Yu et al. BMC Genomics (2018) 19:478
https://doi.org/10.1186/s12864-018-4849-9

BMC Genomics

METHODOLOGY ARTICLE Open Access

Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction

Bin Yu[1,2,3*†], Shan Li[1,2†], Wenying Qiu[1,2], Minghui Wang[1,2], Junwei Du[4], Yusen Zhang[5] and Xing Chen[6]

Abstract

Background: Apoptosis is associated with some human diseases, including cancer, autoimmune disease, neurodegenerative disease and ischemic damage, etc. Apoptosis proteins subcellular localization information is very important for understanding the mechanism of programmed cell death and the development of drugs. Therefore, the prediction of subcellular localization of apoptosis protein is still a challenging task.

Results: In this paper, we propose a novel method for predicting apoptosis protein subcellular localization, called PsePSSM-DCCA-LFDA. Firstly, the protein sequences are extracted by combining pseudo-position specific scoring matrix (PsePSSM) and detrended cross-correlation analysis coefficient (DCCA coefficient), then the extracted feature information is reduced dimensionality by LFDA (local Fisher discriminant analysis). Finally, the optimal feature vectors are input to the SVM classifier to predict subcellular location of the apoptosis proteins. The overall prediction accuracy of 99.7, 99.6 and 100% are achieved respectively on the three benchmark datasets by the most rigorous jackknife test, which is better than other state-of-the-art methods.

Conclusion: The experimental results indicate that our method can significantly improve the prediction accuracy of subcellular localization of apoptosis proteins, which is quite high to be able to become a promising tool for further proteomics studies. The source code and all datasets are available at https://github.com/QUST-BSBRC/PsePSSM-DCCA-LFDA/.

Keywords: Apoptosis proteins, Subcellular localization, Pseudo-position specific scoring matrix, Detrended cross-correlation analysis coefficient, Local fisher discriminant analysis, Support vector machine

Background

Protein maintains a highly ordered operation of the protection of the cell system [1]. At the cellular level, proteins work only in specific locations. It is necessary to fulfill the protein's function that subcellular locations provide a specific chemical environment and set of interaction partners [2]. Apoptosis is cell physiological death which is closely related to intracellular control [3]. Cancer and autoimmune disease occurs when blocking apoptotic protein appears, ischemic damage or neurodegenerative disease occurs when unwanted apoptosis appears [4]. Studying proteins involved in the apoptotic process can help us understand the pathogenesis of the disease and provide a variety of therapeutic targets. It is very valuable to get information on apoptosis protein subcellular localization, which can help us understand the apoptosis proteins function, cell apoptosis mechanisms and drug development [5]. Therefore, it is a challenging task using the machine learning method to construct the protein subcellular location prediction model.

For nearly two decades, the research of protein subcellular localization prediction has been a hotspot in

* Correspondence: yubin@qust.edu.cn
†Bin Yu and Shan Li contributed equally to this work.
[1]College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China
[2]Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao 266061, China
Full list of author information is available at the end of the article

bioinformatics research. Currently, this topic has achieved some successes [6]. As for protein subcellular localization prediction methods, we found that the research focuses on the following two aspects: (1) Feature extraction of protein sequences. The main methods include N-terminal information prediction method [7, 8], protein amino acid composition prediction method [9, 10], pseudo-amino acid composition [11, 12], dipeptide composition [13, 14], grouped weight coding method [15], discrete wavelet transform [16, 17], position specific scoring matrix [18] and protein GO annotation [19, 20] and so on. (2) The choice of classification algorithm. At present, the commonly used prediction algorithms are increment of diversity [21], deep learning [22], K-nearest neighbor [23, 24], neural network [25], hidden Markov model [26, 27], Bayesian classifier [28, 29], support vector machine [30, 31] and ensemble learning [32–35]. Among them, SVM has fast computing speed, excellent ability to extract information, good generalization performance advantages, making the SVM become the first classifier choice for researchers.

Currently, apoptosis protein subcellular localization prediction has made great advancement. Zhou and Doctor [36] constructed 98 protein apoptosis protein dataset, using the amino acid composition and covariance discriminant method, the overall prediction accuracy reached 72.5% by jackknife test. Huang et al. [37] obtained the accuracy rate of 77.6% by combining the protein instability index with the support vector machine. However, the prediction capacity of this method was unbalanced. Especially, for other class proteins (exclude cytoplasmic, membrane and mitochondrial proteins), the prediction accuracy did not exceed 50%. Bulashevska and Eils [38] used Bayesian classifier based on Markov chain model to construct ensemble classifier, and the prediction accuracy of 98 apoptosis proteins was further improved by jackknife test. Zhang et al. [15] constructed a new apoptosis protein dataset of 225 proteins. They used encoding approach with grouped weight as feature extraction method for protein sequences and support vector machine as classifier (named as EBGW_SVM). The overall prediction accuracy was 83.1% using jackknife test. The feature extraction method of the protein sequence takes into account the distribution of residues with the same unique characteristic, but ignores the physical and chemical properties of the protein sequence. Chen and Li [39] constructed a dataset containing 317 apoptosis protein sequences and obtained higher prediction accuracy, which combined support vector machine and increment of diversity (named as ID_SVM) by using jackknife test. Similarly, Ding et al. [40] used the Fuzzy K-nearest neighbor (FKNN) algorithm and the overall prediction accuracy was 90.9% using CL317 dataset. Qiu et al. [41] used the DWT_SVM method to obtain high prediction accuracy rates of 97.5, 87.6 and 88.8% for CL317, ZW225 and ZD98 datasets, respectively by jackknife test. The above methods ignore the biological information of the protein sequence, so the prediction method of homologous similarity based on the protein sequence and protein functional domain is proposed. Yu et al. [42] proposed a novel pseudo-amino acid model which extracted the sequence characteristics of proteins using amino acid substitution matrices and auto covariance transformation and used support vector machine as classifier. The results of prediction accuracy obtained by jackknife test were 90.0 and 87.1% on the CL317 and ZW225 datasets, respectively. Liu et al. [43] used tri-gram encoding based PSSM as feature extraction method, then used SVM-RFE algorithm to reduce feature vectors, finally the best feature vectors were input to the SVM classifier. The prediction accuracy were 95.9, 97.8 and 96.9% on the CL317, ZW225 and ZD98 datasets, respectively. Dai et al. [44] treated the difference between the N-segment and C-segment of the protein in subcellular location prediction, and proposed a model based on golden ratio segmentation to improve subcellular localization prediction, and achieved a better predictive effect. Xiang et al. [45] introduced evolutionary-conservative information to represent protein sequences. Meanwhile, according to the proportion of golden section in mathematics, the position-specific scoring matrix (PSSM) is divided into several blocks. The overall accuracy of ZD98 and CL317 datasets were 98.98 and 91.11%, respectively by using SVM classifier. Liang et al. [46] combined the Geary autocorrelation function and detrended cross-correlation coefficient methods based on PSSM to extract the protein sequences from the CL317, ZW225 and ZD98 datasets. Under the jackknife test, the overall prediction accuracy were 89.0 84.4 and 91.8%, respectively.

Using only a feature is difficult to have a big breakthrough in the prediction of subcellular localization. At present, researchers usually combined multiple feature extraction methods of protein sequences to obtain more comprehensive protein sequence information. However, the feature vectors of the protein sequences obtained by fusing a variety of features are usually very high. High-dimensional data contains a lot of redundant information, which may seriously affect the performance of the classifier. Dimensionality reduction methods can help us eliminate redundant information and are widely used in data classification and pattern recognition. At present, many researchers introduce a variety of methods to reduce dimension in the subcellular localization prediction, such as SVD (singular value decomposition) [47], Backward feature selection [48], CFS (correlation-based feature selection) [49], Forward selection [50], PSO (particle swarm optimization) [51], mLASSO (multi-label least absolute shrinkage and selection operator) [52], GA (Genetic algorithm) [53] and so on.

In this paper, we presents a new method for predicting subcellular localization of apoptosis proteins, called PsePSSM-DCCA-LFDA. Firstly, obtain sequence information from apoptotsis protein sequences by combining PsePSSM algorithm and DCCA coefficient. Then, the LFDA method is used to reduce the dimension and noise information in the original high-dimensional space. Finally, using SVM as classifier to predict protein subcellular localization. By jackknife test, the optimal parameters of the model are determined under different ξ values, S values, different dimensionality reduction methods and selection of different dimensions, and established PsePSSM-DCCA-LFDA prediction model. Using the most rigorous jackknife test, the overall prediction accuracy are 99.7, 99.6 and 100%, respectively for CL317 dataset, ZW225 dataset and ZD98 dataset. The results show that the PsePSSM-DCCA-LFDA method can get better prediction effect than other existing methods.

Methods

Datasets

In this study, we use CL317 dataset and ZW225 dataset as the training datasets, which select optimal parameters of the prediction model. In addition, the ZD98 dataset is selected as an independent testing dataset that used to test the applicability of the prediction model. The CL317 dataset was constructed by Chen and Li [39], which contained 317 proteins classified into six compartments as cytoplasm proteins, mitochondrion proteins, nucleus proteins, membrane proteins, secreted proteins and endoplasmic reticulum proteins, each class containing 112, 34, 52, 55, 17, and 47 proteins, respectively. For CL317 dataset, distribution of sequence identity percentage are 40.1, 15.5, 18.9 and 25.6% with ≤40%, 41%-80%, 81%-90% and ≥91% sequence identity, respectively. The ZW225 dataset was constructed by Zhang et al. [15], which contained 225 proteins classified into four compartments as nuclear proteins, cytoplasmic proteins, mitochondrial proteins and membrane proteins, each class containing 41, 70, 25 and 89 proteins, respectively. For ZW225 dataset, distribution of sequence identity percentage are 52.9, 16, 16 and 15.1% with ≤40%, 41%-80%, 81%-90% and ≥91% sequence identity, respectively. The ZD98 dataset was constructed by Zhou and Doctor [36], which contained 98 proteins classified into four compartments as cytoplasmic proteins, plasma membrane-bound proteins, mitochondrial proteins and other proteins, each class containing 43, 30, 13 and 12 proteins, respectively. For ZD98 dataset, distribution of sequence identity percentage are 34.7, 30.6, 17.4 and 17.4% with ≤40%, 41%-80%, 81%-90% and ≥91% sequence identity, respectively. The protein sequences are extracted from SWISS-PROT database (version 49.5) in the three datasets, and we can find the accession numbers in the literatures [15, 36, 39].

