Computers in Biology and Medicine 42 (2012) 180–187

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

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

A novel algorithm combining support vector machine with the discrete wavelet transform for the prediction of protein subcellular localization

Ru-Ping Liang, Shu-Yun Huang, Shao-Ping Shi, Xing-Yu Sun, Sheng-Bao Suo, Jian-Ding Qiu *

Department of Chemistry, Nanchang University, Nanchang 330031, PR China

ARTICLE INFO

Article history:
Received 19 January 2011
Accepted 15 November 2011

Keywords:
Subcellular localization
Amino acid polarity
Discrete wavelet transform
Support vector machine
Jackknife test

ABSTRACT

Knowing the subcellular localization of proteins within the cell is an important step in elucidating its role in biological processes, its function and its potential as a drug target for disease diagnosis. As the number of complete genomes rapidly increases, accurate and efficient methods that automatically predict the subcellular localizations become more urgent. In the current paper, we developed a novel method that coupled the discrete wavelet transform with support vector machine based on the amino acid polarity to predict the subcellular localizations of prokaryotic and eukaryotic proteins. The results obtained by the jackknife test were quite promising, and indicated that the proposed method remarkably improved the prediction accuracy of subcellular locations, and could be as an effective and promising high-throughput method in the subcellular localization research.

1. Introduction

As the number of new genomes has risen sharply in recent years, it has once again brought to the forefront problem of protein function prediction. Subcellular localization is a key functional characteristic of potential gene products such as proteins [1]. Currently, the prediction of protein subcellular location is a very hot topic in molecular biology because it has involved the three essential features of a protein: its biological objective, its biochemical activity, as well as its place in the cell where a gene product is active. Therefore, comprehensive knowledge on the subcellular localization of proteins is essential for understanding their roles and interacting partners in cellular metabolism. However, the traditional way to annotate protein subcellular localization in a cell is by biochemical experiments, which are not able to keep up with the large number of sequences that continue to emerge from the genome sequencing projects due to both time-consuming and expensive. To bridge this gap, it is necessary to develop faster, accurate and genomescale computational methods for predicting the subcellular localization of proteins.

Several theoretical and computational methods have been developed over the past decade for predicting the subcellular localization of proteins. Most of the existing prediction methods are broadly classified into four categories according to their input data [2]: (1) methods based on the sorting signals, which rely on the presence of protein targeting or signal peptides [3]; (2) methods based on lexical analysis of keywords (LOCkey) from the functional annotation of proteins [4]; (3) methods based on the uses of phylogenetic profiles [5], domain projection [6] or a combination of evolutionary and structural information; (4) methods based on the differences in the amino acid composition or amino acid properties of proteins [7–10]. In this paper, our interest was focused on the researches about the last type.

Previously, much progress using sequence-based information has been made in computational prediction of protein subcellular localization. Originally, Nakashima and Nishikawa first proposed a method based on amino acid composition and residue-pair frequencies to discriminate between intracellular and extracellular proteins [7]. Subsequently, Chou and Elrod also used the amino acid composition but the covariant discriminant algorithm was employed in their method [13]. The other studies using different algorithms, such as neural network model [14], Markov chain model [12] and support vector machine [11], showed that amino acid composition was closely related to protein subcellular localizations. For further improving the predictive quality, Chou proposed the pseudo-amino acid composition to take the effect of the amino acid order into account [8]. Furthermore, Cai and Chou suggested a hybrid approach integrating the pseudo-amino acid composition, the functional domain composition [15,16], and the information of gene ontology [17]. It indicated that incorporating an amino acid order as well as the amino acid composition made it possible to improve prediction performance. Recently, a sequence representation method using multi-scale energy was established to predict the subcellular location based on the concept of Chou's pseudo-amino acid composition [18]. However,

* Corresponding author. Tel.: +86 791 3969518.
E-mail address: jdqiu@ncu.edu.cn (J.-D. Qiu).

0010-4825/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2011.11.006

developing an more effective method to predict the subcellular location attributes based on their sequence information can not only save time, but can also be helpful to the design of drugs in treating certain diseases that are related to subcellular location attribute defects. Hence, it has become a crucial issue to complement the existing methods and enhance the quality of predicting protein subcellular localization by selecting more informative features. In this paper, a novel model (DWT–SVM) was proposed by combining the discrete wavelet transform (DWT) with support vector machine (SVM) based on the amino acid polarity to predict the subcellular localization of proteins. First and foremost, amino acids of protein subcellular localization were transformed into sequences of polarity energies per residue. Then, the polarity profile was decomposed into wavelet coefficients using DWT. Subsequently, using the statistical method, a series of statistical feature vectors were constructed to represent the protein sequences. Finally, SVM was applied to deal with the problem of multi-classification. The predictive results of the jackknife test show significant improvement compared with the previous algorithms, and hence the methodology presented in the current study could effectively complement the existing subcellular localization prediction methods and assist in the development of automated genome annotation tools.

