Amino Acids (2008) 35: 383–388
DOI 10.1007/s00726-007-0608-y
Printed in The Netherlands

Amino Acids

Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine

X.-B. Zhou, C. Chen, Z.-C. Li, and **X.-Y. Zou**

School of Chemistry and Chemical Engineering, Sun Yat-Sen University, Guangzhou, China

Received July 20, 2007
Accepted September 18, 2007
Published online December 21, 2007;

Summary. Apoptosis proteins play an important role in the development and homeostasis of an organism. The accurate prediction of subcellular location for apoptosis proteins is very helpful for understanding the mechanism of apoptosis and their biological functions. However, most of the existing predictive methods are designed by utilizing a single classifier, which would limit the further improvement of their performances. In this paper, a novel predictive method, which is essentially a multi-classifier system, has been proposed by combing a dual-layer support vector machine (SVM) with multiple compositions including amino acid composition (AAC), dipeptide composition (DPC) and amphiphilic pseudo amino acid composition (Am-Pse-AAC). As a demonstration, the predictive performance of our method was evaluated on two datasets of apoptosis proteins, involving the standard dataset ZD98 generated by Zhou and Doctor, and a larger dataset ZW225 generated by Zhang et al. With the jackknife test, the overall accuracies of our method on the two datasets reach 94.90% and 88.44%, respectively. The promising results indicate that our method can be a complementary tool for the prediction of subcellular location.

Keywords: Subcellular location – Apoptosis protein – Dual-layer support vector machine – Amino acid composition – Dipeptide composition – Amphiphilic pseudo amino acid composition

1. Introduction

Subcellular location is a key functional characteristic of proteins (Hua and Sun, 2001). The knowledge of the subcellular location of a protein is very helpful for understanding its biological function (Reinhardt and Hubbard, 1998), because newly synthesized proteins must be localized to the appropriate subcellular compartments to perform their biological functions. Even for proteins of basic functions known, information about their localizations may give insights about their involvements in specific metabolic pathways (Garg et al., 2005). Although subcellular location of unknown proteins can be determined by experimental methods such as cell fractionation, electron microscopy and fluorescence microscopy, they are both time-consuming and expensive (Chou and Cai, 2002; Feng, 2002). Therefore, it is very urgent to develop an automatic and reliable prediction system for protein subcellular location.

Actually, many predictive methods have been developed, which in general, can be roughly divided into two categories. One is based on N-terminal sorting signal (Nakai and Kanehisa, 1992; Emanuelsson et al., 2000), whose predictive result is inaccurate when the sorting signals are missing or partially included (Hua and Sun, 2001). The other is based on amino acid composition (AAC) (Reinhardt and Hubbard, 1998; Hua and Sun, 2001; Zhou and Doctor, 2003), which, in fact, is absent of any information of sequence order. Subsequently, some new protein features were proposed in order to incorporate sequence order effects of proteins, including pseudo amino acid compositions (Chou, 2001; Chou and Cai, 2003), dipeptide composition (DPC) (Bhasin and Raghava, 2004; Huang and Li, 2004), amino acid pairs composition (Park and Kanehisa, 2003), Markov chains model (Bulashevska and Eils, 2006) and so on. Additionally, a new concept of functional domain composition was proposed by Chou and Cai (2002), who also presented a new method incorporating gene ontology (Chou and Cai, 2004b). However, the methods mentioned above can only deal with proteins that just exist at one subcellular location. For proteins that may simultaneously exist at, or move between, two or more different subcellular locations, two predictors called ''Hum-mPLoc'' (Shen and Chou, 2007b) and ''Euk-mPLoc'' (Chou and Shen, 2007b) were proposed. For a comprehensive description in this area, read-

ers can refer to two recent reviews (Chou and Shen, 2007c; Shen et al., 2007b).

