IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON FEATURE REPRESENTATION AND LEARNING METHODS WITH APPLICATIONS IN LARGE-SCALE BIOLOGICAL SEQUENCE ANALYSIS

Received March 7, 2020, accepted March 14, 2020, date of publication March 20, 2020, date of current version March 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2982160

Using Evolutionary Information and Multi-Label Linear Discriminant Analysis to Predict the Subcellular Location of Multi-Site Bacterial Proteins via Chou's 5-Steps Rule

LEI DU, QINGFANG MENG, HUI JIANG, AND YANG LI
School of Information Science and Engineering, University of Jinan, Jinan 250022, China
Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan 250022, China

Corresponding author: Qingfang Meng (ise_mengqf@ujn.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61671220, Grant 61701192, Grant 51679058 and Grant 61640218, and in part by the Natural Science Foundation of Shandong Province, China under Grant No. ZR2017QF004.

ABSTRACT The function of a protein is closely tied to its subcellular location. Identifying the subcellular location of proteins is a crucial step to understand their functions. However, determining the subcellular location of proteins experimentally is time-consuming and costly. Therefore, developing effective computational methods to predict the subcellular positions of proteins is a hotspot in bioinformatics. Though many models have been proposed to improve the prediction accuracy of protein subcellular localization, there are still several shortcomings: (1) numerous methods ignore the multi-site proteins; (2) high dimensional features bring the burden to the construction of the prediction model. In this work, we proposed a method to predict the subcellular location of bacterial proteins with both single and multiple locations. Two features based on evolutionary information are extracted to solve the multi-site prediction problem, of which one is a 190-dimensional feature vector from absolute entropy correlation analysis (AECA-PSSM) and another is a 480-dimensional feature vector extracted using discrete wavelet transform (PSSM-DWT). After combining both proposed features, multi-label linear discriminant analysis (MLDA) is employed to transform the high-dimensional feature space into a lower-dimensional space. Multi-label k-nearest neighbors algorithm (ML-KNN) is utilized to predict the subcellular location of both single-site and multi-site proteins. Experimental results on Gram-positive dataset and Gram-negative dataset show the effectiveness of the proposed method.

INDEX TERMS Subcellular location, absolute entropy correlation analysis, discrete wavelet transform, multi-label linear discriminant analysis, multi-label k-nearest neighbors.

I. INTRODUCTION

The knowledge of subcellular location of proteins is very important which is closely associate with their function [1]. Only in specific subcellular location can a protein work and identifying the protein subcellular location can help to drug design and medical science. Biochemical experiments are the initial way to determine the relevant information of proteins and label them, but the process is time-consuming and costly. Meanwhile, facing the explosive growth of protein sequences discovered in the post-genomic age, it is impractical to receive the information of proteins only by biochemical experiments. To timely acquire useful information from these sequences for drug design, a lot of information have been successfully predicted using computational approaches, such as subcellular location of proteins [2], 3D structures of membrane proteins [3] and post-translational modification sites in proteins [4], [5]. Actually, the rapid development of bioinformatics has driven the medicinal chemistry undergoing an unprecedented revolution, in which the computational biology has played increasingly important roles in stimulating the development of finding novel drugs [6]. In view of this, this paper focus on developing a computational approach to identify the subcellular location of proteins.

