaches. The best overall prediction accuracy was 87.0% in a jackknife test using a PSSM profile of pattern with width = 5. We also demonstrate that the present GO annotation is far from complete or deep enough for annotating proteins with a specific function.

Peroxidases are ubiquitous enzymes that catalyze a number of oxidative reactions by using various peroxides as electron acceptors [1,2]. These peroxidase proteins are central elements of the antioxidant defense system, which are extremely widespread in almost all microorganisms and higher organisms. They are essential for the regulation of reactive oxygen species levels and for the promotion of various substrates' oxidation [3–5]. There has been increased interest in them over the past few years; for example, the mammalian heme peroxidase enzymes play major roles in both disease prevention and human pathology defense [6,7]. Therefore, knowing the localization of peroxidase proteins will be important for disease prevention and human pathologies.

Proteins in various subcellular locations play distinct roles in biological processes, such as triggering programmed cell death. Protein localization may be used as a starting point for function prediction systems. Knowing a protein's localization is an important step toward understanding its function [8,9]. Experimental and computational methods are two very important methods for annotating protein functional information. During the past 2 decades, a substantial amount of bioinformatics work for predicting protein subcellular location has been carried out and rapidly developed; significant progress has been achieved with the establishment of various organism-specific benchmark datasets [10–15]. However, to the best of our knowledge, there are few theoretical methods for localization prediction for proteins of specific function.

Therefore, it is becoming crucial to develop a reliable automatic subcellular localizer for identifying the locations of functional proteins. In this study we first attempted to annotate the subcellular localization of a specific oxidoreductase, peroxidase, by using a computational method based on state-of-the-art features. Several different descriptors of the Chou' pseudo amino acid pattern have been discussed for localization prediction [16–21], including amino acid composition (AAC) [22], dipeptide composition (DC) [23,24], split amino acid composition (SAAC) [25], evolutionary information (PSSM) [10,26–28], and gene ontology (GO) of homologous proteins [29–32]. All of the above features were selected as input parameters to established an automatic subcellular classifier. The best overall prediction accuracy achieved 87.0% in a jackknife test for eight locations by using a PSSM profile with width = 5. The GO-homology annotation with different sequence identities was also discussed; the evaluation results showed the present GO annotation is far from complete or deep enough for accurately annotating the localization of peroxidase proteins.

* Corresponding authors. Fax: +86 471 5227683.
E-mail addresses: yczuo@imu.edu.cn (Y.-C. Zuo), yanglei_hmu@163.com (L. Yang), eeguoliangfan@sina.com (G.-L. Fan).

http://dx.doi.org/10.1016/j.ab.2014.04.032
0003-2697/

Materials and methods

Benchmark datasets

The data of peroxidase proteins used in this research were extracted from the PeroxiBase database [33]. PeroxiBase is a unique specialized database, which is devoted to established comprehensive peroxidase families and superfamilies from both eukaryotes and prokaryotes. More than 10,000 peroxidase-encoding sequences come from 940 organisms, and each sequence is individually annotated in this database. Since the number of multiplex proteins in the existing database is not large enough to construct a statistically meaningful benchmark dataset for studying a case of multiple locations, only the proteins with singleplex locations were used in this experiment, and every protein is characterized by an expert sequence annotation procedure, with manual curation, which is a guarantee of quality necessary for performing subcellular localization analysis. After the redundant sequences were removed using the CD-HIT algorithm [34], 586 nonredundant peroxidase proteins were obtained. According to the annotation information, these defensin sequences can be classified into eight subcellular locations: apoplastic (30), chloroplastic (44), cytosolic (265), mitochondrial (44), peroxisomal (107), secreted (23), stromal (37), and thylakoid (37). After measuring by the CD-HIT program, most of the protein similarity scores in each family were lower than 80%.

