Hindawi
BioMed Research International
Volume 2019, Article ID 2436924, 9 pages
https://doi.org/10.1155/2019/2436924

Research Article

Prediction of Apoptosis Protein Subcellular Localization with Multilayer Sparse Coding and Oversampling Approach

Xingjian Chen, Xuejiao Hu, Wenxin Yi, Xiang Zou, and Wei Xue

College of Information Science and Technology, Nanjing Agricultural University, Nanjing, 210095, China

Correspondence should be addressed to Wei Xue; xwsky@njau.edu.cn

Received 19 September 2018; Revised 4 January 2019; Accepted 20 January 2019; Published 30 January 2019

Academic Editor: Xudong Huang

The prediction of apoptosis protein subcellular localization plays an important role in understanding the progress in cell proliferation and death. Recently computational approaches to this issue have become very popular, since the traditional biological experiments are so costly and time-consuming that they cannot catch up with the growth rate of sequence data anymore. In order to improve the prediction accuracy of apoptosis protein subcellular localization, we proposed a sparse coding method combined with traditional feature extraction algorithm to complete the sparse representation of apoptosis protein sequences, using multilayer pooling based on different sizes of dictionaries to integrate the processed features, as well as oversampling approach to decrease the influences caused by unbalanced data sets. Then the extracted features were input to a support vector machine to predict the subcellular localization of the apoptosis protein. The experiment results obtained by Jackknife test on two benchmark data sets indicate that our method can significantly improve the accuracy of the apoptosis protein subcellular localization prediction.

1. Introduction

As a basic constituent of organisms, proteins play a critical role in life activities such as metabolism, breeding, growth, and development, especially for the apoptosis protein, which are crucial in the proteomics. Since the functions of an apoptosis protein are closely related to its subcellular location and different kinds of apoptosis proteins can only function in specific subcellular location, it is important to predict the subcellular location of certain apoptosis protein by existing methods, which could not only help us to understand the interactions and properties of apoptosis proteins but also realize the biological pathway involved [1–3]. With the application of high-throughput sequencing techniques and the explosion of sequence data volumes, developing an accurate and reliable computational method to predict apoptosis protein subcellular location has been a great challenge for bioinformaticians, accordingly promoting the development of machine learning in this field [4–8].

By the analysis of research status, the improved directions of using machine learning to predict apoptosis protein subcellular location in the past decade can be roughly categorized into two classes: sequence feature extraction and prediction model [5–10]. Currently the widely used methods for feature extraction are amino acid composition (AAC) [11, 12], pseudo amino acid composition (PseAAC) [13, 14], gene ontology (GO)[15, 16], position specific scoring matrix (PSSM)[17, 18], feature fusion [19, 20], and so on. For example, Zhou et al. used the covariant discriminant function based on Mahalanobis distance and Chou's invariance theorem; combining with traditional AAC feature to predict apoptosis protein subcellular location, the prediction result by Jackknife test on data set ZD98 achieved about 72.5% [21]; Wan et al. proposed GOASVM algorithm based on the information of GO term frequencies and distant homologs to represent a protein in general form of PseAAC and got a high accuracy [22]; Chen et al. used the increment of diversity to fuse N-terminal, C-terminal, and hydrophobic features of apoptosis protein sequences, and the accuracies on ZD98 and CH317 were 90.8% and 82.7%, respectively [23]; Zhao et al. combined the bag of words model with PseAAC method, using K-Means to construct the dictionary of sequence features, and obtained a great predictive effect [24]. At the same time, there are also many efforts for the development of prediction model. For example, Wan et al. proposed an adaptive-decision support vector machine classifier through the annotation information of GO database and realized the prediction of membrane proteins as well as their multifunctional types [25]; Ali et al. extracted the PseAAC features of protein sequences, combining with location voting, k-nearest neighbor and probabilistic neural network to predict the subcellular localization of membrane proteins [26]. Besides, there are also some other prediction models used in this filed such as logistic regression, bayesian classifiers, and long short-term memory [27–29].

In the last decade or so, a recent review [30] pointed that a number of web-servers were developed for predicting the subcellular localization of proteins with both single site and multiple sites [31–36]. In general, proteins can simultaneously exist in multiple sites. In this study, given that the number of multilabel proteins in the existing apoptosis protein database is not large enough to construct a benchmark data set meaningfully in statistics and for the case of multiple locations, the sequence information is more complex and various than single locations, using oversampling approach to copy sequence feature may generate the inaccurate results, so here we did not consider the situation of multilabel proteins.

To summarize the previous research results, it is not difficult to find that the prediction accuracy is relatively low if merely using simple method such as AAC or PseAAC to extract sequence features for classification; as for the other feature extraction methods, like PSSM or feature fusion, though the prediction effect is better, the extraction process is too complicated and time-consuming for practical application. Given that many former researches have proved that support vector machine is one of the best classifiers for the prediction of protein subcellular localization [5, 9, 10, 14, 17, 22], in this study, we focus on obtaining a higher prediction accuracy on the premises of simple feature extraction method and support vector machine to predict the subcellular localization of the apoptosis protein, therefore, finding an efficient approach to optimize the traditional sequence-based feature is the key problem to be solved in this paper.

In the study, we proposed a sparse coding method combined with traditional sequence feature extraction algorithm to extract low-level features of the apoptosis protein sequence, using multilayer pooling based on different sizes of dictionaries to integrate the local and holistic features of the sparse representation. Then the support vector machine was used to complete the final prediction. Given that our adopted benchmark data sets are unbalanced which may influence the classification effects of support vector machine [37], we used an oversampling approach to balance the data sets in the study. Compared with other experimental results with the same support vector machine classifier, the experimental results show that the proposed method can not only simplify the feature extraction process and reduce the time and space complexity of the classifier but also reflect the sequence features more comprehensively and improve the classification performance. The detailed descriptions are shown in the following sections.

