Amino Acids (2012) 42:1619–1625
DOI 10.1007/s00726-011-0848-8

ORIGINAL ARTICLE

Predicting subcellular location of apoptosis proteins with pseudo amino acid composition: approach from amino acid substitution matrix and auto covariance transformation

Xiaoqing Yu · Xiaoqi Zheng · Taigang Liu · Yongchao Dou · Jun Wang

Received: 10 September 2010 / Accepted: 9 February 2011 / Published online: 23 February 2011

Abstract Apoptosis proteins are very important for understanding the mechanism of programmed cell death. Obtaining information on subcellular location of apoptosis proteins is very helpful to understand the apoptosis mechanism. In this paper, based on amino acid substitution matrix and auto covariance transformation, we introduce a new sequence-based model, which not only quantitatively describes the differences between amino acids, but also partially incorporates the sequence-order information. This method is applied to predict the apoptosis proteins' subcellular location of two widely used datasets by the support vector machine classifier. The results obtained by jackknife test are quite promising, indicating that the proposed method might serve as a potential and efficient prediction model for apoptosis protein subcellular location prediction.

Keywords Apoptosis proteins · Subcellular location · Substitution matrix · Auto covariance transformation · Support vector machine

X. Yu · X. Zheng (✉) · J. Wang
Department of Mathematics, Shanghai Normal University, 200234 Shanghai, China
e-mail: xqzheng@shnu.edu.cn

X. Zheng · J. Wang
Scientific Computing Key Laboratory of Shanghai Universities, 200234 Shanghai, China

T. Liu
College of Information Sciences and Engineering, Shandong Agricultural University, 271018 Taian, China

Y. Dou
School of Mathematical Sciences, Dalian University of Technology, 116024 Dalian, China

Introduction

Apoptosis, or programmed cell death, is a fundamental process controlling normal tissue homeostasis by regulating a balance between cell proliferation and death (Chou et al. 1997, 1999, 2000; Chou 2004a, b, c, d, 2005a, b, c; Jacobson et al. 1997). When apoptosis malfunctions, a variety of formidable diseases can ensue: blocking apoptosis is associated with cancer (Adams and Cory 1998; Evan and Littlewood 1998) and autoimmune diseases, while unwanted apoptosis can possibly lead to ischemic damage (Reed and Paternostro 1999) or neurodegenerative disease (Schulz et al. 1999). Apoptosis proteins play a central role in the mechanism of programmed cell death (Raff 1998; Steller 1995). The function of a protein is closely correlated with its subcellular location (Chou 2001; Chou and Cai 2002; Chou and Elrod 1999). To understand the apoptosis mechanism and functions of various apoptosis proteins, it is helpful to know about the subcellular location of apoptosis proteins. Therefore, the study of subcellular location of apoptosis proteins is very important in biology.

During the last decade, much work had been done in an attempt to predict the proteins' subcellular location, which are mainly focused on how to effectively represent a protein sequence and obtain the feature space of the sequence (Cedano et al. 1997; Chen and Li 2004; Chou 2001; Dubchak et al. 1995; Feng 2001; Garg et al. 2005; Gu et al. 2010; Huang and Li 2004; Nakashima and Nishikawa 1994; Zhou and Doctor 2003). Recently, some topics on the impact of the feature space were discussed (Assfalg et al. 2009, 2010). Most of the work obtained the feature space of the protein sequence based on amino acid composition (AAC; Cedano et al. 1997; Feng 2001; Nakashima and Nishikawa 1994; Zhou and Doctor 2003), dipeptide

