Biochimica et Biophysica Acta 1813 (2011) 424–430

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

Identify submitochondria and subchloroplast locations with pseudo amino acid composition: Approach from the strategy of discrete wavelet transform feature extraction

Shao-Ping Shi [a,b], Jian-Ding Qiu [a,*], Xing-Yu Sun [a], Jian-Hua Huang [a], Shu-Yun Huang [a], Sheng-Bao Suo [a], Ru-Ping Liang [a], Li Zhang [a]

[a] *Department of Chemistry, Nanchang University, Nanchang 330031, China*
[b] *Department of Mathematics, Nanchang University, Nanchang 330031, China*

ARTICLE INFO

Article history:
Received 30 October 2010
Received in revised form 6 January 2011
Accepted 10 January 2011
Available online 19 January 2011

Keywords:
Submitochondria locations
Subchloroplast locations
Discrete wavelet transform
Support vector machine
Pseudo amino acid composition

ABSTRACT

It is very challenging and complicated to predict protein locations at the sub-subcellular level. The key to enhancing the prediction quality for protein sub-subcellular locations is to grasp the core features of a protein that can discriminate among proteins with different subcompartment locations. In this study, a different formulation of pseudoamino acid composition by the approach of discrete wavelet transform feature extraction was developed to predict submitochondria and subchloroplast locations. As a result of jackknife cross-validation, with our method, it can efficiently distinguish mitochondrial proteins from chloroplast proteins with total accuracy of 98.8% and obtained a promising total accuracy of 93.38% for predicting submitochondria locations. Especially the predictive accuracy for mitochondrial outer membrane and chloroplast thylakoid lumen were 82.93% and 82.22%, respectively, showing an improvement of 4.88% and 27.22% when other existing methods were compared. The results indicated that the proposed method might be employed as a useful assistant technique for identifying sub-subcellular locations. We have implemented our algorithm as an online service called SubIdent (http://bioinfo.ncu.edu.cn/services.aspx).

1. Introduction

Mitochondria are essential subcellular organelles of eukaryotes, while chloroplasts are typical plant cell organelles. Mitochondria and chloroplasts originated from endosymbiotic bacteria and last shared common ancestry since 2 billion years ago [1]. Both are surrounded by two layers of membrane, mitochondria can be further subdivided into three subcompartments: mitochondrial inner membrane, outer membrane and matrix [2], meanwhile chloroplasts can be subdivided into the envelope membrane, stroma, and thylakoid [3]. Core functions of mitochondria include oxidative phosphorylation, amino acid metabolism, fatty acid oxidation, and ion hemostasis [4]. These and other biochemical pathways intersect in the mitochondrion, making it the key organelle of energy and intermediary metabolism, as well as biosynthetic processes [4]. Chloroplasts perform essential metabolic and biosynthetic functions of global significance, including photosynthesis and amino acid biosynthesis [5]. Photosynthesis in the chloroplasts of plants is of fundamental importance for the existence of biosystems on the Earth [6].

It is clear that proteins located in different subcompartments play distinctive roles in various biological processes. So knowledge of their subcompartment locations can provide useful hints for revealing their functions and understanding how they interact with each other in cellular networking. Moreover, it has been proven that mitochondrial defects are implicated in a spectrum of human diseases, ranging from rare monogenic to common multifactorial disorders [7,8]. Thus, the knowledge of their submitochondria locations can be very helpful for the drug designs of many diseases related with mitochondrial defects. However, it is both expensive and time-consuming to conduct various experiments to obtain information about the protein subcompartment locations. Hence it is becoming a crucial issue to develop some reliable computational methods for identifying protein subcompartment locations.

Actually, many different methods have been developed to predict protein subcellular locations from primary protein sequences [9–12], and have made great progress. However, the prediction of protein localizations at sub-subcellular level is challenging compared with that at the subcellular level [13–16]. To the best of our knowledge, only three computational systems have been reported in literatures for predicting protein submitochondria locations [2,17,18], and only one computational method has been proposed to predict protein subchloroplast locations so far [19]. These methods were mainly based on different pseudo amino acid compositions (PseAAC) to extract feature vectors [2,17–19].

To avoid competently losing the sequence-order information in representing protein samples with the classical amino acid composition [20], PseAAC was introduced [21,22]. For a brief introduction about

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

0167-4889/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbamcr.2011.01.011

Chou's PseAAC, visit the Wikipedia web-page at http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition. Since the concept of PseAAC was proposed, varieties of PseAAC approaches have been widely used in many research studies [23]. Du and Li [17] used the occurrence frequencies of different residues, dipeptide composition and Chou's PseAAC to construct the feature vectors. Nanni and Lumini [18] applied genetic programming extracting 15 "artificial" features as Chou's PseAAC. Zeng et al. [2] constructed a substitution model based on the augmented Chou's PseAAC, which was composed of amino acid composition and auto covariance (AC) variables obtained by using AC to transform numerical vectors of eight physicochemical properties of amino acids into uniform matrices. Du et al. [19] created the PseAAC by computing the correlation function of the physicochemical properties of two residues for subchloroplast location prediction. In these methods, the most total accuracy was 89.7% for predicting submitochondria locations [2], but the most predictive accuracy was 78.05% for mitochondrial outer membrane [18], only 55% and 43.18% for chloroplast thylakoid lumen on the RAW dataset and S60 dataset, respectively [19].