2. Materials and Methods

2.1. Datasets. Two widely used benchmark data sets are adopted in this study: ZD98 and CH317, respectively. The data set ZD98 was constructed by Zhou and Doctor [21]. There are 98 apoptosis protein sequences divided into four kinds of subcellular locations, which are cytoplasmic proteins (Cy), mitochondrial proteins (Mi), membrane proteins (Me), and other proteins (Other). The data set CH317 was constructed by Chen and Li [23] and contains a total of 317 apoptosis protein sequences, in 6 classes of subcellular locations that are secreted proteins (Se), nuclear proteins (Nu), cytoplasmic proteins (Cy), endoplasmic reticulum proteins (En), membrane proteins (Me), and mitochondrial proteins (Mi). Considering that the above data sets are old, we update ZD98 and CH317 data sets with reference to Wang et al. [38] and remove some of the duplicates and error sequences. The specific method is not repeated here. After processing, there were 96 protein sequences remaining in ZD98 data set and 314 protein sequences remaining in CH317 data set. All protein sequences in the above two data sets are from the latest version of the UniProt database (Release 2018_12), and the number of protein sequences in each class of 2 data sets is shown in Table 1.

Table 1: Numbers of protein sequences in different class of 2 datasets.

Dataset	Number of sequences in each class						Total
ZD98	Cy	Me	Mi	Other			
	43	30	13	12			98
CH317	Cy	En	Me	Mi	Nu	Se	
	112	47	55	34	52	17	317

2.2. Feature Extraction. In order to set up a more accurate mapping relationship between each protein sequence and its corresponding feature vector, multilayer sparse coding was introduced in this study to find the most essential feature of original protein sequence based on simple feature extraction method. The algorithm mainly includes the following steps: local feature extraction, sparse coding, and pooling. And the process of sparse coding is divided into 2 sections: dictionary learning and sparse representation. Firstly, the protein sequence is segmented into some fragments, and the traditional protein feature extraction algorithm will be used to extract the features of these fragments, which could be applied for the step of dictionary learning. Then these local features are trained to construct a dictionary and the feature representation of original sequence would be sparsely reconstructed by it. The mean pooling is used to reduce the dimensions of the feature matrix, and finally the pooled vectors based on different dictionary sizes would be integrated as the ultimate features of protein sequences. The flow chart of extraction progress is shown in Figure 1.

2.2.1. Local Feature Extraction. Before the step of sparse coding, it is necessary to extract the local features of protein sequence to constitute a training sample set for dictionary learning. Since every protein sequence has the different length and the critical features may be distributed in different positions of the sequence, in this paper, we adopted sliding

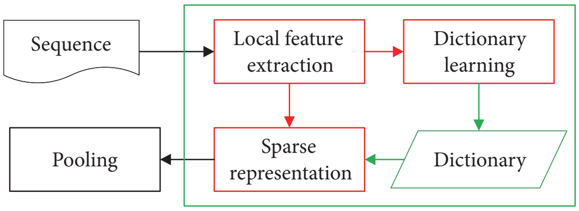

Figure 1: The flow of feature extraction process.

window segmentation method inspired by Noor to cut all the protein sequences into pieces [39], generating a number of sequence fragments afterwards. The size of sliding window represents the segment length of each protein sequence, and the reference formula is

$$L_{min} = \min\left\{L_1, L_2, \ldots, L_{num}\right\}, \qquad \frac{L_{min}}{2} \le s \le L_{min}, \ m \in Z \tag{1}$$

where $L_1, L_2, \ldots, L_{num}$ represent the length of each protein sequence in the whole data set, L_{min} is the shortest sequence of it, and s is the size of sliding window, which indicates that the value of segment length is between $L_{min}/2$ and L_{min}, and the exact value will be selected by the experimental experience.

After the step of segmentation, the existing sequence feature extraction method is used to statistically analyze the component information of sequence fragments and to transform the character sequences into numerical vectors as the local features of the protein. Effective feature extraction method can remarkably increase the final prediction accuracy. Nakashima and Nishikawa [47] firstly associated the amino acid composition (AAC) with the prediction of protein subcellular location in 1994. The AAC coding method was proposed to count the occurrence frequency of each amino acid in the protein sequence, described as follows:

$$P_{AAC} = \left[f_1 f_2 f_3 \cdots f_{20}\right]^T \tag{2}$$

where $f_1, f_2, f_3, \ldots, f_{20}$ represent the number of each amino acid in the protein sequence, respectively and the specific explanation is

$$f_u = \frac{1}{L}\sum_{i=1}^{L} F_i, \quad F_i = \begin{cases} 1, & if\ R_i = A(u) \\ 0, & if\ R_i \neq A(u) \end{cases} \tag{3}$$

L represents the length of each protein sequence, that is, the total number of all the amino acid residues contained. Firstly, 20 amino acids are numbered from 1 to 20, and f_u $(u = 1, 2, 3, \ldots, 20)$ describes the frequency of corresponding number appeared in the sequence. R_i represents each amino acid residue in original sequence, and $A(u)$ represents the amino acid residue which corresponds to the number u.