Pseudo-position specific scoring matrix (PsePSSM)

In order to obtain the evolutionary information of the protein sequences, the protein sequences of the CL317, ZW225 and ZD98 datasets are aligned with the non-redundant (NR) database (ftp://ftp.ncbi.nih.gov/blast/db/) using the PSI-BLAST program [54], and obtain the position specific scoring matrix (PSSM) [55] of the corresponding protein sequences. The NR database contains 85,107,862 protein sequences. We use three iterations and E-value is 0.001 in PSI-BLAST program. The BLOSUM62 matrix is used as substitution matrix for generating the PSSM. PSSM can be expressed for a protein sequence P as the following Eq. (1).

$$P_{PSSM} = \begin{pmatrix} E_{1,1} & E_{1,2} & \cdots & E_{1,20} \\ \vdots & \vdots & \vdots & \vdots \\ E_{i,1} & E_{i,2} & \cdots & E_{i,20} \\ \vdots & \vdots & \vdots & \vdots \\ E_{L,1,} & E_{L,2} & \cdots & E_{L,20} \end{pmatrix} \tag{1}$$

where L is total number of amino acids in the protein sequence, $E_{i,\ j}$ represents the evolution information of amino acids in protein sequences. The rows of PSSM represent the corresponding amino acids positions in protein sequences, and columns of PSSM indicate the 20 amino acid types that may be mutated. The PSSM value ranges from − 9 to 11.

Since the length of the protein sequence in the CL317, ZW225 and ZD98 datasets is inconsistent, the corresponding PSSM dimension for the protein sequence in the dataset is different, which is difficult for our subsequent study. In this paper, PsePSSM [56] algorithm is used to extract the features of protein sequences, and the PSSM of different protein sequences is transformed into a uniform vector.

First, the elements of PSSM are normalized by Eq. (2), whose PSSM value ranges from 0 to 1.

$$f(x) = 1/(1 + e^{-x}) \tag{2}$$

where x is the original PSSM value.

Then, a protein sequence can be expressed using PsePSSM as follows:

$$P_{PsePSSM} = \left(\overline{P_1}, \overline{P_2}, \cdots, \overline{P_{20}}, \theta_1^1, \theta_2^1, \cdots, \theta_{20}^1, \cdots, \theta_1^\xi, \theta_2^\xi, \cdots, \theta_{20}^\xi\right)^T \tag{3}$$

where $\overline{P_j} = \sum_{i=1}^{L} P_{i,j}/L \quad (j = 1, 2, \cdots, 20)$, $\overline{P}_j$ represents the average score of the all amino acid residues which are mutated to j amino acid type in the protein P. $\theta_j^\xi = \frac{1}{L-\xi} \sum_{i=1}^{L-\xi} \left(P_{i,j} - P_{(i+\xi),j}\right)^2 \quad (j = 1, 2, \cdots 20;\ \xi < L,\ \xi \neq 0)$, θ_j^ξ is order information of protein sequences, j is amino acid type, ξ is contiguous distance.

From the above, a protein sequence generates 20+20×ξ dimension feature vector using PsePSSM algorithm.

Detrended cross-correlation analysis coefficient

According to the evolutionary information expressed by the protein sequence, we can obtain the corresponding position score-specific matrix (PSSM), as shown in eq. (1). In order to extract more protein sequence information from the PSSM matrix, the protein sequence information is extracted from the PSSM using the detrended cross-correlation analysis coefficient (DCCA coefficient) method [57–59]. DCCA coefficient is a method based on the trend covariance method, and the least squares linear fitting and trend elimination are carried out for nonstationary signals. The evolutionary information expressed in the form of PSSM is used as the attribute, and each amino acid is considered as one property. PSSM is considered to be the time series of all attributes. Since the size of the PSSM matrix for each protein sequence is L×20, we calculate the 20 columns in the PSSM matrix as 20 non-stationary time series [46, 60].

After normalizing the PSSM matrix using the eq. (2), for any two different columns $\{m_i\}$ and $\{n_1\}$ of PSSM (i=1,2,···,L), L is the length of protein sequence. First we use the Eq. (4) to calculate the new time series M_k and N_k.

$$\begin{cases} M_k = \sum_{i=1}^{k} m_i & k = 1, 2, \cdots, L \\ N_k = \sum_{i=1}^{k} n_i & k = 1, 2, \cdots, L \end{cases} \tag{4}$$

Then the time series M_k and N_k are divided into L-S segments which can be overlapped, each segment contains S+1 data, and then the least squares linearly fitting for each segment of the data to obtain the fitting values $\tilde{M}_{i,k}$ and $\tilde{N}_{i,k}$. Use the Eq. (5) to calculate the covariance of each segment.

$$f_{xy}^2(S,i) = \frac{1}{S+1}\sum_{k=i}^{i+S}\left(M_k - \tilde{M}_{i,k}\right)\left(N_k - \tilde{N}_{i,k}\right) \tag{5}$$

In particular, there are $f_{xx}^2(S,i) = \frac{1}{S+1}\sum_{k=i}^{i+S}\left(M_k - \tilde{M}_{i,k}\right)^2$,$f_{yy}^2(S,i) = \frac{1}{S+1}\sum_{k=i}^{i+S}\left(N_k - \tilde{N}_{i,k}\right)^2$.

Next, the covariance of the L-S segments (whole time series) calculated by using the Eq. (6) is:

$$f_{xy}^2(S) = \frac{1}{L-S}\sum_{i=1}^{L-S} f_{xy}^2(S,i) \tag{6}$$

In particular, there are $f_{xx}^2(S) = \frac{1}{L-S}\sum_{i=1}^{L-S} f_{xx}^2(S,i)$, $f_{yy}^2(S) = \frac{1}{L-S}\sum_{i=1}^{L-S} f_{yy}^2(S,i)$.

Finally, the DCCA coefficients of two different time series $\{m_i\}$ and $\{n_1\}$ are calculated using Eq. (7).

$$\rho_{DCCA} = \frac{f_{xy}^2(S)}{f_{xx}(S) f_{yy}(S)} \tag{7}$$

As can be seen from Eq. (7), ρ_{DCCA} depends on the length L of the protein sequence and the length S+1 of the overlapping portion of each segment. Its value ranges from -1≤ ρ_{DCCA} ≤1, where 1 represents perfect cross-correlation, 0 indicates no cross-correlation, and – 1 represents perfect anti-cross-correlation [61]. Finally, the DCCA coefficient algorithm will generate a 190-dimensional feature vector for a protein sequence.

Local fisher discriminant analysis

This paper uses a supervised dimensionality reduction method, local Fisher discriminant analysis (LFDA) [62]. LFDA has the form of embedded transformation, and it can be easily calculated by solving the generalized eigenvalue problem. Let the protein data matrix be $X = [x_1, x_2, \cdots x_n], x_i \in R^d$, where n is the number of samples of the protein, d is the dimension of the protein sequence feature extraction. $y_i \in \{1, 2\cdots, c\}$, n_ℓ is the number of samples of the categoryℓ, $\sum_{\ell=1}^{c} n_\ell = n$. The local within-class scatter matrix $S^{(w)}$and the local between-class scatter matrix $S^{(b)}$are calculated using Eqs. (8) and (9).

$$S^{(w)} = \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j}^{(w)}\left(x_i - x_j\right)\left(x_i - x_j\right)^T \tag{8}$$

$$S^{(b)} = \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j}^{(b)}\left(x_i - x_j\right)\left(x_i - x_j\right)^T \tag{9}$$

where

$$W_{i,j}^{(w)} = \begin{cases} A_{i,j}/n_\ell & if\ \ y_i = y_j = \ell \\ 0 & if\ \ y_i \neq y_j \end{cases}$$

$$W_{i,j}^{(b)} = \begin{cases} A_{i,j}\left((1/n) - (1/n_\ell)\right) & if\ \ y_i = y_j = \ell \\ 1/n & if\ \ y_i \neq y_j \end{cases}$$

It is worth noting that A is an affinity matrix, $A_{i,\,j} \in A$ is the affinity betweenx_i and x_j. In this paper, we use the affinity matrix$A_{i,\,j} = \exp(-\|x_i - x_j\|/\sigma_i\sigma_j)$ defined by Zelnik-Manor and Perona [63]. $\sigma_i = \|x_i - x_i^{(K)}\|$ represents the local scaling of the surrounding x_i data samples, where $x_i^{(K)}$ is the K nearest neighbor of x_i. The literature [63] proved that in the experiment for high-dimensional data, when K = 7, better results can be obtained, so this article selected K = 7.

Solve LFDA transformation matrixT_{LFDA}

$$T_{LFDA} = \underset{T \in R^{d \times r}}{\arg\max} \left[tr\left(\left(T^T S^{(w)} T \right)^{-1} T^T S^{(b)} T \right) \right] \tag{10}$$

Matrix after the dimension reduction becomes:

$$Z = T'_{LFDA} X \tag{11}$$

Therefore, through the Eq. (11), we eliminate the redundant information contained in the high-dimensional data obtained after the original protein sequence feature extraction. In other words, the fusion PsePSSM algorithm and DCCA coefficient algorithm on the apoptosis protein sequence after the feature extraction matrix $\underline{X}$, through the transformation matrix T_{LFDA}, matrix Z is obtained after dimensionality reduction.

Support vector machine

Support vector machine (SVM) is a supervised machine learning method based on statistical learning theory, which is proposed by Vapnik et al. [64]. Because of its excellent learning and generalization ability, especially the ability to deal with high dimensional sparse vector, it has become a hotspot in the field of data mining and machine learning. In recent years, SVM has also been widely used in the field of bioinformatics. In the field of proteomics research, it has been widely used to predict membrane protein types [65, 66], G protein-coupled receptors [67, 68], protein structure [69–73], protein-protein interaction [74–76], protein subcellular localization [77–80], protein post-translational modification sites [81–84] and other protein structure and function of the study.

SVM is used to solve a two-class classification problem. Set$D = \{(x_i, y_i) | \, i = 1, 2, \cdots, n\}$ is a training set, where$x_i \in R^d$represent sample i, which has d dimension feature vectors, $y_i \in \{+1, -1\}$is class labels of sample i. SVM transforms a linearly indivisible sample of low-dimensional input space into high-dimensional feature space to make it linearly separable.

In this study, we choose the radial basis function (RBF) to perform prediction. Because RBF kernel function is the most widely used kernel function and its superiority for solving nonlinear problem [17, 18, 41–46], which is defined as follows:

$$K\left(x_i, y_j\right) = \exp\left(-\gamma \left\| x_i - y_j \right\|^2 \right) \tag{12}$$

where γ is the kernel width parameter, x_i and y_j are the feature vectors of the i-th and j-th protein sequences, respectively. The egularization parameter C and the kernel parameter γ are optimized based on CL317 and ZW225 datasets by K-fold cross validation using a grid search strategy to obtain the highest overall prediction accuracy by using the LIBSVM software [85], which can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/. In this paper, C is allowed to take a value only between 2^{-5} and 2^{15}, and γ only between $2^{-15}$and 2^{5}.

SVM is originally designed for two-class classification, but CL317, ZW225 and ZD98 are multi-class classification data. At present, three kinds of strategies can be solved multi-classification: one-versus-one (OVO), one-versus-rest (OVR) [86] and direct acyclic graph SVM (DAGSVM) [87]. LIBSVM software implements the "one-versu-one" (OVO) strategy for multi-class classification. The OVO strategy sets up a classifier between any two categories,so if k is the number of classes, then $k(k-1)/2$ classifiers are constructed. During the testing phase, the test samples are submitted to all classifiers, $k(k-1)/2$ classification results are obtained, and the final result is generated by voting. That is to say, the most voting category is the final class. It is worth noting that when there are two categories of voting the same results, we choose the class appearing first of the vote as the final category for the sake of simple operation.