In this study, to improve the quality of predicting submitochondria and subchloroplast locations, especially for mitochondrial outer membrane and chloroplast thylakoid lumen, we proposed a different formulation of PseAAC by the approach of discrete wavelet transform (DWT) feature extraction. 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 provide local structures of sequences which can more effectively reflect the sequence order effects [24]. This method was composed of three main steps. First the protein sequences were transformed into numerical signals by using physicochemical properties of amino acids. Then, these numerical sequences were further processed by DWT to extract salient frequency-band features from signals. Following this, using the statistical method, a series of statistical feature vectors were constructed to represent the protein sequences. Finally, support vector machine (SVM) was applied to deal with the problem of multi-classification. By using the jackknife cross-validation, our method obtained a promising total accuracy of 93.38% for predicting submitochondria locations. Especially the method yielded the predictive accuracies of 82.93% for mitochondrial outer membrane, 82.22% and 64.39% for chloroplast thylakoid lumen on the RAW dataset and S60 dataset, respectively, which indicated that the current method might play a complementary role to the existing methods.

2. Materials and methods

2.1. Protein datasets

Some proteins can simultaneously exist at more than one subcellular location site. This kind of multiplex proteins may have special functions and hence are particularly interesting [25,26]. Here, for simplicity in demonstrating our new method, we just use the training dataset containing single-location proteins only. Nevertheless, with more experimental data available for submitochondria and subchloroplast proteins in future, by using the similar approach as elaborated in Ref. [26], the current method can also be extended to deal with the multiplex proteins as well.

Three datasets used in published works were adopted to validate the performance of the proposed approach. The first dataset included 317 proteins classified into three submitochondria locations: 131 inner membrane proteins, 41 outer membrane proteins, 145 matrix proteins [17]. The identity cut off was set to 40%, i.e., none of the proteins had greater than 40% sequence identity with any other one in the dataset in order to get a balance between the homologous bias and the size of the training set [17]. The second dataset was the RAW dataset [19], which had 737 highly sequences identify protein sequences localized in 4 subchloroplast compartments: 71 stroma, 60 thylakoid lumen, 516 thylakoid membrane, 90 envelopes. The third dataset was the S60 dataset [19] which was constructed with sequence identity cut off value 60% by using the CD-HIT [27] program to remove the highly homologous sequences from the RAW dataset. It contained 262 chloroplast protein sequences: 49 stroma, 44 thylakoid lumen, 129 thylakoid membrane, and 40 envelopes. Moreover, to examine the robustness of this prediction model, two independent test datasets taken from the Swiss-Prot database (December-2010,http://www.expasy.org/sprot/) were constructed. The first independent test dataset included 86 human mitochondria proteins with sequence identity cut off value 40% by using the CD-HIT in a same submitochondria location: 23 inner membrane, 15 outer membrane, 48 matrix (see supplementary materials). The second independent test dataset contained 77 chloroplast proteins with sequence identity cut off value 60% in a same subchloroplast location: 12 stroma, 12 thylakoid lumen, 32 thylakoid membrane, 21 envelopes (see supplementary materials). None of independent test proteins was included in the training dataset.

As suggested by some research[28,29], to remove the homologous sequences from the benchmark dataset, a cutoff threshold of 25% was imposed to exclude those proteins from the benchmark datasets that have equal to or greater than 25% sequence identity to any other in a same subset. However, in this study we did not use such a stringent criterion because otherwise the samples for some subsets would be too few to have statistical significance.

2.2. Protein representation based on DWT

Andrade et al. [30] have proposed that each subcellular location has maintained a characteristic physiochemical environment, and that proteins in each location have adapted to these environments. So we selected physicochemical properties of amino acids to capture the truly specific localizations information of mitochondrial and chloroplast proteins. The hydrophobicity and polarity are the two most important properties among various physicochemical properties, where mutual interaction determines the stability of a protein structure. Thus we took into account hydrophobicity value [31] and polarity [32] to convert these protein sequences into numerical series. After obtaining the numerical sequences of mitochondrial and chloroplast proteins, the feature wavelet coefficients of each protein were extracted by using DWT [33,34].

The most attractive character of DWT is the ability to elucidate simultaneously both spectral and temporal information and is particularly helpful in detecting subtle time localized changes [35]. 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 [36]. 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 [37–39]. Using the low-frequency wavelet coefficients to formulate the sample of a protein can better reflect its overall sequence order effect. In this work, a digital signal of the protein sequence obtained by polarity or hydrophobicity values was decomposed to j scales with details from scale 1 to scale j and an approximation at scale j by the DWT, and $(j+1)$ scales wavelet coefficients were obtained. With the increase of decomposition level j, more feature vectors of the signal can be observed. To further decrease the dimensionality of the extracted feature vectors, statistics over the set of the wavelet coefficients were used [40]. The following statistical features calculated from the approximation coefficients and detail coefficients were used for the classification of subcompartments: (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 sequence can be characterized as a 4(j+1) dimension feature vector. In this study, the decomposition level 5 was chosen, and the obtained 24 dimension feature vectors were then inputted to SVM for classification.