By using AAC to calculate the fragment features of protein sequence P, we can obtain a feature matrix for each original protein sequence constituted by all the AAC features of corresponding fragments. The matrix is shown in

$$V = \begin{bmatrix} v_{11} & v_{1n} & \cdots & v_{1n} \\ v_{21} & \cdots & \cdots & v_{2n} \\ \vdots & \cdots & \cdots & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mn} \end{bmatrix} \tag{4}$$

where m represents the number of fragments cut by a protein sequence, n is the feature dimension processed by AAC algorithm, and v_{mn} represents the probability of occurrence of different amino acid residues. At this time, n is 20. Each row of the matrix represents the feature vectors of different sequence fragments in a protein sequence. Generally we choose some of the fragment features as the local features to construct the dictionary, in this paper; since the number of fragments obtained is not very large, in order to get a better feature representation in spar coding, we chose the local features of all the sequence fragments to form a training sample set $X = [x_1, x_2, x_3, \ldots, x_N]$ for dictionary learning, where $x_i \in R^n$ $(i = 1, 2, 3, \ldots, N)$, x_i represents the feature vector of different protein sequences, that is, the vector in each row in V, and N is the number of fragments belonged to all of the protein sequences in the data set.

2.2.2. Sparse Coding. Sparse coding is a branch of deep neural networks, and it contains 2 main steps: dictionary learning and sparse representation, respectively [48]. It can extract the detailed features of original data set and decompose the input sample set into a linear combination of multiple primitives. The coefficients of the primitives are the features of input sample. The description can be formulated as

$$X = UD \tag{5}$$

where X is the matrix of training sample composed by fragment features; $D = [d_1, d_2, d_3, \ldots, d_K] \in R^{K*n}$ is the primitive matrix named the dictionary, d_i represents the feature elements of dictionary, K is the size of dictionary, n is feature dimension *20* processed by AAC algorithm,; $U = [u_1, u_2, u_3, \ldots, u_N] \in R^{N*K}$ is the sparse representation of original sample, and u_i represents the sparse coefficient of the i-th feature block in the sparse feature space, that is, the projection of x_i in sparse feature space. N is the number of fragments belonged to all of the protein sequences in the data set. The solution of dictionary D can be expressed as

$$\min_{D,U} \sum_{i=1}^{N} \left\|x_i - Du_i\right\|_2^2 \qquad s.t \quad \left\|u_i\right\|_0 \ll T_0, \ i = 1, 2, 3, \ldots, N. \tag{6}$$

where $\| \bullet \|_2$ represents L_2 norm of a vector and $\| \bullet \|_0$ is L_0 norm of a vector. The constraint in formula above means that the number of nonzero elements in u_i needs to be less than or equal to T_0, which is preset and related to the sparse rate. Equation (6) is essentially a nonconvex optimization problem. There are mainly two common solutions for it: the first is to transform it into a convex optimization problem to relax the constraint of equation and then transforms it into the following form:

$$\min_{D,U} \sum_{i=1}^{N} \left(\left\| x_i - Du_i \right\|_2^2 + \lambda \left\| u_i \right\|_1 \right). \tag{7}$$

where λ is the balance factor and $\| \bullet \|_1$ represents L_1 norm of a vector. Equation (7) can usually be solved by regression algorithm, such as LASSO [49]. The second is to solve it by using the heuristic greedy algorithm [50]. The algorithms for the second solution are MOD and K-SVD [51]. In this study, in view of the efficiency and operability of the algorithm, we choose K-SVD as our solution to learn the dictionary; that is, the second solution.K-SVD is an expansion of K-means algorithm proposed by Aharon and Elad [52]. It adopts the method of iterative alternating learning and uses the singular value decomposition to perform K times iterations to optimize the primitives of dictionary, which can better fit the original data. K-SVD is mainly divided into the following steps:

(1) Initialize the dictionary D, and set the terminal condition of iteration;

(2) Fix D, solve the sparse representation U;

(3) Fix U, solve the dictionary D;

(4) Perform steps (2) and (3) alternately until the end of the iteration.

After obtaining the dictionary, the orthogonal matching pursuit (OMP) algorithm is used to complete the sparse representation of the fragment features of the original protein sequence [53]. The basic theory of OMP is to select one of the most matching primitives from the dictionary to perform a sparse approximation with the primitives of original samples and to obtain the residual between them. Then, it continues to select the next proper primitive which is best matched with this signal residual and iterates in this way over and over until the residual and sparse rate meets the fixed terminal conditions. Samples can be approximately presented by a linear combination of these derived primitives. All primitives selected in each process must be orthogonalized first, which would make the convergence speed faster [54]. Constituting the sparse features of all the encoded fragments, we can obtain an $m * K$ sparse matrix Z to represent the feature of each protein sequence, where m is the number of sequence segments in each sequence and K is the size of dictionary, that is, the sparse representation of a protein sequence.

2.2.3. Multilayer Pooling. The dimension of the feature matrix obtained by sparse coding is very high, if we expand it directly, the huge data volume will cause redundant space and time complexities of classification, and it is prone to overfitting. Therefore, it is necessary to reduce the dimensions of the feature matrix. The method of pooling can map the collection of feature vectors into a single vector. There are two different common pooling methods that are the max pooling and mean pooling, respectively. The aggregation statistics of features in different positions can extract the effective information and reduce the calculated amount of numerical matrix [55]. Max pooling takes the maximum value of the feature points in the neighborhood and retains the edge information of the feature matrix more, while mean pooling takes the average value of the feature points in the neighborhood and more to extract the background information [56]. Given that the characters of sequence data are different from images, we chose the mean pooling as the final dimension-reduced method. The formula is shown as follows:

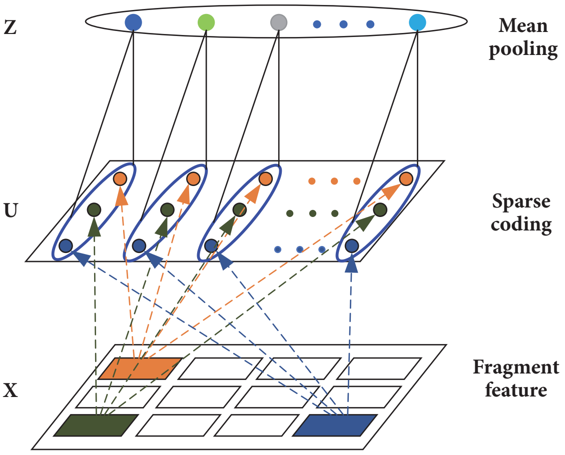

FIGURE 2: The progress of multilayer pooling.

$$\begin{gathered} Z = \left[z_1, z_2, z_3, \ldots, z_K \right]^T \\ z_i = mean \left\{ \left| z_{i1} \right|, \left| z_{i2} \right|, \left| z_{i3} \right|, \ldots \left| z_{im} \right| \right\} \end{gathered} \tag{8}$$

where $i = 1, 2, 3, \ldots K$, z_i being obtained by averaging the m elements in the i-th row of the matrix Z. After being processed by mean pooling, each protein sequence is represented as a K dimensional feature vector, K is the size of dictionary.

In order to obtain a more overall feature representation of original protein sequence, multilayer pooling based on different sizes of dictionary is performed, and several pooling results will be integrated to help extract the local and holistic features severally. The specific description is as follows: in the process of sparse coding, the values of dictionary sizes are set to K_1, K_2, and K_3 respectively; thus 3 different levels of dictionary could be obtained by K-SVD algorithm. Then the OMP algorithm is used to complete the sparse representation of fragment features based on different dictionary sizes, and the sparse features are combined to obtain the feature matrix of original sequence. Finally the sparse matrix will be mean pooled to extract different levels of feature vectors. The vectors in each pooled block are concatenated together to obtain a $K_1 + K_2 + K_3$ dimensional vector as the final feature representation. In this paper, the values of K were set to 30, 50, and 70, respectively, generating a 150 dimensional vector to be selected by principal component analysis (PCA) and sent to the classifier for prediction. The general descriptions of spare coding and pooling can be shown in Figure 2.

Table 2: Numbers of protein sequences in different class of 2 datasets.

Dataset	Number of sequences in each class						Total
ZD98	Cy	Me	Mi	Other			
	43	30	26	24			123
CH317	Cy	En	Me	Mi	Nu	Se	
	112	47	55	51	52	51	368

2.3. Oversampling Method. Since the data sets used in this paper are not balanced, which may cause the low accuracy of prediction, we referred to some similar papers used the oversampling to balance the data set [16, 30, 43]. In order to further illustrate the effect of our method, a simple oversampling method called synthetic minority oversampling technique (SMOTE) was applied in the study to decrease the imbalance of our data set. SMOTE is a classical oversampling method proposed by Chawla et al. [57]. It is widely used for its good classification effect and simple operation. The basic principle of SMOTE algorithm is to synthesize new minority samples between a few neighbouring samples and to reduce the imbalance of the data distribution. The details are as follows:

(1) For each sample X in the class of smaller number of data set, calculate the Euclidean distance from other samples in the minority class to obtain the K nearest neighbor samples.

(2) Assuming that the sampling magnification is N, for each of the few classes of samples X, $n(K > n)$ samples are randomly selected from their K nearest neighbor samples and these n samples are recorded as $y_1, y_2, y_3, \ldots, y_n$.

(3) According to the following, combine each sample X with n samples to perform random interpolation operations to synthesize n interpolated samples Pi:

$$Pi = X + \text{rand}(0,1) * (yi - X), \quad i = 1,2,3,\ldots,n \tag{9}$$

where rand$(0,1)$ represents a random number within $(0,1)$ and yi represents the i-th nearest neighbor sample of X.

(4) Finally, the interpolated sample Pi is added to the original sample set to form a new sample set.

The imbalance degree of the data set determines the value of N, and the imbalanced level (IL) between majority and minority of the data set is calculated according to

$$N = round(IL) \tag{10}$$

where $round(IL)$ represents the value obtained by rounding up IL. Through the above interpolation operation, the majority and the minority samples can be effectively balanced to improve the accuracy of classification. In this study, the minority classes of 2 data sets are balanced by SMOTE, and the processed results after are as Table 2.

2.4. Classifier and Performance Measures. In order to facilitate the comparison with other feature extraction algorithms, we used support vector machine (SVM) as the classification model in this study. After the feature extraction of protein sequences, the universal package of LIBSVM developed by Lin was applied to construct the SVM multiclass classifier [58]. The Jackknife test was also adopted to examine the effectiveness of classifier in our experiment. Jackknife test has the least arbitrary that can always yield a unique result for a given benchmark dataset [59]. Furthermore, in order to have a more comprehensive evaluation, sensitivity (Se), specificity (Sp), Matthew's correlation coefficient (MCC), and the overall accuracy (OA) over the entire data set are applied as the evaluation index [20, 21, 60]. These parameters are detailed in

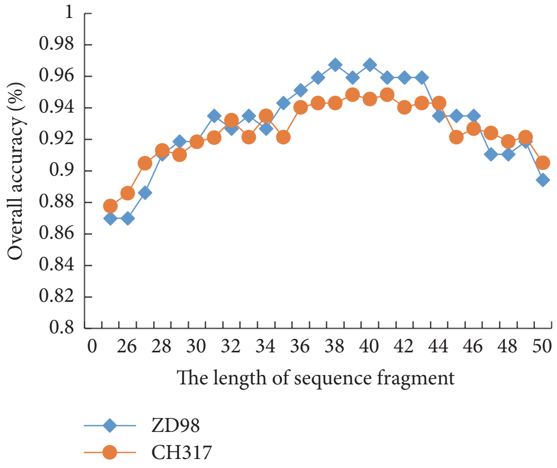

Figure 3: The prediction accuracy based on different lengths of sequence fragments.

$$Se = \frac{TP}{TP + FN} \tag{11}$$

$$Sp = \frac{TN}{TN + FP} \tag{12}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN) * (TP + FP) * (TN + FN) * (TN + FP)}} \tag{13}$$

$$OA = \frac{\sum_{i=1}^{k} TP_i}{N} \tag{14}$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively; N is the total number of protein sequences and k is the class number.