Performance evaluation and model building

In statistical prediction, there are four validation tests: self-consistency test, independent dataset test, k-fold cross-validation and jackknife test, which are often used to evaluate the prediction performance [78, 80]. In this paper, the jackknife test [88, 89] is used to examine the performance of the prediction model. The jackknife test requires testing each sample in the dataset. Specifically, each time one sample is selected as an independent test sample in the dataset, and the remaining samples are used as a training set to establish a prediction model until all the samples have been tested in the dataset.

We use four standard performance measures to evaluate the model performance, including sensitivity (Sens), specificity (Spec), Matthews correlation coefficient (MCC) and overall accuracy (OA), as follows:

$$Sens = \frac{TP}{TP + FN} \tag{13}$$

$$Spec = \frac{TN}{TN + FP} \tag{14}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{15}$$

$$OA = \frac{TP + TN}{TP + FN + FP + TN} \tag{16}$$

where TP represents the numbers of the correctly identified positives, TN represents the numbers of correctly identified negatives, FP represents the numbers of the negatives identified as positives, FN represents the numbers of the positives identified as negatives. In addition,

to assess the generalization performance of the model, the receiver operating characteristic (ROC) curves were used. The AUC is the area calculated under ROC curve plotted by FP rate vs TP rate, which is a quantitative indicator of the robustness of the model. Its values range from 0 to 1.

For convenience, the method is called PsePSSM-DCCA-LFDA in this paper, which is used to predict apoptosis protein subcellular localization. To provide an intuitive picture, the flowchart of PsePSSM-DCCA-LFDA method is shown in Fig. 1. We have implemented it in MATLAB R2014a.

The PsePSSM-DCCA-LFDA prediction model is detailed below:

1) Input CL317 dataset and ZW225 dataset, respectively, which contain apoptosis protein sequences and the class label corresponding to all kinds of proteins;
2) The $20 + 20 \times \xi$dimension feature vector is generated by PsePSSM algorithm. Using DCCA coefficient, the protein sequence is extracted to generate 190 dimension feature vectors. By combining these two methods, the two different apoptosis protein datasets generate the corresponding feature extraction matrices of$X = 317 \times (190 + (20 + 20 \times \xi))$and $X = 225 \times (190 + (20 + 20 \times \xi))$, respectively;
3) Using the LFDA method to solve $T_{LFDA} = \arg\max_{T \in R^{d \times r}} [tr((T^T S^{(w)} T)^{-1} T^T S^{(b)} T)]$, the numerical matrix X extracted in 2) is reduced dimension by $Z = T'_{LFDA} X$, and the matrix Z is obtained by removing the redundant information in the apoptosis protein sequences;
4) The matrix Z after dimensionality reduction are input into the SVM classifier, and the protein subcellular localization prediction is performed by jackknife test;
5) According to the accuracy of prediction, the optimal parameters of the model are selected, including the ξ values and$S$values of parameters, the selection of the dimension reduction algorithm and the dimensionality;
6) Calculate the Sens, Spec, MCC, OA and AUC values of the model, and evaluate prediction performance of the model;
7) Using the independent testing dataset ZD98 to test the PsePSSM-DCCA-LFDA prediction model.

Results

Selection of optimal parameter ξ and$S$

In this study, the apoptosis protein sequences are extracted by the fusion PsePSSM algorithm and the DCCA coefficient algorithm, and obtain the feature information in the protein sequences. It is worth noting that both the PsePSSM algorithm and the DCCA coefficient algorithm can control the validity of the algorithm to extract the feature information of the protein sequence by adjusting some of the parameters in the algorithm. How to get the best parameters of these two feature extraction algorithms is very important for us to construct a protein subcellular localization prediction model. In order to discover the merits of the feature parameters, we use CL317 and ZW225 datasets as the research object, the best parameters of the model are selected by the prediction accuracy under different parameters. In this paper, the PsePSSM algorithm is used to carry out feature extraction on protein sequences, and the ξ value indicates the sequence-order information of the amino acid residues in the protein sequence. If the ξ value is set too large, the feature vector dimension of the protein sequence is too high, resulting in more redundant information, which affects the prediction effect. If the ξ value is set too small, the feature vector contains very little sequence information, and the features of the protein sequence of the apoptosis protein dataset cannot be extracted comprehensively. To find the optimal

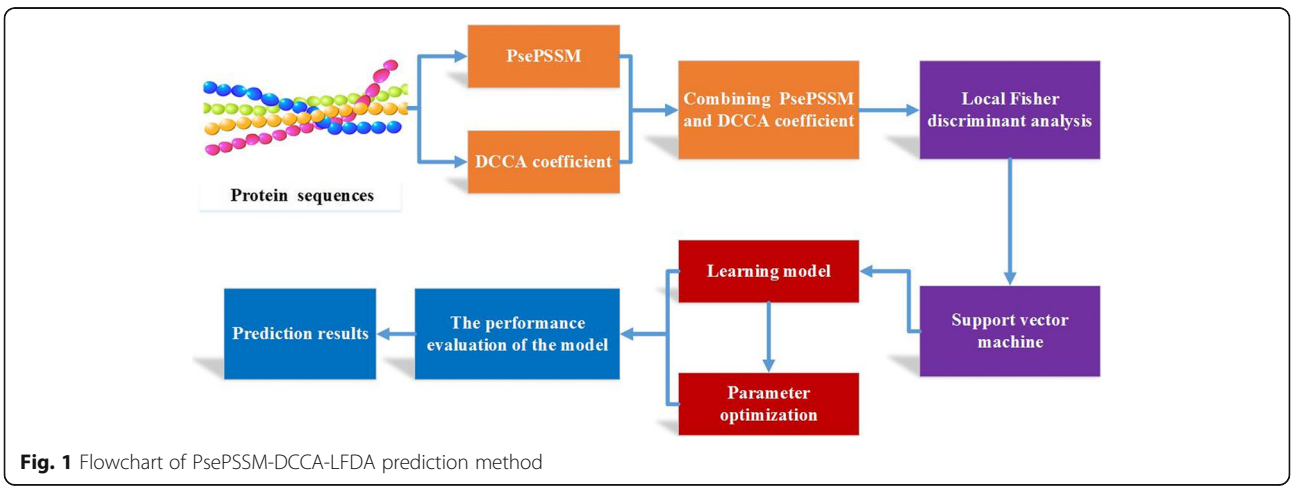

Fig. 1 Flowchart of PsePSSM-DCCA-LFDA prediction method

ξ value in the model, set the ξ values from 0 to 10 in turn. For the different ξ values, the apoptosis protein datasets CL317 and ZW225 are classified by SVM respectively. The SVM is used to select the radial basis function (RBF) and the results are tested by jackknife method. The overall prediction accuracy of each class protein and overall prediction accuracy in the apoptosis protein datasets are obtained, as are shown in Tables 1 and 2.

Table 1 shows that the OA of CL317 dataset are different with constant change of ξ value. The highest prediction accuracy of mitochondrial proteins reach 70.6% when $\xi = 3$, which is 32.4 and 14.7% higher than when ξ values are 0 and 1, respectively. The prediction accuracy of membrane proteins is 87.3% when $\xi = 10$. The OA of CL317 dataset reach 83.6 and 83.9% when $\xi = 3$ and $\xi = 10$, respectively, higher than that when ξ values are taking other values.

Table 2 shows that the OA of ZW225 dataset are different with constant change of ξ value. For cytoplasmic proteins, the highest predictive accuracy is 81.4% when ξ values are 1, 3, 4 and 6, respectively. For membrane proteins, the highest prediction accuracy is 93.3% when $\xi = 0$, which is 7.9% higher than when $\xi = 5$. From the overall prediction accuracy, when ξ values are 3 and 6, the OA of ZW225 dataset reach 77.3 and 77.8%, respectively, which is 6.6 and 7.1% higher than that when $\xi = 0$.

To select the optimal parameters of the PsePSSM algorithm in the subcellular prediction model of apoptosis proteins, CL317 and ZW225 datasets are selected as the training datasets. Fig. 2 shows the OA changes when different ξ values are chosen in CL317 and ZW225 datasets. It can be seen from Fig. 2 that the prediction accuracy of the two datasets is changing with the change of the ξ value. In addition, CL317 and ZW225 datasets reach the highest accuracy, when$\xi = 3$ and$\xi = 6$, respectively. But in order to unify the model parameters, $\xi = 3$ is chosen in the model. Therefore, the PsePSSM algorithm is used to extract the protein sequence, and each protein sequence to obtain $20 + 20 \times \xi = 20 + 20 \times 3 = 80$ dimension feature vector.

In the feature extraction process by using DCCA coefficient, the selection of *S*value has a crucial influence on the construction of the model. *S*is used to determine the length of each overlapping portion of the detrended cross-correlation analysis. Because the length of the shortest protein sequence in the benchmark dataset is 50, the maximum value allowed for S is 49. To find the optimal *S*value in the model, set *S*values from 5 to 49 in turn. For the different *S*values, the apoptosis protein datasets CL317 and ZW225 are classified by SVM respectively. The SVM is used to select RBF and the results are tested by jackknife method. The prediction accuracy of each class protein and the overall prediction accuracy in the apoptosis protein datasets are obtained, as are shown in Tables 3 and 4.

Table 3 shows that the OA of CL317 dataset are different with constant change of S value. The accuracy of cytoplasmic proteins reach 97.3% when $S = 30$ and $S = 35$, respectively. The highest prediction accuracy of mitochondrial proteins reach 92.7%, when $S = 45$ and $S = 49$, respectively, which is 12.7% higher than when$S = 5$. The accuracy of nuclear proteins is 67.3% when $S = 49$, which is 21.1% higher than when$S = 5$. The accuracy of secreted proteins is 88.2% when $S = 25$. From overall prediction accuracy, the OA of CL317 dataset is 85.8% when $S = 35$, which is 9.8% higher than when $S = 5$.

Table 4 shows that the OA of ZW225 dataset are different with constant change of *S*value. The accuracy of cytoplasmic proteins reach 87.1% when$S = 49$. The highest prediction accuracy of membrane proteins reach 91.0% when S values are 20, 25 and 40, respectively. The accuracy of nuclear proteins reach 75.6% when $S = 40$ and $S = 45$, respectively. From the overall prediction accuracy, the OA of ZW225 dataset is 82.7% when $S = 40$, which is higher than other parameters.

