Du *et al. BMC Bioinformatics* (2020) 21:212
https://doi.org/10.1186/s12859-020-3539-1

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Subcellular location prediction of apoptosis proteins using two novel feature extraction methods based on evolutionary information and LDA

Lei Du[1,2], Qingfang Meng[1,2*], Yuehui Chen[1,2] and Peng Wu[1,2]

*Correspondence:
ise_mengqf@ujn.edu.cn
[1]School of Information Science and Engineering, University of Jinan, Jinan 250022, China
[2]Shandong Provincial Key laboratory of Network Based Intelligent Computing, Jinan 250022, China

Abstract

Background: Apoptosis, also called programmed cell death, refers to the spontaneous and orderly death of cells controlled by genes in order to maintain a stable internal environment. Identifying the subcellular location of apoptosis proteins is very helpful in understanding the mechanism of apoptosis and designing drugs. Therefore, the subcellular localization of apoptosis proteins has attracted increased attention in computational biology. Effective feature extraction methods play a critical role in predicting the subcellular location of proteins.

Results: In this paper, we proposed two novel feature extraction methods based on evolutionary information. One of the features obtained the evolutionary information via the transition matrix of the consensus sequence (CTM). And the other utilized the evolutionary information from PSSM based on absolute entropy correlation analysis (AECA-PSSM). After fusing the two kinds of features, linear discriminant analysis (LDA) was used to reduce the dimension of the proposed features. Finally, the support vector machine (SVM) was adopted to predict the protein subcellular locations. The proposed CTM-AECA-PSSM-LDA subcellular location prediction method was evaluated using the CL317 dataset and ZW225 dataset. By jackknife test, the overall accuracy was 99.7% (CL317) and 95.6% (ZW225) respectively.

Conclusions: The experimental results show that the proposed method which is hopefully to be a complementary tool for the existing methods of subcellular localization, can effectively extract more abundant features of protein sequence and is feasible in predicting the subcellular location of apoptosis proteins.

Keywords: Subcellular location, Position-specific scoring matrix, Consensus sequence, Absolute entropy correlation analysis, Linear discriminant analysis

Background

Apoptosis, also known as programmed cell death, is a basic biological phenomenon that is associated with the occurrence of a wide variety of diseases, such as a tumor, autoimmune disease, Alzheimer's disease and so on. It plays an important role in animal development and homeostasis [1]. Studies have shown that apoptosis proteins are essential in this process. The subcellular location of a protein is closely related to its function. And only in the specific subcellular location, can the protein work [2]. Subcellular localization information of apoptosis proteins contributes to exploring the internal mechanism of programmed cell death and designing new drugs [3]. However, the number of apoptosis proteins with clear subcellular location markers is limited in the database, and it is time-consuming and costly to label samples by traditional experimental methods. Therefore, exploring feasible computational methods to predict the subcellular location of the given protein has been a hotspot for nearly two decades.

Over the years, various computational methods have been proposed in apoptosis protein subcellular localization. In 2006, Zhang et al. [4] built an apoptosis protein dataset with 225 proteins in total (ZW225). They proposed using a grouped weight of protein sequence and support vector machine (SVM) to predict the subcellular location of apoptosis proteins (EBGW_SVM). The overall accuracy achieved 83.1% by jackknife test. Chen et al. [5] constructed a different apoptosis protein dataset in 2007, which contains 317 proteins (CL317). By using the increment of diversity (ID), the highest jackknife predictive result was 82.7%. In the same year, they proposed another method called ID_SVM which combined ID with SVM to predict the subcellular location of apoptosis proteins [6]. The ID_SVM algorithm achieved a higher prediction accuracy by jackknife test, on CL317 dataset is 84.2% and ZW225 dataset is 85.8%. Zhang et al. [7] applied the concept of distance frequency and SVM to obtain the highest overall accuracy of 88.0% and 84.0% on CL317 dataset and ZW225 dataset, respectively. Liu et al. [8] extracted the evolutionary information embedded in the position-specific scoring matrix (PSSM) and combined it with auto covariance transformation to establish a PSSM-AC model. The overall accuracy achieved 91.5% on CL317 dataset and 84.0% on ZW225 dataset. Nearly, a series of advances have been achieved in the prediction of apoptosis protein subcellular location [9, 10]. Liang et al. [11] fused two feature descriptors named the frequency of triplet codons in the RNA sequence (FTC) and detrended forward moving-average cross-correlation analysis (DFMCA) to predict the subcellular location of apoptosis proteins, which reached the overall accuracy of 89.0% and 85.3% on CL317 dataset and ZW225 dataset, respectively. Li et al. [12] proposed two feature extraction methods namely generalized chaos game representation (GCGR) and novel statistics and information theory (NSI), they also combined them with other features including PseAAC and dipeptide composition. The jackknife prediction accuracy of CL317 dataset and ZW225 dataset was 92.7% and 87.1%, respectively by using SVM.

In summary, to identify the subcellular location of proteins, diverse methods have been proposed which mainly focus on the following two aspects: feature extraction and classification techniques. Feature extraction methods of protein sequences are the key to the prediction of subcellular location. Existing methods include protein amino acid composition (AAC) [13–15], pseudo-amino acid composition (PseAAC) [16, 17], physicochemical properties [18, 19], position-specific scoring matrix (PSSM) [20], Gene Ontology (GO) [21, 22] and so on. As for the classifier, numerous classifiers have been applied to solve the problem of protein subcellular localization, such as support vector machine (SVM) [23, 24], KNN [25, 26] and neural network [15, 27]. Among them, SVM is used extensively for its good classification performance and fast computing speed.

Some early studies used GO information as the feature to solve this problem and achieved the most significant improvement. However, for new proteins, because of the lack of GO information, it is hard to use the GO terms. Recently, evolutionary information based features extracted from position-specific scoring matrix (PSSM) have shown their effectiveness in subcellular localization [19]. Xie et al. [28] proposed a model named LOCSVMPSI which utilized the PSSM and four-part amino acid compositions as the feature vector. Huang et al. [29] formulated a protein sequence with the pseudo position-specific scoring matrix (PsePSSM). Dehzangi et al. [30] proposed a feature extraction method named PSSM-S to predict the subcellular location of Gram-positive and Gram-negative proteins. Wan et al. [31] combined the profile-alignment features and PseAA features to predict the localization of chloroplast proteins. Liang et al. [32] constructed a PSSM-based model by using Geary autocorrelation function and DCCA coefficient for apoptosis protein subcellular localization prediction. Xiang et al. [33] utilized the proportion of the golden section to split PSSM and proposed segmented evolutionary information to represent protein sequences. Wang et al. [34] proposed segmented amino acid composition in PSSM (PSSM-SAA) to tackle the subcellular localization problem. All these methods have shown that based on evolutionary information, discriminative features can be extracted for classification. Therefore, using PSSM to extract effective features to represent protein sequences is still an outstanding problem.