Meanwhile, methods only using a single classifier have limitations in the prediction of subcellular location (Chou and Shen, 2006a, c). Consequently, many attempts have been made to enhance the prediction quality by utilizing multi-classifier system (Park and Kanehisa, 2003; Chou and Cai, 2004a; Yu et al., 2004; Chou and Shen, 2006a, b, c, 2007a). Yu et al. (2004) proposed a predictive method called CELLO which used multiple SVM classifiers based on n-peptide composition. Chou and Shen (2006a) developed a predictor by fusing multiple basic optimized evidence-theoretic k-nearest neighbor classifiers for the prediction of eukaryotic protein subcellular location. Both of them gave the final decision through a voting system. In our previous study (Chen et al., 2006b), a dual-layer support vector machine (SVM) was developed containing multiple SVM classifiers for the prediction of protein structure class. Instead of the voting system, a second SVM classifier was used to give the final decision, and significant enhancement in success rate was achieved. Encouraged by this, we extend this idea down the prediction of subcellular location.

Apoptosis, also known as programmed cell death, plays a central role in normal tissue homeostasis by regulating a balance between cell proliferation and death (Chou et al., 1997, 2000; Chou, 2004, 2006). Unregulated excessive apoptosis may cause various degenerative and autoimmune diseases. Conversely, an inappropriately low rate of apoptosis may promote survival and accumulation of abnormal cells that can give rise to tumor formation and prolonged autoimmune stimulation such as in cancers and Graves disease (Peter et al., 1997). The study on apoptosis proteins can help us to understand the mechanism of apoptosis and provide many targets for therapeutic intervention (Chou et al., 1997, 2000; Chou, 2004, 2006).

In 2003, Zhou and Doctor (2003) constructed a standard dataset of apoptosis proteins including four major subcellular localization sites. On the basis of the dataset, they developed a predictive method by associating covariant discriminant algorithm (CDA) with AAC. The overall accuracy of their method reached 72.5%. Bulashevska and Eils (2006) proposed a predictor called HensBC by using hierarchical ensemble of Bayesian classifiers based on Markov chains model of primary protein sequence. On the same dataset, HensBC yielded an overall accuracy of 89.8%. More recently, a predictive method called EBGW_SVM which combined a new protein descriptor EBGW (Encoding Based on Grouped Weight) with SVM has been developed by Zhang et al. (2006b), and they got a higher overall accuracy of 92.9%. Despite of the great enhancement in accuracy, EBGW_SVM would limit the further enhancement of predictive performance because it was essentially a single classifier based on single protein feature.

In view of the above facts, this article presents a framework with a dual-layer SVM system. In its first layer, there are three SVM classifiers trained by various protein features. Then the computational results are combined and input into the second layer, where another SVM classifier makes the final decisions adaptively. It is demonstrated through two different working data sets that the predictive performance is improved significantly.

2. Materials and methods

2.1 Dataset

Two datasets were applied to examine the effectiveness of our method. One is the standard dataset ZD98 generated by Zhou and Doctor (2003). It involves 98 apoptosis protein sequences classified into four location categories, of which 43 are cytoplasmic proteins, 30 Plasma membrane-bound proteins, 13 mitochondrial proteins and 12 other proteins (exclude the former three classes of proteins).

The other is a larger dataset ZW225 provided by Zhang et al. (2006b). It consists of 225 apoptosis proteins divided into four subcellular locations with 41 nuclear proteins, 70 cytoplasmic proteins, 25 mitochondrial proteins and 89 membrane proteins. All the protein sequences in the two datasets were extracted from SWISS-PROT (Bairoch and Apweiler, 2000), and the accession numbers can be found in the literatures (Zhou and Doctor, 2003; Zhang et al., 2006a).