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

Various remarkable computational methods have been proposed for prediction of protein subcellular location [7]–[13]. Many studies have shown that the design of feature extraction methods is the key of protein subcellular localization prediction. Amino acid composition (AAC) [14] is an early sequential-based feature extraction method which takes advantages of the information about the frequency of each type of amino acid in the protein sequence. In 2001, to obtain more information of the protein sequence, pseudo-amino acid composition (PseAAC) method [15] has been proposed by using the physicochemical properties. Later, extracted features based on evolutionary information which mainly obtained from the position-specific scoring matrix (PSSM) have gained attention. Liu *et al.* [16] combined the PSSM with an auto covariance transformation and called it PSSM-AC to predict the subcellular location of proteins in 2010. In 2015, Dehzangi *et al.* [17] proposed two segmentation based feature extraction methods from PSSM to predict the subcellular location of Gram-positive and Gram-negative proteins. In 2017, Xiang *et al.* [18] utilized the golden ratio to spit PSSM and segmented evolutionary information was calculated to represent protein sequences. Zhang and Liang [19] integrated Moran autocorrelation and cross correlation with PSSM as proposed in MACC-PSSM model. Over the years, features extracted from PSSM have been widely used in protein subcellular localization prediction and other domains [20]–[22]. All of these methods have shown that the potential features extracted based on evolutionary information can be employed in bioinformatics. Using Gene Ontogy (GO) information as feature extraction methods to predict the subcellular location of proteins has also been obtained a series of results [23]–[26]. Many of the methods mentioned above can only handle single-label systems where each protein just occurs in one subcellular location. However, with more experimental data uncovered, some proteins may simultaneously occur or move between two or more locations. In fact, multi-site proteins are widely found in living organisms and usually have some special functions worth noting, which are of great research value.

In recent years, some methods have been proposed to determine the subcellular location of bacterial proteins using Gene Ontology [27]–[32]. The GO describes the properties of genes and gene products in organisms, and it covers three aspects: cellular component, molecular function and biological process. However, using GO as a feature extraction method, it will generate a high-dimensional feature vector and the dimension of the feature will increase continuously with the update of the GO database [33]. Moreover, for new proteins, they have no GO information. So, using the GO terms as features to predict the subcellular location of proteins will lead to a heavy burden on the classifier and the prediction results may be inaccurate. Therefore, in this paper, we extracted features based on evolutionary information not employ GO terms to predict the subcellular location of proteins.

Characterization of proteins using multiple feature sets can solve the defect of insufficient information in a single feature set, but the dimension of the features will become much higher. And for protein sequences, sometimes even using a single feature set to represent a protein can generate a higher-dimensional feature vector, such as dipeptide composition (400-dimensional) [34] and GO (11118-dimensional) [35]. The high dimensional features often contain a lot of redundant and irrelevant information which may cause a degradation in classification performance and an increment of training time for building the model. Dimensionality reduction is an effective way to solve this problem. At present, there have been some works solving the subcellular location prediction problem with consideration of dimensionality reduction. Nogami *et al.* [36] utilized principal component analysis (PCA) to process protein data. Tang *et al.* [37] proposed a method named iAPSL-IF to identify the subcelular location of apoptosis protein using the SVM-RFE feature selection method. Yu *et al.* [38] proposed a model for the prediction of subcellular location of apoptosis proteins and in their work, local fisher discriminant analysis (LFDA) was employed to reduce the dimension of the features. Wang *et al.* [2] considered four global algorithms of dimensional reduction, including linear discriminant analysis (LDA), median LDA (MDA), generalized Fisher discriminant analysis (GDA), and median–mean line-based discriminant analysis (MMLDA) to map the high-dimensional data into a low-dimensional spaces. Though these above mentioned methods considered reducing the complexity of feature space, they did not take multi-site proteins into account. In other words, the dimensionality reduction method mentioned above may not be effective or applicable when solving the problem of subcellular localization prediction for multi-site proteins.

Considering the above mentioned problems, in this study, we aim to predict the subcellular location of bacterial proteins that contain both single-site and multi-site proteins. First, two discriminant features are extracted to explore evolutionary information embedded in position-specific scoring matrix. One is a 190-dimensional feature named absolute entropy correlation analysis (AECA-PSSM) which represents the relationship between each two attributes in PSSM. And another is a 480-dimensional feature named PSSM-DWT obtained by employing discrete wavelet transform which analyzes the time-frequency distribution of PSSM signal. Then, after combining the two features to generate a 670-dimensional fusion feature vector, multi-label linear discriminant analysis (MLDA) is used to eliminate the noise and reduce the dimension of the extracted features. Finally, all the samples are predicted by ML-KNN algorithm. The evaluation results indicate that our proposed method performs better than other existing models on the subcellular location prediction of proteins.