In our current study, two apoptosis protein datasets CL317 and ZW225 are selected as the training datasets. To determine the optimal parameters of DCCA coefficient algorithm in the model, Fig. 3 shows the change of OA in CL317 dataset and ZW225 dataset by choosing

Table 1 Prediction results of selecting different ξ on CL317 by jackknife test

Locations	ξ										
	Jackknife test (%)										
	0	1	2	3	4	5	6	7	8	9	10
Cy	88.4	90.2	90.2	92.9	92.0	92.9	92.9	92.9	92.9	92.9	92.9
Me	76.4	83.6	81.8	83.6	83.6	81.8	83.6	83.6	87.3	87.3	87.3
Mi	38.2	55.9	58.8	70.6	64.7	64.7	64.7	64.7	64.7	64.7	67.7
Se	47.1	76.5	76.5	76.5	76.5	76.5	76.5	76.5	76.5	76.5	76.5
Nu	51.9	57.7	63.5	73.1	76.9	73.1	73.1	71.2	71.2	71.2	73.1
En	85.1	85.1	85.1	85.1	85.1	85.1	85.1	85.1	85.1	85.1	85.1
OA	72.2	78.5	79.5	83.6	83.3	82.6	83.0	82.6	83.3	83.3	83.9

Table 2 Prediction results of selecting different ξ on ZW225 by jackknife test

Locations	ξ										
	Jackknife test (%)										
	0	1	2	3	4	5	6	7	8	9	10
Cy	74.3	81.4	80.0	81.4	81.4	80.0	81.4	80.0	80.0	80.0	80.0
Me	93.3	91.0	87.6	88.8	86.5	85.4	85.4	85.4	85.4	85.4	85.4
Mi	16.0	32.0	24.0	40.0	44.0	44.0	44.0	40.0	36.0	36.0	36.0
Nu	48.8	63.4	61.0	68.3	68.3	75.6	75.6	75.6	75.6	75.6	75.6
OA	70.7	76.4	73.3	77.3	76.9	77.3	77.8	76.9	76.4	76.4	76.4

different$S$. It can be seen from Fig. 3 that$S$values are different for the highest overall prediction accuracy of two datasets. The average overall prediction accuracy of CL317 and ZW225 datasets is the highest when $S = 40$. That is, the DCCA coefficient algorithm chooses optimal parameter$S = 40$. At this time, the 190 dimension feature vector can be obtained by extracting each protein sequence by DCCA coefficient method.

Selection of dimensionality reduction method and optimal dimension

The increasing dimension of the dataset makes the classification more difficult and the development to a certain extent can cause curse of dimensionality. For high-dimensional data, firstly, dimensionality reduction is carried out, and then data after dimensionality reduction is input into the learning system. In order to achieve the ideal protein subcellular localization prediction accuracy, the PsePSSM and DCCA coefficients are first fused to extract features of the protein sequences. In the discussion of section 3.1, in the PsePSSM algorithm, select $\xi = 3$. In the DCCA coefficient algorithm, select $S = 40$. At this time, each protein sequence in the dataset generates a $(20 + 20 \times 3) + 190 = 80 + 190 = 270$ dimension feature vector.

Then, PCA (Principal Component Analysis) [90], Laplacian Eigenmaps [91], AKPCA (Adaptive Kernel Principal Component Analysis) [92] and LFDA (Local Fisher Discriminant Analysis) dimensionality reduction method are used to compare the effect of protein subcellular localization overall prediction accuracy by using these four dimensionality reduction methods. In this study, we use the SVM to classify with the radial basis kernel function, and the results are tested by jackknife method. The overall prediction accuracy of subcellular localization of two apoptosis protein datasets are obtained with different dimensionality reduction methods and under different dimensions, as shown in Tables 5 and 6.

As can be seen from Table 5, for the CL317 dataset, choosing different dimensionality reduction methods and dimensions have a significant effect on the accuracy

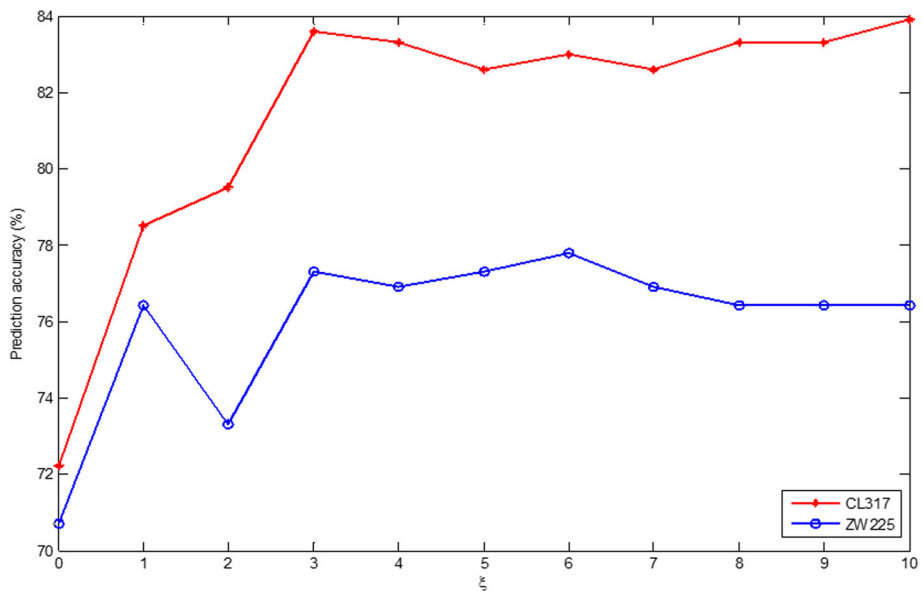

Fig. 2 Effect of selecting different values of ξ on CL317 and ZW225 datasets by jackknife test

Table 3 Prediction results of selecting different *S* on CL317 by jackknife test

Locations	*S*									
	Jackknife test (%)									
	5	10	15	20	25	30	35	40	45	49
Cy	96.4	90.2	92.0	90.2	95.5	97.3	97.3	95.5	95.5	96.4
Me	83.0	83.0	83.0	83.0	83.0	83.0	83.0	83.0	83.0	83.0
Mi	80.0	87.3	83.6	87.3	83.6	87.3	89.1	90.9	92.7	92.7
Se	44.1	52.9	82.4	79.4	88.2	85.3	85.3	82.4	82.4	82.4
Nu	46.2	63.5	63.5	65.4	61.5	61.5	63.5	65.4	65.4	67.3
En	64.7	52.9	52.9	52.9	70.6	70.6	76.5	76.5	64.7	58.8
OA	76.0	78.2	81.4	81.4	83.9	84.9	85.8	85.5	85.2	85.5

of protein subcellular prediction. When Laplacian Eigenmaps and AKPCA method are used to reduce dimension, the dimension is 50, and CL317 dataset obtains the highest overall prediction accuracy, which is 84.9 and 89.6%, respectively. When PCA method is used to reduce dimension and dimensionality chooses 40, the highest overall prediction accuracy of the CL317 dataset is 89.3%. When LFDA method is used to reduce dimension and dimensionality chooses 10, the overall prediction accuracy is the highest, which is 99.7%. It shows that when choosing different dimensionality reduction methods, getting the best dimension for the CL317 dataset is different. By comparing the overall prediction accuracy by different dimensionality reduction methods with different dimensions, we can find that when the LFDA dimensionality reduction method is adopted, the dimensionality is 10, the overall prediction accuracy is the highest, 10.1% higher than when AKPCA dimensionality reduction method is used and the dimension is 50. It can be more intuitively found in Fig. 4 for the CL317 dataset, when the LFDA dimensionality reduction method is selected, the highest overall prediction accuracy of the model is achieved when dimension is 10.

As can be seen from Table 6, for the ZW225 dataset, choosing different dimensionality reduction methods and dimensions has a significant effect on the accuracy of protein subcellular prediction. When PCA method is selected to reduce the dimensionality and dimension chooses 30, the highest overall prediction accuracy of 85.8% is achieved in the ZW225 dataset. When using the Laplacian Eigenmaps method to reduce dimension, and dimensionality chooses 30 or 50, the overall prediction accuracy of the dataset is the highest, which is 82.2%. When using the AKPCA method to reduce dimension, and dimensionality chooses 40, the highest overall prediction accuracy is 86.2%. When using the LFDA method to reduce dimension, and dimensionality chooses 10, 20, 30, 40, 50, 60, 70 or 80, the highest overall prediction accuracy is 99.6%. It indicates that the choice of the optimal dimension is closely related to the use of dimensionality reduction methods. In this paper, by comparing the overall prediction accuracy by different dimensionality reduction methods with different dimensions, it can be found that when the LFDA dimensionality reduction method is adopted and the dimension is 10, the overall prediction accuracy is the highest, 17.4% higher than when Laplacian Eigenmaps dimensionality reduction method is used and dimension is 30. It can be more intuitively found in Fig. 5 for the ZW225 dataset, when the LFDA dimensionality reduction method is selected, the highest overall prediction accuracy of the model when dimension is 10, 20, 30, 40, 50, 60, 70 or 80. Since the two apoptosis protein datasets CL317 and ZW225 are selected as the training set, in order to unify the parameters of the model, the LFDA dimensionality reduction method is adopted in this paper, and the optimal dimension is 10-dimensional.

Table 4 Prediction results of selecting different *S* on ZW225 by jackknife test

Locations	*S*									
	Jackknife test (%)									
	5	10	15	20	25	30	35	40	45	49
Cy	85.7	81.4	80.0	85.7	84.3	85.7	84.3	85.7	85.7	87.1
Me	84.3	86.5	86.5	91.0	91.0	89.9	89.9	91.0	89.9	89.9
Mi	28.0	36.0	48.0	52.0	56.0	56.0	56.0	56.0	56.0	56.0
Nu	48.8	68.3	70.7	61.0	58.5	58.5	70.7	75.6	75.6	73.2
OA	72.0	76.0	77.3	79.6	79.1	79.1	80.9	82.7	82.2	82.2

Effect of feature extraction algorithm on results

Feature extraction method converts character representation of a protein sequence into a numerical representation, which uses the corresponding feature vector to represent protein sequence information. PsePSSM method can get homology and sequence information of amino acids in the protein sequences. DCCA coefficient method is an extension of the DCCA and the DFA (detrended fluctuation analysis). Here, only the evolutionary represented in the form of PSSM is adopted as the considered properties. The PsePSSM algorithm and the DCCA coefficient method are combined to obtain more protein sequence information, but this will obtain high-dimensional features to make the model worse, which contain more redundant variables. LDFA dimensionality reduction method use local within-class scatter matrix and local between-class scatter matrix to remove the redundant information based on the feature information of the protein sequences in the dataset and the corresponding class labels. In this paper, the optimal feature extraction algorithm is selected by comparing the influence of different feature extraction methods on the prediction results. Two different predicted results of the two apoptosis protein datasets CL317 and ZW225 are shown in Tables 7 and 8. Furthermore, we