To solve the problem of insufficient information in a single feature set, researchers have paid attention to fuse multiple features to formulate protein sequences in recent years. Zhang et al. [35] fused Moran autocorrelation and cross correlation with PSSM to get protein sequence information, then the principal component analysis was used to reduce redundant and irrelevant information. Wan et al. [36] adopted a linear neighborhood propagation (LNP) classifier ensemble scheme to incorporate both split amino-acid composition (SAAC) features and profile-alignment (PA) features for predicting sub-chloroplast localization. Qu et al. [37] presented a method to predict the subcellular location of multi-site proteins by combining N-terminal signals, pseudo amino acid composition, physicochemical property, stereo-chemical property and amino acid index distribution. However, the fusion feature vectors usually established by splicing multiple different features and the features integrated through this method often have a high dimension [11, 38–40]. The high-dimensional features contain a good deal of redundant information which may have a harmful influence on performing the classifier. Dimensionality reduction algorithms can help to eliminate the redundant data from the original feature space and are widely used in machine learning [41].

In this paper, we focus on extracting features based on evolutionary information embedded in position-specific scoring matrix (PSSM). Two novel feature extraction methods are therefore proposed to predict the subcellular locations of apoptosis proteins. First, according to PSSM, we transform the protein primary sequence into the consensus sequence, and propose a novel feature extraction method based on the consensus sequence, which named consensus sequence-based transition matrix (CTM). The CTM feature reflects distributions information of the amino acid transitions. For each protein sequence, CTM method can obtain a 40-dimensional feature vector. Then we propose another feature extraction method calculated from PSSM matrix directly. It can establish a 190-dimensional feature named absolute entropy correlation analysis (AECA-PSSM). AECA-PSSM derives from relative entropy or KL divergence, and it reflects the relationship between each two columns of PSSM. Thus, for a given protein sequence, we can generate a 230-dimensional fusion feature. Next linear discriminant analysis (LDA) is used to eliminate the noise and reduce the dimension of the proposed features. Finally, the feature vector after dimensionality reduction is fed into SVM to identify the subcellular location. And the proposed CTM-AECA-PSSM-LDA method reaches a higher classification performance in identifying subcellular locations of apoptosis proteins on CL317 and ZW225 datasets.

Results

Performance of the two proposed feature extraction methods

In this paper, the protein samples are formulated with two evolutionary information based feature extraction methods: consensus sequence-based transition matrix (CTM) and absolute entropy correction analysis (AECA-PSSM). To investigate the effectiveness of the proposed method, we first test the performance of the two novel feature extraction methods.

In the consensus sequence-based transition matrix method, we utilize the consensus sequence transformed from the protein primary sequence to integrate evolutionary information embedded in PSSM. Then, for the consensus sequence enriched with evolutionary information, we construct a transition matrix to extract features. Figure 1 shows the comparison between features extracted from the transition matrix of protein primary sequence (PTM) and the transition matrix of consensus sequence (CTM). On the whole, the CTM method performs better than PTM method. That's because compared with the protein primary sequence, the consensus sequence obtained by PSSM contains the evolutionary information of the protein. Therefore, the proposed consensus sequence-based transition matrix method provides more discriminating information than the protein primary sequence.

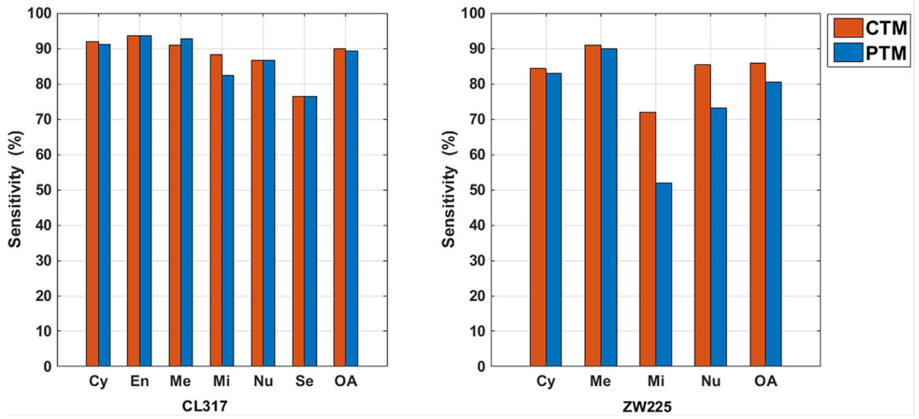

Fig. 1 Comparison between primary sequence-based transition matrix (PTM) feature and consensus sequence-based transition matrix (CTM) feature for the two datasets

Tables 1 and 2 show the classification results of CTM feature extraction method on CL317 and ZW225, respectively. In Table 1, we find that on CL317 dataset, the overall accuracy is 89.91% and the sensitivity of secreted proteins is lower than other five subcellular locations. As can be seen from Table 2, on ZW225 dataset, the overall accuracy obtained by CTM method achieves 85.78% and expect mitochondrion proteins, the sensitivity of other subcellular locations is 84.29%-91.01%.

For the absolute entropy correlation analysis method, we use a novel analytical method derived from KL divergence to measure the relationship between each two columns in PSSM. Tables 3 and 4 show the classification results of AECA-PSSM feature extraction method on CL317 dataset and ZW225 dataset, respectively. It can be seen from Table 3 that on CL317 dataset, except the secreted proteins, the sensitivity in other subcellular locations is 84.62%-94.55% and the overall accuracy is 89.91%. From Table 4, we can find that on ZW225 dataset, the sensitivity of mitochondrion proteins is lower than other subcellular locations and the overall accuracy of the entire dataset reaches of 85.78%.

The sequence similarity has a significant influence for the prediction performance and the lower the sequence similarity is, the more difficult the prediction is. According to the above experimental results, we can find that the classification results of the two proposed feature extraction methods are lower in ZW225 dataset. The possible reason is that the ZW225 dataset has lower sequence similarity than the CL317 dataset, but the results are acceptable. The secreted proteins in CL317 dataset and mitochondrion proteins in ZW225 dataset have the lower sensitivity. The reason may be that the sample sizes of these two subcellular locations in the datasets are small and the classifier tends to predict samples as majority classes.

Effect of different feature extraction methods

One of the most important but also most difficult problems in computational biology is to convert the protein sequence into an effective numerical representation, which is known as feature extraction. In this paper, two novel evolutionary information based feature extraction methods are proposed to represent protein sequence information. After getting the CTM feature and AECA-PSSM feature, we combined them to form a 230 dimensional fusion feature vector: CTM-AECA-PSSM. However, as more sequence information is obtained by combining the two features, it also brings more noise, which has a negative impact on the predictor. Then we project the 230-dimensional fusion feature into a $p = C - 1$ dimensional feature space by LDA dimensionality reduction method. Table 5 shows the contributions of different feature extraction methods on CL317 dataset and ZW225 dataset.