2.2 Protein features

AAC is the occurrence frequency of each amino acid residue in a protein. A protein can be represented as a 20-*D* (dimension) vector according to AAC. DPC is the occurrence frequency of each two adjacent amino acid residues. It is used to encapsulate the global information about each protein sequence and a protein can be represented as a 400-*D* vector by means of DPC. The Amphiphilic pseudo amino acid composition (Am-Pse-AAC) originally proposed by Chou (2005) is designed to reflect the sequence-order effects by using the hydrophobicity (Tanford, 1962) and hydrophilicity (Hopp and Woods, 1981) of the constitute amino acids in a protein. By using it, a protein sample can be represented as follows:

$$P = [P_1, \ldots, P_{20}, P_{20+1}, \ldots, P_{20+\lambda}, \ldots, P_{20+2\lambda}], \quad (1)$$

where the first 20 numbers in Eq. (1) represent the classic AAC, and the next 2λ discrete numbers are described as sequence-correlation factor, which can be calculated according to the literature (Chou, 2005). For different problems, the optimal value of λ is variable. In this study, the optimal value of λ was selected as the one that yielded the highest overall accuracy through the jackknife test. Detailed descriptions about the Am-Pse-AAC can refer to Chou's paper (2005). Recently, a very flexible PseAA web-server has been established at the website http://chou.med.harvard.edu/bioinf/PseAA/, where readers can easily calculate various kinds of pseudo amino acid composition.

2.3 SVM

SVM has been widely used in biological sequence analysis (Yang, 2004). The main idea of SVM (Cortes and Vapnik, 1995) is to map the samples

from input space into a high dimensional feature space by the so-called kernel function and to seek a separating hyper-plane with the maximal margin (i.e. Optimal Separating Hyper-plane (OSH)). It has many attractive features including effective avoidance of over-fitting, capability of handling large feature space, and absence of local minima.

The key step of training SVM is the selection of parameters. A few parameters such as the penalty parameter C and the parameters of the kernel function must be determined in advance. Herein we use the jackknife test to select parameters. In the jackknife test, each protein is singled out in turn as a query protein with the remaining proteins for training SVM.

It is a multi-classification problem for the prediction of subcellular location. The simple solution is to reduce the multi-classification to a series of binary classifications. In this study, we adopted the one-versus-rest method (Brown et al. 2000; Ding and Dubchak, 2001) to transfer it into a series of two-class problems. For example, for a K-classes problem, there are K two-class subclassifiers needed to be constructed by the one-versus-rest method. The ith subclassifier is trained by considering all the proteins in the ith class as positive samples and all other classes as negative samples. In practice, however, this method will lead to the so-called 'False Positive' problem (Ding and Dubchak, 2001) which would cause ambiguous prediction results. To avoid the drawback, the 'winner-takes-all' scheme (Angulo et al., 2003) was utilized, in which the output of each binary classifier will be specific numerical values instead of label $+1$ or -1, and the final results of prediction will be given by considering the maximum of the output values. The software used to implement SVM is SVM-lite – a MATLAB MEX-interface to SVM$^{\text{light}}$ written by Tom Briggs and can be freely downloaded from http://webspace.ship.edu/thbrig/mexsvm/ for academic purpose.

2.4 Dual-layer SVM

The structure of dual-layer SVM containing multi-classifiers based on multiple protein compositions is shown in Fig. 1. The first layer is made up of three SVM classifiers: SVM1, SVM2 and SVM3, which is based on AAC, DPC and Am-Pse-AAC, respectively. The corresponding dimensions of input vector of three SVM classifiers are 20, 400 and $20+2\lambda$ (the optimized value of λ is 2 for ZD98 and 4 for ZW225). For the first layer, there are totally $4\times 3=12$ SVM binary classifiers needed to construct by using the one-versus-rest method. preliminary tests show the RBF kernel can give best results. Therefore, RBF kernel was selected for all the 12 SVM binary classifiers.