2.3. Prediction algorithm

SVM is a kind of machine learning algorithm based on statistical learning theory which was introduced by Vapnik [41]. It searches for an optimal separating hyper plane which maximizes the margin in feature space. Due to its ability to handle noise, large datasets, and large input spaces [42,43], SVM has been widely used to predict membrane protein type [44], protein structural class [42], specificity of GalNAc-transferase [45], HIV protease cleavage sites in protein [46], beta-turn types [47], protein signal sequences and their cleavage sites [48], alpha-turn types [49], catalytic triads of serine hydrolases [50], B-cell epitope prediction [51], as well as protein subcellular location [52], among many other protein attribute. Details about the theory of SVM can be found in the literature [53,54]. Here, a radial basis function (RBF) was chosen as the kernel function, and two parameters the penalty parameter C and the kernel width parameter γ were tuned based on the training set using the grid search strategy in LIBSVM [55]. For actual implementation we used the LIBSVM package (version 2.81) which can be free downloaded from: http://www.Csie.Ntu.Edu.Tw/~cjlin/libsvm/.

In this work, the classification models consist of dual-layer SVMs: the first layer SVMs were implemented to identify mitochondria and chloroplasts, if a test protein was predicted to be mitochondria, then the second layer SVMs were implemented to identify which submitochondria locations the test mitochondrial protein belongs to. After a test sequence was predicted to be chloroplast, we continued to predict which subchloroplast locations the test chloroplast belongs by using the second layer SVMs. SVM was originally designed for binary classification [41], whereas prediction of submitochondria and subchloroplast locations is a multiclass classification problem. In this study, the one versus one (o-v-o) SVM training strategy was adopted to decompose multiclass into a series of binary SVMs to solve this problem [56]. For an N class classification, $N^*(N-1)/2$classifier needs to be trained, covering all possible different pairwise combinations (i,j), $i<j$, such that when training the (i,j) classifier patterns belonging to class i are used as positive samples and those from class j are taken as negative samples. Finally, one unknown sample is classified into the class obtained by accumulating the binary decisions and selecting as the winning class the one with more votes [56].

2.4. Assessment of predictive performances

In statistical prediction, the independent dataset test, subsampling test, and jackknife test are often used in literatures for examining the accuracy of a predictor [57]. However, as elucidated in Ref. [9] and demonstrated by Eq.1 of Ref. [29], the jackknife test is deemed to be the most objective one that can always yield a unique result for a given benchmark dataset. Therefore, the jackknife test has been increasingly and widely used to test the powers of various statistical predictors (see, e.g., [58–76]). Accordingly we also adopted jackknife test to evaluate the powers of the prediction method proposed in this study.

Accuracy, total accuracy and Matthews correlation coefficient (MCC) were utilized to assess the performance of prediction system. The MCC takes into account not only the number of true positives but also the number of false positives, false negatives and true negatives, and it is generally regarded as a balanced measure which can be used even if the classes are of very different sizes, for these reasons the MCC is more reliable than the accuracy. All of the above measurements are defined as follows:

$$\mathrm{Accuracy}(i) = \frac{\mathrm{TP}(i)}{\mathrm{TP}(i) + \mathrm{FN}(i)}, \quad \mathrm{Total\ Accuracy} = \frac{1}{N}\sum_{i=1}^{k} TP(i),$$

$$\mathrm{MCC}(i) = \frac{\mathrm{TP}(i)\mathrm{TN}(i) - \mathrm{FP}(i)\mathrm{FN}(i)}{\sqrt{(\mathrm{TP}(i) + \mathrm{FP}(i))(\mathrm{TP}(i) + \mathrm{FN}(i))(\mathrm{TN}(i)) + \mathrm{FN}(i))(\mathrm{TN}(i) + \mathrm{FP}(i))}}$$

where: N is the total number of the sequences in training data set, k is the number of classes; $\mathrm{TP}(i), \mathrm{TN}(i), \mathrm{FP}(i), \mathrm{FN}(i)$denote the number of true positives, true negatives, false positives and false negatives of the ith location, respectively.

3. Results and discussion

3.1. Selecting wavelet functions

Wavelet transform (WT) is based on the idea of mapping a signal onto a set of basis functions. Based on different basis functions, the wavelet functions have different families; every wavelet family has its quality fitting for different signals and has different results [33]. As the characteristics of the analyzing wavelet control the performance of the WT, the better the analyzing wavelet function matches the underlying structure in the signal, the more concise and sparse the WT representation. So the selection of wavelet functions becomes an important stage to achieve optimal performance in signal processing. For the Bior3.1, Bior3.9, Rbio3.3, Rbio3.5 and Rbio3.7 functions have better performance in analyzing the protein sequences in the 46 wavelet functions [77], we have tried these five wavelet functions for testing in the research. The performances for submitochondria locations with different types of wavelet functions were summarized in Table 1. As can be seen from the Table 1, the predictive accuracy of mitochondrial outer membrane was significantly lower than those of other compartments in each wavelet function. One possible cause of this could be that outer membrane proteins contained β-barrel as their membrane spanning segments [78], and yet the prediction of β-barrel membrane spanning segments was more difficult due to the lack of a clear pattern in their membrane spanning strands [79]. By using the Bior3.1 wavelet function, the predictive accuracy for mitochondrial outer membrane reached 63.41%, which was superior to other wavelet functions. Since the aim of this paper was to enhance the prediction performance for mitochondrial outer membrane, the Bior3.1 wavelet function was selected as the appropriate wavelet function in this study.