composition (DPC) (Chen and Li 2004; Huang and Li 2004). However, these approaches treated each peptide or polypeptide separately, and their relationships were ignored. Actually, some amino acids have similar properties and thus can be substituted for each other without changing either the structure or the function of the proteins. To partially incorporate this effect, some sequence feature space models based on classifications of amino acids were proposed. For example, based on the concept of coarse-grained description and grouping, Zhang et al. (2006) presented a new encoding method with grouped weight for protein sequence (encoding based on grouped weight, named as EBGW). Recently, Zhang et al. (2009) introduced a novel representation method of protein sequence for prediction of subcellular location on the basis of distance frequency and used a novel way to calculate distance frequency. Chen and Li (2007a) also proposed a new algorithm by using a distinctive set of information parameters derived from primary sequences. By an attempt on different classifications of 20 amino acids, the prediction accuracy was greatly improved. Though the overall prediction accuracy had been improved for apoptosis proteins using existing methods, they still have some disadvantages. For example, amino acids of the same group also have discrepancies in some properties, but they could not be distinguished in the above methods. In other words, the above methods failed to describe the differences between amino acids quantitatively. In addition, the sequence-order information of protein sequences was ignored.

Actually, many studies have indicated that sequence-based prediction approaches, such as protein subcellular location prediction (Chou and Shen 2007a, 2010b), protein quaternary attribute prediction (Xiao et al. 2009), identification of proteases and their types (Chou and Shen 2008b), and signal peptide prediction (Chou and Shen 2007b; Hiss and Schneider 2009), can timely provide very useful information and insights for both basic research and drug design and hence are widely welcome by science community. The present study is attempted to develop a novel sequence-based method for predicting apoptosis protein subcellular localization in hopes that it may become a useful complementary tool to the existing methods in the relevant areas.

To avoid losing many important information hidden in protein sequences, the pseudo amino acid composition (PseAAC) was proposed (Chou 2001, 2005a) to replace the simple AAC for representing the sample of a protein. For a summary about its development and applications, such as how to use the concept of Chou's PseAAC to develop 16 different forms of PseAAC, including those that are able to incorporate the functional domain information, gene ontology (GO) information, cellular automaton image information, sequential evolution information, among many others, see a recent comprehensive review (Chou 2009). In this paper, we aim to propose a different model of PseAAC to represent protein samples via the approach of amino acid substitution matrix and auto covariance transformation. This method is applied to predict the apoptosis proteins' subcellular location of two datasets. Based on the amino acid substitution matrix, we first convert a given apoptosis protein sequence with L residues into a $20 \times L$ matrix by representing each peptide with a 20-D vector. Then the auto covariance transformation is used to transform the above representation matrix into a fixed-length vector. Finally, we employ the SVM and the jackknife test to evaluate our method. Our prediction results show that the overall prediction accuracy of apoptosis proteins subcellular location for the two datasets ZW225 and CL317 is 87.1 and 90%, respectively.

Materials and methods

Datasets

In this study, we use the two datasets constructed by Zhang et al. (2006) and Chen and Li (2007b). The former dataset (denoted as ZW225) consists of 225 apoptosis proteins divided into four subcellular locations with 41 nuclear proteins, 70 cytoplasmic proteins, 25 mitochondrial proteins and 89 membrane proteins, while proteins sequences in the second dataset (denoted as CL317) are classified into six types in subcellular locations, including 112 cytoplasmic proteins, 55 membrane proteins, 34 mitochondrial proteins, 17 secreted proteins, 52 nuclear proteins and 47 endoplasmic reticulum proteins. All the protein sequences in the two datasets are extracted from SWISS-PROT, and the accession numbers can be found in the literature (Zhou and Doctor 2003; Zhang et al. 2006).

As is well known, the sequence similarity of a dataset will seriously affect the final evaluation results, and thus should be considered in construction of a dataset. For example, Chou and Shen (2010a, b) constructed a benchmark dataset of eukaryotic proteins using a cutoff similarity threshold of 25%. But in this study we did not use a stringent threshold to cutoff the homologous sequences from the original datasets because the current two datasets, which served as widely used benchmark datasets to evaluate a new proposed method (Chen and Li 2007b; Zhang et al. 2009; Gu et al. 2010), contain too few samples to reduce the identity.