Table 1 Classification results of CTM feature for the CL317 dataset

location	Jackknife test					
	Cy	En	Me	Mi	Nu	Se
Sen(%)	91.96	93.62	90.91	88.24	86.54	76.47
Spe(%)	91.71	100	98.47	98.94	96.98	100
Acc(%)	91.80	99.05	97.16	97.79	95.27	98.74
MCC	0.82	0.96	0.90	0.88	0.83	0.87
F	0.89	0.97	0.92	0.90	0.86	0.87
OA(%)	89.91					

Table 2 Classification results of CTM feature for the ZW225 dataset

location	Jackknife test			
	Cy	Me	Mi	Nu
Sen(%)	84.29	91.01	72.00	85.37
Spe(%)	92.90	92.65	99.00	95.11
Acc(%)	90.22	92.00	96.00	93.33
MCC	0.77	0.83	0.78	0.78
F	0.84	0.90	0.80	0.82
OA(%)	85.78			

We listed in Table 5 the sensitivity of each subcellular location and the overall accuracy of different feature extraction methods in CL317 and ZW225 datasets. On CL317 dataset, we can get a better prediction results which reach the overall accuracy of 90.22% by using CTM-AECA-PSSM method. And on ZW225 dataset, the overall accuracy after fusing the CTM algorithm and AECA-PSSM algorithm is 85.33%, which is 0.45% lower than using the two algorithms alone. The reason may be that when the two features are fused by generating a higher-dimensional feature vector, we can only get more information from the protein sequence, but the noise caused by the redundant and irrelevant information can not be eliminated which may make the performance of the classifier worse. When the two features are fused, the role of them is not fully played. Therefore, the prediction accuracy of CTM-AECA-PSSM on ZW225 dataset decreased slightly. From Table 5, we can also find that after dimensionality reduction using LDA, the prediction results are significantly improved on both CL317 dataset and ZW225 dataset. It indicates that LDA can effectively eliminate the redundant and irrelevant information and improve the accuracy of subcellular localization. To further assess the robustness of the model using different feature extraction methods, Figs. 2 and 3 show the ROC curves using the four different feature extraction algorithms on CL317 dataset and ZW225 dataset, respectively.

Prediction results of different classification algorithms

In this paper, we consider four different classification algorithms, including extreme learning machine (ELM) [42], K-nearest neighbors (KNN) [34], logistic regression (LR) [43] and support vector machine (SVM). The prediction results under the four classifiers by jackknife test on the CL317 dataset and ZW225 dataset are shown in Table 6. It can be seen from Table 6 that the four classifiers have ideal prediction results on the two datasets which shows the effectiveness of our extracted features. On CL317 dataset, all of these four classifiers achieve the overall accuracy more than 99%. SVM, ELM and LR all achieve the highest overall accuracy of 99.68%. On ZW225 dataset, the overall accuracy

Table 3 Classification results of AECA-PSSM feature for the CL317 dataset

location	Jackknife test					
	Cy	En	Me	Mi	Nu	Se
Sen(%)	91.07	91.49	94.55	91.18	84.62	76.47
Spe(%)	95.12	99.63	96.95	98.59	96.98	99.67
Acc(%)	93.69	98.42	96.53	97.79	94.95	98.42
MCC	0.86	0.94	0.88	0.89	0.82	0.83
F	0.91	0.95	0.90	0.90	0.85	0.84
OA(%)	89.91					

Table 4 Classification results of AECA-PSSM feature for the ZW225 dataset

location	Jackknife test			
	Cy	Me	Mi	Nu
Sen(%)	87.14	94.38	68.00	75.61
Spe(%)	92.90	90.44	98.00	97.83
Acc(%)	91.11	92.00	94.67	93.78
MCC	0.79	0.84	0.71	0.78
F	0.86	0.90	0.74	0.82
OA(%)	85.78			

achieved by the four classifiers are more than 93% and SVM achieves the highest overall accuracy 95.56%. Therefore, in this paper, we choose SVM as the final classification prediction algorithm.

Classification results of the proposed method

Prediction of apoptosis protein subcellular localization is an important research content in bioinformatics. In this work, we propose a method named CTM-AECA-PSSM-LDA to identify the subcellular location of apoptosis proteins. First the two proposed feature extraction methods CTM and AECA-PSSM are employed to represent the protein sequence. Then dimensionality reduction is performed by LDA. Finally, the sample data after dimensionality reduction are classified by SVM. The results of jackknife test on CL317 and ZW225 datasets are presented in Table 7. From Table 7, we can see that the OA for CL317 dataset and ZW225 dataset by our method achieve 99.68% and 95.56%, respectively. The experimental results indicate that the method can effectively predict the subcellular location of apoptosis proteins.

Comparison with the other prediction methods

In this section, to further evaluate the effectiveness of the proposed method, we compare it with some previous methods on the same apoptosis protein datasets. Tables 8 and 9 show the prediction results of different methods on CL317 dataset and ZW225 dataset, respectively. All the results are obtained using jackknife test. The OA of the two datasets and the sensitivity of each subcellular class are listed.

Based on CL317 dataset, the performance of the proposed CTM-AECA-PSSM-LDA model is compared with ten previous predictors. The OA of these methods range from

Table 5 Classification results of different feature extraction methods

Dataset	Feature extraction method	Jackknife test (%)						
		Sensitivity						
		Cy	En	Me	Mi	Nu	Se	OA
CL317	CTM	91.96	93.62	90.91	88.24	86.54	76.47	89.91
	AECA-PSSM	91.07	91.49	94.55	91.18	84.62	76.47	89.91
	CTM-AECA-PSSM	92.86	91.49	92.73	85.29	88.46	76.47	90.22
	CTM-AECA-PSSM-LDA	99.11	100	100	100	100	100	99.68
ZW225	CTM	84.29	\	90.01	72.00	85.37	\	85.78
	AECA-PSSM	87.14	\	94.38	68.00	75.61	\	85.78
	CTM-AECA-PSSM	87.14	\	88.76	76.00	80.49	\	85.33
	CTM-AECA-PSSM-LDA	97.14	\	91.01	100	100	\	95.56

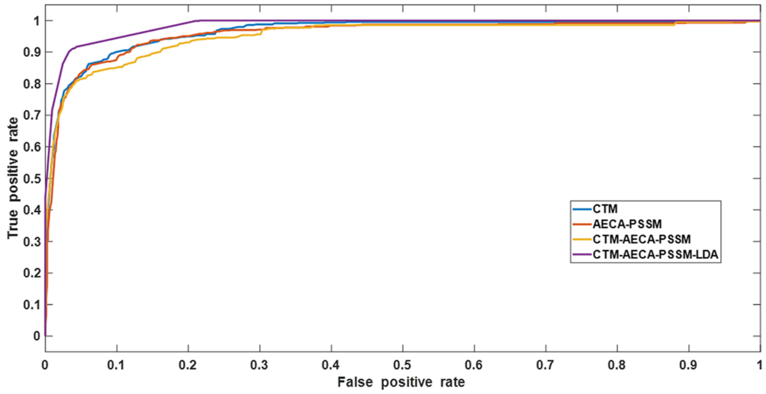

Fig. 2 The ROC curves of different feature extraction methods on CL317 dataset

82.7% to 99.7%, among which CTM-AECA-PSSM-LDA achieves the highest prediction accuracy (99.7%). The sensitivity of endoplasm proteins, membrane proteins, mitochondrion proteins, nucleus proteins and secreted proteins achieve 100% in our method. And for the cytoplasm proteins, the sensitivity reaches of 99.1% which is also the highest.

Similarly, based on ZW225 dataset, the proposed CTM-AECA-PSSM-LDA prediction model is compared with nine other existing methods. The OA of our method (95.6%) is higher than other predictors test in this study. The sensitivity of mitochondrion proteins and nucleus proteins (both 100%) are the highest among all the methods which shows the excellent ability of our method in identifying mitochondrion and nucleus proteins. Moreover, the highest sensitivity of cytoplasm proteins (97.1%) also achieves by our method.

