Amino Acids (2010) 38:1201–1208
DOI 10.1007/s00726-009-0331-y

ORIGINAL ARTICLE

Predicting subcellular location of apoptosis proteins based on wavelet transform and support vector machine

Jian-Ding Qiu · San-Hua Luo · Jian-Hua Huang · Xing-Yu Sun · Ru-Ping Liang

Received: 6 February 2009 / Accepted: 20 July 2009 / Published online: 4 August 2009

Abstract Apoptosis proteins have a central role in the development and homeostasis of an organism. These proteins are very important for understanding the mechanism of programmed cell death. As a result of genome and other sequencing projects, the gap between the number of known apoptosis protein sequences and the number of known apoptosis protein structures is widening rapidly. Because of this extremely unbalanced state, it would be worthwhile to develop a fast and reliable method to identify their subcellular locations so as to gain better insight into their biological functions. In view of this, a new method, in which the support vector machine combines with discrete wavelet transform, has been developed to predict the subcellular location of apoptosis proteins. The results obtained by the jackknife test were quite promising, and indicated that the proposed method can remarkably improve the prediction accuracy of subcellular locations, and might also become a useful high-throughput tool in characterizing other attributes of proteins, such as enzyme class, membrane protein type, and nuclear receptor subfamily according to their sequences.

Keywords Apoptosis protein · Subcellular location · Discrete wavelet transform · Support vector machines · Hydrophobicity

Introduction

Apoptosis, or programmed cell death, is a fundamental process controlling normal tissue homeostasis by regulating the balance between cell proliferation and cell death (Jacobson et al. 1997; Vaux et al. 1994; Zhou et al. 1999). This process entails the autolytic degradation of cellular components, and is characterized by blebbing of cell membranes, shrinkage of cell volumes, and condensation of nuclei (Kerr et al. 1972), and is currently an area of intense investigation. Cell death and renewal are responsible for maintaining the proper turnover of cells, which ensures a constant controlled flux of fresh cells. Programmed cell death and cell proliferation are tightly coupled. When apoptosis malfunctions, a variety of formidable diseases will ensue, such as cancer (Adams and Cory 1998; Evan and Littlewood 1998) and autoimmune diseases caused by blocking apoptosis, ischemic damage (Reed and Paternostro 1999) and neurodegenerative diseases (Schulz et al. 1999) possibly caused by unwanted apoptosis. Apoptosis is considered to play a key role in these devastating diseases and, in principle, provides many targets for therapeutic intervention (Barinaga 1998; Chou et al. 1998, 1999).

Obtaining information about subcellular locations of apoptosis proteins is very helpful in understanding the apoptosis mechanism and functions of proteins (Schulz et al. 1999; Suzuki et al. 2000). However, it is both expensive and time-consuming to conduct various experiments to obtain relevant information. With the explosion of

Electronic supplementary material The online version of this article (doi:10.1007/s00726-009-0331-y) contains supplementary material, which is available to authorized users.

J.-D. Qiu (✉) · S.-H. Luo · J.-H. Huang · X.-Y. Sun · R.-P. Liang
Department of Chemistry, Nanchang University, 330031 Nanchang, People's Republic of China
e-mail: jdqiu@ncu.edu.cn

J.-D. Qiu
Department of Chemical Engineering, Pingxiang College, 337055 Pingxiang, People's Republic of China

protein sequences generated in the post-genomic era, it is both challenging and indispensable to develop an automated method to quickly and reliably annotate the subcellular attributes of uncharacterized proteins. The knowledge thus obtained can help us timely utilize these newly found protein sequences for both basic research and drug discovery (Chou 2004). Therefore, it is urgent to develop an automatic and reliable prediction system for protein subcellular location.