2. Materials methods

2.1. Data sets

Fig. 1 shows the flowchart of the proposed approach combining the DWT with SVM algorithm to predict protein subcellular localization. As presented in Fig. 1, in this study, two datasets of proteins as a benchmark have been used. The first dataset, NNPSL dataset as the training dataset, was originally constructed by Reinhardt and Hubbard [14]. It included 997 prokaryotic sequences, which were classified into three location categories (688 cytoplasm, 107 extracellular and 202 periplasm) and 2427 eukaryotic sequences belonging to four location categories (684 cytoplasm, 325 extracellular, 321 mitochondrion and 1097 nucleus). All the protein sequences in the dataset were extracted from Swiss-Prot 33.0 and no transmembrane proteins were included as they could be quite reliably predicted by some known methods [19–21]. Within each subcellular location, none of the sequences has more than 90% identity to any other sequences. Though there is probably somewhat out of date, this data set is frequently used as a typical training and test set by some investigators and the experimental results can still exhibit some prediction capabilities of various methods [10–15]. The second dataset named TargetP dataset, as an independent testing dataset, was constructed by Chou and Shen in plant cell [9]. It contained 978 different protein sequences, which were distributed among 12 subcellular locations: 56 cell membrane, 32 cell wall, 286 chloroplast, 182 cytoplasm, 42 endoplasmic reticulum, 22 extracellular, 21 golgi apparatus, 150 mitochondrion, 152 nucleus, 21 peroxisome, 39 plastid and 52 vacuole. To investigate the impact of homology on estimation of the classification accuracy in this work, none of the proteins in the dataset had $\geq 25\%$ pairwise sequence identity to any other in the same subset.

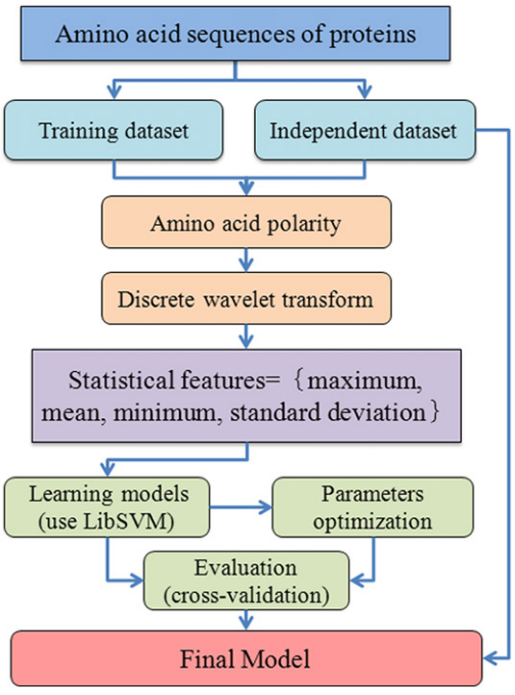

Fig. 1. System flowchart of proposed method based on discrete wavelet transform incorporating SVM algorithm.

2.2. Discrete wavelet transform

In recent years, wavelets analysis has been applied to a large variety of biological signals [22,23], and there is a growing interest in using wavelet functions in the analysis of sequence and the investigation of protein structure [24–26]. The most attractive character of wavelet transform (WT) is the ability to elucidate simultaneously both spectral and temporal information, in contrast to the Fourier transform that only elucidates spectral information [27,28]. The coefficients of the DWT can be divided into two parts: one is the approximation coefficient, which represents the high-scale and low-frequency components of the signal, and the other is the detail coefficient, which represents the low-scale and high-frequency components of the signal [29]. According to both experimental and theoretical progress in protein dynamics, it is clear that low-frequency internal motions do exist in protein and DNA molecules and indeed play a significant role in biological functions [30–32]. Using the low-frequency wavelet coefficients to formulate the sample of a protein can better reflect its overall sequence-order effect.

To apply the DWT directly, protein amino acid sequence must be transformed into real numbers. Because the polarity [33] takes vital role in the protein synthesis process [34], we first map protein amino acid sequences into protein polarity sequences, and then process these polarity sequences by DWT. The polarity value of 20 kinds of amino acids was presented in Table S1 (please see Supplementary materials) [33]. Here, we chose a eukaryotic protein (Swiss-prot ID: A33_PLEWA) as an example to describe the process of extracting the hidden information using DWT, as given in Fig. 2. Fig. 2 shows the decomposition process. However, in order to further decrease the dimensionality and noise of the extracted feature vectors, the statistics over the set of the wavelet coefficients were employed [35]. The following statistical features, calculated from the approximation coefficients and detail coefficients, were used for the subcellular location of protein: (i) maximum of the wavelet coefficients in each sub-band, (ii) mean of the wavelet coefficients in each sub-band, (iii) minimum of the wavelet coefficients in each sub-band and (iv) standard deviation of the wavelet coefficients in each sub-band. So, a protein x can be characterized as a $4(j+1)$ dimension feature vector. In this study, the decomposition level $j=4$ was chosen to predict the subcellular localization [36], and the obtained 20 dimension feature vectors were then fed to classifiers.