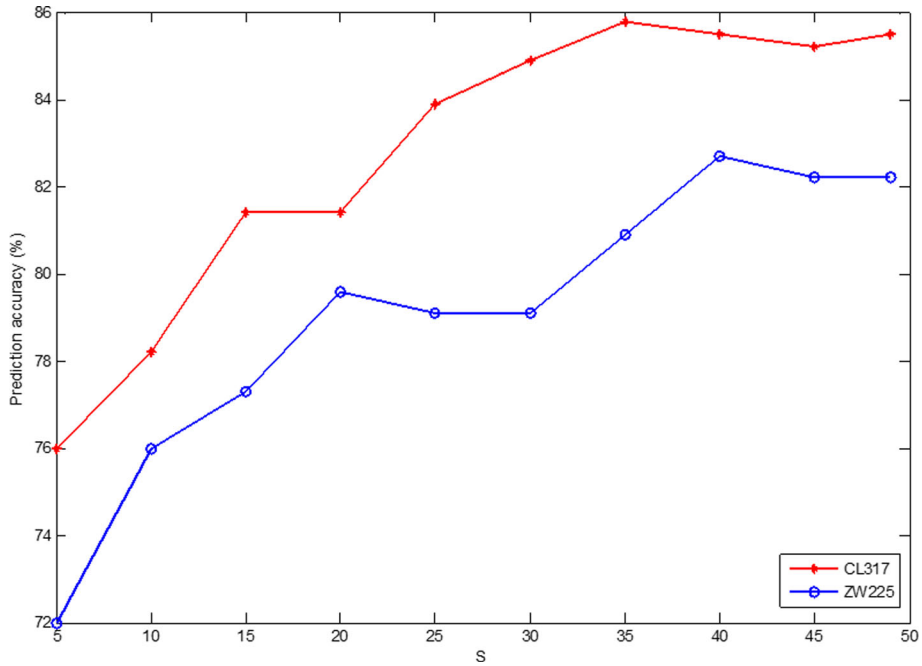

Fig. 3 Effect of selecting different values of *S* on CL317 and ZW225 datasets by jackknife test

analyze the robustness of the model under different feature extraction algorithms, which use ROC curve. As we know, the ROC curve is used in positive vs negative (two classes) classification. But apoptosis proteins subcellular localization prediction is a multi-class prediction problem. We first use the one-versus-rest (OVR) strategy to transform the multi-classification problem into two-classification problems. One of the classes is selected as positive samples i.e. "positive" one and other classes as negative samples [69]. Then for these two-classification true positive rate and false positive rate, the average of them was taken as the final result [43]. Figures 6 and 7 are the ROC curves obtained by four different feature extraction methods for the CL317 dataset and ZW225 dataset, respectively.

Table 7 shows that the OA of CL317 dataset are different, which use different feature extraction algorithms. The OA of PsePSSM algorithm reach 83.6%, which is 1.9% lower than DCCA coefficient algorithm. The OA of PsePSSM-DCCA algorithm is 86.8%, which is 3.2, 1.3% higher than PsePSSM and DCCA coefficient algorithm, respectively. The LFDA algorithm is used to reduce the dimensionality after two algorithms. The accuracy of each class has been obviously improved by using LFDA algorithm and OA of CL317 dataset reach 99.7%. For PsePSSM algorithm, the accuracy of secreted proteins reach 70.6%, which is lower than DCCA coefficient, PsePSSM-DCCA and PsePSSM-DCCA-LFDA algorithm, respectively. The accuracy of secreted proteins is 100% by PsePSSM-DCCA-LFDA algorithm, which is 29.4% higher than the PsePSSM method. Fig. 6 shows that PsePSSM-DCCA-LFDA reach largest coverage area of the ROC curve, whose AUC value is 0.9842. In addition, the AUC values of PsePSSM, DCCA coefficient and PsePSSM-DCCA are 0.9591, 0.9520 and 0.9587, respectively.

Table 5 Prediction results of subcellular localization of the CL317 dataset by selecting different dimensionality reduction methods and different dimensions

Algorithms	Dimensions									
	Jackknife test (%)									
	10	20	30	40	50	60	70	80	90	100
PCA	79.5	84.9	89.0	89.3	87.7	86.1	85.8	83.9	82.6	80.4
Laplacian	63.4	73.8	79.2	82.6	84.9	84.5	82.3	80.8	78.5	74.4
AKPCA	78.2	84.5	87.7	89.0	89.6	88.6	87.4	88.3	86.8	86.4
LFDA	99.7	98.7	98.7	98.4	98.4	97.8	97.5	97.5	97.2	97.2

Table 6 Prediction results of subcellular localization of the ZW225 dataset by selecting different dimensionality reduction methods and different dimensions

Algorithms	Dimensions									
	Jackknife test (%)									
	10	20	30	40	50	60	70	80	90	100
PCA	74.2	79.1	85.8	84.9	83.6	81.3	78.7	78.2	77.8	74.7
Laplacian	69.3	80.0	82.2	80.9	82.2	77.8	73.8	72.0	68.0	67.1
AKPCA	74.7	82.2	84.0	86.2	84.9	83.1	80.9	80.9	77.3	78.2
LFDA	99.6	99.6	99.6	99.6	99.6	99.6	99.6	99.6	99.1	99.1

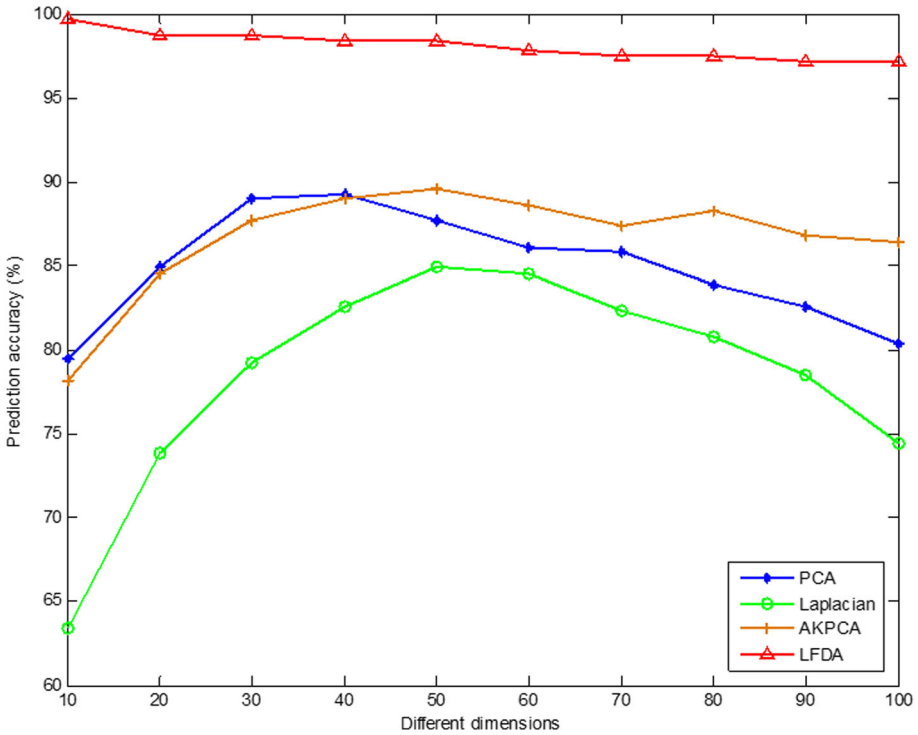

Fig. 4 Effects of selecting four different dimensionality reduction methods and different dimensions on the overall prediction results of subcellular localization in CL317 dataset

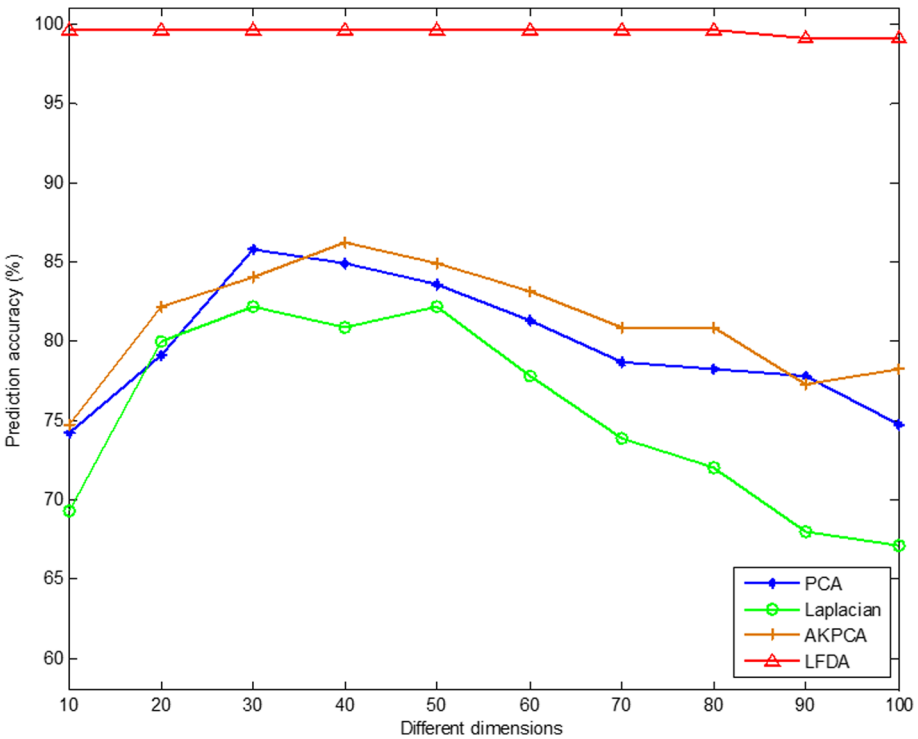

Fig. 5 Effects of selecting four different dimensionality reduction methods and different dimensions on the overall prediction results of subcellular localization in ZW225 dataset

Table 7 Prediction results of different feature extraction methods on CL317 by jackknife test

Algorithms	Locations						
	Jackknife test (%)						
	Cy	Me	Mi	Se	Nu	En	OA
PsePSSM	92.9	85.1	83.6	70.6	73.1	76.5	83.6
DCCA	95.5	83.0	90.9	82.4	65.4	76.5	85.5
PsePSSM-DCCA	93.8	85.1	90.9	82.4	76.9	70.6	86.8
PsePSSM-DCCA-LFDA	99.1	100	100	100	100	100	99.7

Table 8 shows that the OA, accuracy of each class are different for ZW225 dataset with different feature extraction algorithms. The OA of PsePSSM algorithm reach 77.3%, which is 5.4, 7.6 and 22.3% lower than DCCA coefficient algorithm, PsePSSM-DCCA algorithm and PsePSSM-DCCA-LFDA algorithm, respectively. The OA of PsePSSM-DCCA algorithm is 84.9%, which is 7.6, 2.2% higher than PsePSSM and DCCA coefficient algorithm, respectively. Using the LFDA algorithm to reduce the dimensionality, the PsePSSM-DCCA algorithm as feature extraction method, the prediction accuracy of the four kinds of proteins in the ZW225 dataset has been improved remarkably, and the OA of the model has reach 99.6%. Fig. 7 shows that PsePSSM-DCCA-LFDA reach largest coverage area of the ROC curve, whose AUC value is 0.9805. In addition, the AUC values of PsePSSM, DCCA coefficient and PsePSSM-DCCA are 0.9380, 0.9386 and 0.9464, respectively. Analyzing and comparing the prediction results and robustness of prediction model on CL317 and ZW225 datasets by using four different feature extraction methods, we choose PsePSSM-DCCA-LFDA as feature extraction method in this paper.