Features and modules

Support vector machine (SVM), as a strong machine learning technique, is used to evaluate various alternative features of our work. SVM is a machine learning algorithm based on statistical learning theory, which has been successfully used for classification [35]. The basic idea of SVM is to transform the data into a high-dimensional feature space and then determine the optimal separating hyperplane by using a kernel function. In this work, we used the free software LIBSVM to predict peroxidase protein location. A radial basis function (RBF) was chosen as the kernel function. For multiclassification, SVM uses a one-versus-one strategy and constructs $k \times (k-1)/2$ classifiers and voting strategy to assign the class for an arbitrary protein sequence. Here various features of a protein sequence were utilized to perform a comprehensive study and achieve maximum accuracy.

PSSM profile of patterns

Evolutionary conservation usually reflects important biological function. An amino acid at a conserved site of a protein is preferred to locate at a functionally important region [36]. PSI-BLAST is a robust measure of residue conservation in a given location. Evolutionary information on protein sequences like PSSM can be created using a PSI-BLAST search. Compared to the compositional information, the PSSM profile provides more important information of evolutionary significance about residue conservation at a given position in a protein sequence [31,37]. In this study, the PSSM was generated using the PSI-BLAST search with a cutoff E value of 0.001 against the Swiss-Prot database.

The PSSM provides a matrix of dimension L rows and 20 columns for a protein chain with L amino acid residues, where 20 columns represent the occurrence/substitution of each type of 20 amino acids [38]. We summed all of the rows in the PSSM corresponding to the same amino acid in the sequence and then divided each element by the length of the sequence. In the prediction of peroxidase location, we used PSSM profiles with different similarities to generate 400 dimension (20×20 residue pairs) input vectors as parameters.

Composition profile of patterns

The aim of calculating the protein composition is to transform the variable lengths of the protein sequence to fixed-length vectors. This is an important and crucial step for protein classification using a computational approach because it requires a fixed-length pattern.

Amino acid and dipeptide compositions

The AAC representation of a given sequence is composed of 20 different amino acids with a variety of shapes, sizes, and chemical properties. A protein can be represented as a 20-dimensional (20D) vector according to AAC [22]. DC is the occurrence frequency of each of 2 adjacent amino acid residues. It is used to encapsulate the global information of each protein sequence, and a protein can be represented as a 400D vector by means of DC [39–41]. In this study, the AAC and DC of the N-part split amino acid composition were selected as classification vectors.

Split amino acid composition

In simple amino acid-, dipeptide-, and pseudo amino acid-based compositions, the composition is taken at once for the whole sequence, whereas in the split amino acid composition model, the protein sequence is divided into different parts and the composition of each part is calculated separately [25]. The composition is taken independently for the N parts of the protein sequence [42]. Hence, the advantage of SAAC over standard AAC is that it provides a greater weight of compositional biasness to proteins that have a signal at different sequence regions. In our SAAC model each protein is divided into 1 to 10 parts to train the optimal parameter combination for the SVM program.

Gene ontology profile of patterns

Gene Ontology is one of the databases that describes molecular function, and the molecular function of the GO database is correlated to the subcellular location [43]. Accordingly, protein sequences formulated in the GO database space would be clustered in a way that better reflects their subcellular locations [29]. However, to incorporate more information, instead of using only 0 and 1 element, as done in Ref. [44], here let us use a different approach as described below.

First, we searched for the homologous proteins of protein **P** from the Swiss-Prot database (released on 5 September 2012) using the PSI-BLAST method, with the expected value $E \leqslant 0.001$ for the BLAST parameter [31]. Second, we collected those proteins that had $\geqslant$60% pairwise sequence identity with protein **P** into a subset, $\mathbf{P}^{homo}$, called the "homology set" of **P**. All the elements in $\mathbf{P}^{homo}$ could be deemed the "representative proteins" of **P,** sharing some similar attributes such as structural conformation and biological function. These representative proteins retrieved from the Swiss-Prot database must each have their own accession number. Third, we searched each of the accession numbers collected in the second step against the GO database to find the corresponding GO number. Last, we statistically analyzed each coordinate of the vector and found that many of the coordinates were equal to 0. This denoted that certain GOs did not belong to any protein; these GOs were eliminated, and the dimension of the GO feature vector was decreased in this manner.