3.2. Selection of optimal decomposition scale

A WT decomposes a signal into several vectors of coefficients. 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 would introduce ineluctable redundancy in the decomposing process [77]. On the other hand, decomposing a longer sequence with too low a decomposition level would omit much detailed information [77]. In order to gain the highest predictive accuracy, an appropriate decomposition scale was selected. Considering that most mitochondrial and chloroplast sequences contain 150–600 amino acids [80], 3–6 scales were chosen to decompose the test sequences, separately. Because we mainly wanted to improve the quality of predicting for mitochondrial outer membrane and chloroplast thylakoid lumen, the performance of mitochondrial outer membrane, chloroplast thylakoid lumen (RAW) and chloroplast thylakoid lumen (S60) under four decomposition scales were shown in Fig. 1. A significant increase in accuracy can be observed for those proteins with decomposition scale 5, and the accuracies were about 82.93% for mitochondrial outer

Table 1
Prediction accuracies of submitochondria locations with different wavelet functions by using the decomposition scale with 4 and polarity values.

Submitochondria locations	Different wavelet functions (%)				
	Bior3.1	Bior3.9	Rbio3.3	Rbio3.5	Rbio3.7
Inner membrane	91.22	87.02	86.64	99.62	81.68
Outer membrane	63.41	60.98	62.20	62.20	37.80
Matrix	96.90	96.55	93.45	100	95.86
Total accuracy	90.22	88.00	86.59	94.95	82.49

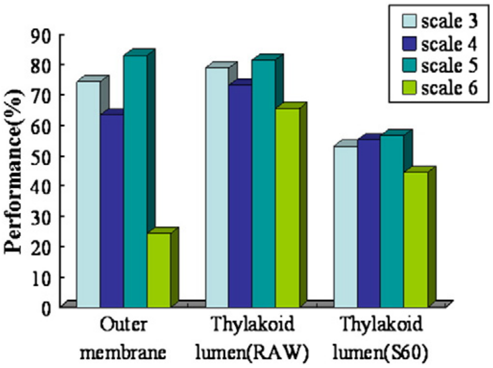

Fig. 1. The performance of different decomposition scales by using the Bior3.1 wavelet and polarity values.

membrane, 81.67% for chloroplast thylakoid lumen (RAW), and 56.82% for chloroplast thylakoid lumen (S60), which were higher than those of other decomposition scales. Therefore, scale 5 was selected as the appropriate decomposition scale in this study.

3.3. Comparison with different physicochemical properties

In Section 2.2, we selected the hydrophobicity and polarity of amino acids to capture the truly specific localizations information of mitochondrial and chloroplast proteins. Hence, we discussed the hydrophobicity value and polarity to impact the results of forecasts. The comparison results were shown in Tables 2 and 3, respectively. We found that the accuracies for submitochondria locations by using hydrophobicity value were significant decrease than those by using polarity, especially the accuracy at outer membrane location. Conversely, for subchloroplast locations, the accuracies by using hydrophobicity value were higher than those by using polarity, especially the accuracy at thylakoid lumen location. Andrade et al. [30] have pointed out that the average physicochemical properties of the molecular surface were correlated with the amino acid composition of the sequence, for this reason we tried to explain the predictive results from amino acid composition. Fig. 2 gives the amino acid average composition of 41 mitochondrial outer membrane sequences and 60 chloroplast thylakoid lumen sequences by COPid [81], respectively. Twenty amino acids were divided into three groups using their individual hydropathies according to the ranges of the hydropathy scale: polar or strongly hydrophilic, strongly hydrophobic, weakly hydrophilic or weakly hydrophobic [82]. It can be seen from Fig. 2 that polar amino acids and strongly hydrophobic amino acids were 46.71% versus 36.91% in mitochondrial outer membrane sequences, while in chloroplast thylakoid lumen sequences polar amino acids and strongly hydrophobic amino acids accounted for 31.8% and 35.3%, respectively. According to literature reports, there existed many charged and polar residues in the membrane of outer membrane proteins [83,84], which was consistent with our calculation results. This indicated that polar characteristic tended to be greater for mitochondrial outer membrane, and hydrophobic characteristic was more prominent than polar characteristic for chloroplast thylakoid lumen. Consequently, we used polarity and hydrophobicity value to predict submitochondria and subchloroplasts locations, respectively.

Table 2
Comparison of polarity with hydrophobicity value for submitochondria locations by using the Bior3.1 wavelet function and decomposition scale 4.

Submitochondria locations	Physicochemical properties (%)	
	Polarity	Hydrophobicity value
Inner membrane	91.22	84.73
Outer membrane	63.41	23.17
Matrix	96.90	94.83
Total accuracy	90.22	81.39

Table 3
Comparison of polarity with hydrophobicity value for subchloroplast locations by using the Bior3.1 wavelet function and decomposition scale 4.