Actually, many efforts have been made in this regard (Bulashevska and Eils 2006; Cedano et al. 1997; Chen and Li 2004; Chen and Li 2007a, b; Chou 2001; Chou and Shen 2006; Ding and Zhang 2008; Feng 2001; Huang et al. 2005; Nakashima and Nishikawa 1994; Park and Kanehisa 2003; Zhang et al. 2006; Zhou and Doctor 2003), most of which were based on amino acid composition (Cedano et al. 1997; Feng 2001; Nakashima and Nishikawa 1994; Park and Kanehisa 2003), where the sample of a protein was represented by 20 discrete numbers, with each representing the occurrence frequency of 20 different constituent native amino acids. Obviously, if one uses the conventional amino acid composition to represent the sample of a protein, all the effects of sequence order and length will be lost. Recently, some new protein features have been proposed in order to incorporate sequence order effects of proteins, including pseudo amino acid composition (Chen and Li 2007b; Chou 2001; Ding and Zhang 2008; Chou and Shen 2007; Shen and Chou 2007; Shi et al. 2007; Zhang et al. 2006, 2008; Chou and Cai 2003, 2004), dipeptide composition (DPC) (Chen and Li 2004; Bhasin and Raghava 2004a, b; Huang and Li 2004; Zhou et al. 2008), Markov chains model (Bulashevska and Eils 2006) and so on. Stimulated by these new representation methods, the present study was initiated in an attempt to introduce a novel approach—the wavelet transform analysis to formulate the protein features.

In this paper, a novel hybridization classifier (DWT_SVM) was developed by fusing discrete wavelet transform (DWT) with support vector machine (SVM) based on the amino acid hydrophobicity in order to predict subcellular locations of apoptosis proteins. First, the amino acids of apoptosis proteins are transformed into sequences of hydrophobic free energies per residue. Second, the hydrophobic profile is decomposed into wavelet coefficients using DWT. Following this, using the statistical method, a series of statistical feature vectors are constructed to represent the apoptosis proteins. Finally, SVM is used to model with these statistics feature vectors. The influence of amino acid hydrophobic values, wavelet functions and decomposition scales on the result has been discussed. The prediction results of the jackknife cross-validation test show significant improvement compared with the previous algorithms, and hence the methodology presented here might be useful for other studies of protein structure and function.

Materials and methods

Datasets

To have a critical comparison between different approaches, three datasets were adopted in our work. Proteins in those datasets were extracted from SWISS-PROT (version 49.5). The ZD98 dataset consists of 98 apoptosis protein sequences, which include 43 cytoplasmic proteins, 30 plasma membrane-bound proteins, 13 mitochondrial proteins and 12 other proteins (Zhou and Doctor 2003). The ZW225 dataset consists of 41 nuclear proteins, 70 cytoplasmic proteins, 25 mitochondrial proteins and 89 membrane proteins (Zhang et al. 2006). The dataset CL317 constructed by Chen and Li, consists of 112 cytoplasmic proteins, 55 membrane proteins, 34 mitochondrial proteins, 17 secreted proteins, 52 nuclear proteins and 47 endoplasmic reticulum proteins (Chen and Li 2007a).

Discrete wavelet transform

A protein sequence can be represented as a series of amino acids by their single-character codes A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y, formulated as

$$R_1R_2R_3R_4R_5R_6R_7R_8\ldots R_L \quad (1)$$

Suppose $H(R_1)$ is the hydrophobic value of the 1st residue R_1, $H(R_2)$ that of the 2nd residue R_2, and so forth. In terms of these hydrophobic values the protein sequence of Eq. 1 can be converted to a digit signal (Qiu et al. 2003, 2004), from which we can generate several groups of wavelet coefficients using WT. A WT is defined as the projection of a function or a signal $f(t)$ onto the wavelet function.

$$T(a,b) = \frac{1}{\sqrt{a}} \int_0^t f(t)\psi\left(\frac{t-b}{a}\right)\mathrm{d}t, \quad (2)$$

where a is a scale variable and b is a translation variable, they belong to the real number $R(n)$, and $a > 0$. $\psi\left(\frac{t-b}{a}\right)$ is the analyzing wavelet function. It is used to plot $T(a,b)$ against a and b in a surface plot known as a scalogram, which is particularly suited to the detection of singularities. DWT analysis can decompose the amino acid sequences into coefficients at different dilations and then remove the noise component from the profiles, so it can give us local structures of sequences which can more effectively reflect the sequence order effects. In addition, the DWT is an economical way to compute the WT, because it is