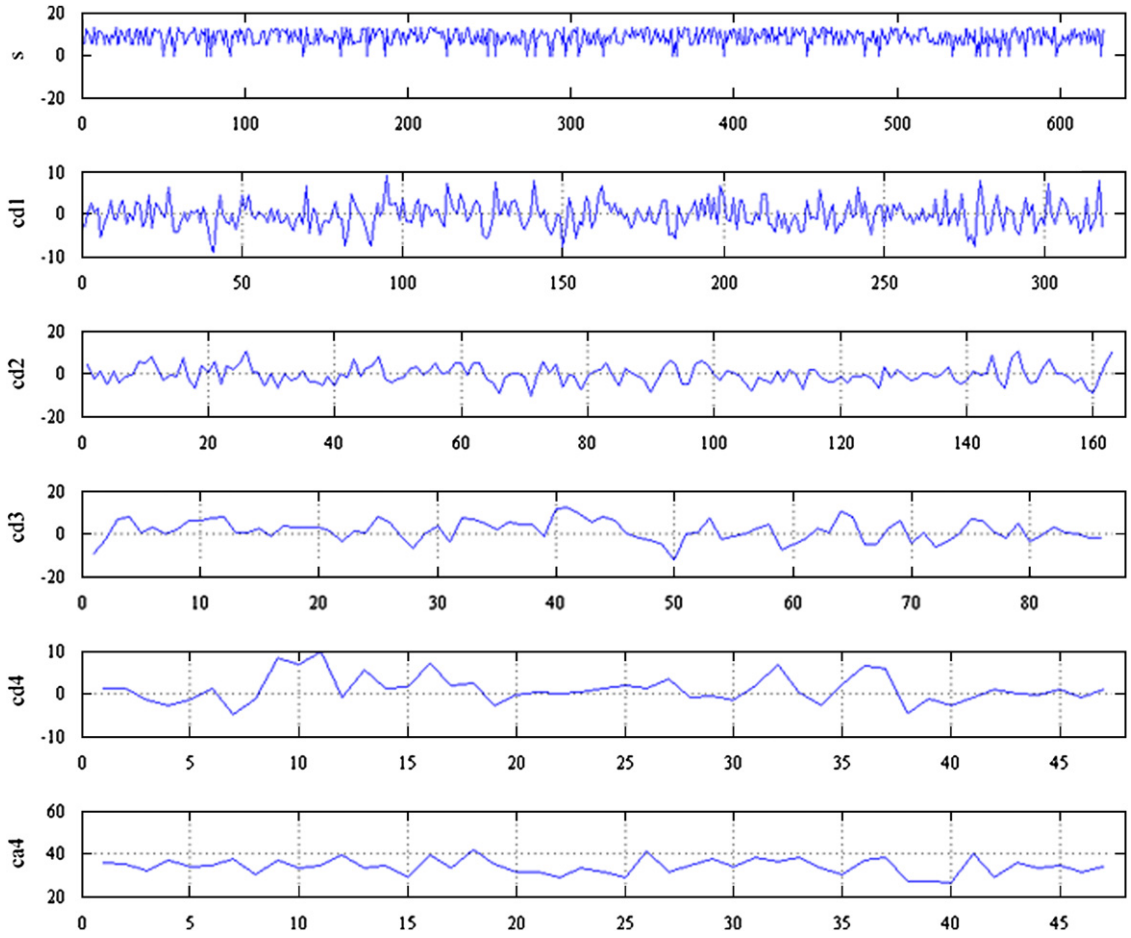

Fig. 2. DWT coefficient plot using Bior2.4 wavelet function. The *y* axis indicates the intensity of signal; the *x* axis indicates the residue position along the sequence. *s* denotes the polar plot of the protein A33_PLEWA; cd1, cd2, cd3 and cd4 are four detail scales for levels from $j=1$ to 4; ca4 denotes the coarse scale for level $j=4$. In the DWT, coefficients at coarse scale (ca4) capture gross and global features of the signal while coefficients at four scales (cd1, cd2, cd3 and cd4) contain local details.

2.3. Multi-class SVM

SVM is rigorously based on Vapnik's statistical learning theory [37,38]. The SVM is particularly attractive to biological sequence analysis due to its ability to handle noise, large datasets and large input spaces, which has been extensively used to solve various biological problems, such as protein classfication [39], specificity of GalNAc-transferase [40] and HIV protease cleavage sites in protein [41]. For a two-class problem, samples are described by the feature vectors x_i $(i=1,2,\ldots,k)$ with corresponding labels $y_i=\{+1, -1\}(i=1,2,\ldots,k)$, where $+1$ and -1 are used to stand for the two classes. To classify them, SVM maps the input vector into a high dimensional feature space using a kernel function $K(x_i, y_j)$. As the kernel, the radial basis function (RBF) was selected because this function outperformed linear and polynomial kernels in terms of the overall predictive accuracy [42]. The RBF kernel is defined by the following equation:

$$K(x_i,y_j) = \exp(-\gamma \|x_i - y_j\|^2) \quad (1)$$

where x_i and y_j are feature vectors representing protein sequences. The kernel width parameter γ in Eq. (1) and penalty parameter C are automatically tuned based on the training set using the grid search strategy in the LIBSVM software [43]. The parameters used for optimal training models of prokaryotic and eukaryotic proteins based on Bior2.4 wavelet function by the jackknife test were presented in Tables S2 and S3 (see Supplementary materials), respectively.