2.5. Parameters Selection. There are two key parameters in this study. One is the length of sequence fragment in the local feature extraction. The shortest protein sequence length in the data set is 50, and the fragment length is selected between 25 and 50. Figure 3 shows the prediction accuracy of the data set ZD98 and CH317, respectively, when taking different slice lengths.

As shown in Figure 3, when the sequence length is between 35 and 40, the prediction accuracies on the data sets ZD98 and CH317 are the highest and tend to be stable, and the

Table 3: The experimental results of data sets.

Dataset		Jackknife test(%)			
	Location	Sn(%)	Sp(%)	MMC(%)	OA(%)
ZD98	Cy	100	95.6	95.9	96.7
	Me	96.7	96.7	95.2	
	Mi	92.3	96.0	86.4	
	Other	95.9	95.8	90.5	
CH317	Cy	95.5	93.8	90.9	94.8
	Me	93.6	92.7	91.1	
	Mi	96.4	94.6	96.7	
	Se	94.1	92.3	83.4	
	Nu	94.2	92.5	89.6	
	En	94.1	90.5	91.5	

Table 4: Comparison of the accuracy of ZD98 data set.

Methods	Jackknife test(%)				
	Cyto	Memb	Mito	Other	OA(%)
DCC_SVM [40]	93.0	86.7	92.3	75.0	88.9
OF_SVM [41]	97.7	86.3	92.3	66.7	90.8
DE_SVM [42]	95.4	93.3	76.9	83.3	90.8
BOW_SVM [24]	97.7	92.9	76.9	83.3	91.7
GA_SVM [17]	95.4	90.0	92.3	83.3	91.8
OA_SVM [43]	95.3	88.9	97.4	91.7	93.2
Our	100	96.7	92.3	95.9	96.7

current length is the optimal value. The optimal values for the two data sets used in this study are 35 and 40, respectively.

When using PCA to select the final feature vectors, the setting of dimension D has an effect on the accuracy of the entire algorithm. The more dimensions are selected and the more features are included, but the training time of the classifier may be too long. The smaller the dimension is, the more likely it is to lose some truly meaningful features and affect the classification effect. Therefore, an optimal D needs to be sought through experiments. Figure 4 shows the prediction accuracy corresponding to the different D taken by the data sets ZD98 and CH317 during the feature selection of PCA.

As shown in Figure 4, when the dimension of the feature vector is low, the prediction accuracy of two data sets is relatively low. When the dimension is higher than a certain value, the prediction accuracy is also reduced. When the dimension is between 60 and 70, the prediction accuracies on the data sets ZD98 and CH317 are the largest and tend to be steady, and the current D is the optimal value. The optimal values for the two data sets used are 60 and 65, respectively.

3. Result and Discussion

The prediction results of our experiments by Jackknife on the data sets ZD98 and CH317 are listed in Tables 3, 4, and 5. To further illustrate the effectiveness of our method, the prediction results in each subcellular location of two data sets are also listed in Tables 3–5, which are sensitivity, specificity, correlation coefficient, and overall accuracy, respectively.

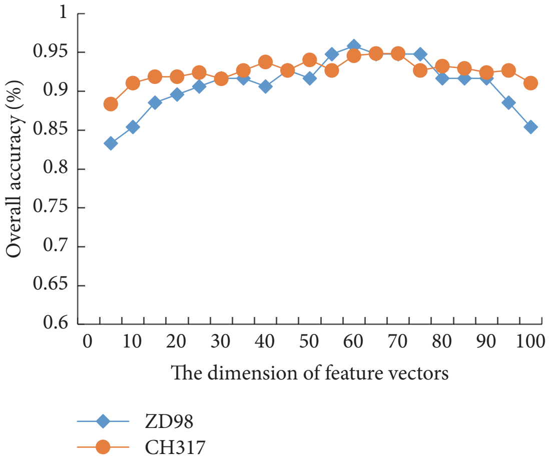

Figure 4: The prediction accuracy based on different dimensions.

It can be seen from Table 3 that the method has obtained good experimental results on both two data sets, and the total accuracies rates are 96.7% and 94.8%, respectively. The experiment proves that the method can effectively increase the accuracy of the prediction of protein subcellular localization. At the same time, in order to facilitate the comparison with other methods, we have listed some experimental results

Table 5: Comparison of the accuracy of CH317 data set.

Methods	Jackknife test(%)						
	Cyto	Memb	Mito	Secr	Nucl	Endo	OA(%)
DCC_SVM [40]	91.1	92.7	82.4	76.5	80.0	93.6	88.3
GA_SVM [17]	92.9	89.1	82.4	76.5	84.6	93.6	89.0
BOW_SVM [24]	94.6	87.3	82.4	82.4	84.3	91.5	89.2
IAC_SVM [44]	96.4	94.5	82.4	76.5	80.8	93.6	90.5
EI_SVM [45]	94.6	95.7	92.7	82.4	90.4	70.6	91.1
CF_SVM [46]	96.4	90.9	92.3	95.7	82.4	64.7	91.5
Our	95.5	93.6	96.4	94.1	94.2	94.1	94.8

based on some improved algorithms of protein sequence feature extraction in the past several years.