Substitution matrix

As is known, the degrees of similarity between 20 amino acids are different. The mutations between them are scored

by a 20 × 20 matrix called substitution matrix (Henikoff and Henikoff 1992; Leslid et al. 2002; Malde 2008). In bioinformatics and evolutionary biology, the substitution matrix describes the rate at which one character in a sequence changes to other character states over time. The substitution matrices are usually used in the context of amino acid or DNA sequence alignment, where the similarity between sequences depends on their divergence time and the substitution rates as represented in the matrix. In this paper, different substitution matrices are used which belong to the two well-known families: Blosum (Henikoff and Henikoff 1992) and Pam (Dayhoff et al. 1978), i.e., Blosum40, Blosum62, Blosum100, Pam40, Pam80, and Pam160.

Representation of protein sequence

We denote a given 20 × 20 substitution matrix as M, and the element $M_{i,j}$ represents the probability of amino acid i mutating to amino acid j during the evolution process (i, $j = 1, 2,\ldots,20$). The matrix M could be denoted as a 20-D vector, that is, $M = (V_1, V_2,\ldots,V_{20})$, where $V_i = (M_{1,i}, M_{2,i},\ldots,M_{20,i})^{\mathrm{T}}$. For a given protein sequence $S = s_1 s_{2\ldots} s_{\mathrm{L}}$, s_i represents the ith amino acid of the protein sequence and could be substituted by a vector V_{s_i} of the substitution matrix M. Then, we can easily obtain a 20 × L matrix D. For convenience, let us denote

$$D = (V_{s_1}, V_{s_2}, \ldots, V_{s_L})$$

to describe the given protein sequence.

In order to employ SVM classifier to perform our method, the protein sequences should be converted into fixed-length vectors. AAC is a conventional feature construction method, which refers to the occurrence frequency of each of these 20 components in a given protein sequence. Since the information in the primary sequence is greatly reduced by considering the AAC alone, other informative features should be taken into account within our studies. Here, the auto covariance (AC) transformation is introduced to convert the above matrix D into a fixed-length vector. As a statistical tool for analyzing sequences of vectors developed by Wold et al. (1993), AC transformation has been successfully used for protein family classification (Guo et al. 2006; Lapinsh et al. 2002), protein interaction prediction (Guo et al. 2008) and prediction of secondary structure content (Lin and Pan 2001; Zhang et al. 1998, 2001). Here, the AC variable measures the average correlation between two residues separated by a distance of lg along the sequence S, which can be calculated by

$$AC(i, \mathrm{lg}) = \sum_{j=1}^{L-\mathrm{lg}} \left(D_{i,j} - \bar{D}_i\right)\left(D_{i,j+\mathrm{lg}} - \bar{D}_i\right)/(L - \mathrm{lg})$$

where i denotes the ith amino acid, L is the length of the protein sequence, $D_{i,j}$ is the matrix score of amino acid i at position j, $\bar{D}_i$ is the average score for amino acid i along the whole sequence:

$$\overline{D_i} = \sum_{j=1}^{L} D_{i,j}/L$$

in such way, the number of AC variables can be calculated as 20 × LG, where LG is the maximum of lg (lg = 1, 2,…,LG). Combining the 20 AAC and the 20 × LG AC variables, each given protein sequence is characterized by a (20 + 20 × LG)-D feature vector.

Support vector machine

In recent years, SVM-based machine learning algorithm has been used for predicting various protein attributes tasks, such as membrane protein type (Cai et al. 2004), protein structural class (Cai et al. 2002a; Ding et al. 2007), specificity of GalNAc-transferase (Cai et al. 2002c), HIV protease cleavage sites in protein (Cai et al. 2002b), and so on. The algorithm often obtains higher prediction accuracy compared with other classification approaches, when the invariant feature vectors are used (Cai et al. 2002a; Hua and Sun 2001; Huang and Shi 2005; Zhou et al. 2007). The basic idea of applying SVM to pattern classification can be stated briefly: first, map the input vectors into one feature space; then, within this feature space, construct a hyperplane which can separate the two classes. SVM training always seeks a global optimized solution and avoids overfitting, so it has the ability to deal with a large number of features.