The multi-class classification problem is commonly solved by a decomposing and reconstructing procedure when the binary class SVM is applied. There are several methods to extend the SVM for classifying multi-class problems, for example 'One-Versus-Rest (OVR)' [44], 'One-Versus-One (OVO)' [45], and DAGSVM [46]. In this paper, we used the OVO strategy, which can scale well to a large number of classes and is a simple method to deal with multi-class classification. 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. So, multi-class problems change the problems with two classes so that we can utilize SVM to classify several classes. The software used to implement the SVM in this paper is LIBSVM written by Chang and Lin [43] and Platt et al. [45] and can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

2.4. Performance measure

Predictive quality was examined with three approaches, one based on self-consistency test and the other two upon k-fold cross-validation. In self-consistency test, the subcellular localization protein in a given dataset was predicted using the rules derived from the same dataset. For 10-fold cross-validation, the dataset was randomly divided into k subsets. Each time, one of the k subsets was used as the test set and the other $k-1$ subsets were assembled to form a training set. In the jackknife cross-validation, each protein in the learning dataset was singled out in turn as a test protein and the predictor was trained by the

Table 1
Predictive performance of different wavelet functions for prokaryotic sequences in the NNPSL dataset based on SVM by the jackknife cross-validation.

Wavelet functions	The predictive performance (%)			
	Sensitivity	Specificity	Overall accuracy	MCC
Haar	98.2	97.6	96.7	88.9
Db4	95.7	98.4	97.6	93.4
Coif4	99.0	98.5	97.9	92.9
Bior3.3	96.0	93.7	91.1	74.9
Bior2.4	98.9	98.7	98.4	95.8
Sym4	94.1	91.9	94.3	72.2
Sym6	93.3	76.3	92.4	68.4

remaining proteins. In other words, the subcellular location of each protein was predicted by the rules derived using all the other proteins except the one that was being predicted.

To evaluate the prediction performance of classification, the sensitivity, specificity, accuracy, Matthew's Correlation Coefficient (MCC) [47] and overall accuracy were utilized to assess the performance of prediction system. The definitions of these measures are as follows:

$$\text{specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \tag{2}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3}$$

$$\text{accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \tag{4}$$

$$\text{overall accuracy} = \frac{\sum_{i=1}^{k} \text{Acc}(i)}{N} \tag{5}$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FN})\times(\text{TP}+\text{FP})\times(\text{TN}+\text{FN})\times(\text{TN}+\text{FP})}} \tag{6}$$

TP, TN, FP and FN are the number of true positives, true negatives, false positives and false negatives, respectively, and N is the total number of sequences, where k is the class number, and Acc(i) is the proportion of correct predictions of localization i, containing positive and negative samples. Furthermore, we evaluated the predictive power using a receiver operating curve (ROC) [48], that is, the plot of sensitivity as a function of specificity based on various thresholds. The area under the ROC curve, referred to AUC, summarizes the predictor behavior: a random predictor has AUC=0.5, while a perfect predictor has AUC=1.0, so that a larger ROC indicates better predictive power.

Besides, Welch's t-test is used to evaluate the statistical significance of differences [49]. This test assumes that the two populations are normally distributed with potentially unequal variance and calculates a P-value that shows whether the differences of the two populations are significant [50,51]. The significance level is set at $P=0.05$. When $P \le 0.05$, there is statistically significantly different, or else there is no significant difference.

3. Results and discussion

3.1. Selection of wavelet function

Wavelet transform is based on the idea of mapping a signal into a set of basis function. Based on different basis functions, the wavelets have different families, and every family has its quality fitting for different signals and emerges different results [36,52]. As the characteristics of the analyzing wavelet influence the performance of DWT, the better the analyzing wavelet matches the underlying structure in the signal, the better feature values can be extracted from the sequences. Hence, selection of a suitable wavelet basis that possesses desirable properties, such as compactly support, orthogonality, symmetry, smoothness and high order of vanishing moments, is necessary for the signal processing [53]. It is well known that for wavelet selection there are many conflict conditions that restrict the selection of a wavelet function, and none of wavelet basis shares simultaneously all of these desirable properties.

In order to investigate the effect of the wavelets on the prediction of protein subcellular location, seven wavelet functions: Haar, Daubechies of number 4 (Db4), Biorthogonals of number 2.4 (Bior2.4) and number 3.3 (Bior3.3), Coiflet of number 4 (Coif4) and Symlets of number 4(Sym4) and number 6 (Sym6) were chosen for testing in the research. The predictive performance for prokaryotic and eukaryotic sequences in the NNPSL dataset based on different wavelet functions are presented in Tables 1 and 2. As seen in Table 1, the sensitivity of Bior2.4 wavelet function for prokaryotic sequences was 0.1% lower than that of Coif4 wavelet function. However, the specificity, overall accuracy and MCC based on Bior2.4 wavelet function were higher than those of other wavelet functions. For eukaryotic sequences, in Table 2, the sensitivity, specificity, overall accuracy and MCC based on Bior2.4 wavelet function achieved 97.2%, 96.5%, 96.7% and 92.5%, respectively, which were superior to other wavelet functions. In order to further determine if the performance differences between seven wavelet functions, the paired Welch's t-test was used. As seen in Table S4 (see Supplementary materials), most of P-value was more than 0.05 in prokaryotic and eukaryotic proteins, indicating that there was no statistical significance between different wavelet functions. However, we can find significant difference between Bior2.4 and other wavelet functions with $P \le 0.05$ in eukaryotic proteins. This is because Bior2.4 wavelet function is a symmetric and orthogonal function and can be through the Mallat algorithm to select the suitable approximate coefficient [54]. Moreover, the Bior2.4 wavelet function can also effectively remove redundant information and extract useful sequence information. Eventually, the Bior2.4 wavelet function makes a large distinction of feature vectors between different protein categories so that the prediction performance has been improved and the difference is statistically significant. Therefore, the Bior2.4 wavelet function was applied as the appropriate wavelet function for the prediction of protein subcellular localization in this study.