In Tables 4 and 5, DCC_SVM comes from Liang [40], by using detrended cross-correlation coefficient(2016); OF_SVM comes from Zhang [41], by using λ-Order Factor and principal component analysis(2017); DE_SVM comes from Liang [42], by fusing two different descriptors based on evolutionary information(2018); BOW_SVM comes from Zhao [24], by using bag of words(2017); GA_SVM comes from Liang [17], by using geary autocorrelation and DCCA coefficient(2017); OA_SVM comes from Zhang [43], by using oversampling and pseudo amino acid composition(2018); IAC_SVM comes from Zhang [44], by using integrating auto-cross correlation and PSSM(2018); EI_SVM comes from Xiang [45], by using evolutionary information(2017); CF_SVM comes from Chen [46], by using a set of discrete sequence correlation factors(2015); all the methods use SVM as the final classifier.

It can be seen from Table 4 that the result on the data set ZD98 has a maximum improvement of the overall prediction accuracy, increasing by about 6 to 8 percentage points compared with traditional protein sequence feature extraction algorithms such as DCC_SVM, OF_SVM, and DE_SVM. In the subcellular class of cytoplasmic proteins, the prediction accuracy rate is 100%, which means that all the sequences in this class are predicted correctly, and the overall prediction accuracy is better than other methods as well. Compared the experimental results with other improved feature extraction algorithms such as BOW_SVM, GA_SVM, and OA_SVM, the accuracy on the same data set is also improved by about 3 to 5 percentage points. Experiments show that the proposed method indeed provides a better source of information for protein sequences and have significant advantages than other similar feature extraction methods. From the comparison in Table 5, we can see that the prediction result on mitochondrial proteins of data set CH317 is up to 96.4%, which is about 4.1% to 14% higher than other algorithms. The accuracy rate in the class of Nuclear has also increased by 14.2% maximally, improving the total prediction accuracy by 3.3 to 4.3 percentage points compared with the improved algorithms such as IAC_SVM, EI_SVM, and CF_SVM, which further demonstrates that the method can optimize the underlying features of the sequence and effectively improve the prediction accuracy of apoptosis protein subcellular localization. Compared with the traditional protein sequence feature extraction and their improved methods, the time complexity of our algorithm is not only low but can also achieve better results based on the simple AAC feature. The background information of the feature representation can also be extracted by mean pooling and comprehensively reflect the distribution of sequence features more, as well as improving the classification accuracy.

4. Conclusions

Prediction of apoptosis protein subcellular localization has always been the hotspot of bioinformaticians all over the world. Based on the traditional protein sequence feature extraction algorithm AAC, this paper introduced sparse coding to optimize sequence features and proposed a feature fusion method based on multilevel dictionary. The main contribution includes firstly using sliding window segmentation to extract the sequence fragments of protein sequences, and the traditional feature extraction algorithm was used to encode them. Then the K-SVD algorithm was used to learn the dictionary, and the sequence feature matrix was sparsely represented by the OMP algorithm. The feature representation based on different sizes of dictionaries is mean-pooled to help extract the overall and local feature information. Finally the SVM multiclass classifier is used to predict the subcellular location of the proteins. Experiments show that the proposed method can obtain better results in the prediction success rate of most subcellular classes and have important guiding significance for improving the feature expression of traditional apoptosis protein sequence feature extraction algorithms. Generally speaking, it is a relatively effective method for predicting the subcellular localization of apoptosis proteins.

Data Availability

The data used to support the findings of this study is available from the corresponding author upon request, and you can also find it from https://github.com/Multisc/Multi_sc_subloc.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was partially supported by National Key R&D Program of China (2017YFD0800204) and the Fundamental Research Funds for the Central Universities (No. KYZ201600175).

References

[1] M. Breker and M. Schuldiner, "The emergence of proteome-wide technologies: Systematic analysis of proteins comes of age," *Nature Reviews Molecular Cell Biology*, vol. 15, no. 7, pp. 453–464, 2014.

[2] S. Wang and Y. Yue, "Protein subnuclear localization based on a new effective representation and intelligent kernel linear discriminant analysis by dichotomous greedy genetic algorithm," *PLoS ONE*, vol. 13, no. 4, Article ID e0195636, 2018.

[3] T. J. Burns, A. P. Frei, P. F. Gherardini et al., "High-throughput precision measurement of subcellular localization in single cells," *Cytometry Part A*, vol. 91, no. 2, pp. 180–189, 2017.

[4] M. Salvatore, N. Shu, and A. Elofsson, "The SubCons webserver: A user friendly web interface for state-of-the-art subcellular localization prediction," *Protein Science*, vol. 27, no. 1, pp. 195–201, 2017.

[5] X. Cheng, X. Xiao, and K.-C. Chou, "PLoc-mHum: Predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information," *Bioinformatics*, vol. 34, no. 9, pp. 1448–1456, 2017.

[6] Q. Xiang, B. Liao, X. Li et al., "Subcellular localization prediction of apoptosis proteins based on evolutionary information and support vector machine," *Artificial Intelligence in Medicine*, vol. 78, no. 9, pp. 41–46, 2017.

[7] M. R. Uddin, A. Sharma, D. M. Farid et al., "EvoStruct-SUB: An accurate Gram-positive protein subcellular localization predictor using evolutionary and structural features," *Journal of Theoretical Biology*, vol. 443, pp. 138–146, 2018.

[8] S. Qiao, B. Yan, and J. Li, "Ensemble learning for protein multiplex subcellular localization prediction based on weighted KNN with different features," *Applied Intelligence*, vol. 48, no. 7, pp. 1813–1824, 2018.