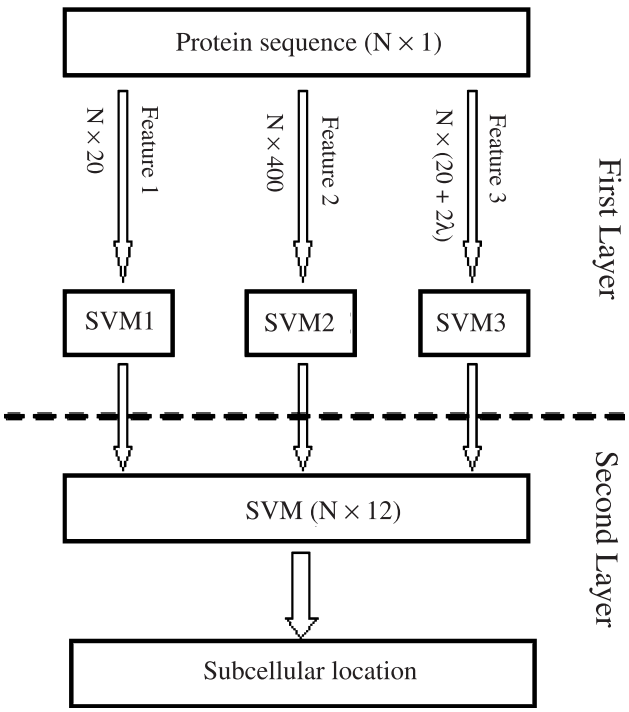

Fig. 1. Structure of the dual-layer SVM

To give the final decision, the second layer SVM classifier is designed by merging the outputs from the first layer as input. Different from the first layer SVM classifiers, the relationship between the inputs and the outputs in the second layer SVM may be linear. Accordingly, we selected the linear kernel function for the second layer SVM.

2.5 Assessment of predictive performances

The prediction quality is examined by the jackknife test. Among the independent dataset test, sub-sampling (e.g., 5-fold sub-sampling) test, and jackknife test, which are often used for examining the accuracy of a statistical prediction method, the jackknife test is deemed the most rigorous and objective as analyzed by a comprehensive review (Chou and Zhang, 1995) and has been increasingly adopted by leading investigators to test the power of various prediction methods (Zhou, 1998; Zhou and Assa-Munt, 2001; Gao et al., 2005; Wang et al., 2005; Xiao et al., 2005, 2006; Du and Li, 2006; Guo et al., 2006; Chen et al., 2006a, b; Mondal et al., 2006; Niu et al., 2006; Sun and Huang, 2006; Wen et al., 2007; Zhang et al., 2006; Chou and Shen, 2007d, e; Lin and Li, 2007a,b; Shen and Chou, 2007c; Chen et al., 2007; Ding et al., 2007; Liu et al., 2007; Shen and Chou, 2007a; Shen et al., 2007a; Shi et al., 2007; Zhang and Ding, 2007).

The overall prediction accuracy, the prediction accuracy and Matthew's correlation coefficient (Matthews, 1975) (MCC) for each subcellular location calculated for assessment of the prediction system are given by Eqs. (2)–(4).

$$\text{overall accuracy} = \frac{\sum_{i=1}^{k} p(i)}{N}, \tag{2}$$

$$\text{accuracy } (i) = \frac{p(i)}{\text{obs}(i)}, \tag{3}$$

$$\text{MCC } (i) = \frac{p(i)n(i) - u(i)o(i)}{\sqrt{(p(i)+u(i))(p(i)+o(i))(n(i)+u(i))(n(i)+o(i))}}, \tag{4}$$

where N is the total number of sequences, k is the class number, obs (i) is the number of sequences observed in location i, $p(i)$ is the number of correctly predicted sequences of location i, $n(i)$ is the number of correctly predicted sequences not of location i, $u(i)$ is the number of under-predicted sequences and $o(i)$ is the number of over-predicted sequences.

3. Results and discussion

3.1 Predictive performance of the dual-layer SVM

Firstly the standard dataset ZD98 was used to evaluate the performance of the dual-layer SVM and the results of jackknife test are summarized in Table 1. As shown in Table 1, SVM1 based on AAC and SVM2 based on DPC give the same accuracy of 90.82%. SVM3 based on Am-Pse-AAC reach the highest overall accuracy of 93.88%, indicating that Am-Pse-AAC can capsulate more information of protein sequence. To further enhance the predictive performance, the dual-layer SVM was devised on the ba-

Table 1. Performance of various SVMs based on different compositions on dataset ZD98 by the Jackknife test