3.2. Comparison with machine learning methods

For a query protein, how can we identify which category it belongs? Many different prediction algorithms have been developed to address this problem. Here the performance of the three

Table 2
Predictive performance of different wavelet functions for eukaryotic sequences in the NNPSL dataset based on SVM by the jackknife cross-validation.

Wavelet functions	The predictive performance (%)			
	Sensitivity	Specificity	Overall accuracy	MCC
Haar	90.3	88.4	89.3	75.4
Db4	95.0	92.9	93.5	84.7
Coif4	95.2	95.5	94.6	87.9
Bior3.3	94.5	91.9	92.8	83.5
Bior2.4	97.2	96.5	96.7	92.5
Sym4	92.5	89.4	91.0	78.8
Sym6	92.2	89.3	80.7	78.3

following classifier algorithm was discussed: *K* nearest neighbor (KNN) [55], Bayes [56] and SVM algorithms. The algorithms of KNN and Bayes were available in MATLAB programming environment. The distance used in KNN was the Euclidean distance, and the number of nearest neighbor was 5. The default parameters of Bayes were used [56]. The results of three classifiers for prokaryotic and eukaryotic sequences in the NNPSL dataset based on Bior2.4 wavelet function by the jackknife test were summarized in Tables 3 and 4, respectively. In Table 3, the accuracy by SVM method for prokaryotic proteins was 15.4% and 26.8% higher than those of KNN and Bayes algorithms, respectively. For eukaryotic proteins, in Table 4, the sensitivity, specificity, accuracy and MCC by SVM method reached 97.2%, 96.5%, 96.7% and 92.4%, respectively, which outperformed the other algorithms. Moreover, the ROC curves for the assessment of the three classifier performance of eukaryotic proteins were plotted in Fig. 3, where different curves denoted corresponding prediction performances of SVM, KNN and Bayes. Better overall performances of the model were represented with larger values. As shown in Fig. 3, SVM classifier was superior to the others classifier. Furthermore, the paired Welch's *t*-test was also employed to examine the statistical difference between three machine learning methods (Table S5, see Supplementary materials). For prokaryotic proteins, as showed in Table S5 (panel a), there was statistically different between KNN and SVM with a significant level ($P=1.079\times10^{-7}$). It was the same result between Bayes and SVM ($P=0.007$), whereas no significant difference between KNN and Bayes ($P=0.915$) can be found. Similarly, for eukaryotic proteins (Table S5, panel b), the significant level between KNN and SVM was 3.370×10^{-8} while between KNN and Bayes was 0.051, indicating that the SVM method could effectively avoid overfitting and deal with large feature spaces and the absence of local minima. Therefore, the SVM method was selected as the appropriate classifier and incorporated Bior2.4 wavelet function to construct the DWT–SVM model for predicting protein subcellular localizations in this study.

3.3. The power of statistical prediction method

In statistical prediction, the three most commonly used test procedures with respect to protein subcellular localization prediction are 10-fold cross-validation, jackknife cross-validation and resubstitution test. The resubstitution test is used to examine the self-consistency of a prediction model and the cross-validation really reflects the effectiveness of a predictor. The predictive accuracy for training models of prokaryotic and eukaryotic proteins using SVM algorithm based on Bior2.4 wavelet function was presented in Table 5. As seen in Table 5, the overall accuracies for prokaryotic and eukaryotic proteins in self-consistency test achieved 100% and 99.6%, respectively, indicating that the DWT–SVM model has grasped the complicated relationship with subcellular localization categories after being trained.

Previous studies have demonstrated that the jackknife test is deemed the most objective and rigorous way for cross-validation [57,58]. During the process of jackknife analysis, both the training and testing datasets are actually open, and each protein in the dataset is in turn moving from one to the other. The overall accuracies of the jackknife test for prokaryotic and eukaryotic proteins by the DWT–SVM method achieved 98.4% and 96.7%, respectively.

Table 3
Performance comparisons of different classifier methods for prokaryotic sequences based on Bior2.4 wavelet function by the jackknife cross-validation in the NNPSL dataset.

Classifiers	The performance of evaluating method results (%)				
	Sensitivity	Specificity	Accuracy	AUC	MCC
SVM	98.9	98.7	98.4	75.9	95.8
KNN	76.3	78.6	83.0	72.6	44.8
Bayes	64.3	71.7	72.8	69.9	41.5

Table 4
Performance comparisons of different classifier methods for eukaryotic sequences on Bior2.4 wavelet function by the jackknife cross-validation in the NNPSL dataset.