Subchloroplast locations	RAW (%)		S60 (%)	
	Polarity	Hydrophobicity value	Polarity	Hydrophobicity value
Stroma	81.22	84.04	83.67	85.71
Thylakoid lumen	73.33	77.22	55.30	60.61
Thylakoid membrane	99.22	98.90	94.57	96.90
Envelopes	89.26	91.11	79.17	80.83
Total accuracy	94.17	94.75	83.59	86.25

3.4. Results and comparison with different methods

As mentioned above, mitochondria and chloroplasts are all essential subcellular organelles, and have many resemblances in the structure and function. So we also predicted submitochondria locations and subchloroplast locations based on DWT feature extraction in this study. First, we put 317 mitochondrial proteins and 737 chloroplast proteins all together for distinction between mitochondria and chloroplasts. By using the Bior3.1 wavelet, scale 5 and polarity values, 310 out of 317 mitochondrial proteins were predicted correctly, the success rate was about 97.7%; 732 out of the 737 chloroplast proteins were predicted correctly, making the success rate about 99.3%. This showed that the method we used could effectively distinguish mitochondrial proteins from chloroplast proteins. Despite the high similarities in the structure and function, mitochondria and chloroplasts do exhibit some differences [85,86]. Then, if a test protein was predicted to be mitochondrion, we further identified which submitochondria location it belongs to; also, after a test sequence was predicted to be chloroplast, we continued to predict its subchloroplast location. The optimal parameters combination (C and γ) used for training models were given in supplementary materials Table S1 and Table S2. The results of submitochondria locations and subchloroplast locations were listed in Tables 4 and 5, respectively.

To evaluate the prediction performance of the current method objectively, we made comparisons with existing methods. As shown in Table 4, the total accuracy for submitochondria locations by the current approach was 93.38%, which was higher than those by other three methods. Especially the predictive accuracy of mitochondrial outer membrane was 82.93%, which had been remarkably enhanced. Furthermore, the MCC range of 0.79 to 0.88 showed that our method had good prediction performance. It can be seen from Table 5 that the

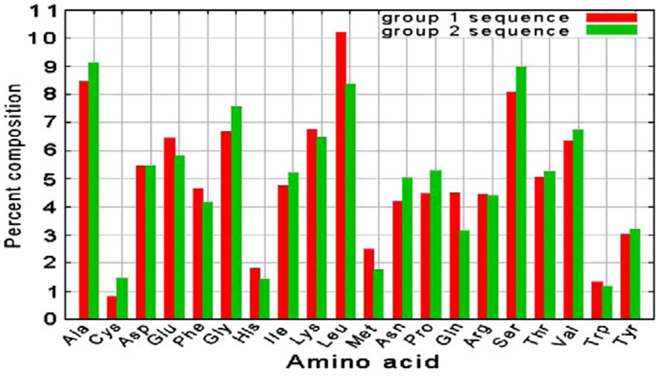

Fig. 2. Amino acid average composition of 41 mitochondrial outer membrane sequences and 60 chloroplast thylakoid lumen. Group 1 sequence is 41 mitochondrial outer membrane, group 2 sequence is 60 chloroplast thylakoid lumen.

Table 4
Comparison of different methods for submitochondria locations.

Assessment methods	Submitochondria locations	Prediction methods			
		SUBMITO [17]	GP-LOC [18]	AC [2]	Our method[a]
Jackknife test (%)	Inner membrane	85.50	83.21	91.80	91.60
	Outer membrane	51.20	78.05	66.10	82.93
	Matrix	94.50	97.24	96.40	97.93
	Total accuracy	85.20	89.00	89.70	93.38
MCC	Inner membrane	0.79	0.80	0.79	0.86
	Outer membrane	0.64	0.77	0.63	0.88
	Matrix	0.77	0.85	0.79	0.79

[a] By polarity in the decomposition scale 5 and Bior3.1 wavelet function.

total accuracies on the RAW dataset and S60 dataset reached 97.96% and 89.31%, respectively, which were remarkably higher than those of the SubChlo [19]. Besides, the predictive accuracies of chloroplast thylakoid lumen on the RAW dataset and S60 dataset were 82.22% and 64.39%, respectively, about 27.22% and 21.21% higher than those of the SubChlo [19]. The predictive accuracies of chloroplast envelopes on the RAW dataset and S60 dataset were 100% and 80%, respectively, about 15.56% and 40% higher than the results in SubChlo [19]. Moreover, we observed that the performance on RAW dataset was better than on S60 dataset. This observation indicates that the performance on RAW dataset may be over estimated, as the RAW dataset contains much more homologous and redundant sequences. Although homology is known to significantly impact the prediction accuracy, no standards are imposed when it comes to performing tests [87]. Du et al. [19] also investigated the homologous and redundancy problem based on RAW and S60 dataset, and drew the conclusion that the only way to get rid of such problem is to preprocess the dataset to remove those highly homologous sequences before training the predictor.