Performance of prediction model

In PsePSSM-DCCA-LFDA prediction model, protein sequence information is extracted by fusing the PsePSSM and DCCA coefficient methods, and then the subcellular localization of apoptosis protein datasets is predicted by SVM based on LFDA dimensionality reduction method. According to the above analysis, when using PsePSSM, $\xi = 3$ is selected, when using DCCA coefficient, $S = 40$is selected. Using the LFDA method to reduce the dimension of the dataset, the optimal dimension chooses 10. The RBF is selected as the kernel function of SVM. In this paper, the most rigorous jackknife test methods are used to test the datasets CL317 and ZW225, the main results are shown in Table 9.

Table 8 Prediction results of different feature extraction methods on ZW225 by jackknife test

Algorithms	Locations				
	Jackknife test (%)				
	Cy	Me	Mi	Nu	OA
PsePSSM	81.4	88.8	40.0	68.3	77.3
DCCA	85.7	91.0	56.0	75.6	82.7
PsePSSM-DCCA	88.6	88.8	56.0	87.8	84.9
PsePSSM-DCCA-LFDA	100	98.9	100	100	99.6

As can be seen from Table 9, the OA of CL317 dataset is 99.7% by using jackknife test. The sensitivity of each class is 100% except cytoplasmic proteins. The sensitivity of cytoplasmic proteins is 99.1%. The specificity of each class is 100% except mitochondrial proteins. The OA of ZW225 dataset is 99.6% by using jackknife test. The sensitivity, specificity and MCC of mitochondrial and nuclear proteins are 100, 100% and 1, respectively. The sensitivity of cytoplasmic proteins is 100%, the specificity and MCC are 99.4% and 0.99, respectively.

Comparison with other methods

In this section, to demonstrate the effectiveness of the proposed method PsePSSM-DCCA-LFDA, we compared with other recently reported prediction methods on the same apoptosis proteins datasets. All the methods are performed using jackknife cross-validation test. Tables 10 and 11 details the comparison of the proposed method and other prediction methods on the CL317 and ZW225 datasets, respectively.

As can be seen from Table 10, the OA of CL317 dataset is 99.7% by using PsePSSM-DCCA-LFDA, which is 2.2–17% higher than other prediction methods. We can find that the overall accuracy by our method is higher than that of ID [93], ID_SVM [39], DF_SVM [21], FKNN [40] and so on. The value of sensitivity for each protein class is listed. For example, the sensitivity of mitochondrial proteins, nuclear proteins, secreted proteins, endoplasmic proteins and membrane proteins eached 100% by our method, while the ID [93] are 85.3, 82.7, 88.2, 83.0 and 81.8%, respectively. For the cytoplasmic proteins, the sensitivity of our method is 99.1%, which is also the highest, which is 12.7% higher than that of the Auto_Cova [42] method. For the CL317 dataset, our proposed method has achieved satisfactory prediction results.

As can be seen from Table 11, the OA of ZW225 dataset is 99.6% using PsePSSM-DCCA-LFDA, which is almost 16.5, 15.6, 13.8 and 12.5% higher than EBGW_SVM [15], DF_SVM [21], FKNN [40], Auto_Cova [42], respectively. Especially for the most difficult case-mitochondrial proteins, the predictive accuracy has improved to 100% by our method, which is 40% higher than that of the EBGW_SVM [15], 36% higher than the prediction accuracy of DF_SVM [21]. It indicates that the model of this paper has excellent properties for the prediction of mitochondrial proteins in apoptosis proteins. In general, for the ZW225 dataset, our proposed method has achieved satisfactory prediction results.

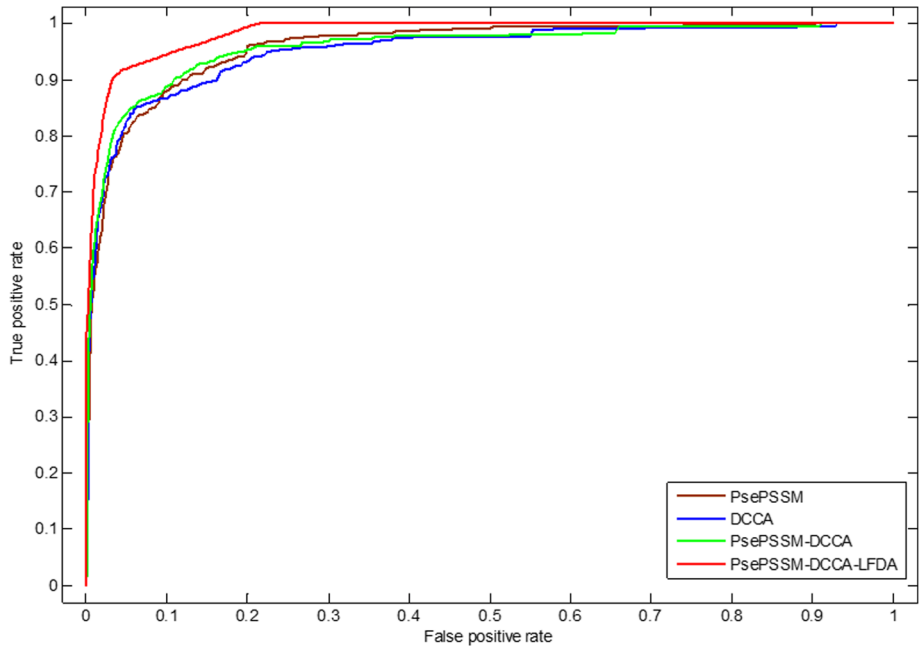

Fig. 6 The ROC curves of four different feature extraction methods on dataset CL317

In order to further validate the actual prediction ability of the model, we use the independent testing dataset ZD98 to test the model. When using PsePSSM, $\xi = 3$ is selected. When using DCCA coefficient, $S = 40$ is selected. Using LFDA method to reduce the dimension of the dataset, the optimal dimension chooses 10. The RBF is selected as the kernel function of SVM. The results of the ZD98 dataset are tested by the jackknife cross-validation method and compared with other reported prediction methods. Table 12 shows the predictive results of the subcellular localization on the ZD98 dataset.

As can be seen from Table 12, the OA of ZD98 dataset is 3.1–11.2% higher than other methods by using PsePSSM-DCCA-LFDA, which is 9.2% higher than ID

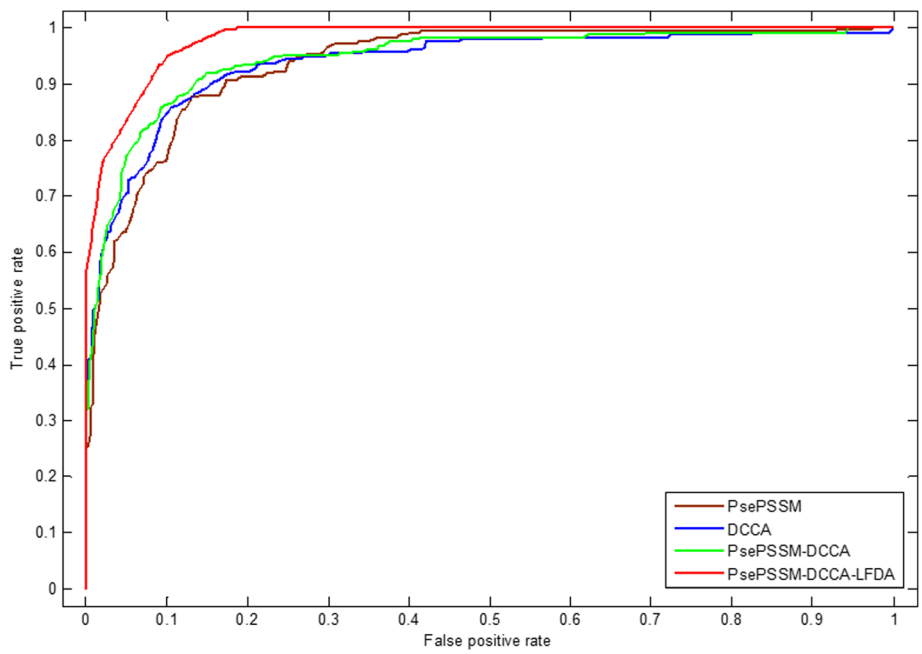

Fig. 7 The ROC curves of four different feature extraction methods on dataset ZW225

Table 9 Prediction performance of datasets CL317 and ZW225 protein subcellular localization on the jackknife test method

Locations	Jackknife test					
	CL317			ZW225		
	Sens (%)	Spec (%)	MCC	Sens (%)	Spec (%)	MCC
Cy	99.1	100	0.99	100	99.4	0.99
Me	100	100	1	98.9	100	0.99
Mi	100	99.6	0.99	100	100	1
Se	100	100	1	–	–	–
Nu	100	100	1	100	100	1
En	100	100	1	–	–	–
OA (%)	99.7			99.6		

[93], 11.2% higher than ID_SVM [39] and DWT_SVM [41]. In addition, the sensitivity of mitochondrial proteins, cytoplasmic proteins and membrane proteins reached 100% by our method, while the ID_SVM [39] are 84.6, 95.3 and 93.3%, respectively. Especially for mitochondrial proteins, the prediction accuracy of DWT_SVM [41] is 53.9%, which is 46.1% lower than that of our method. It shows that our method has achieved good results of the mitochondrial proteins prediction. For the accuracy of the other proteins by the algorithm proposed in this paper is 100%, which is 41.7% higher than the ID_SVM method [39]. In conclusion, the above results indicate that the prediction model we construct can significantly improve the prediction accuracy of protein subcellular localization and has achieved satisfactory prediction results.