Approach	Cytoplasmic		Plasma		Mitochondrial		Other		Overall
	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC
SVM1	93.02	0.86	90.00	0.90	100	0.89	75.00	0.80	90.82
SVM2	95.35	0.88	90.00	0.86	92.31	0.87	75.00	0.85	90.82
SVM3	95.35	0.90	96.67	0.93	100	0.96	75.00	0.85	93.88
Dual-layer SVM	95.35	0.92	96.67	0.93	92.31	0.91	91.67	0.95	94.90

* *ACC* Accuracy; *MCC* Matthew's correlation coefficient. SVM1 is based on AAC, SVM2 based on DPC and SVM3 based on Am-Pse-AAC

Table 2. Comparison with other methods on dataset ZD98 by the Jackknife test

Subcellular location	CDA[a]		HensBC[b]		EBGW_SVM[c]		Dual-layer SVM[d]	
	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
Cytoplasmic	97.7	–	95.3	0.89	97.67	0.90	95.35	0.92
Plasma membrane	73.3	–	90.0	0.83	90.00	0.88	96.67	0.93
Mitochondrial	30.8	–	92.3	0.83	92.31	0.91	92.31	0.91
Other	25.0	–	66.7	0.80	83.33	0.90	91.67	0.95
Overall accuracy	72.5	–	89.8	–	92.86	–	94.90	–

[a] Data was taken from Zhou and Doctor (2003)
[b] Data was taken from Bulashevska and Eils (2006)
[c] Data was taken from Zhang et al. (2006b)
[d] Data from current paper

sis of various features described above. The final overall accuracy of the dual-layer SVM is 94.90%, which is better than any individual feature-based module. It implies that the dual-layer SVM based on multiple features can take better advantage of the sequence information of a protein than the single-layer SVM based on individual feature. The detailed performance of the dual-layer SVM is shown in the last row of Table 1.

3.2 Comparison with existing methods

The predictive performance of the dual-layer SVM was compared with that of existing prediction methods. The dataset ZW98 was also tested with CDA (Zhou and Doctor, 2003), HensBC approach (Bulashevska and Eils, 2006) and EBGW_SVM (Zhang et al., 2006b). CDA was based on AAC, HensBC approach based on the Markov chains, and EBGW_SVM based on EBGW. The results of CDA, HensBC, EBGW_SVM and the dual-layer SVM were obtained through the jackknife test, and listed in Table 2.

As can be seen from Table 2, the overall accuracy of the dual-layer SVM is 94.90%, which is almost 22%, 5%, and 2% higher than that of CDA, HensBC, EBGW_SVM, respectively. Especially for the most difficult case – the forth class (i.e. other subcellular location), the predictive accuracy is improved to 91.67% by our method. Although the predictive accuracy is a convenient measure for predictive performance, it is still not enough to draw a conclusion, because it overlooks over-predictions. As taking into account of both under- and over-prediction, *MCC* can offer a complementary measure for the predictive performances (Yu et al., 2004). The value of *MCC* is 1 for a perfect prediction and 0 for completely random assignment. The *MCCs* of the dual-layer SVM range from 0.91 to 0.95 which is higher than those of HensBC and EBGW_SVM (see Table 2). These results further indicate that our method can significantly improve the predictive performance by using multiple features with a more powerful machine learning method.

A much large dataset ZW225, constructed by Zhang et al. (2006a), was also utilized to evaluate the generalization ability of our method. And the results of comparison between our method and EBGW_SVM (Zhang et al., 2006b) with jackknife test are listed in Table 3. As can be seen, in contrast to EBGW_SVM, the predictive performance is remarkably improved by our method. The overall accuracy of 88.44% has been obtained, which is appropriately 5% higher than that of EBGW_SVM. Especially for the nuclear and mitochondrial proteins, the accuracies are improved nearly 15% and 16%. All the