Performance measures

The prediction performance was evaluated by the sensitivity (Sn), specificity (Sp), positive predictive value (PPV), and overall accuracy (OA), which were defined as follows:

(a) sensitivity or coverage of positive examples, $Sn = TP_i/(TP_i + FN_i)$;
(b) specificity or coverage of negative examples, $Sp = TN_i/(TN_i + FP_i)$;
(c) positive predictive value or confidence of positive examples, $PPV = TP_i/(TP_i + FP_i)$;
(d) overall accuracy, $OA = \sum_i(TP_i + TN_i)/\sum_i(TP_i + FN_i + TN_i + FP_i)$;

where TP_i is the number of observed positive samples predicted to be positive samples, TN_i is the number of observed negative samples predicted to be negative samples, FN_i is the number of observed positive samples predicted to be negative samples, and FP_i is the number of observed negative samples predicted to be positive samples.

Result and discussion

Evaluation

In statistical prediction, three cross-validation methods, independent dataset test, subsampling test, and jackknife test, are often used to examine a predictor for its effectiveness in practical application [45]. The jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset and hence has been increasingly used by investigators to examine the accuracy of various predictors [10–15,46]. During the jackknife test, each protein is singled out in turn as a test sample, the remaining proteins are used as a training set to calculate the test sample's membership and predict the class. Therefore, we adopted the jackknife validation in this study.

Performance evaluation of evolutionary information-based PSSM patterns

Each protein in the dataset of peroxidase proteins can be translated into a group of numerical vector representations. In this section, our initial vector set was based on evolutionary information for the PSSM 20 amino acid (PSSM_20) and the PSSM 400 dipeptide (PSSM_400). We trained the SVM classifier on the dataset with different sequence similarities (80–90%). The lower similarity criterion was not to be accepted, because the currently available data do not allow us to do so. The numbers of proteins for some subsets would have been too few to have statistical significance.

To achieve the best performance for predicting subcellular location, the sliding window size was optimized with respect to the overall accuracy. The optimized sliding window size was obtained by testing the performance of various sliding window sizes with the default parameters in SVM. Fig. 1 shows the results of various sliding window sizes on the original dataset. It was found that a sliding window of five amino acids achieved the best predicting performance for the PSSM_400 pattern. It is indicated that consideration of correlation between neighboring residues can significantly enhance prediction accuracy. The evaluation details are described in Table 1. From the results shown in Table 1, we can see that the evolutionary information is indeed a good sequence feature for describing peroxidase subcellular location. The prediction overall accuracy based on the 400 dipeptide profile achieved 87.0%. For the protein localization of cytosolic, mitochondrial, and thylakoid, both the Sn and the PPV were higher than 90%.

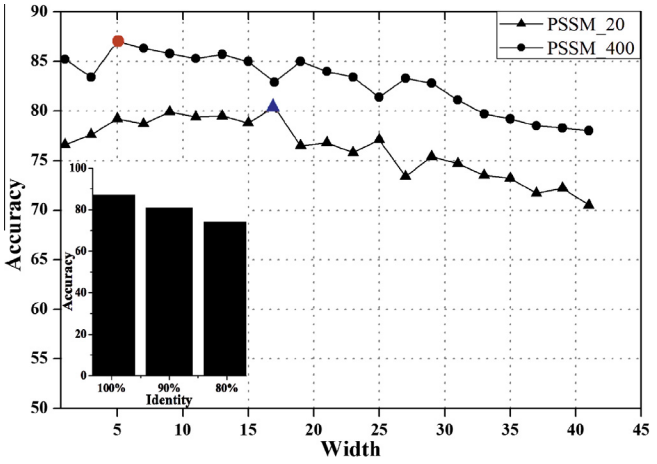

Fig.1. The prediction result with respect to various window sizes based on the jackknife test. The blue triangle indicates the accuracy based on the PSSM profile of 20 amino acids (PSSM_20). The red dot indicates the accuracy based on the PSSM profile of 400 dipeptides (PSSM_400). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
The performance of jackknife validation for the PSSM profile of patterns.