[9] S. Wang, B. Nie, K. Yue et al., "Protein subcellular localization with gaussian kernel discriminant analysis and its kernel parameter selection," *International Journal of Molecular Sciences*, vol. 18, no. 12, pp. 2718–2733, 2017.

[10] W. Xue, X.-Y. Hong, N. Zhao et al., "Predicting protein subcellular localization by approximate nearest neighbor searching," in *Proceedings of the Chinese Control and Decision Conference*, pp. 2842–2846, IEEE, 2017.

[11] Y. Huang and G. Huang, "A homology and pseudo amino acid composition-based multi-label model for predicting human membrane protein types," *Current Proteomics*, vol. 15, no. 2, pp. 135–141, 2018.

[12] M. Saidijam, S. Azizpour, and S. G. Patching, "Amino acid composition analysis of human secondary transport proteins and implications for reliable membrane topology prediction," *Journal of Biomolecular Structure and Dynamics*, vol. 35, no. 5, pp. 929–949, 2017.

[13] M. Rahimi, M. R. Bakhtiarizadeh, and A. Mohammadi-Sangcheshmeh, "OOgenesis_Pred: A sequence-based method for predicting oogenesis proteins by six different modes of Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 414, pp. 128–136, 2017.

[14] K. Ahmad, M. Waris, and M. Hayat, "Prediction of protein submitochondrial locations by incorporating dipeptide composition into chou's general pseudo amino acid composition," *Journal of Membrane Biology*, vol. 249, no. 3, pp. 293–304, 2016.

[15] X. Xiao, X. Cheng, S. Su et al., "PLoc-mGpos: Incorporate key gene ontology information into general PseAAC for predicting subcellular localization of gram-positive bacterial proteins," *Natural Science*, vol. 09, no. 09, pp. 330–349, 2017.

[16] M. A. M. Hasan, S. Ahmad, M. N. I. Mondal et al., "PredMultiLoc-Gneg: Predicting subcellular localization of gram-negative bacterial proteins using feature selection in gene ontology space and resolving the data imbalance issue," in *Proceedings of the IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pp. 109–112, IEEE, 2017.

[17] Y. Liang, S. Liu, and S. Zhang, "Geary autocorrelation and DCCA coefficient: Application to predict apoptosis protein subcellular localization via PSSM," *Physica A: Statistical Mechanics and its Applications*, vol. 467, pp. 296–306, 2017.

[18] Y.-H. Yao, Z.-X. Shi, and Q. Dai, "Apoptosis protein subcellular location prediction based on position-specific scoring matrix," *Journal of Computational and Theoretical Nanoscience*, vol. 11, no. 10, pp. 2073–2078, 2014.

[19] S. Wan, M.-W. Mak, and S.-Y. Kung, "HybridGO-Loc: Mining hybrid features on gene ontology for predicting subcellular localization of multi-location proteins," *PLoS ONE*, vol. 9, no. 3, Article ID e89545, 2014.

[20] Z. Wang, Q. Zou, Y. Jiang et al., "Review of protein subcellular localization prediction," *Current Bioinformatics*, vol. 9, no. 3, pp. 331–342, 2014.

[21] G.-P. Zhou and K. Doctor, "Subcellular location prediction of apoptosis proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 50, no. 1, pp. 44–48, 2003.

[22] S. Wan, M. Mak, and S. Kung, "GOASVM: A subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 323, pp. 40–48, 2013.

[23] Y.-L. Chen and Q.-Z. Li, "Prediction of the subcellular location of apoptosis proteins," *Journal of Theoretical Biology*, vol. 245, no. 4, pp. 775–783, 2007.

[24] N. Zhao, L. Zhang, W. Xue et al., "Application of bag of words model in the prediction of protein subcellular location," *Journal of Food Science and Biotechnology*, vol. 36, no. 3, pp. 296–301, 2017.

[25] S. Wan, M.-W. Mak, and S.-Y. Kung, "Mem-ADSVM: A two-layer multi-label predictor for identifying multi-functional types of membrane proteins," *Journal of Theoretical Biology*, vol. 398, pp. 32–42, 2016.

[26] F. Ali and M. Hayat, "Classification of membrane protein types using voting feature interval in combination with Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 384, pp. 78–83, 2015.

[27] S. Wan, M.-W. Mak, and S.-Y. Kung, "MPLR-Loc: An adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction," *Analytical Biochemistry*, vol. 473, pp. 14–27, 2015.

[28] S. Sáez-Atienzar, J. Martínez-Gómez, J. I. Alonso-Barba et al., "Automatic quantification of the subcellular localization of chimeric GFP protein supported by a two-level Naive Bayes classifier," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1531–1537, 2015.

[29] S. r. Sønderby, C. K. Sønderby, H. Nielsen, and O. Winther, "Convolutional LSTM networks for subcellular localization

of proteins," in *International Conference on Algorithms for Computational Biology*, vol. 9199 of *Lecture Notes in Computer Science*, pp. 68–80, 2015.

[30] K. Chou, "Impacts of bioinformatics to medicinal chemistry," *Medicinal Chemistry*, vol. 11, no. 3, pp. 218–234, 2015.

[31] X. Cheng, X. Xiao, and K.-C. Chou, "PLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC," *Genomics*, vol. 110, no. 1, pp. 50–58, 2018.

[32] S. Wan, M.-W. Mak, and S.-Y. Kung, "FUEL-mLoc: Feature-unified prediction and explanation of multi-localization of cellular proteins in multiple organisms," *Bioinformatics*, vol. 33, no. 5, pp. 749-750, 2017.