Table 3. Comparison with EBGW_SVM on dataset ZW225 by the Jackknife test

Subcellular location	EBGW_SVM[a] ACC	Dual-layer SVM[b] ACC
Nuclear	63.41	78.05
Cytoplasm	90.00	91.43
Mitochondrial	60.00	76.00
Membrane	93.26	94.38
Overall accuracy	83.11	88.44

[a] Data was taken from Zhang et al. (2006b)
[b] Data from current paper

results show that our method is superior to EBGW_SVM. The reason may be that multi-classifier system based on multiple protein features can make use of more information than single classifier based on single protein feature, and therefore can enhance predictive performance significantly. Certainly, the cost of the adoption of multi-classifier system is its complicated framework and time consuming in model training and testing. Consequently, it is the major task to select the protein features and to optimize the framework of multi-classifier system in our future work.

4. Conclusion

With the rapid increment of protein sequence data, it is indispensable to develop an automated and reliable predictive method for subcellular location. In this paper, a prediction system is provided for apoptosis proteins by constructing the dual-layer SVM based on multiple protein features. The results from two different datasets show that our method can significantly improve the predictive performance. It is anticipated that our method can be a complementary tool for the prediction of subcellular location.

Acknowledgments

The authors wish to thank Dr. Zhang Z. H. for providing the dataset ZW225. The financial supports from the National Natural Science Foundation of China (No. 20475068, 20575082), the Natural Science Foundation of Guangdong Province (No. 7003714) and the Scientific Technology Project of Guangdong Province (No. 2005B30101003) are acknowledged.

References

Angulo C, Parra X, Català A (2003) K-SVCR. A support vector machine for multi-class classification. Neurocomputing 55: 57–77

Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28: 45–48

Bhasin M, Raghava GPS (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res 32: W414–W419

Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Haussler D (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 97: 262–267

Bulashevska A, Eils R (2006) Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. BMC Bioinformatics 7: 298

Chen C, Tian YX, Zou XY, Cai PX, Mo JY (2006a) Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol 243: 444–448

Chen C, Zhou XB, Tian YX, Zhou XY, Cai PX (2006b) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357: 116–121

Chen J, Liu H, Yang J, Chou KC (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33: 423–428

Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins Struct Funct Genet 43: 246–255

Chou KC (2004) Review: structural bioinformatics and its impact to biomedical science. Curr Med Chem 11: 2105–2134

Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19

Chou KC (2006) Frontiers in medicinal chemistry. In: Atta-ur-Rahman, Reitz AB (eds) Bentham Science Publishers, The Netherlands, pp 455–502

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 227: 45765–45769

Chou KC, Cai YD (2003) Prediction and classification of protein subcellular location – sequence-order effect and pseudo amino acid composition. J Cell Biochem 90: 1250–1260

Chou KC, Cai YD (2004a) Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. J Cell Biochem 91: 1197–1203

Chou KC, Cai YD (2004b) Prediction of protein subcellular locations by GO-FunD-PseAA predictor. Biochem Biophys Res Commun 320: 1236–1239

Chou KC, Jones D, Heinrikson RL (1997) Prediction of the tertiary structure and substrate binding site of caspase-8. FEBS Lett 419: 49–54

Chou KC, Shen HB (2006a) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic k-nearest neighbor classifiers. J Proteome Res 5: 1888–1897

Chou KC, Shen HB (2006b) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347: 150–157

Chou KC, Shen HB (2006c) Predicting protein subcellular location by fusing multiple classifiers. J Cell Biochem 99: 517–527

Chou KC, Shen HB (2007a) Large-scale plant protein subcellular location prediction. J Cell Biochem 100: 665–678

Chou KC, Shen HB (2007b) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 6: 1728–1734

Chou KC, Shen HB (2007c) Review: recent progresses in protein subcellular location prediction. Anal Biochem doi: 10.1016/j.ab.2007.07.006

Chou KC, Shen HB (2007d) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Commun 357: 633–640

Chou KC, Shen HB (2007e) MemType-2L: a Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360: 339–345

Chou KC, Tomasselli AG, Heinrikson RL (2000) Prediction of the tertiary structure of a caspase-9/inhibitor complex. FEBS Lett 470: 249–256

Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol 30: 275–349

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20: 273–297

Ding CHQ, Dubchak I (2001) Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17: 349–358

Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Peptide Lett 14: 811–815

Du P, Li Y (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinformatics 7: 518

Emanuelsson O, Nielsen H, Brunak S, Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005–1016

Feng ZP (2002) An overview on predicting the subcellular location of a protein. In Silico Biol 2: 291–303

Garg A, Bhasin M, Raghava GPS (2005) Support vector machine-based method for subcellular location of human proteins using amino acid compositions, their order and similarity search. J Biol Chem 280: 14427–14432

Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, bessel function, and Chebyshev filter. Amino Acids 28: 373–376

Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J (2006) Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30: 397–402

Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 78: 3824–3828

Hua SJ, Sun ZR (2001) Support vector machine approach for protein subcellular location prediction. Bioinformatics 17: 721–728

Huang Y, Li Y (2004) Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20: 21–28

Lin H, Li QZ (2007a) Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun 354: 548–551

Lin H, Li QZ (2007b) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 28: 1463–1466

Liu DQ, Liu H, Shen HB, Yang J, Chou KC (2007) Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments. Amino Acids 32: 493–496

Matthews BW (1975) Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405: 442–451

Mondal S, Bhavna R, Mohan Babu R, Ramakumar S (2006) Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol 243: 252–260

Nakai K, Kanehisa M (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 14: 897–911

Niu B, Cai YD, Lu WC, Zheng GY, Chou KC (2006) Predicting protein structural class with AdaBoost learner. Protein Peptide Lett 13: 489–492

Park KJ, Kanehisa M (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics 19: 1656–1663

Peter ME, Heufelder AE, Hengartner MO (1997) Advances in apoptosis research. Proc Natl Acad Sci USA 94: 12736–12737

Reinhardt A, Hubbard T (1998) Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 26: 2230–2236

Shen HB, Chou KC (2007a) Using ensemble classifier to identify membrane protein types. Amino Acids 32: 483–488

Shen HB, Chou KC (2007b) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355: 1006–1011

Shen HB, Chou KC (2007c) Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85: 233–240

Shen HB, Yang J, Chou KC (2007a) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33: 57–67

Shen HB, Yang J, Chou KC (2007b) Review: methodology development for predicting subcellular localization and other attributes of proteins. Expert Rev Proteomic 4: 453–463

Shi JY, Zhang SW, Pan Q, Cheng YM, Xie J (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33: 69–74

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30: 469–475

Tanford C (1962) Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. J Am Chem Soc 84: 4240–4247

Wang M, Yang J, Chou KC (2005) Using string kernel to predict signal peptide cleavage site based on subsite coupling model. Amino Acids (Erratum, ibid. 2005, 29: 301) 28, 395–402

Wen Z, Li M, Li Y, Guo Y, Wang K (2007) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32: 277–283

Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61

Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC (2006) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30: 49–54

Yang ZR (2004) Biological applications of support vector machines. Brief Bioinform 5: 328–338

Yu CS, Lin CJ, Huwang JK (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on *n*-peptide compositions. Protein Sci 13: 1402–1406

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006) Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30: 461–468

Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids, doi: 10.1007/s00726-007-0496-1

Zhang ZH, Wang ZH, Wang YX (2006a) Prediction of the subcellular location of apoptosis-related proteins with encoding based on grouped weight for protein sequence. Acta Biophys Sin 22: 275–282

Zhang ZH, Wang ZH, Zhang ZR, Wang YX (2006b) A novel method for apoptosis protein subcellular localization prediction combing encoding based on grouped weight and support vector machine. FEBS Lett 580: 6169–6174

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17: 729–738

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins Struct Func Genet 44: 57–59

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins Struct Func Genet 50: 44–48

Authors' address: Xiao-Zou, School of Chemistry and Chemical Engineering, Sun Yat-Sen University, Guangzhou 510275, P. R. China, Fax: +86-20-84112245, E-mail: ceszxy@mail.sysu.edu.cn