Discussion

Apoptosis is a kind of elementary life phenomenon that exists widely in the biological world and apoptosis proteins play a significant role in this process. The function of apoptosis proteins is strongly related to their subcellular localization information. Facing the explosive growth of protein sequences in the post-genome era, to timely obtain useful

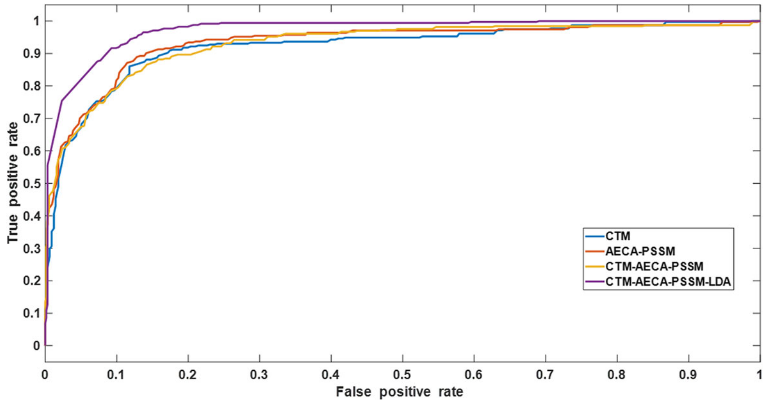

Fig. 3 The ROC curves of different feature extraction methods on ZW225 dataset

Table 6 Prediction results of different classifiers

Dataset	classifier	Jackknife test (%)						
		Sensitivity						
		Cy	En	Me	Mi	Nu	Se	OA
CL317	ELM	99.11	100	100	100	100	100	99.68
	KNN	99.11	100	98.18	100	100	100	99.37
	LR	99.11	100	100	100	100	100	99.68
	SVM	99.11	100	100	100	100	100	99.68
ZW225	ELM	91.43	\	92.13	92.00	100	\	93.33
	KNN	91.43	\	91.01	100	100	\	93.78
	LR	92.86	\	92.13	100	100	\	94.67
	SVM	97.14	\	91.01	100	100	\	95.56

information of sequences for drug design, it is an urgent need to develop computational methods for predicting the subcellular location of apoptosis proteins.

In this work, we propose a novel method to predict the subcellular location of apoptosis proteins named CTM-AECA-PSSM-LDA. Two novel feature extraction methods based on evolutionary information embedded in PSSM are designed. Firstly, the consensus sequence-based transition matrix (CTM) feature is extracted to reflect the amino acid transition information, and secondly, absolute entropy correlation analysis (AECA-PSSM) is proposed to obtain the relationship between each two columns of PSSM. After the two features mentioned above are fused together, LDA is adopted to reduce the dimension of the feature vector and eliminate the redundant information. Finally, SVM is regarded as a classifier to predict the subcellular location of apoptosis proteins. The overall accuracy by jackknife test is 99.7% and 95.6% for CL317 dataset and ZW225 dataset, respectively. Compared with other existing methods test in this paper, the overall accuracy is 3.7%-17% and 3.4%-12.5% higher on CL317 dataset and ZW225 dataset, respectively. The proposed CTM-AECA-PSSM-LDA method not only generates more discriminative features but also obtains satisfactory predictive performance on CL317 and ZW225 datasets, showing its potential for predicting apoptosis protein subcellular locations.

However, though our proposed CTM-AECA-PSSM-LDA method can effectively raise the prediction accuracy in two widely used benchmark datasets ZW225 and CL317, there are some disadvantages. Our method is trained on the two commonly used dataset and mainly consider the situation of single-site proteins that are ubiquitous in the existing apoptotic protein database. A series of recent publications [46–48] in demonstrating new

Table 7 Classification results of the proposed CTM-AECA-PSSM-LDA method

Locations	Jackknife test									
	CL317					ZW225				
	Sen (%)	Spe (%)	Acc (%)	MCC	F	Sen (%)	Spe (%)	Acc (%)	MCC	F
Cy	99.11	100	99.68	0.99	1	97.14	94.84	95.56	0.90	0.93
En	100	100	100	1	1	\	\	\	\	\
Me	100	99.62	99.68	0.99	0.99	91.01	98.53	95.56	0.91	0.94
Mi	100	100	100	1	1	100	100	100	1	1
Nu	100	100	100	1	1	100	100	100	1	1
Se	100	100	100	1	1	\	\	\	\	\
OA (%)			99.68					95.56		

Table 8 Comparison from different methods on CL317 dataset by jackknife test

Methods	Jackknife test (%)						
	Sensitivity						OA
	Cy	En	Me	Mi	Nu	Se	
ID [5]	81.3	83.0	81.8	85.3	82.7	88.2	82.7
ID_SVM [6]	91.1	87.2	89.1	79.4	73.1	58.8	84.2
DF_SVM [7]	92.9	86.5	85.5	76.5	93.6	76.5	88.0
PSSM-AC [8]	93.8	95.7	90.9	91.2	86.5	82.4	91.5
Liang et al. [32]	92.9	93.6	89.1	82.4	84.6	76.5	89.0
Zhang et al. [44]	96.1	100	95.7	93.9	95.5	98.0	96.0
FTC-DFMCA-PSSM [11]	92.9	93.6	89.1	82.4	86.5	93.6	89.0
MACC-PSSM [35]	96.4	93.6	94.5	82.4	80.8	76.5	90.5
Chen et al. [10]	95.5	94.1	93.6	96.4	94.2	94.1	94.8
ERT-ECT-PSSM-IS [45]	93.8	94.1	100	97.9	96.2	92.7	95.0
CTM-AECA-PSSM-LDA	99.1	100	100	100	100	100	99.7

findings or approaches pointed out that some proteins can simultaneously exist in or move between multiple subcellular locations. And these multi-site proteins usually have special functions and are of great search value. Given that it is difficult to collect a large enough multi-site apoptosis protein benchmark dataset meaningfully in statistics, similar to those of CL317 and ZW225 at present, which are the most widely used in previous studies, therefore, our method is still verified on those two datasets. In our future research, we will consider proteins with both single- and multi-site.

Conclusion

The purpose of studying the subcellular location of apoptosis proteins is to further explore the intrinsic mechanism of programmed cell death and better understand the nature of life. Base on the evolutionary information, we propose two novel feature extraction methods to generate sequence feature of proteins. Then linear discriminant analysis algorithm is used to reduce the dimension of the extracted features. Finally SVM classifier is employed to predict the subcellular location of proteins. By jackknife test, on the two benchmark datasets CL317 and ZW225, the OA reach 99.7% and 95.6%, respectively.