A series of publications and comprehensive review papers [22], [39]–[43] have demonstrated and summarized that, to establish a useful predictor for a biological system,

TABLE 1. Details of gram-position dataset.

Order	Subcellular locations	Number of proteins
1	Cell membrane	174
2	Cell wall	18
3	Cytoplasm	208
4	Extracellular	123
total number of locative proteins		523
tolal number of different proteins		515

TABLE 2. Details of gram-negative dataset.

Order	Subcellular locations	Number of proteins
1	Cell inner membrane	557
2	Cell outer membrane	124
3	Cytoplasm	410
4	Extracellular	133
5	Fimbrium	32
6	Flagellum	12
7	Nucleoid	8
8	Periplasm	180
total number of locative proteins		1456
tolal number of different proteins		1392

one needs to follow the Chou's 5-steps rule [44] to accomplish the following procedures: (1) construct or select a benchmark dataset to train and test the model; (2) formulate the samples with an effective expression that can truly reflect their intrinsic correlation with the target; (3) introduce or develop a powerful algorithm to construct the model; (4) properly perform cross-validation tests to objectively evaluate the performance of the model; (5) establish a user-friendly web-server for the proposed model. Below we will elaborate how to deal with these five steps one by one.

II. MATERIALS AND METHODS

A. DATA SET

According to Chou's 5-steps rule, the first step is how to construct or select a benchmark dataset to effectively train and test your model [44]. In this study, two widely used benchmark datasets are applied to predict the subcellular location of bacterial proteins. One is established for Gram-positive bacterial proteins [45] and another is Gram-negative bacterial proteins [32].

The Gram-positive dataset contains 519 different protein sequences located in four locations, of which 515 proteins belong to one subcellular location and 4 to two subcellular locations. Hence, there are 523 locative protein sequences in total. The details of the Gram-position bacteria dataset are given in Table 1.

The Gram-negative dataset contains 1392 different protein sequences located in eight locations, where 1328 proteins belong to one subcellular location and 64 to two subcellular locations. And there are 1456 locative protein sequences in total. The details of the Gram-position bacteria dataset are shown in Table 2.

It is worth noting that none of the proteins contained in the two datasets have $\geq 25\%$ paired sequences identity to any other proteins in the same location.

B. FEATURE EXTRACTION METHOD

One of the most important but also most difficult problems in computational biology is to formulate the biological sequence with a discrete model or a vector which known as feature extraction, because the classifiers can only handle the vectors rather than the sequence samples directly. To avoid completely lose of sequence-order information, PseAAC was proposed [15] and it has been widely used in nearly all the areas of computational proteomics [17], [46], [47]. Because of the widespread use of the concept of Chou's PseAAC, four powerful web-servers were established, including 'PseAAC' [48], 'PseAAC-Builder' [49], 'propy' [50] and 'PseAAC-General' [51]. Encouraged by the success of using PseAAC to deal with protein sequences, PseKNC (Pseudo K-tuple Nucleotide Composition) [52] was developed to extract features from DNA\RNA sequences. Particularly, a very powerful tool called 'Pse-in-One' [53] and its updated version 'Pse-in-One 2.0' [54] have been established to generate feature vectors for protein\peptide and DNA\RNA sequences. It is essential to develop effective feature extraction algorithms to express the protein sequence. Consider the 2nd rule of Chou's 5-steps rule [44], in this study, we propose to utilize two novel features extracted from PSSM to predict subcellular location of bacteria proteins.