In our study, the LIBSVM package is used to implement the SVM classifier (Chang and Lin 2009). The radial basis function (RBF) is chosen as the kernel function, which is defined as $K(x, x') = \exp\left(-\gamma|x - x'|^2\right)$. Two parameters, the regularization parameter C and the kernel width parameter γ are optimized on the training set using a grid search strategy in the LIBSVM.

Evaluation methods

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 jackknife test (Chou and Zhang 1995). However, as elucidated in (Chou and Shen 2008a) and demonstrated by Eq. 50 of (Chou and Shen 2007a), among the three cross-validation methods, 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 (Chen et al. 2009; Ding et al.

2009; Jiang et al. 2008; Li and Li 2008; Lin 2008; Lin et al. 2008; Zeng et al. 2009; Zhou 1998; Zhou et al. 2007). So, in this paper, jackknife test is employed to evaluate the prediction performance of our method. Each protein sequence in the samples is singled out in turn as a test sample, and the remaining protein sequences are used as training samples. To evaluate the performance of the test, the overall prediction accuracy A_c, individual sensitivity S_{in}, individual specificity S_{ip} and Matthews's correlation coefficient MCC_i are discussed, and they are calculated as follows:

$$S_{in} = \frac{TP_i}{TP_i + FN_i}$$

$$S_{ip} = \frac{TN_i}{TN_i + FP_i}$$

$$MCC_i = \frac{TP_i TN_i - FP_i FN_i}{\sqrt{(TP_i + FP_i)(TP_i + FN_i)(TN_i + FP_i)(TN_i + FN_i)}}$$

$$A_c = \frac{\sum_i TP_i}{N}$$

where TP_i denotes the numbers of the ith subcellular location correctly recognized positives, FN_i denotes the numbers of the ith subcellular location recognized as other subcellular location, FP_i denotes the numbers of other subcellular location recognized as the ith subcellular location, TN_i denotes the numbers of other subcellular location correctly recognized. N is the number of all protein sequences.

Results and discussion

Firstly, the dataset CL317 is applied to validate the proposed method. We transform each apoptosis protein sequence into a fixed-length vector through the substitution matrix and auto covariance transformation. Then these feature vectors are fed to the SVM classifier to perform our prediction. In this paper, we select radial basis kernel function to build the prediction model and the two parameters C and γ are set at $C = 128$, $\gamma = 8$. In order to optimize our prediction accuracy, we try to investigate the effect of the parameter LG value and different substitution matrices (Blosum40, Blosum62, Blosum100, Pam40, Pam80, Pam160) variation on the quality of our method; results are shown in Fig. 1. As is seen from Fig. 1, the accuracy first increases to a maximum value and then slightly goes down as the value of LG increases, but last with a little fluctuation. The best prediction accuracy reaches 90% for the dataset CL317, when the value of LG is 15, and the substitution matrix is Blosum100. It is worth mentioning that the Blosum matrices, which are based on the replacement patterns found in more highly conserved regions of the sequences, were also proved to have better performance in many research (Henikoff and Henikoff 1993; Johnson and Overington 1993).