3.5. Predictive performance of independent test

Moreover, as a demonstration of practical application, predictions were also conducted for two independent test datasets. The results of independent test for submitochondria locations and subchloroplast locations were shown in supplementary materials Table S3 and Table S4, respectively. The overall accuracies of independent test for submitochondria locations and subchloroplast locations were 91.86% and 84.42%, respectively, about 1.52% and 4.89% lower than those of train test. The predictive accuracies of mitochondrial outer membrane and thylakoid lumen were increased to 86.67% and 66.67%, respectively. The predictive accuracies of other subcompartments by train test were slightly higher 1.21% to 4.64% than those of independent test. If the performance of the independent test is much worse than train test, then the trained model may be over-fitting for the training data. Generally, the performance in the independent test was just a little lower than those obtained in train test, which was also acceptable and indicated the robustness of our prediction model.

Why could the prediction accuracy be improved so much by DWT? The high performance of most prediction methods arises mainly from the cooperation between informative features and efficient classifier design. As a machine learning technique, SVM requires a fixed length of pattern, it is impossible to use this technique in case of protein with too small or too large length [88]. These problems can be overcome by DWT. DWT can capture hidden components from biological data, reduce the dimension of the input vector, improve calculating efficiency, and more effectively reflect the overall sequence order feature of a protein. Therefore, DWT methods appear to be a natural way to achieve vast improvements in the quality of prediction of subcompartment locations.

Table 5
Comparison of different methods for subchloroplast locations.

Subchloroplast locations	RAW (%)		S60 (%)	
	SubChlo [19]	Our method[a]	SubChlo[19]	Our method[a]
Stroma	78.87	85.92	67.35	85.71
Thylakoid lumen	55.00	82.22	43.18	64.39
Thylakoid membrane	96.12	100	83.72	98.19
Envelopes	84.44	100	40.00	80.00
Total accuracy	89.69	97.96	67.18	89.31

[a] By hydrophobicity value in the decomposition scale 5 and Bior3.1 wavelet function.

4. Conclusion

Prediction of protein locations at the sub-subcellular level is a very challenging and complicated problem. The key to enhancing the prediction quality for protein sub-subcellular locations is to grasp the core features of a protein that are intimately related to its localization in a cell. In this study, a novel pseudo amino acid approach based on DWT feature extraction has been proposed for the prediction of submitochondria and subchloroplast locations. Our method can efficiently distinguish mitochondrial proteins from chloroplast proteins. In comparison with previous literature methods, the current method can more effectively reflect the sequence order effects, and the predictive performance was significantly enhanced, indicating that the current method is an effective tool for the prediction of protein sub-subcellular locations. Extraction of informative features through DWT plays an important role in designing an accurate system for predicting protein sub-subcellular locations. It is anticipated that the powerful approach may become a useful high throughput tool for many other relevant area in bioinformatics, proteomics, and molecular biology. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods [89], we have implemented our algorithm as an online service called SubIdent (http://bioinfo.ncu.edu.cn/services.aspx.).

Acknowledgments

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.bbamcr.2011.01.011.

References

[1] W.L. Hao, OrgConv: detection of gene conversion using consensus sequences and its application in plant mitochondrial and chloroplast homologs, BMC Bioinform. 11 (2010) 114.

[2] Y.H. Zeng, Y.Z. Guo, R.Q. Xiao, Y. Li, L.Z. Yu, M.L. Li, Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach, J. Theor. Biol. 259 (2009) 366–372.

[3] M. Ferro, D. Salvi, S. Brugiere, S. Miras, S. Kowalski, M. Louwagie, J. Garin, J. Joyard, N. Rolland, Proteomics of the chloroplast envelope membranes from *Arabidopsis thaliana*, Mol. Cell. Proteomics 2 (2003) 325–345.

[4] M. Elstner, C. Andreoli, U. Ahting, I. Tetkol, T. Klopstock, T. Meitinger, H. Prokisch, MitoP2: an integrative tool for the analysis of the mitochondrial proteome, Mol. Biotechnol. 40 (2008) 306–315.

[5] T. Kleffmann, D. Russenberger, A. von Zychlinski, W. Christopher, K. Sjolander, W. Gruissem, S. Baginsky, The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions, Curr. Biol. 14 (2004) 354–362.

[6] A.V. Melkikh, V.D. Seleznev, O.I. Chesnokova, Analytical model of ion transport and conversion of light energy in chloroplasts, J. Theor. Biol. 264 (2010) 702–710.

[7] B.B. Lowell, G.I. Shulman, Mitochondrial dysfunction and type 2 diabetes, Science 307 (2005) 384–387.

[8] M.T. Lin, M.F. Beal, Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases, Nature 443 (2006) 787–795.

[9] K.C. Chou, H.B. Shen, Recent progress in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.

[10] Z.M. Wang, L. Jiang, M.L. Li, L.N. Sun, R.Y. Lin, Fast Fourier transform-based support vector machine for subcellular localization prediction using different substitution models, Acta Bioch. Bioph. Sin. 39 (2007) 715–721.

[11] J.D. Qiu, S.H. Luo, J.H. Huang, X.Y. Sun, R.P. Liang, Predicting subcellular location of apoptosis proteins based on wavelet transform and support vector machine, Amino Acids 38 (2010) 1201–1208.

[12] Q.B. Gao, Z.C. Jin, C. Wu, Y.L. Sun, J. He, X. He, Feature extraction techniques for protein subcellular localization prediction, Curr. Bioinform. 4 (2009) 120–128.