1) POSITION-SPECIFIC SCORING MATRIX

Position-specific scoring matrix (PSSM) which obtained by PSI-BLAST, is a representation of evolutionary information of proteins [55], [56]. In our research, to obtain the PSSM for the proteins in our employed datasets, the E-value and the iterations numbers are set to 0.001 and 3, respectively. For a protein sequence with the length of N, its PSSM is an $N \times 20$ matrix (where N is the length of the protein and the columns represent 20 types of amino acids), which can be expressed as follows:

$$PSSM = \begin{bmatrix} E_{1\to1} & E_{1\to2} & \cdots & E_{1\to20} \\ E_{2\to1} & E_{2\to2} & \cdots & E_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ E_{i\to1} & E_{i\to2} & \cdots & E_{i\to20} \\ \vdots & \vdots & \vdots & \vdots \\ E_{N\to1} & E_{N\to2} & \cdots & E_{N\to20} \end{bmatrix} \quad (1)$$

where $E_{i\to j}$ represents the probability describing how the amino acid at the ith position in the protein sequence mutates into the j type of amino acid during evolution. After getting the PSSM, we can normalized the elements $E_{i\to j}$ by $P_{ij} = \frac{1}{1+e^{-E_{i\to j}}}$.

2) ABSOLUTE ENTROPY CORRELATION ANALYSIS (AECA-PSSM)

The elements P_{ij} in PSSM is the probability that the amino acid in the ith position replaced by a specific amino acid type j. Each column in PSSM is considered as one amino acid property, thus the PSSM can be treated as probability distributions of all properties. Within a PSSM, there are 20 columns in total, so that we can get 20 probability distributions for a PSSM. In order to measure the correlation between different properties, absolute entropy correlation analysis method is proposed.

Relative entropy, also known as Kullback-Leibler divergence (KL divergence or KLD) [57] describes the difference between two probability distributions P and Q, and it is an asymmetric measure. The relative entropy between two different probability distributions P and Q is as follows:

$$D_{KL}(P||Q) = \sum_{i=1}^{N} P(i)log\left(\frac{1}{Q(i)}\right) - \sum_{i=1}^{N} P(i)log\left(\frac{1}{P(i)}\right) = \sum_{i=1}^{N} P(i)log\left(\frac{P(i)}{Q(i)}\right) \quad (2)$$

The relative entropy is always non-negative on the basis of Gibbs inequality, and when it equals to 0, it means that the two probability distributions are the same ones. Because the relative entropy doesn't satisfy the commutative law, that is to say $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, if directly describe the relationship between each two columns in PSSM in terms of relative entropy, we need a $19 \times 20 = 380$-dimensional feature vector to fully extract the information in PSSM. Therefore, absolute entropy which is the modified form of relative entropy is defined as:

$$D(P, Q) = \frac{1}{2}(D_{KL}(P||Q) + D_{KL}(P||Q)) = \frac{1}{2}\sum_{i=1}^{N}(P(i) - Q(i))log\left(\frac{P(i)}{Q(i)}\right) \quad (3)$$

The absolute entropy is symmetric and can reflect the distance between two variables absolutely. The absolute entropy is also always equal or greater than zero, and when it is zero, it also represents that the two distributions are the same.

Since the length of different protein sequences are not the same, we average the final calculation by dividing the sequence length N to eliminate the effect of protein length. For the PSSM which we have stated to consider as 20 probability distributions, the absolute entropy correlation analysis is employed to analyze the pairwise relationship between each two columns in PSSM and the protein sequence information is extracted. By absolute entropy correlation analysis which is a symmetric form, a 190-dimensional feature is established. Compared to using relative entropy, the dimension of the feature is reduced by half and calculation is simpler.

3) DISCRETE WAVELET TRANSFORM, PSSM-DWT

Discrete wavelet transform (DWT) is an adaptable signal processing tool which analyses the signal by decomposing it into a series of coarse approximation and detail information [58]. It can capture information in both frequency and location content. Nanni *et al.* [59], [60] proposed an algorithm to represent a protein as an image and applied DWT to describe a protein by decomposing the matrix of the protein image into coefficients at different levels. Similar to Shen's work [40], in this paper, discrete wavelet transform is implemented to decompose the PSSM. Through experiments, 5-level discrete wavelet transform is applied to analyze the PSSM. Figure 1 is an example of 5-level decomposition of one column in PSSM. In each stage of this scheme, the approximation coefficient is decomposed with high-pass and low-pass filters and then down-sampled. Thus the columns in PSSM can be decomposed into one final approximation $A5$ and details $D1 - D5$.