Table 9 Comparison from different methods on ZW225 dataset by jackknife test

Methods	Jackknife test (%)				
	Sensitivity				OA
	Cy	Me	Mi	Nu	
EBGW_SVM [4]	90.0	93.3	60.0	63.4	83.1
ID_SVM [6]	92.9	91.0	68.0	73.2	85.8
DF_SVM [7]	87.1	92.1	64.0	73.2	84.0
PSSM-AC [8]	82.9	92.1	68.0	78.0	84.0
Liang et al. [32]	87.1	89.1	68.0	75.6	84.4
Zhang et al. [44]	93.5	92.1	96.0	93.5	92.2
MACC-PSSM [35]	88.6	92.1	64.0	75.6	84.9
FTC-DFMCA-PSSM [11]	88.6	93.3	64.0	75.6	85.3
ERT-ECT-PSSM-IS [45]	80.0	91.0	92.0	87.8	87.1
CTM-AECA-PSSM-LDA	97.1	91.0	100	100	95.6

Experimental results show that the proposed method outperforms the previous predictors listed in the literature for most subcellular classes and indicate that it is promising for the recognition of subcellular locations. In general, our method is a relatively effective way to predict the subcellular location of apoptosis proteins. We hope that our method will be used as a complementary tool in the field of subcellular localization for proteins. According to a series of recent publications [49, 50], user-friendly and publicly accessible web-servers make great importance in building predictive system. We will make great efforts to provide a web-server for the proposed method in our future work.

Methods

Dataset

In this study, two benchmark datasets CL317 and ZW225 are applied to test the performance of the proposed method whose purpose is to determine the subcellular location of apoptosis proteins. The CL317 dataset was constructed in 2007 by Chen and Li [5]. It contains 317 protein sequences which located in six different subcellular locations respectively called cytoplasm (Cy), endoplasm (En), membrane (Me), mitochondrion (Mi), nucleus (Nu) and secreted (Se). The ZW225 dataset was established in 2006 by Zhang and Wang et al. [4]. It contains 225 protein sequences which located in four subcellular locations called cytoplasm (Cy), membrane (Me), mitochondrion (Mi) and nucleus (Nu). All of the two datasets are extracted from the SWISS-PROT database. Despite the small size of the two datasets, they are commonly used in the previous investigations [45, 51]. The details of two datasets CL317 and ZW225 are shown in Table 10.

The proposed feature extraction method

Effective feature extraction methods play a critical role in the subcellular location of proteins. In this paper, we propose two novel evolutionary information based feature extraction methods to effectively elucidate protein sequences. One of the feature extraction methods gets evolutionary information via the transition matrix of the consensus sequence (CTM). Another feature extraction method directly utilizes the evolutionary information from PSSM based on absolute entropy correlation analysis (AECA-PSSM).

To obtain the PSSM, PSI-BLAST program [52] is used to deal with the protein primary sequence from CL317 and ZW225 datasets. In our research, the non-redundant (NR) database is utilized, in the meantime, the E-value and the iterations numbers are respectively set to 0.001 and 3 [11]. The PSSM of a protein sequence with the length of L can be

Table 10 Details of the two datasets CL317 and ZW225

Dataset	Order	Subcellular localization	Number of proteins
CL317	1	Cytoplasm	112
	2	Endoplasm	47
	3	Membrane	55
	4	Mitochondrion	34
	5	Nucleus	52
	6	Secreted	17
ZW225	1	Cytoplasm	70
	2	Membrane	89
	3	Mitochondrion	25
	4	Nucleus	41

expressed by:

$$PSSM = \begin{bmatrix} N_{1\to 1} & N_{1\to 2} & \cdots & N_{1\to 20} \\ N_{2\to 1} & N_{2\to 2} & \cdots & N_{2\to 20} \\ \vdots & \vdots & \vdots & \vdots \\ N_{i\to 1} & N_{i\to 2} & \cdots & N_{i\to 20} \\ \vdots & \vdots & \vdots & \vdots \\ N_{L\to 1} & N_{L\to 2} & \cdots & N_{L\to 20} \end{bmatrix} \tag{1}$$

where L is the number of the amino acid residues in the protein sequence and $N_{i\to j}$ indicates the relative probability describing how the ith amino acid position in the protein sequence mutates into the j amino acid type during biological evolution processes. After we obtain the PSSM for a given protein, elements $N_{i\to j}$ in PSSM can be normalized by Eq. (2).

$$P_{ij} = \frac{1}{1 + e^{-N_{i\to j}}} \tag{2}$$

The proposed consensus sequence-based transition matrix (CTM) feature extraction method

Unlike many methods that extract features from the protein primary sequences, to integrate the evolutionary information, we attempt to extract features from the consensus sequences. After getting the PSSM of a given protein sequence S, the consensus sequence S^c enriched with evolutionary information can be obtained using the following formula:

$$index_i = \arg\max(P_{i,j} : 1 \leq j \leq 20), 1 \leq i \leq L \tag{3}$$

Through Eq. (3), we calculate the argument of the amino acid type corresponding to the maximum substitution probability in each row of the PSSM. Then, we can replace the ith amino acid residue located in the original protein sequence by the $index_i$th amino acid type to obtain the consensus sequence. Through this process, we transform the protein primary sequence into a consensus sequence and integrate the evolutionary information.

In order to provide information about the distributions of the 20 amino acids transitions in the protein consensus sequences, we propose a feature extraction method based on the transition matrix of the consensus sequence.

For the consensus sequence $S^c = \left\{S_1^c, S_2^c, \cdots, S_L^c\right\}$ of a given protein sequence, we represent it as a directed graph $G = (V, E)$. $V = v_1, v_2, \cdots, v_n$ is the set of vertices corresponding to the 20 types of amino acids, and $E = e_1, e_2, \cdots, e_m$ is the set of edges which model the pairwise relationship between amino acids. Since there are 20 types of amino acids that make up protein sequences, we can obtain 20 × 20 different combinations of amino acid pairs, which means that there are $m = 20 \times 20$ edges appear in graph G. In this paper, we use the occurrence number of the amino acid pairs to describe the pairwise relationship.

In Fig. 4, we take a short consensus sequence as an example to demonstrate the construction of graph G. For a consensus sequence "CWWRCWWWLWWWRWQWWWWPWWCWDCWWWHCWWQ", we only show the edges starting from node W (amino acid W)in graph G. In the consensus sequence, the occurrence of amino acid pairs "WW" is 12, so that the weight of the self-joining edge of node W is 12. And amino acid pairs "WC" occurs once in the sequence, so the edge starting from W to C weights 1.

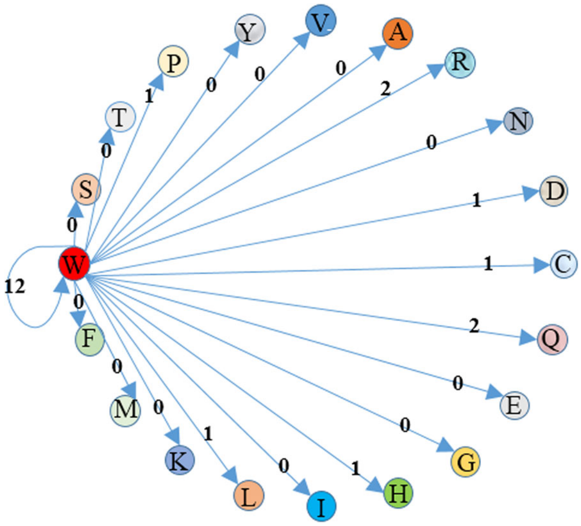

Fig. 4 A sample example of the constructive process of graph *G*

The graph G can be represented as a transition matrix $T(G) = T_{i,j}, (i \leq 20, j \leq 20)$ which is denoted as:

$$T(G) = \begin{bmatrix} T_{1,1} & T_{1,2} & \cdots & T_{1,20} \\ \vdots & \vdots & \vdots & \vdots \\ T_{i,1} & T_{i,2} & \cdots & T_{i,20} \\ \vdots & \vdots & \vdots & \vdots \\ T_{20,1} & T_{20,2} & \cdots & T_{20,20} \end{bmatrix} \tag{4}$$

where $T_{i,j}$ represents the nature of the weighted edges and we specify it as the occurrence number of corresponding amino acid pairs in the consensus sequence. The detail of the feature based on the transition matrix is as follows.