computed only on a dyadic grid of points, where the subsampling is at a different rate for different scales. So, in this work the DWT is the preferred wavelet representation. The DWT uses $a_0 = 2$ and $b_0 = 1$, so that the results can lead to a binary dilation of 2^{-m} and a dyadic translation of $n2^{-m}$. Therefore,

$$\psi_{m,n}(t) = 2^{-m}\psi(2^{-m}t - n) \tag{3}$$

Here, $m = 1, 2,\ldots$, and $n = 0, 1, 2,\ldots$. The wavelet coefficients of the signal $f(t)$ are obtained by following formula:

$$T(a,b) = \langle f(t), \psi_{a,b}(t)\rangle = 2^{-\mathrm{m}/2}\int_0^t f(t)\psi(2^{-m}\cdot t - n) \tag{4}$$

Although Eq. 4 can now be used to estimate the feature, it is redundant and involves extremely large computations. It is often reasonable to assume that only a few coefficients contain information about the underlying function, while other coefficients can be attributed to noise. Therefore, the following exponents of 'maximum line', 'minimum line', 'mean line' and 'standard deviation line' were used to predict the subcellular location of apoptosis proteins. The maximum line, minimum line, mean line and standard deviation line of a protein subcellular location feature are the lines joining maximum, minimum, mean and standard deviation of its wavelet coefficients at different scales (Qiu et al. 2009a, b, c). Consequently, sequence p_k can be characterized as a $4(m + 1)$ dimension feature vector, which can be put in SVM directly.

$$Z(k) = \max(k) + \min(k) + \mathrm{mean}(k) + \mathrm{stdev}(k) \tag{5}$$

Here, $\max(k)$, $\min(k)$, $\mathrm{mean}(k)$, stdev(k) are maximum, minimum, mean and standard deviation of the wavelet coefficients in each sub-band, respectively, and can be defined as (see Appendix):

1. Maximum of the wavelet coefficients in each sub-band.
2. Minimum of the wavelet coefficients in each sub-band.
3. Mean of the wavelet coefficients in each sub-band.
4. Standard deviation of the wavelet coefficients in each sub-band.

Multi-class SVM

The SVM introduced by Vapnik (1995) has proven to be a useful learning machine, especially for classification. A classification problem usually involves training data and testing data that consist of some data instances. Each instance in training data contains one class label and one feature vector. The goal of SVM is to construct a classifier that classifies the data instances in the testing data. For a binary classification problem, assume x_i ($i = 1, 2,\ldots,N$) to be input training vectors and $y_i \varepsilon \{+1,-1\}$be their corresponding target classes. Let N be the total number of input vectors. The SVM classification problem can then be formulated in terms of a convex quadratic optimization problem, formulated as:

$$\max \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \tag{6}$$

Subject to,

$$\text{(I)}\quad 0 \le \alpha_s \le C; \quad i = 1, 2, \ldots, N \quad \text{(II)} \quad \sum_{i=1}^{n} y_i \alpha_i = 0$$

C is a regularization parameter that controls the trade-off between margin and classification error. Function $K(x_i, x_j)$ is called the SV kernel when satisfying the Mercer condition principle. A SV kernel function is shown by the following formula (7).

$$K(x_i, x_j) = \tau(x_i)^T \tau(x_j) \tag{7}$$

In this study, the radial basis function (RBF) was selected as the kernel function, formulated as:

$$K(x, x_i) = \exp\left\{-\frac{1}{2\sigma^2}\|x - x_i\|^2\right\} \tag{8}$$

where, σ is the kernel width parameter which is automatically tuned based on the training set using the grid search strategy in the LIBSVM software (Chang and Lin 2002).