	PSSM (400)			PSSM (20)		
	Sn	Sp	PPV	Sn	Sp	PPV
Apoplastic	76.7	98.7	76.7	63.3	98.2	65.5
Chloroplastic	63.6	98.7	80.0	65.9	97.2	65.9
Cytosolic	95.1	91.3	90.0	92.0	87.3	85.6
Mitochondrial	84.1	99.4	92.5	75.0	98.3	78.6
Peroxisomal	91.6	97.1	87.5	80.4	96.5	83.5
Secreted	56.5	99.5	81.3	56.5	99.5	81.3
Stromal	86.5	98.0	74.4	64.9	98.0	68.6
Thylakoid	75.7	99.5	90.3	64.9	98.4	72.7
Accuracy	510/586 = 87.0%			471/586 = 80.4%		

Therefore, the optimized sliding window size was set to 5 in the following comparison.

Comparison with other methods

It is objective to compare the proposed methods with previously published classifiers using the same dataset and the same parameters. Since there is no published work to predict the peroxidase subcellular location, we cannot provide the comparison analysis with published results to confirm that the predictive model proposed here is superior to other methods. For the purpose of comparison, we compared the performance of the method with some of other classical approaches, Logistic Function, RBFNetwork, NaiveBayes, LogitBoost, and J48, which trained by using Waikato Environment for Knowledge Analysis [47] on the benchmark dataset (Table 2). The results in Table 2 show that prediction accuracy obtained by our method based on the PSSM_400 profile achieved 87.0%, about 10% higher than the other methods, which suggests our proposed method is a state-of-the-art method for subcellular location of peroxidase protein.

Performance evaluation of different SAAC-based approaches

In this section, we investigate how a particular part of the protein sequence affects the prediction accuracy and determine the optimal amount of information needed for peroxidase subcellular location. In our SAAC model, each protein was divided into 1 to 10 parts to discuss the prediction performance of the SVM pro-

Table 2
Comparison of SVM with other machine learning methods based on the PSSM_400 pattern.

	Apo	Chl	Cyt	Mit	Per	Sec	Stro	Thy	Ac
SVM (400, *W* = 5)	**76.7**	**63.6**	**95.1**	**84.1**	**91.6**	**56.5**	**86.5**	**75.7**	**87.0**
SVM (400, *W* = 5)	63.3	65.9	92.0	75.0	80.4	56.5	64.9	64.9	80.4
SVM (400 DP)	76.7	75.0	92.8	75.0	89.7	56.5	86.5	75.7	85.9
SVM (20 DP)	70.0	54.5	92.5	68.2	88.8	56.5	67.6	70.3	81.6
Logistic function (400, *W* = 5)	70.3	59.0	60.0	82.7	75.6	86.1	51.6	65.5	78.0
RBFNetwork (400, *W* = 5)	60.5	63.6	27.8	83.2	73.0	82.1	44.2	60.0	72.0
NaiveBayes (400, *W* = 5)	63.0	65.2	32.4	84.0	67.4	92.8	52.2	45.8	74.1
LogitBoost (400, *W* = 5)	80.0	61.4	60.0	76.3	63.6	78.7	65.8	63.6	74.2
J48 (400, *W* = 5)	70.3	46.5	30.6	69.5	38.0	84.0	45.9	53.6	66.4

The bold values show the best results. *W,* width; Apo, apoplastic; Chl, chloroplastic; Cyt, cytosolic; Mit, mitochondrial; Per, peroxisomal; Sec, secreted; Stro, stromal; Thy, thylakoid; Ac, accuracy.

gram. The amino acid composition and dipeptides were calculated as SVM parameter vectors to predict subcellular location. Fig. 2 shows the prediction results of overall accuracy based on *K*-peptide compositions of *N* split parts (*N, K*).