Firstly, we count the number of edges starting from each vertex so that we get the first feature descriptor based on the transition matrix of the consensus sequence. It can be obtained by:

$$CTM_i^1 = \sum_{j=1}^{20} T_{i,j}, (i = 1, \cdots, 20) \tag{5}$$

The normalized value of $T_{i,j}$ can be calculated by $p_{i,j} = \frac{T_{i,j}}{\sum_{r=1}^{20} \sum_{c=1}^{20} T_{r,c}}$. Then by applying the Shannon entropy to the normalized value of $T_{i,j}$, we can obtain the second feature descriptor based on the transition matrix of consensus sequence. It reflects another attribute of each vertex and is obtained as follows:

$$CTM_i^2 = -\sum_{j=1}^{20} p_{i,j} log(p_{i,j}), (i = 1, \cdots, 20) \tag{6}$$

Finally, by using Eqs. (5) and (6), a 40-dimensional feature vector is established based on the transition matrix of the consensus sequence. The consensus sequence-based transition matrix (CTM) feature extraction method gets the distributions of the 20 amino acids transitions, rather than just the amino acid composition, and it also incorporates the evolutionary information from the amino acid sequence. Compared to the typical dipeptide composition, the dimension of the proposed CTM feature is smaller significantly.

The proposed feature extraction method of absolute entropy correlation analysis based on PSSM (AECA-PSSM)

The element P_{ij} in PSSM indicates the related probability of the amino acid in the *i*th position evolves into a particular amino acid type. Therefore, each column in a PSSM can be regarded as a probability distribution. And for a PSSM, there are 20 columns in total, so that we can obtain 20 probability distributions in a PSSM. To further extract protein sequence information from the position-specific scoring matrix (PSSM), the absolute entropy correlation analysis method (AECA-PSSM) is proposed for the expression of proteins. The AECA-PSSM is a method based on the relative entropy method, and it is used to analyze the pairwise relationship between each two columns of PSSM.

Relative entropy [53], also known as Kullback-Leibler divergence (KL divergence or KLD) or information divergence, is an asymmetric method which is used to measure the difference between two probability distributions. So it is desirable to naturally analyze information in PSSM utilizing the relative entropy based methods. The relative entropy (KL divergence) between two different probability distributions can be described as follows:

$$\begin{aligned} D_{KL}(P||Q) &= \sum_{i=1}^{N} P(i) log\left(\frac{1}{Q(i)}\right) - \sum_{i=1}^{N} P(i) log\left(\frac{1}{P(i)}\right) \\ &= \sum_{i=1}^{N} P(i) log\left(\frac{P(i)}{Q(i)}\right) \end{aligned} \tag{7}$$

According to Gibbs inequality, KL divergence is always non-negative. When it equals to 0, it means that the two distributions are the same. And it is obvious that $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, so the KL divergence doesn't absolutely reflect the distance between two variables. If we directly use the KL divergence to analyze the information embedded in PSSM, we need a $20 \times 19 = 380$ dimensional vector because of the asymmetry of KLD. In order to make the relationship between two variables to satisfy the commutative law, the absolute entropy is calculated by:

$$\begin{aligned} D(P, Q) &= \frac{1}{2}(D_{KL}(P||Q) + D_{KL}(P||Q)) \\ &= \frac{1}{2} \sum_{i=1}^{N} (P(i) - Q(i)) log\left(\frac{P(i)}{Q(i)}\right) \end{aligned} \tag{8}$$

The absolute entropy is also always non-negative and zero also represents that the two distributions are the same ones. Through absolute entropy, the difference between two signals can be uniquely determined.

For the PSSM which we have stated to consider as 20 probability distributions, the absolute entropy correlation analysis is employed between each two probability distributions. Finally, for a protein sequence, a $20 \times 19/2 = 190$ dimensional feature vector is established through AECA-PSSM.

As the above, for each protein sequence, it can be described as a 230-dimensional feature vector by fusing the 40-dimensional consensus sequence-based transition matrix (CTM) feature and 190-dimensional absolute entropy correlation analysis (AECA-PSSM) feature.

Linear discriminant analysis for dimensionality reduction of the proposed features

Though more information can be learned by combining multiple features, it also results in more irrelevant and redundant information which imposes a burden on the classifier. Dimensionality reduction is an effective way to resolve this problem. Hence, linear discriminant analysis (LDA) [54], a supervised dimensionality reduction method is employed to reduce the dimension of the proposed features and eliminate the noise.

LDA [55] is one of the most popular dimensionality reduction methods. Given a data set with n protein samples $\{x_i, y_i\}_{i=1}^{n}$, where $x_i \in R^d$ is the feature vector of the protein sample and $y_i \in \{1, 2, \cdots, C\}$ is the corresponding class label of the sample. Let π_c be the subset corresponding to protein samples with label c and contain n_c data points, $\sum_{c=1}^{C} n_c = n$. We write $X = [x_1, x_2, \ldots, x_n]$. The within-class scatter matrix $S^{(\omega)}$ and the between-class scatter matrix $S^{(b)}$ are separately defined as follows.

$$S^{(\omega)} = \sum_{c=1}^{C} \sum_{x_i \in \pi_c} (x_i - m_c)(x_i - m_c)^T \tag{9}$$

$$S^{(b)} = \sum_{c=1}^{C} n_c(m_c - m)(m_c - m)^T \tag{10}$$

where $m_c = \frac{1}{n_c} \sum_{x_i \in \pi_c} x_i$ is the class mean of π_c, and $m = \frac{1}{n} \sum(x_i)$ is the global mean of all samples. The optimization criteria of LDA is to seek a linear transformation that maps the samples in the high dimensional space to a lower dimensional space, such that the between-class scatter is maximized and the within-class scatter is minimized. Therefore, the optimization objective of LDA is as follows:

$$W^* = \arg\max_{W} \left[tr\left(\frac{W^T S^{(b)} W}{W^T S^{(\omega)} W} \right) \right] \tag{11}$$

where W^* is the projection matrix. And the objective can be solved by generalized eigenvalue problem $S^{(b)} W = \lambda S^{(\omega)} W$. And the optimal projection matrix W^* can be constructed by taking the eigenvectors of $(S^{(\omega)})^{-1} S^{(b)}$ consistent with the $p, (p < C)$ largest eigenvalues.

The projection of LDA can be obtained through Eq. (12):

$$Q = (W^*)^T X \tag{12}$$

Through LDA, the original high-dimensional feature is projected into a lower-dimensional space and the complexity of the classifier is decreased.