The multi-class classification problem is commonly solved by a decomposing and reconstructing procedure when the binary class SVM is implied. There are several methods to extend the SVM for classifying multi-class problems, for example 'One-Versus-Rest (OVR)' (Vapnik 1998), 'One-Versus-One (OVO)' (Kreßel 1999), and DAGSVM (Platt et al. 2000). This paper used the 'One-Versus-One' strategy. For a k-classification problem, the OVO strategy constructs $k*(k - 1)/2$ classifiers with each one trained with the data from two different classes. The software used to implement the SVM in this paper is LIBSVM written by Chang and Lin (Joachims 1999; Chang and Lin 2002) and can be freely downloaded from: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Results and discussion

Effect of decomposition scales

A WT decomposes a signal into several groups (vectors) of coefficients. Different coefficient vectors contain information about characteristics of the sequence at different

scales. Coefficients at coarse scales capture gross and global features of the signal while coefficients at fine scales contain local details (Hirakawa et al. 1999). The curves of wavelet coefficients at four decomposition scales are shown in Fig. 1. The original protein signal is represented by "S", A_m denotes the approximation at level m and D_m denotes the detail at level m. It can be seen that along with the decomposition processed by DWT, the feature information in amino acid sequence of each scale differs.

Restricted by the property of wavelet decomposition, different decomposition scales have different results in analyzing protein sequences. On the one hand, decomposing a shorter sequence with too high a decomposition scale will introduce ineluctable redundancy in the decomposing process. On the other hand, decomposing a longer sequence with too low a decomposition level will omit much detailed information. In order to gain the highest predictive accuracy, an appropriate decomposition scale a is selected (Qiu et al. 2009a, b, c). To choose the appropriate decomposition scale, the test sequences are decomposed with scales from 2 to 8 separately with the aforementioned 317 proteins. Table 1 shows the effect of decomposition scales on CL317 in re-substitution test. From Table 1, we can see that a significant increase in accuracy can be observed for those proteins with decomposition scale 3, and the accuracy is about 99.4%, which is higher than that of other decomposition scales. Therefore, scale 3 is selected as the appropriate decomposition scale for the detection of subcellular locations in this study.

Effect of wavelet functions

In wavelet theory, a function is represented by an expansion of infinite series in terms of a dilated and translated version of a basic function ψ called the 'mother' wavelet (Daubechies 1992; Grunbaum 1992; Mallat 1989). The simplest example of a wavelet basis is the Harr basis; other frequently used wavelet bases are those developed by Daubechies (1992). Several wavelet families, with different properties (orthogonal, biorthogonal, semiorthogonal) have recently been developed (Daubechies 1992; Grunbaum 1992; Mallat 1999). For example, the Daubechies were the first compactly supported orthonormal wavelets. Symlet wavelets are a slightly symmetrical version of the Daubechies. The Coifman wavelets are a more symmetrical version of the Daubechies (Walczak 2000). The Biorthogonal wavelets are designed to overcome the conflict between symmetry and exact reconstruction (except Haar).

Currently, there is no standard method to select a wavelet function in WT. Some criteria have been proposed to select a wavelet. One of them was that the wavelet and signal should have good similarities. Table 2 shows the predictive results performed by different wavelet functions