[33] X. Guo, F. Liu, Y. Ju, Z. Wang, and C. Wang, "Human protein subcellular localization with integrated source and multi-label ensemble classifier," *Scientific Reports*, vol. 6, no. 1, 2016.

[34] S. Wan, M.-W. Mak, and S.-Y. Kung, "Ensemble linear neighborhood propagation for predicting subchloroplast localization of multi-location proteins," *Journal of Proteome Research*, vol. 15, no. 12, pp. 4755–4762, 2016.

[35] J. J. A. Armenteros, C. K. Sønderby, S. K. Sønderby, H. Nielsen, and O. Winther, "DeepLoc: Prediction of protein subcellular localization using deep learning," *Bioinformatics*, vol. 33, no. 21, pp. 3387–3395, 2017.

[36] S. Wan, M.-W. Mak, and S.-Y. Kung, "Gram-LocEN: Interpretable prediction of subcellular multi-localization of gram-positive and gram-negative bacterial proteins," *Chemometrics and Intelligent Laboratory Systems*, vol. 162, pp. 1–9, 2017.

[37] K. Tbarki, S. B. Said, R. Ksantini, and Z. Lachiri, "Landmine detection improvement using one-class SVM for unbalanced data," in *Proceedings of the International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pp. 1–6, IEEE, 2017.

[38] X. Wang, H. Li, Q. Zhang, and R. Wang, "Predicting subcellular localization of apoptosis proteins combining GO features of homologous proteins and distance weighted KNN classifier," *BioMed Research International*, vol. 2016, Article ID 1793272, 8 pages, 2016.

[39] M. H. M. Noor, Z. Salcic, and K. I.-K. Wang, "Adaptive sliding window segmentation for physical activity recognition using a single tri-axial accelerometer," *Pervasive and Mobile Computing*, vol. 38, no. 1, pp. 41–59, 2017.

[40] Y. Liang, S. Liu, and S. Zhang, "Detrended cross-correlation coefficient: Application to predict apoptosis protein subcellular localization," *Mathematical Biosciences*, vol. 282, pp. 61–67, 2016.

[41] S. Zhang and J. Jin, "Prediction of protein subcellular localization by using λ-order factor and principal component analysis," *Letters in Organic Chemistry*, vol. 14, no. 9, pp. 717–724, 2017.

[42] Y. Liang and S. Zhang, "Prediction of apoptosis protein's subcellular localization by fusing two different descriptors based on evolutionary information," *Acta Biotheoretica*, vol. 66, no. 1, pp. 61–78, 2018.

[43] S. Zhang and X. Duan, "Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 437, pp. 239–250, 2018.

[44] S. Zhang and Y. Liang, "Predicting apoptosis protein subcellular localization by integrating auto-cross correlation and PSSM into Chou's PseAAC," *Journal of Theoretical Biology*, vol. 457, pp. 163–169, 2018.

[45] Q. Xiang, B. Liao, X. Li et al., "Subcellular localization prediction of apoptosis proteins based on evolutionary information and support vector machine," *Artificial Intelligence in Medicine*, vol. 78, pp. 41–46, 2017.

[46] H. Chen, X. Chen, Q. Hu, and Z. Cao, "Predicting protein subcellular location based on a novel sequence numerical model," *Journal of Computational and Theoretical Nanoscience*, vol. 12, no. 1, pp. 82–87, 2015.

[47] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *Journal of Molecular Biology*, vol. 238, no. 1, pp. 54–61, 1994.

[48] Y. Quan, Y. Xu, Y. Sun, Y. Huang, and H. Ji, "Sparse coding for classification via discrimination ensemble," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5839–5847, 2016.

[49] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based Fisher discrimination dictionary learning for image classification," *International Journal of Computer Vision*, vol. 109, no. 3, pp. 209–232, 2014.

[50] B. M. Whitaker, P. B. Suresha, C. Liu, G. D. Clifford, and D. V. Anderson, "Combining sparse coding and time-domain features for heart sound classification," *Physiological Measurement*, vol. 38, no. 8, pp. 1701–1729, 2017.

[51] A. Cherian and S. Sra, "Riemannian sparse coding for positive definite matrices," in *Proceedings of the European Conference on Computer Vision*, pp. 299–314, 2014.

[52] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[53] J. Wen, Z. Zhou, J. Wang, X. Tang, and Q. Mo, "A sharp condition for exact support recovery with orthogonal matching pursuit," in *Proceedings of the IEEE Transactions on Signal Processing*, vol. 65, pp. 1370–1382, IEEE, 2017.

[54] A. Cohen, W. Dahmen, and R. DeVore, "Orthogonal matching pursuit under the restricted isometry property," *Constructive Approximation. An International Journal for Approximations and Expansions*, vol. 45, no. 1, pp. 113–127, 2017.

[55] Y. Liu, J. Cheng, Y. Ma, and Y. Chen, "Protein secondary structure prediction based on two dimensional deep convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer and Communications (ICCC)*, pp. 1995–1999, IEEE, 2017.

[56] Y. Chen, "Long sequence feature extraction based on deep learning neural network for protein secondary structure prediction," in *Proceedings of the Information Technology and Mechatronics Engineering Conference (ITOEC)*, pp. 843–847, IEEE, 2017.

[57] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[58] C. Chang and C. Lin, "LIBSVM: A Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article no 27, 2011.

[59] H. Ding, Z. Y. Liang, F. B. Guo, J. Huang, W. Chen, and H. Lin, "Predicting bacteriophage proteins located in host cell with feature selection technique," *Computers in Biology and Medicine*, vol. 71, pp. 156–161, 2016.

[60] K. C. Chou and H. B. Shen, "Recent progress in protein subcellular location prediction," *Analytical Biochemistry*, vol. 370, no. 1, pp. 1–16, 2007.