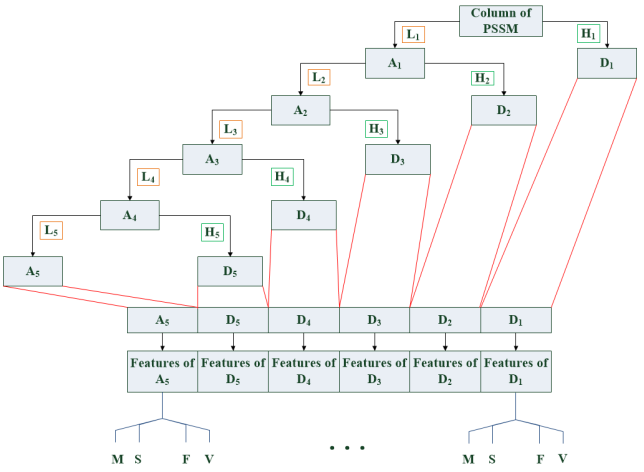


FIGURE 1. Process of the PSSM-DWT;H_n is the high-pass filter,L_n the low-pass filter.

To further describe the time-frequency distribution of the signal, four values are calculated from the different coefficients. The four features are as follows:

1) M: The mean of the absolute values of each sub-band coefficients.

2) S: The standard deviation of each sub-band coefficients.

3) F: The fluctuation index of each sub-band coefficients.

$$F = \frac{1}{N}\sum_{t=1}^{N-1}|c(t+1) - c(t)| \quad (4)$$

4) V: The variation coefficient of the absolute values in each sub-band coefficients.

$$V = \frac{\sigma}{\mu} \quad (5)$$

where

$$\mu = \frac{1}{N}\sum_{t=1}^{N}|c(t)|, \quad \sigma = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(|c(t)| - \mu)^2} \quad (6)$$

Feature M is the frequency distribution of the signal and feature S, F, and V represent the degree of changes in the frequency distribution. Finally, for a PSSM which has 20 columns in total, a 480-dimension feature vector is obtained.

C. MULTI-LABEL LINEAR DISCRIMINANT ANALYSIS FOR DIMENSIONALITY REDUCTION

Feature fusion can solve the defect of insufficient information in using a single feature set, so that fusing features calculated by different algorithms becomes an effective method to improve the accuracy of protein subcellular localization prediction. However, fusion features generally have higher dimensions and contain more redundant and irrelevant information, which may have a negative impact on the prediction.

An effective dimensionality reduction method can remove redundant and irrelevant information from the extracted features and improve the efficiency of classification [61]. Some bacterial proteins are located in more than one subcellular locations, so this is a multi-label problem and in this paper, we employ a multi-label dimensionality reduction method named multi-label linear discriminant analysis (MLDA) to reduce the dimension of the proposed features [62].

Given a dataset with n samples $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ and L classes, where $\mathbf{x}_i \in R^d$ is the features of a protein sample and $\mathbf{y}_i \in \{0,1\}^L$ is the corresponding class label of the sample. If $\mathbf{x}_i$ belongs to the l-th class, $\mathbf{y}_i(l) = 1$, otherwise, $\mathbf{y}_i(l) = 0$. We write $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T = [\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(L)}]$, where $\mathbf{y}_{(l)} \in \{0,1\}^n$.