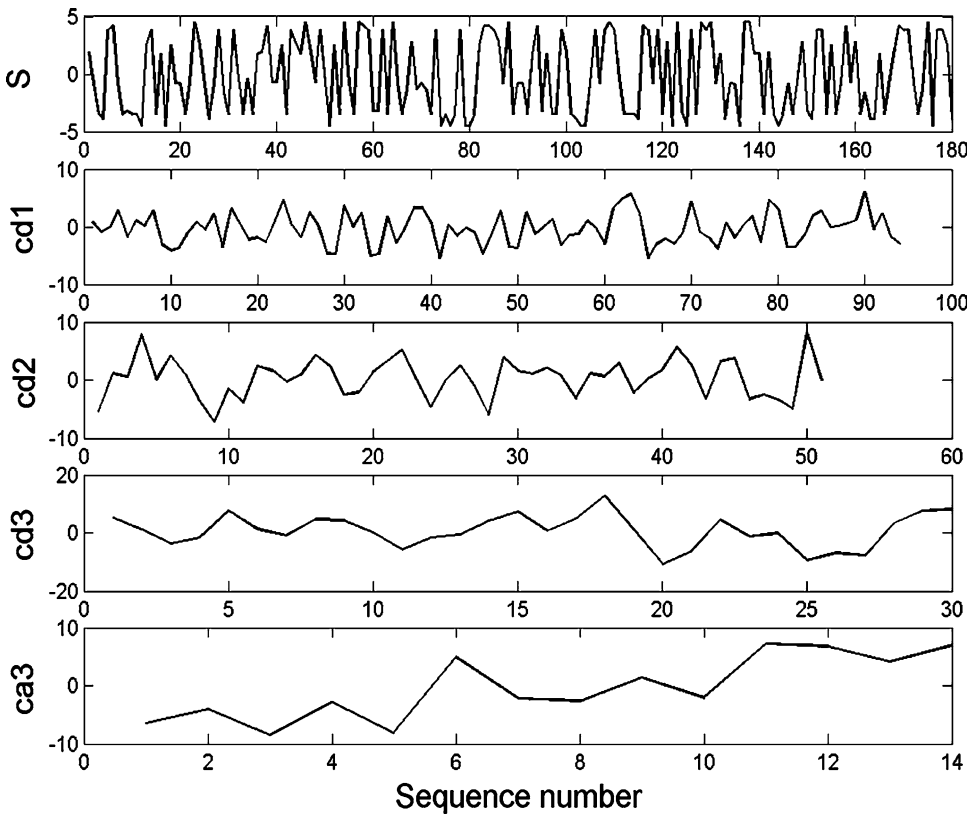

Fig. 1 DWT coefficient plot of *Q8CJ70* (Brookhaven Protein Databank accession: *Q8CJ70_Secreted protein*) protein by using the Bior3.3 wavelet and Kyte-Doolittle hydrophobicity scales

Table 1 Experiment results of different decomposition scales on CL317 in re-substitution test (%)

Decomposition scales	Sensitivity for each class (%)						Overall accuracy
	Cyto	Memb	Mito	Secr	Nucl	Endo	
2	89.3	92.7	100	88.2	92.3	93.6	92.1
3	100	98.2	100	94.1	100	100	99.4
4	100	100	100	41.2	100	100	96.8
5	100	100	100	29.4	100	97.9	95.9
6	100	100	100	88.2	100	97.9	99.1
7	100	100	100	100	100	89.4	98.4
8	52.9	100	100	100	100	100	97.2

Table 2 Experiment results of different wavelets on CL317 in re-substitution test (%)

Wavelets	Sensitivity for each class (%)						Overall accuracy
	Cyto	Memb	Mito	Secr	Nucl	Endo	
Db1	100	52.7	100	52.9	100	97.9	89.0
Db6	100	100	100	52.9	100	100	97.5
Db10	95.5	96.4	100	100	100	68.1	93.1
Meyer	100	100	100	58.8	88.5	100	95.9
Sym10	100	100	100	64.7	75	100	94.0
Coif4	89.3	92.7	94.1	47.1	86.5	95.7	88.6
Bior2.4	100	100	100	47.1	100	100	97.2
Bior3.3	100	98.2	100	94.1	100	100	99.4

at a decomposition level of 3 and Kyte-Doolittle hydrophobicity scales (KDHΦ). The training accuracy can reach 99.4% when using Bior3.3 wavelet ($m = 3$, $n = 8$) to extract feature. However, when using the other wavelet functions, the training accuracy only ranges from 88.6 to 97.5%. Furthermore, the training accuracies of mitochondrial proteins, endoplasmic reticulum proteins, nuclear proteins and cytoplasmic proteins can reach 100%. So, in this paper, Bior3.3 wavelet function is selected as the decomposition wavelet.

Influence of hydrophobicity data types

Most of the methods are based on the physicochemical properties that contribute to the function of proteins. Among various properties, the most important one is hydrophobicity, which determines the stability of the protein structure (Eisenberg et al. 1984; Eisenberg and McLachlan 1986). Quantitative estimates of the hydrophobicity can be derived from their relative concentrations in organic versus aqueous bulk phase of a binary solution. Different experimental conditions, solvents, and computational schemes have led to different sets of estimates of

Table 3 Experiment results of different hydrophobicity data types on CL317 in re-substitution test (%)