[13] H.B. Shen, K.C. Chou, Predicting protein subnuclear location with optimized evidence—theoretic K-nearest classifier and pseudo amino acid composition, Biochem. Biophys. Res. Commun. 337 (2005) 752–756.

[14] Z. Lei, Y. Dai, An SVM-based system for predicting protein subnuclear localizations, BMC Bioinform. 6 (2005) 291.

[15] W.L. Huang, C.W. Tung, H.L. Huang, S.F. Hwang, S.Y. Ho, ProLoc: prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features, BioSystems 90 (2007) 573–581.

[16] F.M. Li, Q.Z. Li, Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach, Amino Acids 34 (2008) 119–125.

[17] P.F. Du, Y.D. Li, Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical, BMC Bioinform. 7 (2006) 518.

[18] L. Nanni, A. Lumini, Genetic programming for creating Chou's pseudo amino acid based features for submitochondria location, Amino Acids 34 (2008) 653–660.

[19] P.F. Du, S.J. Cao, Y.D. Li, SubChlo: predicting protein subchloroplast locations with pseudo- amino acid composition and the evidence-theoretic K-nearest neighbor (ET-KNN) algorithm, J. Theor. Biol. 261 (2009) 330–335.

[20] K.C. Chou, A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space, Proteins Struct. Funct. Genet. 21 (1995) 319–344.

[21] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins Struct. Funct. Genet. 43 (2001) 246–255.

[22] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.

[23] K.C. Chou, Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology, Curr. Proteomics 6 (2009) 262–274.

[24] 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 discretewavelet transform, Anal. Biochem. 390 (2009) 68–73.

[25] K.C. Chou, H.B. Shen, A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0, PLoS ONE 5 (2010) e9931.

[26] K.C. Chou, H.B. Shen, Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites, J. Proteome Res. 6 (2007) 1728–1734.

[27] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.

[28] K.C. Chou, H.B. Shen, Cell-Ploc: a package of web servers for predicting subcellular localization of proteins in various organisms, Nat. Protoc. 3 (2008) 153–162.

[29] K.C. Chou, H.B. Shen, Cell-Ploc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms, Nat. Sci. 2 (2010) 1090–1103.

[30] M.A. Andrade, S.I. O'Donoghue, B. Rost, Adaption of protein surfaces to subcellular location, J. Mol. Biol. 276 (1998) 517–525.

[31] J.L. Fauchereand, V. Pliska, Transformational homologies in amino acid sequence, Eur. J. Med. Chem. 18 (1983) 369–375.

[32] R. Grantham, Amino acid difference formulae to help explain protein evolution, Science 185 (1974) 862–864.

[33] I. Daubechies, Ten lectures on wavelets, SIAM: CBMS-NSF regional conference series in applied mathematics, 1992.

[34] 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.

[35] K. Mori, N. Kasashima, T. Yoshioka, Y. Ueno, Prediction of spalling on a ball bearing by applying the discrete wavelet transform to vibration signals, Wear 195 (1996) 162–168.

[36] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 674–693.

[37] K.C. Chou, Low-frequency collective motion in biomacromolecules and its biological functions, Biophys. Chem. 30 (1988) 3–48.

[38] K.C. Chou, Low-frequency vibration of DNA molecules, Biochem. J. 221 (1984) 27–31.

[39] K.C. Chou, Low-frequency motions in protein molecules: beta-sheet and beta-barrel, Biophys. J. 48 (1985) 289–297.

[40] A. Kandaswamy, C.S. Kumar, R.P. Ramanathan, S. Jayaraman, N. Malmurugan, Neural classification of lung sounds using wavelet coefficients, Comput. Biol. Med. 34 (2004) 523–537.

[41] V.N. Vapnik, Statistical Learning Theory, Wiley-Interscience, New York, 1998.

[42] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Prediction of protein structural classes by support vector machines, Comput. Chem. 26 (2002) 293–296.

[43] 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.

[44] J.D. Qiu, X.Y. Sun, J.H. Huang, R.P. Liang, Prediction of the types of membrane proteins based on discrete wavelet transform and support vector machines, Protein J. 29 (2010) 114–119.

[45] 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.

[46] 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.

[47] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Support vector machines for the classification and prediction of beta-turn types, J. Pept. Sci. 8 (2002) 297–301.

[48] Y.D. Cai, S. Lin, K.C. Chou, Support vector machines for prediction of protein signal sequences and their cleavage sites, Peptides 24 (2003) 159–161.

[49] Y.D. Cai, K.Y. Feng, Y.X. Li, K.C. Chou, Support vector machine for predicting alpha-turn types, Peptides 24 (2003) 629–630.

[50] Y.D. Cai, G.P. Zhou, C.H. Jen, S.L. Lin, K.C. Chou, Identify catalytic triads of serine hydrolases by support vector machines, J. Theor. Biol. 228 (2004) 551–557.

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

[52] K.C. Chou, Y.D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, J. Biol. Chem. 277 (2002) 45765–45769.

[53] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (1998) 121–167.

[54] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[55] C.C. Chang, C.J. Lin, LIBSVM: a library for support machines [software], http://www.csie.ntu.edu.tw/cjlin/libsvm (2001).