As is shown in Fig. 2, the overall accuracy reached a maximum of 85.9% based on a 2-peptide composition of *N* = 4 or *N* = 5. When *N* = 2, the prediction overall accuracy using amino acid composition performed the best (*N* = 2, *K* = 1). For different values of *N,* it was shown that the prediction ability increases along with the *N* increase, up to the peak at which *N* equals 4 or 5, and decreases when *N* > 6. This indicates that the split amino acid composition indeed provides greater weight of compositional biasness to proteins that have a signal at different sequence regions.

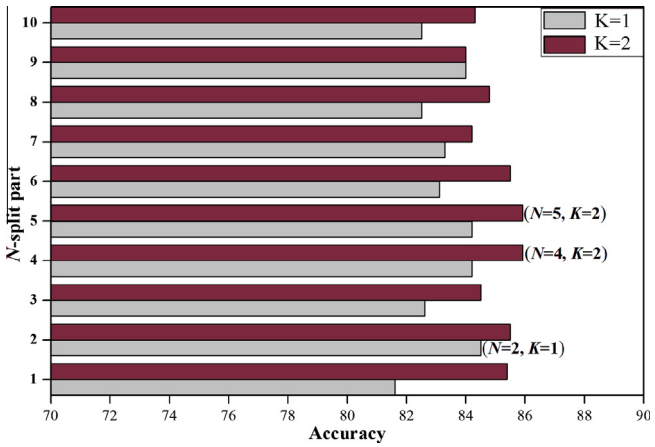

Fig.2. Prediction results of different SAAC-based approaches. The gray bar indicates the accuracy based on 1-peptide compositions (*K* = 1). The red bar indicates the accuracy based on 2-peptide compositions (*K* = 2). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Performance evaluation based on gene ontology patterns of homologous protein

The property derived from the amino acid sequence performs poorly for a dataset with low-similarity sequences [31]. GO is another very important feature for prediction of subcellular localization. The GO-based methods make use of the well-organized biological knowledge about genes and gene products in the GO database [48]. Recently GO annotation has been used successfully to solve various sequence-based prediction problems and to extract many other important features of proteins. In this study, we propose an efficient GO method called GO-homology to represent a protein in the general form of Chou' pseudo amino acid composition [30]. The GO-homology method first selects a subset of relevant GO terms to form a GO vector space. Then for each protein, the method calculates the occurrence accession number in a subset of the selected homologous proteins as a means to construct GO vectors for predicting subcellular location.

This study also evaluated how the sequence identity annotated in the Swiss-Prot database affects the prediction performance of the GO-homology method. Table 3 shows the prediction results of different sequence identities (80–90%). The GO patterns of homologous proteins with 50 and 70% similarity performed better than the other homologous GO patterns. The best overall accuracy based on 375 GO annotations achieved 71.5%. However, the prediction ability was far lower than that of the PSSM pattern. There may be two reasons for this result: (1) for the specific function proteins, most of the peroxidase proteins with different subcellular localizations contain GO annotations that are too similar to discriminate. So the GO patterns of homologous protein methods do not have the ability to clearly classify the different subcellular localizations. (2) For the accession numbers available in the Swiss-Prot database, the current GO annotation is far from complete for specific functions. Most of the proteins still cannot be clearly formulated in the GO space. The above results demonstrate that the GO pattern method of homologous proteins is not suitable for annotating the localization of peroxidase proteins.

Table 3
Performance evaluation based on GO-homology patterns with different sequence similarities.

	40% (424 GO)	50% (375 GO)	60% (330 GO)	70% (375 GO)	80% (277 GO)	90% (261 GO)
Apoplastic	70.0	**73.3**	73.3	40.0	33.3	40.0
Chloroplastic	52.3	**52.3**	36.4	40.9	27.3	29.6
Cytosolic	90.8	**84.7**	93.1	96.6	95.0	96.2
Mitochondrial	56.8	**84.1**	84.1	72.7	56.8	50.0
Peroxisomal	22.4	**65.4**	25.2	72.9	53.3	18.7
Secreted	21.7	**21.7**	43.5	30.4	0.0	0.0
Stromal	5.6	**58.3**	2.8	0.0	52.8	16.7
Thylakoid	0.0	**46.0**	94.6	24.3	51.4	5.4
Accuracy	58.0	**71.5**	67.2	70.2	67.1	56.1

The bold values show the best results.