Classifiers	The performance of evaluating method results (%)				
	Sensitivity	Specificity	Accuracy	AUC	MCC
SVM	97.2	96.5	96.7	75.8	92.4
KNN	77.6	79.8	81.0	72.9	55.0
Bayes	64.5	70.1	68.1	70.1	36.5

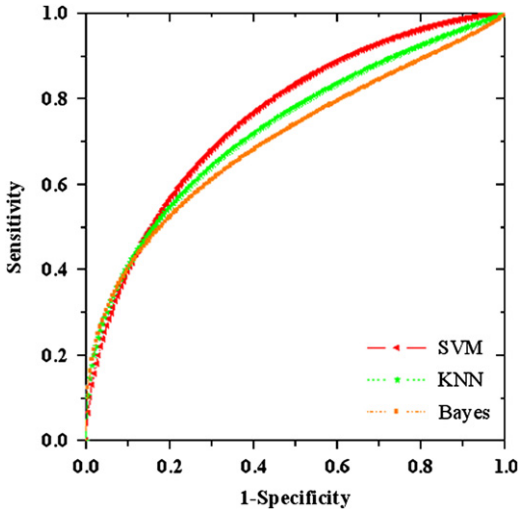

Fig. 3. ROC curves of different classifiers for eukaryotic proteins. The horizontal axes are the false positive rates (1-specificity) and vertical axes are the true positive rate (sensitivity). For specific threshold for discriminating multi-class problem, there will be a pair of these two values. A pair of values is then represented by a point in this two-dimensional space.

Table 5
Overall accuracy among 10-fold cross-validation, jackknife test and self-consistency test for training models of prokaryotic and eukaryotic proteins used SVM algorithm based on Bior2.4 wavelet function.

Proteins	The overall accuracy (%)		
	10-Fold cross-validation	Jackknife test	Self-consistency test
Prokaryotic	100	98.4	100
Eukaryotic	96.9	96.7	99.6

3.4. Comparison of the prediction programs

To demonstrate the reliability and efficiency of the presented method, the performance of the DWT–SVM method was compared with those of the previous methods that used the same training dataset. Tables 6 and 7 show the comparison of predictive accuracy with existing methods for prokaryotic and eukaryotic proteins in the NNPSL dataset by the jackknife cross-validation, respectively. In Table 6, for prokaryotic sequences, the overall accuracy of DWT–SVM method based on Bior2.4 wavelet function was 98.4%, which was 3.7–11.9% higher than those of available methods [10–13,17]. In particular, the accuracy for

Table 6
Performance comparison of predictive accuracies for prokaryotic proteins by the jackknife cross-validation in the NNPSL dataset.

Predictor	Algorithm	Accuracy (%)			Overall accuracy (%)
		Cytoplasm	Extracellular	Periplasm	
Our method	DWT–SVM	98.4	99.5	97.9	98.4
Niu et al. [10]	AdaBoost	98.8	73.8	77.2	91.8
Hua and Sun [11]	SVM	97.5	76.6	78.2	91.4
Yuan [12]	MM	93.6	77.6	79.7	89.1
Chou and Elrod [13]	CD	91.6	80.4	72.3	86.5
Chou and Cai [17]	ISort	N/A	N/A	N/A	94.7

Table 7
Performance comparison of predictive accuracies for eukaryotic proteins by the jackknife cross-validation in the NNPSL dataset.

Predictor	Algorithm	Accuracy (%)				Overall accuracy (%)
		Cytoplasm	Extracellular	Mitochondria	Nuclear	
Our method	DWT–SVM	97.7	97.7	97.9	95.2	96.7
Niu et al. [10]	AdaBoost	84.5	76.3	49.2	89.2	80.8
Hua and Sun [11]	SVM	76.9	80.0	56.7	87.4	79.4
Yuan [12]	MM	78.1	62.2	69.2	74.1	73.0
Chou and Cai [17]	ISort	N/A	N/A	N/A	N/A	92.9

Table 8
Performance comparison of the independent testing dataset by the jackknife test in the TargerP dataset.

Subcellular location	The results of DWT–SVM model (%)					The results of Chou and Shen's method (%) [9]
	Sensitivity	Specificity	Accuracy	AUC	MCC	Accuracy
Cell membrane	86.7	87.1	85.5	74.3	64.8	42.9
Cell wall	88.0	88.0	86.4	74.6	58.0	25.0
Chloroplast	89.0	96.3	89.5	73.9	56.6	86.7
Cytoplasm	83.5	88.9	83.2	73.4	45.4	39.6
Endoplasmic reticulum	89.7	87.3	85.9	75.0	64.5	40.5
Extracellular	96.1	93.6	93.0	75.9	71.0	13.6
Golgi apparatus	92.6	90.4	89.7	75.4	69.6	28.6
Mitochondrion	85.6	90.5	83.4	73.7	49.5	76.0
Nucleus	87.6	87.5	86.7	74.5	61.5	89.5
Peroxisome	92.0	88.5	89.3	75.5	70.1	66.7
Plastid	87.4	82.0	84.5	75.0	59.7	10.3
Vacuole	87.3	86.2	84.6	74.6	61.3	63.7

extracellular and periplasm sequences were 19.1–25.7% and 18.2–25.6% higher than those of the other methods [10–13], respectively. For eukaryotic sequences, in Table 7, the overall accuracy by the present method (DWT–SVM) was 96.7%, which was 3.8%, 15.8%, 17.3% and 25.7% higher than those of the ISort [17], AdaBoost [10], SVM [11] and MM algorithms [12], respectively. The accuracies for cytoplasm, extracellular, mitochondria and nuclear reached 97.7%, 97.7%, 97.9% and 95.2%, respectively, which were remarkably superior when compared to the existing methods that are listed in Table 7 [10–12,17]. Moreover, Table S6 (see Supplementary materials) showed the comparison of feature dimensions for these existing algorithms. The experimental results show that the DWT–SVM approach is convenient and effective to extract valuable information from protein sequences.