Table 10 Prediction results of different methods on CL317 dataset by jackknife test

Methods	Jackknife test (%)						
	Sensitivity for each class (%)						OA (%)
	Cy	Me	Mi	Se	Nu	En	
ID [93]	81.3	81.8	85.3	88.2	82.7	83.0	82.7
ID_SVM [39]	91.1	89.1	79.4	58.8	73.1	87.2	84.2
DF_SVM [21]	92.9	85.5	76.5	76.5	93.6	86.5	88.0
FKNN [40]	93.8	92.7	82.4	76.5	90.4	93.6	90.9
PseAAC_SVM [95]	93.8	90.9	85.3	76.5	90.4	95.7	91.1
EN_FKNN [96]	98.2	83.6	79.4	82.4	90.4	97.9	91.5
DWT_SVM [41]	100	98.2	82.4	94.1	100	100	97.5
Auto_Cova [42]	86.4	90.7	93.8	85.7	92.1	93.8	90.0
APSLAP [97]	99.1	89.1	85.3	88.2	84.3	95.8	92.4
Liu et al. [43]	98.2	96.4	94.1	82.4	96.2	95.7	95.9
PSSM_AC [98]	93.8	90.9	91.2	82.4	86.5	95.7	91.5
DCCA coefficient [46]	91.1	92.7	82.4	76.5	80.8	93.6	88.3
PsePSSM-DCCA-LFDA	99.1	100	100	100	100	100	99.7

Table 11 Prediction results of different methods on ZW225 dataset by jackknife test

Methods	Jackknife test (%)				
	Sensitivity for each class (%)				OA (%)
	Cy	Me	Mi	Nu	
EBGW_SVM [15]	90.0	93.3	60.0	63.4	83.1
ID_SVM [39]	92.9	91.0	68.0	73.2	85.8
DF_SVM [21]	87.1	92.1	64.0	73.2	84.0
FKNN [40]	84.3	93.3	72	85.5	85.8
EN_FKNN [96]	94.3	94.4	60.0	80.5	88.0
DWT_SVM [41]	87.1	93.2	64	90.2	87.6
Liu et al. [43]	97.1	98.9	96.0	97.6	97.8
PSSM_AC [98]	82.9	92.1	68.0	78.0	84.0
Auto_Cova [42]	81.3	93.3	85.7	84.6	87.1
PsePSSM-DCCA-LFDA	100	98.9	100	100	99.6

Discussion

In this paper, we propose a novel method for predicting apoptosis protein subcellular localization, called PsePSSM-DCCA-LFDA. When using PsePSSM, $\xi = 3$ is selected. When using DCCA coefficient, $S = 40$ is selected. Using LFDA method to reduce the dimension of the dataset, the optimal dimension chooses 10. The RBF is selected as the kernel function of SVM. The overall prediction accuracy are 99.7, 99.6 and 100% for CL317 dataset, ZW225 dataset and ZD98 dataset by the most rigorous jackknife test, respectively, which is better than other state-of-the-art methods. The OA of CL317 dataset is 99.7% by using PsePSSM-DCCA-LFDA, which is 2.2–17% higher than other prediction methods. The OA of ZW225 dataset is 99.6% by using

Table 12 Prediction results of different methods on the independent testing dataset ZD98 by jackknife test

Methods	Jackknife test (%)				
	Sensitivity for each class (%)				OA (%)
	Cy	Me	Mi	Other	
ID [93]	90.7	90.0	92.3	91.7	90.8
ID_SVM [39]	95.3	93.3	84.6	58.3	88.8
DF_SVM [21]	97.7	96.7	92.3	75.0	93.9
FKNN [40]	95.3	96.7	100	91.7	95.9
PseAAC_SVM [95]	95.3	93.3	92.3	83.3	92.9
DWT_SVM [41]	95.4	93.3	53.9	91.7	88.8
APSLAP [97]	95.3	90.0	100	91.7	94.9
Liu et al. [43]	95.3	100	100	91.7	96.9
PSSM_AC [98]	97.7	96.7	100	83.3	95.9
EBGW_SVM [15]	97.7	90.0	92.3	83.3	92.9
DCCA coefficient [46]	93.0	86.7	92.3	75.0	88.9
PsePSSM-DCCA-LFDA	100	100	100	100	100

PsePSSM-DCCA-LFDA, which is 1.8–16.5% higher than other prediction methods. The OA of ZD98 dataset is 3.1–11.2% higher than other methods by using PsePSSM-DCCA-LFDA.

PsePSSM-DCCA-LFDA demonstrated good performance on predicting apoptosis protein subcellular localization, which is better than the state-of-the-art methods. It is mainly due to the following reasons:

1. Both the PsePSSM algorithm and the DCCA coefficient method extract feature information from the PSSM corresponding to the protein sequences. Although both algorithms are data mining the evolutionary information of protein sequences in order to obtain the best numerical representation of the protein sequences, the two algorithms are different. PsePSSM feature extraction takes into account the sequence-order information of the protein sequence. The DCCA coefficient uses the columns in the PSSM as the least squares fitting and the trend elimination as the non-stationary time series to remove the PSSM between the cross-correlation.
2. LFDA can effectively remove redundant information in the protein sequences without losing important information in the apoptosis protein sequence.
3. SVM classification algorithm can deal with high-dimensional data, avoiding over-fitting and effectively removing non-support vector.

Protein subcellular localization information can explain the disease mechanism, provide theoretical basis and solution. Medical studies have found that abnormal subcellular localization of proteins occurs, when cells is cytopathic. Further, abnormally localized proteins provide molecular markers for the early diagnosis of diseases and can become molecular targets for the design of new drugs, which achieve the goal of curing diseases.

Currently our method is still trained on small dataset, because CL317, ZW225 and ZD98 datasets are widely used benchmark datasets, it is difficult to collect large-scale experimentally verified. In the next step, we will build a large-scale protein subcellular dataset for prediction research.

Conclusion

With the advent of the big data age, the gap between the number of proteins in the public database and its functional annotations is widening. The critical challenge of bioinformatics is to develop automated methods for fast and accurately determining the structures and functions of proteins [94]. In this paper, a novel method for protein subcellular localization prediction is proposed. We use the LFDA dimensionality reduction method and the SVM algorithm to predict the apoptotic protein subcellular localization. Firstly, we fuse the PsePSSM and DCCA coefficient methods to carry out feature extraction on protein sequences. Then, the extracted feature vectors are used to reduce the dimension using LFDA method, and the subcellular localization of apoptosis proteins are predicted by SVM algorithm. By jackknife test, the OA of the three benchmark datasets reach 99.7, 99.6 and 100%, respectively. The results show that the PsePSSM-DCCA-LFDA method has good performance by comparing with others, which use the same benchmark datasets. Since user-friendly and publicly accessible web-server is one of the important factors in building a practical predictive system [78, 88], in order for the convenience of the researchers, we will develop a web-server or standalone version for the prediction method presented in this paper.

Abbreviations

AKPCA: Adaptive Kernel Principal Component Analysis; CFS: Correlation-based feature selection; Cy: Cytoplasmic proteins; DAGSVM: Direct acyclic graph SVM; DCCA coefficient: Detrended cross-correlation analysis coefficient; DFA: Detrended fluctuation analysis; En: Endoplasmic reticulum proteins; FN: The number of false negatives; FP: The number of false positives; GA: Genetic algorithm; LFDA: Local Fisher discriminant analysis; MCC: Matthews correlation coefficient; Me: Membrane proteins; Mi: Mitochondrial proteins; mLASSO: Multi-label least absolute shrinkage and selection operator; Nu: Nuclear proteins; OA: Overall accuracy; OVO: One-versus-one; OVR: One-versus-rest; PCA: Principal Component Analysis; PsePSSM: Pseudo-position specific scoring matrix; PsePSSM-DCCA-LFDA: Incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction; PSO: Particle swarm optimization; ROC: Receiver operating characteristic; Se: Secreted proteins; Sens: Sensitivity; Spec: Specificity; SVD: Singular value decomposition; SVM: Support vector machine; TN: The number of true negatives; TP: The number of true positives

Acknowledgements

The authors sincerely thank the two anonymous reviewers for their valuable comments.

Funding

This work was supported by the National Natural Science Foundation of China (Nos. 21601103, 51372125, 11701309 and 51572136), the Natural Science Foundation of Shandong Province of China (No. ZR2018MC007), the Project of Shandong Province Higher Educational Science and Technology Program (No. J17KA159), and the Key Laboratory Open Foundation of Shandong Province.

Availability of data and materials

All the benchmark datasets and source code can be downloaded from GitHub (https://github.com/QUST-BSBRC/PsePSSM-DCCA-LFDA/).

Authors' contributions

BY and SL conceived the algorithm, prepared the datasets, carried out experiments, and wrote the manuscript. WYQ, MHW designed, performed and analyzed experiments. JWD, YSZ and XC provided some suggestions for the manuscript writing and revising it critically for important intellectual content. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China. [2]Artificial Intelligence and Biomedical

Big Data Research Center, Qingdao University of Science and Technology, Qingdao 266061, China. [3]School of Life Sciences, University of Science and Technology of China, Hefei 230027, China. [4]College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China. [5]School of Mathematics and Statistics, Shandong University at Weihai, Weihai 264209, China. [6]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 21116, China.