Conclusion

Using bioinformatics techniques to identify the subcellular localization from a given primary sequence is one of the active areas in protein classification [49]. However, only a few computational methods have been developed for specific functions of proteins. In this study, a benchmark dataset of the subcellular location of peroxidase proteins was first constructed. Then we proposed an SVM-based method to predict subcellular localization using many different descriptors of the Chou' pseudo amino acid profile of patterns. Evaluation results also showed that the prediction performance of smoothed PSSM encoding performed better than the state-of-the-art approaches on the benchmark datasets. The PSSM profile of patterns achieved the best performance. We also demonstrated that the present GO annotation is far from complete or deep enough for classifying proteins with specific functions, such as peroxidase proteins.

Acknowledgments

The authors thank Christophe Dunand for sharing the datasets. This work was supported by the Program of Higher-Level Talents of Inner Mongolia University (Nos. 115115, 135147), the Specialized Research Fund for the Doctoral Program of Higher Education (20131501120009), and the Natural Science Foundation of Inner Mongolia Autonomous Region (Nos. 2013MS0503, 2013MS0504).

References

[1] N. Tokunaga, T. Kaneta, S. Sato, Y. Sato, Analysis of expression profiles of three peroxidase genes associated with lignification in *Arabidopsis thaliana*, Physiol. Plant. 136 (2009) 237–249.
[2] M. Tognolli, C. Penel, H. Greppin, P. Simon, Analysis and expression of the class III peroxidase large gene family in *Arabidopsis thaliana*, Gene 288 (2002) 129–138.
[3] L.V. Bindschedler, J. Dewdney, K.A. Blee, J.M. Stone, T. Asai, J. Plotnikov, C. Denoux, T. Hayes, C. Gerrish, D.R. Davies, Peroxidase dependent apoplastic oxidative burst in *Arabidopsis* required for pathogen resistance, Plant J. 47 (2006) 851–863.
[4] R. Mittler, S. Vanderauwera, M. Gollery, F. Van Breusegem, Reactive oxygen gene network of plants, Trends Plant Sci. 9 (2004) 490–498.
[5] L. Almagro, L.G. Ros, S. Belchi-Navarro, R. Bru, A.R. Barceló, M. Pedreño, Class III peroxidases in plant defence reactions, J. Exp. Bot. 60 (2009) 377–390.
[6] L. Flohé, F. Ursini, Peroxidase: a term of many meanings, Antioxid. Redox Signal. 10 (2008) 1485–1490.
[7] M.J. Davies, C.L. Hawkins, D.I. Pattison, M.D. Rees, Mammalian heme peroxidases: from molecular mechanisms to health implications, Antioxid. Redox Signal. 10 (2008) 1199–1234.
[8] K.C. Chou, H.B. Shen, Recent progress in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.
[9] K.C. Chou, H.B. Shen, Recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (2009) 63–92.
[10] W.Z. Lin, J.A. Fang, X. Xiao, K.C. Chou, ILoc-animal: a multi-label learning classifier for predicting subcellular localization of animal proteins, Mol. BioSyst. 9 (2013) 634–644.
[11] G.P. Zhou, K. Doctor, Subcellular location prediction of apoptosis proteins, Proteins Struct. Funct. Bioinform. 50 (2003) 44–48.
[12] K.C. Chou, H.B. Shen, Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites, J. Proteome Res. 6 (2007) 1728–1734.
[13] K.C. Chou, H.B. Shen, Large scale plant protein subcellular location prediction, J. Cell. Biochem. 100 (2007) 665–678.