[56] C.H.Q. Ding, I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks, Bioinformatics 17 (2001) 349–358.

[57] K.C. Chou, C.T. Zhang, Review: prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.

[58] G.P. Zhou, An intriguing controversy over protein structural class prediction, J. Protein Chem. 17 (1998) 729–738.

[59] F.M. Li, Q.Z. Li, Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach, Protein Pept. Lett. 15 (2008) 612–616.

[60] J.D. Qiu, J.H. Huang, S.P. Shi, R.P. Liang, Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform, Protein Pept. Lett. 17 (2010) 715–722.

[61] X. Xiao, P. Wang, K.C. Chou, GPCR-CA: a cellular automaton image approach for predicting G-protein-coupled receptor functional classes, J. Comput. Chem. 30 (2009) 1414–1423.

[62] K.C. Chou, H.B. Shen, Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides, Biochem. Biophys. Res. Comm. 357 (2007) 633–640.

[63] 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.

[64] J.D. Qiu, S.H. Luo, J.H. Huang, R.P. Liang, Using support vector machines to distinguish enzymes: approached by incorporating wavelet transform, J. Theor. Biol. 256 (2009) 625–631.

[65] G.P. Zhou, N. Assa-Munt, Some insights into protein structural class prediction, Proteins Struct. Funct. Genet. 44 (2001) 57–59.

[66] G.P. Zhou, K. Doctor, Subcellular location prediction of apoptosis proteins, Proteins Struct. Funct. Genet. 50 (2003) 44–48.

[67] M. Esmaeili, H. Mohabatkar, S. Mohsenzadeh, Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses, J. Theor. Biol. 263 (2010) 203–209.

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

[69] G.Y. Zhang, B.S. Fang, Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition, J. Theor. Biol. 253 (2008) 310–315.

[70] X.B. Zhou, C. Chen, Z.C. Li, X.Y. Zou, Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes, J. Theor. Biol. 248 (2007) 546–551.

[71] Y.S. Ding, T.L. Zhang, 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 (2008) 1887–1892.

[72] C. Chen, L. Chen, X. Zou, P. Cai, 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 (2009) 27–31.

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

[74] X. Jiang, R. Wei, T.L. Zhang, Q. Gu, 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 (2008) 392–396.

[75] H. Mohabatkar, Prediction of cyclin proteins using Chou's pseudo amino acid composition, Protein Pept. Lett. 17 (2010) 1207–1214.

[76] Q. Gu, Y.S. Ding, T.L. Zhang, Prediction of G-protein-coupled receptor classes in low homology using Chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns, Protein Pept. Lett. 17 (2010) 559–567.

[77] Z.N. Wen, K.L. Wang, M.L. Li, F.S. Nie, Y. Yang, Analyzing functional similarity of protein sequences with discrete wavelet transform, Comput. Biol. Chem. 29 (2005) 220–228.

[78] G.E. Schulz, The structure of bacterial outer membrane proteins, BBA Biomembr. 1565 (2002) 308–317.

[79] P.G. Bagos, T.D. Liakopoulos, I.C. Spyropoulos, S.J. Hamodrakas, A hidden Markov model method, capable of predicting and discriminating β-barrel outer membrane proteins, BMC Bioinform. 5 (2004) 29.

[80] L. Jiang, M.L. Li, Prediction of mitotochondrial proteins using discrete wavelet transform, Protein J. 25 (2006) 241–249.

[81] M. Kumar, V. Thakur, G.P. Raghava, COPid: composition based protein identification, In Silico Biol. 8 (2008) 121–128.

[82] Y.L. Chen, Q.Z. Li, Prediction of the subcellular location of apoptosis proteins, J. Theor. Biol. 245 (2007) 775–783.

[83] M.M. Gromiha, R. Majumdar, P.K. Ponnuswamy, Identification of membrane spanning beta strands in bacterial porins, Protein Eng. 10 (1997) 497–500.

[84] M.M. Gromiha, A simple method for predicting transmembrane alpha helices with better accuracy, Protein Eng. 12 (1999) 557–561.

[85] E. Glaser, J. Soll, in: H. Daniell, C. Chase (Eds.), Targeting signals and import machinery of plastids and plant mitochondria, Molecular Biology and Biotechnology of Plant Organelles: Chloroplasts and Mitochondria, Springer, Dordrecht, The Netherlands, 2004, pp. 385–418.

[86] A.K. Berglund, C. Pujol, A.M. Duchene, E. Glaser, Defining the determinants for dual targeting of amino acyl-tRNA synthetases to mitochondria and chloroplasts, J. Mol. Biol. 393 (2009) 803–814.

[87] L.A. Kurgan, L. Homaeian, Prediction of structural classes for protein sequences and domains—impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy, Pattern Recogn. 39 (2006) 2323–2343.

[88] D. Zhu, B. Ji, C. Meng, B. Shi, Z. Tu, Z. Qing, Study of wavelet denoising in apple's charge-coupled device near-infrared spectroscopy, J. Agric. Food Chem. 55 (2007) 5423–5428.

[89] K.C. Chou, H.B. Shen, Review: recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 2 (2009) 63–92.