Received: 15 December 2017 Accepted: 1 June 2018
Published online: 19 June 2018

References

1. Eisenhaber F, Bork P. Wanted: subcellular localization of proteins based on sequence. Trends Cell Biology. 1998;8:169–70.
2. Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Blal HA, Alm T, Asplnud A, et al. A subcellular map of the human proteome. Science. 2017;356:eaal3321.
3. Jacobson MD, Weil M, Raff MC. Programmed cell death in animal development. Cell. 1997;88:347–54.
4. Kaufmann SH, Hengarther MO. Programmed cell death: alive and well in the new millennium. Trends Cell Biology. 2001;11:526–34.
5. Suzuki M, Ypule RJ, Tjandra N. Structure of box: coregulation of dimmer formation and intracellular location. Cell. 2000;103:645–54.
6. Chou KC, Shen HB. Recent progress in protein subcellular location prediction. Anal Biochem. 2007;370:1–16.
7. Nakai K, Kanehisa DM. Expert system for predicting protein localization sites in gramnegative bacteria. Proteins Struct Funct Bioinform. 1991;11:95–110.
8. Emanuelsson O, Nielsen H, Brunak S, Von HG. Predicting subcelluar localization of proteins based on their N-terminal amino acids sequences. J Mol Biol. 2000;300:1005–16.
9. Nakashima H, Nishikawa K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol. 1994;238:54–61.
10. Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of ptoteins. Nucleic Acids Res. 1998;26:2230–6.
11. Chou KC. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins Struct Funct Bioinform. 2001;43:246–55.
12. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics. 2005;21:10–9.
13. Bhasin M, Raghava GPS. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res. 2004;32:414–9.
14. Khan A, Majid A, Hayat M. CE-PLoc: an ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition. Comput Biol Chem. 2011;35:218–29.
15. Zhang ZH, Wang ZH, Zhang ZR, Wang YX. A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. FEBS Lett. 2006;580:6169–74.
16. Liang RP, Huang SY, Shi SP, Sun XY, Luo SB, Qiu JD. A novel algorithm combining support vector machine with the discrete wavelet transform for the prediction of protein subcellular localization. Comput Biol Med. 2012;42:180–7.
17. Shi JY, Zhang SW, Pan Q, Chen YM, Xie J. Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids. 2007;33:69–74.
18. 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.
19. Wan SB, Mak MW, Kung SY. mPLR-Loc: an adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction. Anal Biochem. 2015;473:14–27.
20. Wan SB, Mak MW, Kung SY. R3P-Loc: a compact multi-label predictor using ridge regression and random projection for protein subcellular localization. J Theor Biol. 2014;360:34–45.
21. Zhang L, Liao B, Li DC, Zhu W. A novel representation for apoptosis protein subcellular localization prediction using support vector machine. J Theor Biol. 2009;259:361–5.
22. Armenteros JJA, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics. 2017;33:3387–95.
23. Huang Y, Li YD. Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics. 2004;20:21–8.
24. Chou KC, Shen HB. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res. 2006;5:1888–97.
25. Emanuelsson O, Nielsen H, Heijne GV. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci. 1999;8:978–84.
26. Lin TH, Murphy RF, Barjoseph Z. Discriminative motif finding for predicting protein subcellular localization. IEEE/ACM Trans Comput Biol Bioinforma. 2011;8:441–51.
27. Saini H, Raicar G, Dehzangi A, Lal S, Sharma A. Subcellular localization for gram positive and gram negative bacterial proteins using linear interpolation smoothing model. J Theor Biol. 2015;386:25–33.
28. Scott MS, Thomas DY, Hallett MT. Predicting subcellular localization via protein motif co-occurrence. Genome Res. 2014;14:1957–66.
29. King BR, Vural S, Pandey S, Barteau A, Guda C. ngLOC: software and web server for predicting protein subcellular localization in prokaryotes and eukaryotes. BMC Res Notes. 2012;5:351–7.
30. Cai YD, Zhou GP, Chou KC. Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J. 2003;84:3257–63.
31. Ali F, Hayat M. Classification of membrane protein types using voting feature interval in combination with Chou's pseudo amino acid composition. J Theor Biol. 2015;384:78–83.
32. Zhang L, Zhou WD, Li FZ. Kernel sparse representation-based classifier ensemble for face recognition. IEEE Trans Signal Proces. 2015;74:123–37.
33. Chou KC, Shen HB. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun. 2006;347:150–7.
34. Shen HB, Chou KC. Gpos-PLoc: an ensemble classifier for predicting subcellular localization of gram-positive bacterial proteins. Protein Eng Des Sel. 2007;20:39–46.
35. Shen HB, Chou KC. Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers. 2007;85:233–40.
36. Zhou GP, Doctor K. Subcellular location prediction of apoptosis proteins. Proteins Struct Funct Bioinform. 2002;50:44–8.
37. Huang J, Shi F, Zhou HB. Support vector machine for predicting apoptosis proteins types by incorporating protein instability index. Bioinformatics. 2005;3:121–3.
38. Bulashevska A, Eils R. Predicting protein subcellular locations using hierarchical ensemble of bayesian classiners based on markov chains. BMC Bioinforma. 2006;7:298.
39. Chen YL, Li QZ. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. J Theor Biol. 2007;248:377–81.
40. Ding YS, Zhang TL. Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. Pattern Recog Lett. 2008;29:1887–92.
41. Qiu JD, Luo SH, Huang JH, Sun XY, Liang RP. Predicting subcellular location of apoptosis proteins based on wavelet transform and support vector machine. Amino Acids. 2010;38:1201–8.
42. Yu XQ, Zheng XQ, Liu TG, Dou YC, Wang J. Predicting subcellular location of apoptosis proteins with pseudo amino acid composition: approach from amino acid substitution matrix and auto covariance transformation. Amino Acids. 2012;42:1619–25.
43. Liu TG, Tao PY, Li XW, Wang CH. Prediction of subcellular location of apoptosis proteins combining tri-gram encoding based in PSSM and recursive feature elimination. J Theor Biol. 2015;366:8–12.
44. Dai Q, Ma S, Hai YB, Yao YH, Liu XQ. A segmentation based model for subcellular location prediction of apoptosis protein. Chem Intell Lab Syst. 2016;158:146–54.
45. Xiang QL, Liao B, Li XH, Xu HM, Chen J, Shi ZX, Dai Q, Yao YH. Subcellular localization prediction of apoptosis proteins based on evolutionary information and support vector machine. Artif Intell Med. 2017;78:41–6.
46. Liang YY, Liu SY, Zhang SL. Geary autocorrelation and DCCA coefficient application to predict apoptosis protein subcellular localization via PSSM. Physica A. 2017;467:296–306.

47. Pacharawongsakda E, Theeramunkong T. Predict subcellular locations of singleplex and multiplex proteins by semi-supervised learning and dimension-reducing general mode of Chou's PseAAC. IEEE Trans Nanobioscience. 2014;12:311–20.
48. Li LQ, Yu SJ, Xiao WD, Li YS, Li ML, Huang L, Zheng XQ, Zhou SW, Yang H. Prediction of bacterial protein subcellular localization by incorporating various features into Chou's PseAAC and a backward feature selection approach. Biochimie. 2014;104:100–7.
49. Briesemeister S, Rahnenführer J, Kohlbacher O. Going from where to why-interpretable prediction of protein subcellular localization. Bioinformatics. 2010;26:1232–8.
50. Lin H, Ding H, Guo FB, Huang J. Prediction of subcellular location of mycobacterial protein using feature selection techniques. Mol Divers. 2010;14:667–71.
51. Mandal M, Mukhopadhyay A, Maulik U. Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou's PseAAC. Med Biol Eng Comput. 2015;53:331–44.
52. Wan SB, Mak MW, Kung SY. mLASSO-hum: a LASSO-based interpretable human-protein subcellular localization predictor. J Theor Biol. 2015;382:223–34.
53. Kandaswamy KK, Pugalenthi G, Möller S, Hartmann E, Kalies KU, Suganthan PN, Martinetz T. Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition. Protein Pept Lett. 2010;17:1473–9.
54. Altschul SF, Madden TL, Schäffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
55. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999;292:195–202.
56. Shen HB, Chou KC. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng Des Sel. 2007;20:561–7.
57. Zebende GF. DCCA cross-correlation coefficient: quantifying level of cross-correlation. Physica A. 2011;390:614–8.
58. Chen YY, Cai LH, Wang RF, Song ZX, Deng B, Wang J, Yu HT. DCCA cross-correlation coefficients reveals the change of both synchronization and oscillation in EEG of Alzheimer disease patients. Physica A. 2018;490:171–84.
59. Vassoler RT, Zebende GF. DCCA cross-correlation coefficient apply in time series of air temperature and air relative humidity. Physica A. 2012;391:2438–43.
60. Podobnik B, Stanley HE. Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series. Phys Rev Lett. 2008;100:084102.
61. Podobnik B, Jiang ZQ, Zhou WX, Stanley HE. Statistical tests for power-law cross-correlated processes. Phys Rev E. 2011;84:066118.
62. Sugiyama M. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. J Mach Learn Res. 2007;8:1027–61.
63. Zelnik-Manor L, Perona P. Self-tuning spectral clustering. Adv Neural Inf Proces Syst. 2004;17:1601–8.
64. Vapnik V. The nature of statistical learning theory. New York: Springer-Verlag; 1995.
65. Khan M, Hayat M, Khan SA, Iqbal N. Unb-DPC: identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC. J Theor Biol. 2017;415:13–9.
66. Wan SB, Mak MW, Kung SY. Mem-ADSVM: a two-layer multi-label predictor for identifying multi-functional types of membrane proteins. J Theor Biol. 2016;398:32–42.
67. Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. Bioinformatics. 2002;18:147–59.
68. Li ZC, Zhou X, Dai Z, Zou XY. Classification of G-protein coupled receptors based on support vector machine with maximum relevance minimum redundancy and genetic algorithm. BMC Bioinform. 2010;11:325.
69. Liang YY, Liu SY, Zhang SL. Prediction of protein structural class based on different autocorrelation descriptors of position-specific scoring matrix. MATCH Commum Math Comput Chem. 2015;73:765–84.
70. Zhang LC, Kong L, Han XD, Lv JF. Structural class prediction of protein using novel feature extraction method from chaos game representation of predicted secondary structure. J Theor Biol. 2016;400:1–10.
71. Hua S. Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol. 2001;308:397–407.
72. Yu B, Lou LF, Li S, Zhang YS, Qiu WY, Wu X, Wang MH, Tian BG. Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising. J Mole Graph Model. 2017;76:260–73.
73. Li Z, Wang J, Zhang SP, Zhang QF, Wu WM. A new hybrid coding for protein secondary structure prediction based on primary structure similarity. Gene. 2017;618:8–13.
74. Guo YZ, Yu LZ, Wen ZN, Li ML. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res. 2008;36:3025–30.
75. Vyas R, Bapat S, Jain E, Karthikey M, Tambe S, Kulkarni BD. Building and analysis of protein-protein interactions related to diabetes mellitus using support vector machine, biomedical text mining and network analysis. Comput Biol Chem. 2016;65:37–44.
76. Zhang SB, Tang QR. Protein-protein interaction inference based on semantic similarity of gene ontology terms. J Theor Biol. 2016;401:30–7.
77. Zhang SB, Tang QR. Predicting protein subcellular localization based on information content of gene ontology terms. Comput Biol Chem. 2016;65:1–7.
78. Yu B, Li S, Chen C, Xu JM, Qiu WY, Wu X, Chen RX. Prediction subcellular localization of gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino composition. Chem Intell Lab Syst. 2017;167:102–12.
79. Hasan MA, Ahmad S, Molla MK. Protein subcellular localization prediction using multiple kernel learning based support vector machine. Mol BioSyst. 2017;13:785–95.
80. Yu B, Li S, Qiu WY, Chen C, Chen RX, Wang L, Wang MH, Zhang Y. Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising. Oncotarget. 2017;8:107640–65.
81. Wang LN, Shi SP, Xu HD, Wen PP, Qiu JD. Computational prediction of species-specific malonylation sites via enhanced characteristic strategy. Bioinformatics. 2017;33:1457–63.
82. Chen Z, Zhou Y, Zhang ZD, Song JG. Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features. Briefings in Bioinforma. 2015;16:640–57.
83. Lee TY, Chen SA, Huang HY, Ou YY. Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. PLoS One 2011;6:el7331.
84. Kim JH, Lee J, Oh B, Kimm K, Koh I. Prediction of phosphorylation sites using SVMs. Bioinformatics. 2004;20:3179–84.
85. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol. 2011;2:1–27.
86. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes. Pattern Recogn. 2011;44:1761–76.
87. Platt JC, Cristianini N, Shawe-Taylor J. Large margin DAGs for multiclass classification. Adv Neural Inf Process syst. 2000;12:547–53.
88. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol. 2011;273:236–47.
89. Chen W, Yang H, Feng PM, Ding H, Lin H. iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics. 2017;33:3518–23.
90. Wold S, Esbensen K, Geladi P. Principal component analysis. Chem Intell Lab Syst. 1987;2:37–52.
91. Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 2003;15:1373–96.
92. Zhang DQ, Zhou ZH, Chen SC. Adaptive kernel principal component analysis with unsupervised learning of kernels. Proceedings of the 6th IEEE international conference on data mining (ICDM'06). Hong Kong. 2006:1178–82.
93. Chen YL, Li QZ. Prediction of the subcellular location of apoptosis proteins. J Theor Biol. 2007;245:775–83.
94. Wang YX, Mao H, Yi Z. Protein secondary structure prediction by using deep learning method. Knowl Based Syst. 2017;118:115–23.
95. Lin H, Wang H, Ding H, Chen YL, Li QZ. Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition. Acta Biotheor. 2009;57:321–30.
96. Gu Q, Ding YS, Jiang XY, Zhang TL. Prediction of subcellular location apoptosis proteins with ensemble classifier and feature selection. Amino Acids. 2010;38:975–83.
97. Saravanan V, Lakshmi PT. APSLAP: an adaptive boosting technique for predicting subcellular localization of apoptosis protein. Acta Biotheor. 2013; 61:481–97.
98. Liu TG, Zheng XQ, Wang CH, Wang J. Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation. Protein Pept Lett. 2010;17:1263–9.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com