[14] G.L. Fan, Q.Z. Li, Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition, Amino Acids 43 (2012) 545–555.
[15] C. Ding, L.F. Yuan, S.H. Guo, H. Lin, W. Chen, Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions, J. Proteomics 77 (2012) 321–328.
[16] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (2011) 236–247.
[17] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[18] K.C. Chou, Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology, Curr. Proteomics 6 (2009) 262–274.
[19] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins Struct. Funct. Bioinform. 43 (2001) 246–255.
[20] W. Chen, P.M. Feng, H. Lin, Prediction of ketoacyl synthase family using reduced amino acid alphabets, J. Ind. Microbiol. Biotechnol. 39 (2012) 579–584.
[21] G.L. Fan, Q.Z. Li, Y.C. Zuo, Predicting acidic and alkaline enzymes by incorporating the average chemical shift and gene ontology informations into the general form of Chou's PseAAC, Process Biochem. 48 (2013) 1048–1053.
[22] Y.C. Zuo, W. Chen, G.L. Fan, Q.Z. Li, A similarity distance of diversity measure for discriminating mesophilic and thermophilic proteins, Amino Acids 44 (2012) 573–580.
[23] Y.C. Zuo, Q.Z. Li, Using K-minimum increment of diversity to predict secretory proteins of malaria parasite based on groupings of amino acids, Amino acids 38 (2009) 859–867.
[24] Y.C. Zuo, Q.Z. Li, Using reduced amino acid composition to predict defensin family and subfamily: integrating similarity measure and structural alphabet, Peptides 30 (2009) 1788–1793.
[25] B. Panwar, G.P.S. Raghava, Predicting sub-cellular localization of tRNA synthetases from their primary structures, Amino Acids 42 (2011) 1703–1713.
[26] W.L. Huang, C.W. Tung, H.L. Huang, S.Y. Ho, Predicting protein subnuclear localization using GO amino acid composition features, Biosystems 98 (2009) 73–79.
[27] Z. Lei, Y. Dai, Assessing protein similarity with gene ontology and its use in subnuclear localization prediction, BMC Bioinformatics 7 (2006) 491.
[28] H.B. Shen, K.C. Chou, Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites, Biochem. Biophys. Res. Commun. 355 (2007) 1006–1011.
[29] X. Xiao, Z.C. Wu, K.C. Chou, A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites, PLoS One 6 (2011) e20592.
[30] K.C. Chou, Z.C. Wu, X. Xiao, ILoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins, PLoS One 6 (2011) e18258.
[31] X. Wang, G.Z. Li, A multi-label predictor for identifying the subcellular locations of singleplex and multiplex eukaryotic proteins, PLoS One 7 (2012) e36317.
[32] X. Xiao, Z.C. Wu, 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 (2011) 42–51.
[33] N. Fawal, Q. Li, B. Savelli, M. Brette, G. Passaia, M. Fabre, C. Mathe, C. Dunand, PeroxiBase: a database for large-scale evolutionary analysis of peroxidases, Nucleic Acids Res. 41 (2013) D441–444.
[34] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.
[35] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–27 (Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm).
[36] A.A. Schaffer, L. Aravind, T.L. Madden, S. Shavirin, J.L. Spouge, Y.I. Wolf, E.V. Koonin, S.F. Altschul, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, Nucleic Acids Res. 29 (2001) 2994–3005.
[37] L. Li, Y. Zhang, L. Zou, C. Li, B. Yu, X. Zheng, Y. Zhou, An ensemble classifier for eukaryotic protein subcellular location prediction using gene ontology categories and amino acid hydrophobicity, PLoS One 7 (2012) e31057.