Support vector machine (SVM)

Support vector machine (SVM) [56] is a well-known supervised algorithm proposed by Vapnik. The core principle of SVM is to find a classification hyperplane to maximize the distance between positive and negative samples. SVM is built on statistical learning theory. More precisely, it is the approximate realization of minimum structural risk. When faced with samples which are linearly inseparable in low-dimensional space, SVM utilizes the kernel function to render them linearly separable in high-dimensional space. In this work, we choose a radial basis function (RBF) to solve a nonlinear problem. The equation of RBF is defined as:

$$K(x, x_i) = exp\left(-\gamma ||x - x_i||^2\right), \gamma > 0 \tag{13}$$

LIBSVM toolbox [57] is used in this work to train the classification model. SVM is originally designed for two-class classification problems, and when it comes to the multi-class classification problem, such as the protein subcellular location, it is necessary for us to build appropriate multi-class classifiers. LIBSVM toolbox uses one-versus-one (OVO) strategy to solve multi-class classification problems. The specific method is to construct an SVM classifier between any two kinds of samples, so that if there are k categories we will get $k(k-1)/2$ SVM classifiers. When categorizing a sample unlabeled, the category that gets the most votes is the final class of the unknown sample.

Model validation and performance evaluation

In our experiment, the jackknife test is used to evaluate the effectiveness of the classifier [4]. Jackknife test can get a unique result and it is deemed to be the most objective and reasonable. For a given dataset, the jackknife test needs to test every sample in the dataset. The principle of the jackknife test is to select one sample from the dataset as an independent test sample, and use the remaining samples for training until all the samples in the dataset have been tested. For example, as for the CL317 dataset which contains 317 apoptosis proteins, each protein sequence will be treated as a test sequence, and the remaining 316 sequences will be used to train the classification model. After all 317 sequences were tested, the result is achieved.

Furthermore, to evaluate the performance of the model more comprehensively, six evaluation metrics including sensitivity (Sen), specificity (Spe), accuracy (Acc), Matthews correlation coefficient (MCC), F-measure (F) and the overall accuracy (OA) are used in this paper, which can be calculated as follows:

$$Recall_i \ or \ Sen_i = \frac{TP_i}{TP_i + FN_i} \tag{14}$$

$$Spe_i = \frac{TN_i}{TN_i + FP_i} \tag{15}$$

$$MCC_i = \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{(TP_i + FP_i)(TP_i + FN_i)(TN_i + FP_i)(TN_i + FN_i)}} \tag{16}$$

$$Acc_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{17}$$

$$Precision_i = \frac{TP_i}{TN_i + FP_i} \tag{18}$$

$$F_i = 2 \times \frac{Recall_i \times Precision_i}{Recall_i + Precision_i} \tag{19}$$

$$OA = \frac{\sum_{i=1}^{c} TP_i}{\sum_{i=1}^{c} (TP_i + FN_i)} \tag{20}$$

where TP_i, FN_i, TN_i, FP_i represent the true positive rate, false negative rate, true negative rate and false positive rate of category i respectively.

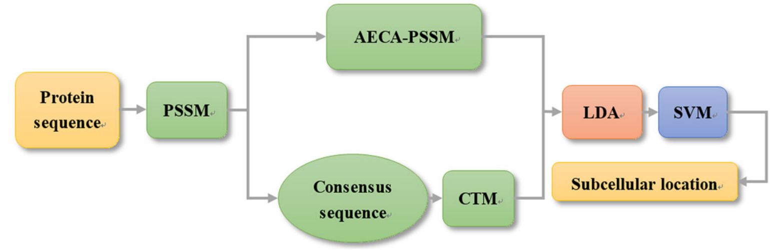

Fig. 5 Pipeline of the proposed CTM-AECA-PSSM-LDA method for predicting apoptosis proteins subcellular location

The detail of the CTM-AECA-PSSM-LDA subcellular location prediction method

The detail of the model which used to predict the subcellular location of apoptosis proteins is as follows. The pipline of this proposed method is shown in Fig. 5. For convenience, the proposed method is called CTM-AECA-PSSM-LDA.

Step 1: Input the protein samples in CL317 dataset and ZW225 dataset, respectively. Using CTM a 40-dimensional feature vector is generated and a 190-dimensional feature vector is extracted by AECA-PSSM. By combining these two different features, a 230-dimensional feature vector is established.

Step 2: Using LDA dimensionality reduction method to reduce the redundancy of the 230-dimensional feature vector.

Step 3: Employing SVM to identify the subcellular locations of apoptosis proteins.

Abbreviations

AECA: Absolute entropy correlation analysis; Acc: Accuracy; CTM: Consensus sequence-based transition matrix; ELM: Extreme learning machine; F: F-measure; KNN: K-nearest neighbors; LDA: Linear discriminant analysis; LR: Logistic regression; MCC: Matthews correlation coefficient; OVO: One-versus-one; OA: Overall accuracy; PSSM: Position-specific scoring matrix; SVM: Support vector machine; Sen: Sensitivity; Spe: Specificity

Acknowledgements

The authors thank the anonymous reviewers for their advisable comments.

Authors' contributions

LD designed and conceived the method of this article, performed the experiments, and wrote the paper. QFM supervised the whole work. QFM, YHC and PW provided some suggestions for the paper writing. All authors read and approved the final version of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 61671220, 61701192, 61640218), the Natural Science Foundation of Shandong Province, China (Grant No. ZR2017QF004).

Availability of data and materials

All the datasets and source code can be downloaded from https://github.com/LEI-SHUN/PSL.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 29 October 2019 Accepted: 6 May 2020
Published online: 24 May 2020

References

1. Yaron F, Hermann S. Programmed cell death in animal development and disease. Cell. 2011;147(4):742–58.
2. Linn F, Charlotte S, Marie S, Martin H, Kalle J, Mikaela W, Annica A, Mathias U, Emma L. Mapping the subcellular protein distribution in three human cell lines. J Proteome Res. 2011;10(8):3766–77.