In order to validate the performance of the proposed approach further, the dataset ZW225 is adopted. We also employ the parameter LG = 15 and substitution matrix Blosum100 on the dataset ZW225. Through SVM classifier

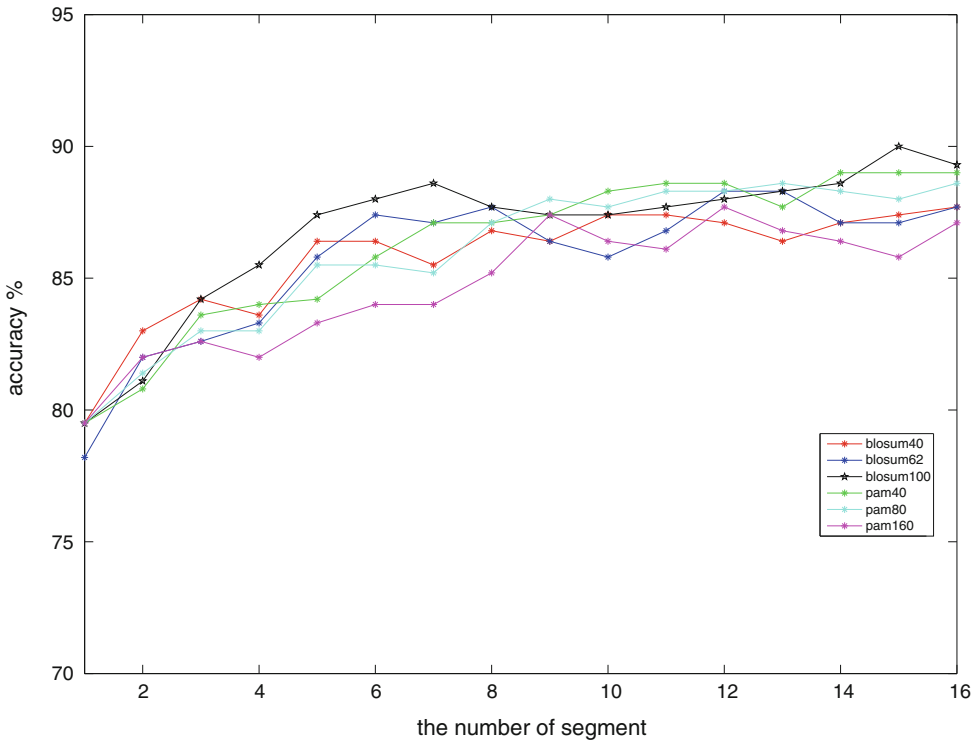

Fig. 1 Effect of the LG values and substitution matrices on dataset CL317 in jackknife test

Table 1 Prediction results on two datasets in jackknife test

Dataset	Subcellular location	Sensitivity	Specificity	MCC	Overall accuracy (%)
ZW225	Cyto	81.3	91.0	76.7	87.1
	Memb	93.3	95.6	89.8	
	Mito	85.7	98.5	76.2	
	Nucl	84.6	96.8	78.8	
CL317	Cyto	86.4	92.2	82.3	90.0
	Memb	90.7	98.1	87.8	
	Mito	93.8	99.3	89.9	
	Nucl	92.2	98.5	89.6	
	Secr	85.7	99.3	76.7	
	Endo	93.8	98.9	89.9	

and the jackknife test, the prediction results of dataset CL317 and ZW225 are listed in Table 1.

From Table 1, we can see that the overall accuracies for ZW225 and CL317 datasets by our method achieve 87.1 and 90%, respectively. Table 2 shows the prediction results of different methods by the jackknife test for the ZW225 dataset. We can find that the overall accuracy by our method is higher than that of EBGW_SVM (Zhang et al. 2006), DF_SVM (Zhang et al. 2009), ID_SVM (Chen and Li 2007b). The value of sensitivity for each protein class is listed. For example, the sensitivity of mitochondrial proteins reaches 85.7% in our method, while the others are 60, 64, 68 and 60%. For the nuclear proteins, the sensitivity of our method is 84.6%, which is also the highest. To evaluate the performance of our method, we also compared other methods with our method on the CL317 dataset in Table 3. The overall accuracy of our method reaches 90%, which is slightly lower than FKNN (Jiang et al. 2008), FKNN (Ding and Zhang 2008), PseAAC_SVM (Lin et al. 2009), EN_FKNN (Gu et al. 2010), but higher than the other three methods (Chen and Li 2007a, b; Zhang et al. 2009). All the results indicate that the proposed method has a good performance for prediction of subcellular locations. The successful performance of our method may be attributed to the following reasons: (1) compared with the conventional classification-based and composition-based methods, our approach making use of the Blosum100 matrix could quantitatively measure various degrees of similarity between amino acids; (2) the parameter value LG = 15 is adopted in the auto covariance transformation, so our method considered correlations between not only neighbor residues but also residues with a long distance in a sequence, which could describe more sequence-order information.