In addition, as an examination for the practical application of the new approach, the TargetP dataset is used as an independent testing dataset. Table 8 shows the predictive results of the subcellular localization of plant cell in the TargerP dataset. As can be seen from Table 8, the accuracy and AUC of each subcellular localization category in the TargerP dataset were over 80% and 70%, respectively. Particularly, the accuracy of extracellular sequences (the number of extracellular sequences is 22) reached 93.0%, which was 79.4% higher than that of Chou and Shen [9]. Moreover, the MCC for most of subcellular locations were about 60%, except those of cytoplasm and mitochondrion sequences were 45.4% and 49.5%, respectively. It can be noticed from the above that none of the proteins in the TargetP dataset had $\geq 25\%$ pairwise sequence identity to any other in the same subset. However, for low sequence identify ($<25\%$) TargetP dataset, the predictive accuracies for each of subcellular localization had been remarkably enhanced compared with those of Chou and Shen [9]. Thus, our proposed method not only enhanced significantly the accuracy of prediction, but also possessed obvious and effective character in the aspect of resistant sequence homology.

4. Conclusion

In this paper, a new method that integrated DWT into SVM for protein subcellular localization prediction is presented. DWT, the novel feature extraction method based on the amino acid polarity,

can reduce the dimension of the input vector, improve calculating efficiency and more effectively reflect the overall sequence order feature of a protein. Furthermore, SVM method can easily deal with high dimensional data and incorporate other useful features. The overall accuracies for prokaryotic and eukaryotic sequences from Reinhardt and Hubbard database by jackknife test reached 98.4% and 96.7%, respectively. The above results not only indicated that the new representation to extracting the information from the primary sequence of proteins is efficient, but also made a further step towards the systematic analysis of genome data. It is expected that the new method may provide new insights to unexplained activities of proteins or may support to attribute new activities to proteins along with *in vivo* experiments.

Conflict of interest statement

The author has no conflict of interests concerning this work.

Acknowledgment

This work was supported by Grants from the National Natural Science Foundation of China (20605010, 21065006 and 21175064).

Appendix A. Supplementary materials

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.compbiomed.2011.11.006.