[38] C.C. Wang, Y. Fang, J. Xiao, M. Li, Identification of RNA-binding sites in proteins by integrating various sequence information, Amino Acids 40 (2011) 239–248.
[39] A. Khan, A. Majid, M. Hayat, CE-PLoc: an ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition, Comput. Biol. Chem. 35 (2011) 218–229.
[40] H. Lin, W. Chen, Prediction of thermophilic proteins using feature selection technique, J. Microbiol. Methods 84 (2011) 67–70.
[41] H. Lin, H. Ding, Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition, J. Theor. Biol. 269 (2011) 64–69.
[42] T.H. Afridi, A. Khan, Y.S. Lee, Mito-GSAAC: mitochondria prediction using genetic ensemble classifier and split amino acid composition, Amino Acids 42 (2012) 1443–1454.
[43] J.A. Blake, M. Dolan, H. Drabkin, D.P. Hill, N. Li, D. Sitnikov, S. Bridges, S. Burgess, T. Buza, F. McCarthy, D. Peddinti, L. Pillai, S. Carbon, H. Dietze, A. Ireland, S.E. Lewis, C.J. Mungall, P. Gaudet, R.L. Chrisholm, P. Fey, W.A. Kibbe, S. Basu, D.A. Siegele, B.K. McIntosh, D.P. Renfro, A.E. Zweifel, J.C. Hu, N.H. Brown, S. Tweedie, Y. Alam-Faruque, R. Apweiler, A. Auchinchloss, K. Axelsen, B. Bely, M. Blatter, C. Bonilla, L. Bouguerleret, E. Boutet, L. Breuza, A. Bridge, W.M. Chan, G. Chavali, E. Coudert, E. Dimmer, A. Estreicher, L. Famiglietti, M. Feuermann, A. Gos, N. Gruaz-Gumowski, R. Hieta, C. Hinz, C. Hulo, R. Huntley, J. James, F. Jungo, G. Keller, K. Laiho, D. Legge, P. Lemercier, D. Lieberherr, M. Magrane, M.J. Martin, P. Masson, P. Mutowo-Muellenet, C. O'Donovan, I. Pedruzzi, K. Pichler, D. Poggioli, P. Porras Millan, S. Poux, C. Rivoire, B. Roechert, T. Sawford, M. Schneider, A. Stutz, S. Sundaram, M. Tognolli, I. Xenarios, R. Foulgar, J. Lomax, P. Roncaglia, V.K. Khodiyar, R.C. Lovering, P.J. Talmud, M. Chibucos, M.G. Giglio, H. Chang, S. Hunter, C. McAnulla, A. Mitchell, A. Sangrador, R. Stephan, M.A. Harris, S.G. Oliver, K. Rutherford, V. Wood, J. Bahler, A. Lock, P.J. Kersey, D.M. McDowall, D.M. Staines, M. Dwinell, M. Shimoyama, S. Laulederkind, T. Hayman, S. Wang, V. Petri, T. Lowry, P. D'Eustachio, L. Matthews, R. Balakrishnan, G. Binkley, J.M. Cherry, M.C. Costanzo, S.S. Dwight, S.R. Engel, D.G. Fisk, B.C. Hitz, E.L. Hong, K. Karra, S.R. Miyasato, R.S. Nash, J. Park, M.S. Skrzypek, S. Weng, E.D. Wong, T.Z. Berardini, E. Huala, H. Mi, P.D. Thomas, J. Chan, R. Kishore, P. Sternberg, K. Van Auken, D.

Howe, M. Westerfield, Gene ontology annotations and resources, Nucleic Acids Res. 41 (2013) D530–535.
[44] K.C. Chou, H.B. Shen, Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization, Biochem. Biophys. Res. Commun. 347 (2006) 150–157.
[45] K.C. Chou, C.T. Zhang, Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.
[46] H. Lin, H. Ding, F.B. Guo, J. Huang, Prediction of subcellular location of mycobacterial protein using feature selection techniques, Mol. Divers. 14 (2010) 667–671.
[47] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explor. Newsl. 11 (2009) 10–18.
[48] S. Wan, M.W. Mak, S.Y. Kung, GOASVM: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition, J. Theor. Biol. 323 (2013) 40–48.
[49] X. Xiao, W.Z. Lin, K.C. Chou, Recent advances in predicting protein classification and their applications to drug development, Curr. Top. Med. Chem. 13 (2013) 1622–1635.