Table 2 Comparison of different methods by the jackknife test on ZW225 dataset

Method	Sensitivity for each class (%)				Overall accuracy (%)
	Cyto	Memb	Mito	Nucl	
EBGW_SVM[a]	90.0	93.3	60.0	63.4	83.1
DF_SVM[b]	87.1	92.1	64.0	73.2	84.0
ID_SVM[c]	92.9	91.0	68.0	73.2	85.8
EN_FKNN[d]	94.3	94.4	60.0	80.5	88.0
Our method	81.3	93.3	**85.7**	**84.6**	87.1

The bold values indicate the highest accuracies achieved by our method

[a] Zhang et al. (2006)

[b] Zhang et al. (2009)

[c] Chen and Li (2007b)

[d] Gu et al. (2010)

Table 3 Comparison of different methods by the jackknife test on CL317 dataset

Method	Sensitivity for each class (%)						Overall accuracy (%)
	Cyto	Memb	Mito	Secr	Nucl	Endo	
ID[a]	81.3	81.8	85.3	88.2	82.7	83.0	82.7
ID_SVM[b]	91.1	89.1	79.4	58.8	73.1	87.2	84.2
DF_SVM[c]	92.9	85.5	76.5	76.5	93.6	86.5	88.0
FKNN[d]	92.0	89.1	85.3	76.5	92.3	93.7	90.2
FKNN[e]	93.8	92.7	82.4	76.5	90.4	93.6	90.9
PseAAC_SVM[f]	93.8	90.9	85.3	76.5	90.4	95.7	91.1
EN_FKNN[g]	98.2	83.6	79.4	82.4	90.4	97.9	91.5
Our method	86.4	90.7	93.8	85.7	92.1	93.8	90.0

[a] Chen and Li (2004)

[b] Chen and Li (2007b)

[c] Zhang et al. (2009)

[d] Jiang et al. (2008)

[e] Ding and Zhang (2008)

[f] Lin et al. (2009)

[g] Gu et al. (2010)

Conclusions

Based on amino acid substitution matrix and auto covariance transformation, a new representation model for

protein sequence was presented, and applied to predict the apoptosis proteins subcellular location. Two datasets CL317 and ZW225 are selected to validate the performance of our proposed method. Comparing with other feature extraction approaches, our model is shown effectively in obtaining information from protein sequences. The experiment results indicated that the proposed method is promising. With the growing amount of the size of the datasets, we hope that our model will be a useful complementary tool to the existing methods for further study in the prediction of apoptosis proteins subcellular location.

Acknowledgments This work was partially supported by the National Natural Science Foundation of China (No. 10731040), Shanghai Leading Academic Discipline Project (No. S30405) and Innovation Program of Shanghai Municipal Education Commission (No. 09zz134).