Hydrophobicity	Sensitivity for each class (%)						Overall accuracy
	Cyto	Memb	Mito	Secr	Nucl	Endo	
FDHΦ	98.2	98.2	79.4	64.7	76.9	97.9	90.9
MHΦ	100	89.1	70.59	64.7	80.8	100	89.9
KDHΦ	100	98.2	100	94.1	100	100	99.4

hydrophobicity. In fact, there are several hydrophobicity data types available for amino acids to transform protein sequences into real numbers (Qiu et al. 2004). Three sets of hydrophobicity data types, Kyte-Doolittle hydrophobicity scales (KDHΦ) (Kyte and Doolittle 1982), Mandell hydrophobicity scales (MHΦ) (Hirakawa et al. 1999), and Fauchereand hydrophobicity scales (FHΦ) (Fauchere and Pliska 1983), have been investigated in this paper (see Appendix). Table 3 shows the prediction results performed by different hydrophobicity data types at a decomposition scale of 3 and bior3.3 wavelet. From Table 3, we can clearly see that compared with the prediction accuracies of the MHΦ and FHΦ, the KDHΦ presents the highest prediction ability.

Comparison of different classifiers

The so-called re-substitution test is an examination for the self-consistency of an identification method. It is necessary but not sufficient for evaluating an identification method. As a complement, a cross-validation test, it is needed. As is well-known, there are three methods usually used for cross-validation in statistical prediction, namely, the subsampling test, independent dataset test and jackknife test. Of these tests, the jackknife test is deemed as the most effective and objective one (Chou and Zhang 1995). During jackknife test, each protein in the data set is in turn singled out as a tested protein and all the rule-parameters are calculated based on the remaining proteins. The prediction accuracies by jackknife test for the 317 proteins classified into 6 subcellular locations are enumerated in Table 4, and are compared with other published results of the same dataset. It can be seen that the overall accuracy of the proposed approach (DWT_SVM) is 97.5%, about 15 and 13% higher than that of ID (Chen and Li 2007a) and ID_SVM (Chen and Li 2007b), respectively. Furthermore, the prediction accuracies of nuclear proteins and endoplasmic reticulum proteins have been remarkably enhanced. The prediction accuracy of the proposed approach for nuclear protein is 100%, about 17 and 27% higher than those by ID (Chen and Li 2007a) and ID_SVM (Chen and Li 2007b), respectively. The success rate of the proposed approach for endoplasmic reticulum protein is

Table 4 Prediction results with different models on CL317 in Jack-Knife test

Model	Sensitivity for each class (%)						Overall accuracy
	Cyto	Memb	Mito	Secr	Nucl	Endo	
ID	81.3	81.8	85.3	88.2	82.7	83.0	82.7
Chen and Li (2007a)							
ID_SVM	91.1	89.1	79.4	58.2	73.1	87.2	84.2
Chen and Li (2007b)							
IEPseAA	90.2	90.9	82.4	88.2	86.5	91.5	89.0
Shi et al. (2007)							
DWT_SVM	100	98.2	82.4	94.1	100	100	97.5

Table 5 Prediction results with different models on ZD98 in Jack-knife test

Model	Sensitivity for each class (%)				Overall accuracy
	Cyto	Memb	Mito	Other	
Instab_SVM	33/43 = 76.8	25/30 = 83.3	12/13 = 92.5	6/12 = 50.0	76/98 = 77.6
Huang et al. (2005)					
Dipep_Diver	38/43 = 88.4	27/30 = 90.0	12/13 = 92.3	6/12 = 50.0	83/98 = 84.7
Chen and Li (2004)					
AAC_CCA	42/43 = 97.7	22/30 = 73.3	4/13 = 30.8	3/12 = 25.0	71/98 = 72.5
Zhou and Doctor (2003)					
EBGW_SVM	42/43 = 97.7	27/30 = 90.0	12/13 = 92.3	10/12 = 83.3	91/98 = 92.9
Zhang et al. (2006)					
DWT_SVM	41/43 = 95.4	28/30 = 93.3	7/13 = 53.9	11/12 = 91.7	87/98 = 88.8