3. Guo-Sheng H, Zu-Guo Y, Vo A. Predicting the subcellular location of apoptosis proteins based on recurrence quantification analysis and the hilbert-huang transform. Chinese Physics B. 2011;20(10):100504.
4. Zhang Z-H, Wang Z-H, Zhang Z-R, Wang Y-X. A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. FEBS Lett. 2006;580(26): 6169–74.
5. Chen Y-L, Li Q-Z. Prediction of the subcellular location of apoptosis proteins. J Theor Biol. 2007;245(4):775–83.
6. Chen Y-L, Li Q-Z. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. J Theor Biol. 2007;248(2):377–81.
7. Zhang L, Liao B, Li D, Zhu W. A novel representation for apoptosis protein subcellular localization prediction using support vector machine. J Theor Biol. 2009;259(2):361–5.
8. Liu T, Zheng X, Wang C, Wang J. Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation. Protein Pept Lett. 2010;17(10):1263–9.
9. Wang X, Li H, Zhang Q, Wang R. Predicting subcellular localization of apoptosis proteins combining GO features of homologous proteins and distance weighted KNN classifier. BioMed Res Int. 2016;2016:1–8.
10. Chen X, Hu X, Yi W, Zou X, Xue W. Prediction of apoptosis protein subcellular localization with multilayer sparse coding and oversampling approach. BioMed Res Int. 2019;2019(4):1–9.
11. Liang Y, Zhang S. Prediction of apoptosis protein's subcellular localization by fusing two different descriptors based on evolutionary information. Acta Biotheor. 2018;66(1):61–78.
12. Li B, Cai L, Liao B, Fu X, Bing P, Yang J. Prediction of protein subcellular localization based on fusion of multi-view features. Molecules. 2019;24(5):919.
13. Habib T, Zhang C, Yang JY, Yang MQ, Deng Y. Supervised learning method for the prediction of subcellular localization of proteins using amino acid and amino acid pair composition. BMC Genomics. 2008;9(1):1–9.
14. Feng Z. Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers. 2015;58(5):491–9.
15. Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res. 1998;26(9):2230–6.
16. Chou K-C. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun. 2000;278(2):477–83.
17. Chou K-C. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics. 2001;43(3):246–55.
18. Sarda D, Chua GH, Li K-B, Krishnan A. pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. BMC Bioinforma. 2005;6(1):152.
19. Dehzangi A, Sohrabi S, Heffernan R, Sharma A, Lyons J, Paliwal K, Sattar A. Gram-positive and gram-negative subcellular localization using rotation forest and physicochemical-based features. BMC Bioinforma. 2015;16(4):1.
20. Uddin MR, Sharma A, Farid DM, Rahman MM, Dehzangi A, Shatabda S. EvoStruct-Sub: An accurate Gram-positive protein subcellular localization predictor using evolutionary and structural features. J Theor Biol. 2018;443:138–46.
21. Wang X, Zhang J, Li G-Z. Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble. BMC Bioinforma. 2015;16(12):1.
22. Wan S, Mak MW, Kung SY. mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines. BMC Bioinforma. 2012;13(1):290.
23. Yao Y-H, Shi Z-X, Dai Q. Apoptosis protein subcellular location prediction based on position-specific scoring matrix. J Comput Theor Nanosci. 2014;11(10):2073–8.
24. Liang Y, Liu S, Zhang S. Detrended cross-correlation coefficient: Application to predict apoptosis protein subcellular localization. Math Biosci. 2016;282:61–7.
25. Huang Y, Li Y. Prediction of protein subcellular locations using fuzzy *k*-NN method. Bioinformatics. 2004;20(1):21–8.
26. Chou K-C, Shen H-B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res. 2006;5(8):1888–97.
27. Cai Y-D, Liu X-J, Chou K-C. Artificial neural network model for predicting protein subcellular location. Comput Chem. 2002;26(2):179–82.
28. Xie D, Li A, Wang M, Fan Z, Feng H. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. Nucleic Acids Res. 2005;33(suppl_2):105–10.
29. Huang C, Yuan J. Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites. Biosystems. 2013;113(1):50–7.
30. Dehzangi A, Heffernan R, Sharma A, Lyons J, Paliwal K, Sattar A. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. J Theor Biol. 2015;364:284–94.
31. Wan S, Mak M-W, Kung S-Y. Transductive learning for multi-label protein subchloroplast localization prediction. IEEE/ACM Trans Comput Biol & Bioinforma. 2016;14(1):212–24.
32. Liang Y, Liu S, Zhang S. Geary autocorrelation and DCCA coefficient: Application to predict apoptosis protein subcellular localization via PSSM. Physica A: Stat Mech & Appl. 2016;467:296–306.
33. Xiang Q, Liao B, Li X, Xu H, Chen J, Shi Z, Dai Q, Yao Y. Subcellular localization prediction of apoptosis proteins based on evolutionary information and support vector machine. Artif Intell Med. 2017;78:41–6.
34. Wang S, Li W, Fei Y, Cao Z, Xu D, Guo H. An improved process for generating uniform PSSMS and its application in protein subcellular localization via various global dimension reduction techniques. IEEE Access. 2019;7:42384–95.
35. Zhang S, Liang Y. Predicting apoptosis protein subcellular localization by integrating auto-cross correlation and PSSM into Chou's PseAAC. J Theor Biol. 2018;457:163–9.
36. Wan S, Mak M-W, Kung S-Y. Ensemble linear neighborhood propagation for predicting subchloroplast localization of multi-location proteins. J Proteome Res. 2016;15(12):4755–62.
37. Qu X, Wang D, Chen Y, Qiao S, Zhao Q. Predicting the subcellular localization of proteins with multiple sites based on multiple features fusion. IEEE/ACM Trans Comput Biol & Bioinforma. 2015;13(1):36–42.
38. Javed F, Hayat M. Predicting subcellular localization of multi-label proteins by incorporating the sequence features into Chou's PseAAC. Genomics. 2019;111(6):1325–32.

39. Wei L, Liao M, Gao X, Wang J, Lin W. mGOF-loc: A novel ensemble learning method for human protein subcellular localization prediction. Neurocomputing. 2016;217:73–82.
40. Chen J, Xu H, He P.-a., Dai Q, Yao Y. A multiple information fusion method for predicting subcellular locations of two different types of bacterial protein simultaneously. BioSystems. 2016;139:37–45.
41. Wang S, Liu S. Protein sub-nuclear localization based on effective fusion representations and dimension reduction algorithm LDA. Int J Mol Sci. 2015;16(12):30343–61.
42. You Z-H, Lei Y-K, Zhu L, Xia J, Wang B. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. In: BMC Bioinformatics; 2013. p. 10, BioMed Central.
43. Wan S, Mak M-W, Kung S-Y. mPLR-Loc: An adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction. Anal Biochem. 2015;473:14–27.
44. Zhang S, Duan X. Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC. J Theor Biol. 2017;437:239.
45. Ruan X, Zhou D, Nie R, Hou R, Cao Z. Prediction of apoptosis protein subcellular location based on position-specific scoring matrix and isometric mapping algorithm. Medical & Biological Engineering & Computing. 2019;57(12):2553–65.
46. Wang X, Zhang W, Zhang Q, Li G-Z. MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier. Bioinformatics. 2015;31(16):2639–45.
47. Wan S, Mak M-W, Kung S-Y. FUEL-mLoc: feature-unified prediction and explanation of multi-localization of cellular proteins in multiple organisms. Bioinformatics. 2017;33(5):749–50.
48. Wan S, Mak M-W, Kung S-Y. Gram-locEN: Interpretable prediction of subcellular multi-localization of Gram-positive and gram-negative bacterial proteins. Chemometr Intell Lab Syst. 2017;162:1–9.
49. Chou K-C, et al. The pLoc_bal-mGneg predictor is a powerful web-server for identifying the subcellular localization of gram-negative bacterial proteins based on their sequences information alone. Int J Sci. 2020;9(01):27–34.
50. Xiao X, Cheng X, Chen G, Mao Q, Chou K-C. pLoc_bal-mVirus: predict subcellular localization of multi-label virus proteins by Chou's general PseAAC and IHTS treatment to balance training dataset. Med Chem. 2019;15(5):496–509.
51. Zhang S, Zhang T, Liu C. Prediction of apoptosis protein subcellular localization via heterogeneous features and hierarchical extreme learning machine. SAR QSAR Environ Res. 2019;30(3):209–28.
52. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
53. Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat. 1951;22(1):79–86.
54. Tong W, Jie Y. Predicting subcellular localization of gram-negative bacterial proteins by linear dimensionality reduction method. Protein & Peptide Letters. 2010;17(1):32–7.
55. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936;7(2):179–88.
56. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
57. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Trans on Intell Syst Technol (TIST). 2011;2(3):27.

Publisher's Note