References

Adams JM, Cory S (1998) The Bcl-2 protein family: arbiters of cell survival. Science 281:1322–1326

Assfalg J, Gong J, Kriegel HP, Pryakhin A, Wei T, Zimek A (2009) Supervised ensembles of prediction methods for subcellular localization. J Bioinform Comput Biol 7(2):269–285

Assfalg J, Gong J, Kriegel HP, Pryakhin A, Wei T, Zimek A (2010) Investigating a correlation between subcellular localization and fold of proteins. J UCS 16(5):604–621

Cai YD, Liu XJ, Xu XB, Chou KC (2002a) Prediction of protein structural classes by support vector machines. Comput Chem 26:293–296

Cai YD, Liu XJ, Xu XB, Chou KC (2002b) Support vector machines for predicting HIV protease cleavage sites in protein. J Comput Chem 23:267–274

Cai YD, Liu XJ, Xu XB, Chou KC (2002c) Support vector machines for predicting the specificity of GalNAc-transferase. Peptides 23:205–208

Cai YD, Liu XJ, Xu XB, Chou KC (2002d) Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J Cell Biochem 84:343–348

Cai YD, Pong Wong R, Feng K, Jen JCH, Chou KC (2004) Application of SVM to predict membrane protein types. J Theor Biol 226:373–376

Cedano J, Aloy P, Pérez-Pons JA, Querol E (1997) Relation between amino acid composition and cellular location of proteins. J Mol Biol 266:594–600

Chang C, Lin CJ (2009) Libsvm: a library for support vector machines. http://www.csie.ntu.edu.tw/cjlin/libsvm

Chen YL, Li QZ (2004) Prediction of the subcellular location apoptosis proteins using the algorithm of measure of diversity. Acta Sci Nat Univ Nei Mong 25:413–417

Chen YL, Li QZ (2007a) Prediction of the subcellular location of apoptosis proteins. J Theor Biol 245:775–783

Chen YL, Li QZ (2007b) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J Theor Biol 248:377–381

Chen C, Chen L, Zou X, Cai P (2009) Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein Pept Lett 16:27–31

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: structure, function, and genetics (Erratum: ibid., 2001, vol. 44, 60) 43:246–255

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

Chou KC (2004b) Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor. Biochem Biophys Res Commun 319:433–438

Chou KC (2004c) Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. Biochem Biophys Res Commun 316:636–642

Chou KC (2004d) Molecular therapeutic target for type-2 diabetes. J Proteome Res 3:1284–1288

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

Chou KC (2005b) Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein. J Proteome Res 4:1681–1686

Chou KC (2005c) Prediction of G-protein-coupled receptor classes. J Proteome Res 4:1413–1418

Chou KC (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr Proteomics 6:262–274

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277:45765–45769

Chou KC, Elrod DW (1999) Protein subcellular location prediction. Protein Engine 12:107–118

Chou KC, Shen HB (2007a) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16

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

Chou KC, Shen HB (2008a) Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3:153–162

Chou KC, Shen HB (2008b) ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochem Biophys Res Comm 376:321–325

Chou KC, Shen HB (2010a) A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS ONE 5:e9931

Chou KC, Shen HB (2010b) Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. PLoS ONE 5:e11335

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

Chou KC, Zhang TC, Maggiora MG (1997) Disposition of amphiphilic helices in heteropolar environments. Proteins 28:99–108

Chou JJ, Li H, Salvessen GS, Yuan J, Wagner G (1999) Solution structure of BID, an intracellular amplifier of apoptotic signalling. Cell 96:615–624

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

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins, vol 5. National Biomedical Research Foundation, Washington, pp 345–352

Ding YS, Zhang TL (2008) Using Chous pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. Pattern Recognit Lett 29:1887–1892

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

Ding H, Luo L, Lin H (2009) Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein Pept Lett 16:351–355

Dubchak I, Muchnik I, Holbrook SR, Kim SH (1995) Prediction of protein folding class using global description of amino acid sequence. PNAS USA 92:8700–8704

Evan G, Littlewood T (1998) A matter of life and cell death. Science 281:1317–1322

Feng ZP (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 58:491–499

Garg A, Bhasin M, Raghava GPS (2005) Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J Biol Chem 280(15):14427–14432