Table 6 Prediction results with different models on ZW225 in Jack-Knife test

Model	Sensitivity for each class (%)				Overall accuracy
	Cyto	Memb	Mito	Other (nucl)	
ID_SVM	65/70 = 92.9	81/89 = 91.0	17/25 = 68.0	30/41 = 73.2	193/225 = 85.8
Chen and Li (2007b)					
FKNN	59/70 = 84.3	83/89 = 93.3	18/25 = 72.0	33/41 = 85.5	193/225 = 85.8
Ding and Zhang (2008)					
EBGW_SVM	63/70 = 90.0	83/89 = 93.3	15/25 = 60.0	26/41 = 63.4	187/225 = 83.1
Zhang et al. (2006)					
DWT_SVM	61/70 = 87.1	83/89 = 93.2	16/25 = 64.0	37/41 = 90.2	197/225 = 87.6

100%, about 17 and 13% higher than those by ID (Chen and Li 2007a) and ID_SVM (Chen and Li 2007b), respectively. Shi et al. have developed the multi-scale energy approach (IEPseAA) to predict the protein subcellular localizations by calculating the root mean square energy of the wavelet transform coefficients which can effectively reflect the sequence order effect, and the remarkably improved accuracy indicates the efficiency of the wavelet transform in the feature extraction (Shi et al. 2007). From Table 4, it can be seen that the overall accuracy of IEPseAA is 89.0%, about 6.3 and 4.8% higher than those by ID and ID_SVM, respectively. However, the overall accuracy of IEPseAA is about 8.5% lower than that by DWT_SVM, which combines the maximum, minimum, mean and standard deviation of the wavelet coefficients in each sub-band based on the amino acid hydrophobicity to code the protein sequence.

In order to further test the performance of the proposed approach, the dataset ZW98 (Zhou and Doctor 2003) and ZW225 (Zhang et al. 2006) were also adopted. The prediction results of the DWT_SVM model are enumerated in Tables 5 and 6, and are compared with other published results of the same dataset. It can be seen from Table 5 that for the dataset ZW98, the overall jackknife accuracy obtained by the proposed approach is 88.8%, which is 16.3% higher than the result of Zhou and Doctor (2003), 4.1% higher than the result of Chen and Li (2004), 11.2% higher than the result of Huang et al. (2005), but a little

lower than the result of Zhang et al. (2006). In addition, the much larger dataset ZW225 was also utilized to evaluate the generalization ability of our method. As shown in Table 6, the overall prediction jackknife rate for the dataset of the 225 apoptosis proteins could still reach 87.6%, which is appropriately 5% higher than that of EBGW_SVM (Zhang et al. 2006) and about 2% higher than that of ID_SVM (Chen and Li 2007b) and FKNN (Ding and Zhang 2008). Especially for nuclear proteins, the accuracy of the proposed approach is 90.2%, about 27, 17 and 5% higher than that of EBGW_SVM (Zhang et al. 2006), ID_SVM (Chen and Li 2007b) and FKNN (Ding and Zhang 2008), respectively. All the results indicate that the proposed method has a good performance for prediction of subcellular locations.

Conclusions

Our results show that the novel DWT_SVM model can successfully predict subcellular localization of apoptosis proteins. The novel feature extraction method, DWT based on the amino acid hydrophobicity, can reduce the dimension of the input vector, improve calculating efficiency, and more effectively reflect the overall sequence order feature of a protein. It is anticipated that the new method could be potentially useful for the classification of G-protein coupled receptors (GPCRs), nuclear receptors, enzyme families and analysis of protein function.

Acknowledgments This work was supported by grants from the National Natural Science Foundation of China (20605010, 20865003, 20805023), the Jiangxi Province Natural Science Foundation (2007JZH2644), the Opening Foundation of State Key Laboratory of Chem/Biosensing and Chemometrics of Hunan University (2006022, 2007012).