References

[1] F. Eisenhaber, P. Bork, Wanted: subcellular localization of proteins based on sequence, Trans. Cell Biol. 8 (1998) 169–170.
[2] C. Guda, S. Subramaniam, pTARGET: a new method for predicting protein subcellular localization in eukaryotes, Bioinformatics 21 (2005) 3963–3969.
[3] K. Nakai, P. Horton, PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization, Trends Biochem. Sci. 24 (1999) 34–36.
[4] R. Nair, B. Rost, Inferring sub-cellular localization through automated lexical analysis, Bioinformatics 18 (2002) S78–S86.
[5] E.M. Marcotte, I. Xenarios, A. van Der Bliek, D. Eisenberg, Localizing proteins in the cell from their phylogenetic profiles, USA Proc. Natl. Acad. Sci. 97 (2000) 12115–12120.
[6] R. Mott, J. Schultz, P. Bork, C.P. Ponting, Predicting protein cellular localization using a domain projection method, Genome Res. 12 (2002) 1168–1174.
[7] H. Nakashima, K. Nishikawa, Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies, J. Mol. Biol. 238 (1994) 54–61.
[8] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins Struct. Funct. Genet. 43 (2001) 246–255.
[9] K.C. Chou, H.B. Shen, Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization, Biochem. Biophys. Res. Commun. 347 (2006) 150–157.
[10] N. Niu, Y.H. Jin, K.Y. Feng, W.C. Lu, Y.D. Cai, G.Z. Li, Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins, Mol. Divers. 12 (2008) 41–45.
[11] S.J. Hua, Z.R. Sun, Support vector machine approach for protein subcellular localization prediction, Bioinformatics 17 (2001) 721–728.
[12] Z. Yuan, Prediction of protein subcellular locations using Markov chain models, FEBS Lett. 451 (1999) 23–26.
[13] K.C. Chou, D.W. Elrod, Using discriminant function for prediction of subcellular location of prokaryotic proteins, Biochem. Biophys. Res. Commun. 252 (1998) 63–68.
[14] A. Reinhardt, T. Hubbard, Using neural networks for prediction of the subcellular location of proteins, Nucleic Acids Res 26 (1998) 2230–2236.
[15] Y.D. Cai, K.C. Chou, Predicting subcellular localization of proteins in a hybridization space, Bioinformatics 20 (2004) 1151–1156.
[16] K.C. Chou, Y.D. Cai, Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition, J. Cell. Biochem. 91 (2004) 1197–1203.
[17] K.C. Chou, Y.D. Cai, A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology, Biochem. Biophys. Res. Commun. 311 (2003) 743–747.
[18] J.Y. Shi, S.W. Zhang, Q. Pan, Y.M. Cheng, J. Xie, SVM-based method for subcellular localization of protein using multi-scale energy and pseudo amino acid composition, Amino Acids 33 (2007) 69–74.
[19] B. Rost, P. Fariselli, R. Casadio, Topology prediction for helical transmembrane proteins at 86% accuracy, Protein Sci. 5 (1996) 1704–1718.
[20] T. Hirokawa, S. Boon-Chieng, M. Shigeki, SOSUI: classification and secondary structure prediction system for membrane proteins, Bioinformatics 14 (1998) 378–379.
[21] P. Liò, M. Vannucci, Wavelet change-point prediction of transmembrane proteins, Bioinformatics 16 (2000) 376–382.
[22] A. Aldroubi, M. Unser, Wavelets in Medicine and Biology, CRC Press, Boca Raton, 1996.
[23] P. Liò, Wavelets in bioinformatics and computational biology: state of art and perspectives, Bioinformatics 19 (2003) 2–9.
[24] A.J. Mandell, K.A. Selz, M.F. Shlesinger, Wavelet transformation of protein hydrophobicity sequences suggests their memberships in structural families, Physica A 244 (1997) 254–262.
[25] K.B. Li, P. Issac, A. Krishnan, Predicting allergenic proteins using wavelet transform, Bioinformatics 20 (2004) 2572–2578.
[26] M.A. Rezaei, P. Abdolmaleki, Z. Karami, E.B. Asadabadi, M.A. Sherafat, H. Abrishami-Moghaddam, et al., Prediction of membrane protein types by means of wavelet analysis and cascaded neural networks, J. Theor. Biol. 254 (2008) 817–820.
[27] J.D. Qiu, R.P. Liang, X.Y. Zou, J.Y. Mo, Prediction of protein secondary structure based on continuous wavelet transform, Talanta 61 (2003) 285–293.
[28] X.Q. Lu, H.D. Liu, Z.H. Xie, Q. Zhang, Maximum spectrum of continuous wavelet transform and its application in resolving an overlapped signal, J. Chem. Inf. Comput. Sci. 44 (2004) 1228–1237.
[29] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 674–693.
[30] K.C. Chou, Low-frequency collective motion in biomacromolecules and its biological functions, Biophys. Chem. 30 (1988) 3–48.
[31] K.C. Chou, Low-frequency vibration of DNA molecules, Biochem. J. 221 (1984) 27–31.
[32] K.C. Chou, Low-frequency motions in protein molecules: beta-sheet and beta-barrel, Biophys. J. 48 (1985) 289–297.
[33] R. Grantham, Amino acid difference formular to help explain protein evolution, Science 185 (1974) 862–864.
[34] X.Q. Li, P. Fan, J.H. Fan, Polarity and hydrophobicity interactions in protein synthesis process, J. Theor. Biol. 240 (2006) 87–97.
[35] A. Kandaswamy, C.S. Kumar, R.P. Ramanathan, Jayaraman S. Jayaraman, N. Malmurugan, Neural classification of lung sounds using wavelet coefficients, Comput. Biol. Med. 34 (2004) 523–537.
[36] J.D. Qiu, S.H. Luo, J.H. Huang, R.P. Liang, Using support vector machines for prediction of protein structural classes based on discrete wavelet transform, J. Comput. Chem. 30 (2009) 1344–1350.
[37] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[38] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[39] N. Zavaljevski, F.J. Stevens, J. Reifman, Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions, Bioinformatics 18 (2002) 689–696.
[40] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Support vector machines for predicting the specificity of GalNAc-transferase, Peptides 23 (2002) 205–208.
[41] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Support vector machines for predicting HIV protease cleavage sites in protein, J. Comput. Chem. 23 (2002) 267–274.
[42] C.J.C. Burges, A tutorial on support vector machine for pattern recognition, Data Min. Knowl. Discovery 2 (1998) 121–167.
[43] C.C. Chang, C.J. Lin, Training nu-support vector regression: theory and algorithms, Neural Comput. 14 (2002) 1959–1977.
[44] U.H. Kreßel, Pairwise classification and support vector machines, in: B. Scholkopf, C.J. Burges, A.J. Smola (Eds.), Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, 1999.
[45] J. Platt, N. Cristianini, J. Shawe-Taylor, Large margin DAGs for multiclass classification, Adv. Neural Inf. Proc. Syst. 12 (2000) 547–553.
[46] T. Joachims, Making large-scale SVM learning practical, in: B. Scholkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods—Support Vector Learning, MIT Press, Cambridge, 1999.
[47] B.W. Matthews, Comparison of predicted and observed secondary structure of T4 phage lysozyme, Biochem. Biophys. Acta 405 (1975) 442–451.
[48] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (2006) 861–874.
[49] B.L. Welch, The generalization of "student's" problem when several different population variances are involved, Biometrika 34 (1947) 28–35.
[50] L. Schmidt, Statistical significance testing and cumulative knowledge in psychology, Psychol. Methods 1 (1966) 115–129.
[51] J. Janez, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[52] J.D. Qiu, J.H. Huang, R.P. Liang, X.Q. Lu, Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform, Anal. Biochem. 390 (2009) 68–73.
[53] D.F. Li, G.C. Wu, Construction of a class of Daubechies type wavelet bases, Chaos Soliton Fractals 42 (2009) 620–625.

[54] L. Holm, C. Sander, Mapping the protein universe, Science 273 (1996) 595–602.
[55] B. Dasarathy, Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, McGraw-Hill Computer Science Series, IEEE Computer Society Press, Las Alamitos, California, 1991.
[56] M. James, Classification Algorithms, Collins, London, 1985.
[57] K.C. Chou, H.B. Shen, Recent progress in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.
[58] K.C. Chou, C.T. Zhang, Review: prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.