Gu Q, Ding YS, Jiang XY, Zhang TL (2010) Prediction of subcellular location apoptosis proteins with ensemble classifier and feature selection. Amino Acids 38(4):975–983

Guo Y, Li M, Lu M, Wen Z, Huang Z (2006) Predicting g-protein coupled receptors-g-protein coupling specificity based on auto-cross-covariance transform. Proteins 65:55–60

Guo YZ, Yu LZ, Wen ZN, Li ML (2008) Using support vector machine combined with auto covariance to predict protein–protein interactions form protein sequences. Nucleic Acids Res 36:3025–3030

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Natl Acad Sci USA 89:10915–10919

Henikoff S, Henikoff JG (1993) Performance evaluation of amino acid substitution matrices. Protein Struct Funct Genet 17:49–61

Hiss JA, Schneider G (2009) Architecture, function and prediction of long signal peptides. Brief Bioinform 10:569–578

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

Huang Y, Li Y (2004) Prediction of protein subcellular location using fuzzy k-NN method. Bioinformatics 20(1):121–128

Huang J, Shi F (2005) Support vector machines for predicting apoptosis proteins types. Acta Biotheor 53:39–47

Jacobson MD, Weil M, Raff MC (1997) Programmed cell death in animal development. Cell 88:347–354

Jiang X, Wei R, Zhang T, Gu Q (2008) Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Pept Lett 15:392–396

Johnson MS, Overington JP (1993) A structural basis of sequence comparisons: an evaluation of scoring methodologies. J Mol Bio 233:716–738

Lapinsh M, Gutcaits A, Prusis P, Post C, Lundstedt T, Wikberg JE (2002) Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. Protein Sci 11:795–805

Leslid C, Eskin E, Noble WS (2002) The spectrum kernel: a string kernel for SVM protein classification. In: Pacific symposium on biocomputing (PSB), pp 564–575

Li FM, Li QZ (2008) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Pept Lett 15:612–616

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

Lin Z, Pan XM (2001) Accurate prediction of protein secondary structural content. J Protein Chem 20:217–220

Lin H, Ding H, Guo FB, Zhang AY, Huang J (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept Lett 15:739–744

Lin H, Wang H, Ding H, Chen YL, Li QZ (2009) Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition. Acta Biotheor 57:321–330

Malde K (2008) The effect of sequence quality on sequence alignment. Bioinformatics 24(7):897–900

Nakashima H, Nishikawa K (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue pair frequencies. J Mol Biol 238:54–61

Raff M (1998) Cell suicide for beginners. Nature 396:119–122

Reed JC, Paternostro G (1999) Postmitochondrial regulation of apoptosis during heart failure. Proc Natl Acad Sci USA 96:7614–7616

Schulz JB, Weller M, Moskowitz MA (1999) Caspases as treatment targets in stroke and neurodegenerative diseases. Ann Neurol 45:421–429

Steller H (1995) Mechanisms and genes of cellular suicide. Science 267:1445–1449

Wold S, Jonsson J, Ssjörström M, Sandberg M, Rännar S (1993) DNA and peptide sequences and chemical processes multivariately modeled by principal component analysis and partial least-squares projections to latent structures. Anal Chim Acta 277:239–253

Xiao X, Wang P, Chou KC (2009) Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition. J Appl Crystallogr 42:169–173

Zeng YH, Guo YZ, Xiao RQ, Yang L, Yu LZ, Li ML (2009) Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J Theor Biol 259:366–372

Zhang CT, Lin ZS, Zhang ZD, Yan M (1998) Prediction of the helix/strand content of globular proteins based on their primary sequences. Protein Eng 11:971–979

Zhang ZD, Sun ZR, Zhang CT (2001) A new approach to predict the helix/strand content of globular proteins. J Theor Biol 208:65–78

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

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

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

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins 50:40–48

Zhou XB, Chen C, Li ZC, Zou XY (2007) Using Chou's amphiphilic pseudoamino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 248:546–551