References

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

Barinaga M (1998) Stroke-damaged neurons may commit cellular suicide. Science 281:1302–1303

Bhasin M, Raghava GPS (2004a) Classification of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem 279:23262–23266

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

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

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 CC, Lin CJ (2002) Training nu-support vector regression: theory and algorithms. Neural Comput 14:1959–1977

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

Chen YL, Li QZ (2007a) Prediction of 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

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, Cai YD (2003) A new hybrid approach to predict subcellular localization of proteins by incorporating gene oncology composition. Biochem Biophys Res Commun 311:743–747

Chou KC, Cai YD (2004) Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun 321:1007–1009

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

Chou KC, Shen HB (2007) 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, Zhang CT (1995) Prediction of proteins structural classes. Crit Rev Biochem Mol Biol 30:275–349

Chou JJ, Matsuo H, Duan H, Wagner G (1998) Solution structure of the RAIDD CARD and model for CARD/CARD interaction in caspase-2 and caspase-9 recruitment. Cell 94:171–180

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

Daubechies I (1992) Ten lectures on wavelets. In: CBMS-NSF regional conference series in applied mathematics. SIAM

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

Eisenberg D, McLachlan AD (1986) Solvation energy in protein folding and binding. Nature 319:199–203

Eisenberg D, Weiss RM, Terwilliger TC (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci USA 81:140–144

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

Fauchere JL, Pliska V (1983) Transformational homologies in amino acid sequence. Eur J Med Chem 18:369–375

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

Grunbaum FA (1992) An introduction to wavelets. Science 257:821–822

Hirakawa H, Muta S, Kuhara S (1999) The hydrophobic cores of proteins predicted by wavelet analysis. Bioinformatics 15:141–148

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

Huang J, Shi F, Zhou HB (2005) Support vector machine for predicting apoptosis proteins types by incorporating protein instability index. China J Bioinf 3:121–123

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

Joachims T (1999) Making large-scale SVM learning practical. In: Scholkopf B, Burges C, Smola A (eds) Advances in kernel methods—support vector learning. MIT Press, Cambridge

Kerr JF, Wyllie AH, Currie AR (1972) Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. Br J Cancer 26:239–257

Kreßel UH (1999) Pairwise classification and support vector machines. In: Schölkopf B, Burges CJ, Smola AJ (eds) Advances in kernel methods: support vector learning. MIT Press, Cambridge

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132

Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 11:674–693

Mallat SG (1999) A wavelet tour of signal processing. Academic Press, San Diego

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

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

Platt J, Cristianini N, Shawe-Taylor J (2000) Large margin DAGs for multiclass classification. Adv Neural Inf Proc Syst 12:547–553

Qiu JD, Liang RP, Zou XY, Mo JY (2003) Prediction of protein secondary structure based on continuous wavelet transform. Talanta 61:285–293

Qiu JD, Liang RP, Zou XY, Mo JY (2004) Prediction of transmembrane proteins based on the continuous wavelet transform. J Chem Inf Comput Sci 44:741–747

Qiu JD, Huang JH, Liang RP, Luo SH (2009a) Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. Anal Biochem 390:68–73

Qiu JD, Luo SH, Huang JH, Liang RP (2009b) Using support vector machines for prediction of protein structural classes based on discrete wavelet transform. J Comput Chem 30:1344–1350

Qiu JD, Luo SH, Huang JH, Liang RP (2009c) Using support vector machines to distinguish enzymes: approached by incorporating wavelet transform. J Theor Biol 256:625–631

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

Shen HB, Chou KC (2007) 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

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(1):69–74

Suzuki M, Youle RJ, Tjandra N (2000) Structure of Bax: coregulation of dimer formation and intracellular localization. Cell 103:645–654

Vapnik V (1995) The nature of statistical learning theory. Springer, New York

Vapnik V (1998) Statistical learning theory. Wiley, New York

Vaux DL, Heacker G, Strasser A (1994) An evolutionary perspective on apoptosis. Cell 76:77–779

Walczak B (2000) Wavelets in chemistry. Elsevier, Amsterdam

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 SW, Zhang YL, Yang YF, Zhao CH, Pan Q (2008) Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. Amino Acids 34:565–572

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

Zhou P, Chou JJ, Olea RS, Yuan J, Wagner G (1999) Solution structure of Apaf-1 CARD and its interaction with caspase-9 CARD: a structural basis for specific adaptor/caspase interaction. Proc Natl Acad Sci USA 96:11265–11270

Zhou XB, Chen C, Li ZC, Zou XY (2008) Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine. Amino Acids 35:383–388