To improve the classification accuracy, MLDA takes advantages of label interactions through label correlations. The label correlation between two classes is defined as follows:

$$C_{lk} = cos(\mathbf{y}_{(l)}, \mathbf{y}_{(k)}) = \frac{\langle \mathbf{y}_{(l)}, \mathbf{y}_{(k)} \rangle}{\|\mathbf{y}_{(l)}\| \|\mathbf{y}_{(k)}\|} \tag{7}$$

Moreover, in order to solve the over-counted problem of multi-label samples, the following normalized matrix $Z = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n]^T \in R^{n \times L}$ is employed:

$$\mathbf{z}_i = \frac{\mathbf{y}_i C}{\|\mathbf{y}_i\|_{\ell 1}} \tag{8}$$

where $\|.\|_{\ell 1}$ is the $\ell 1$-norm of a vector. The class-wise within-class scatter matrix S^ω and the class-wise between-class scatter matrix S^b are separately defined as:

$$\begin{aligned} S_\omega &= \sum_{l=1}^{L} S_\omega^{(l)} \\ &= \sum_{l=1}^{L} \sum_{i=1}^{n} Z_{il} (\mathbf{x}_i - \mathbf{m}_l)(\mathbf{x}_i - \mathbf{m}_l)^T \end{aligned} \tag{9}$$

$$\begin{aligned} S_b &= \sum_{l=1}^{L} S_b^{(l)} \\ &= \sum_{l=1}^{L} \left(\sum_{i=1}^{n} Z_{il} \right) (\mathbf{m}_l - \mathbf{m})(\mathbf{m}_l - \mathbf{m})^T \end{aligned} \tag{10}$$

where $\mathbf{m}_l = \frac{\sum_{i=1}^n Y_{il} \mathbf{x}_i}{\sum_{i=1}^n Y_{il}}$ is the mean of samples of the class l, and $\mathbf{m} = \frac{\sum_{l=1}^L \sum_{i=1}^n Y_{il} \mathbf{x}_i}{\sum_{l=1}^L \sum_{i=1}^n Y_{il}}$ is the multi-label global mean of all samples. MLDA maps the samples in the high dimensional space to a lower dimensional space, it tries to maximize S_b and minimize S_ω, the optimization objective is as follows:

$$W = \arg\max_{W \in R^{d \times p}} \left[tr \left(\frac{W^T S_b W}{W^T S_\omega W} \right) \right] \tag{11}$$

where W is the projection matrix and is constructed by solving the eigenvalue problem $S_w^{-1} S_b v = \lambda v$, $p < L$. And matrix F after dimensionality reduction can be obtained by $F = W^T X$. Therefore, the dimension of the original high-dimensional features is reduced and the redundant information obtained in the original protein sequence feature is eliminated. In other words, after getting the sample matrix X, through the projection matrix W, matrix F with reduced dimension can be obtained. Then, we get features with more discriminating ability and decrease the computation complexity of the classification model.

D. MULTI-LABEL K-NEAREST NEIGHBOR, ML-KNN

Classification algorithm plays an important role in predicting the subcellular localization of proteins, but most of the classifiers only focus on the protein sequences located in one subcellular location and can not handle multiple sites proteins. Numerous proteins have been found to be located in two or more subcellular locations, so it is vital to explore effective predictive algorithms to identify the subcellular locations for both single- and multi-site proteins. Now according to the 3rd rule of Chou's 5-steps rule [44], in this paper, ML-KNN is chosen to solve this problem [63]. Given a instance x and its corresponding label set $\mathbf{y}$, if x belongs to the l-th class, $\mathbf{y}_x(l) = 1$, otherwise $\mathbf{y}_x(l) = 0$. Let $N(x)$ represents the k nearest neighbors of x in the training set. The membership counting vector can be calculated as:

$$\mathbf{C}_x(l) = \sum_{a \in N(x)} \mathbf{y}_a(l) \tag{12}$$

where $\mathbf{C}_x(l)$ represents the number of neighbors of x which belong to class l.

Let H_1^l denotes that the instance has label l, while H_0^l is not. In addition, E_j^l is defined as the event that in the k nearest neighbors of the instance, there are exactly j cases have the label l. First, calculate the prior probabilities $P(H_b^l)(b \in \{0,1\})$ and the posterior probabilities $P(E_j^l | H_b^l)(j \in 0, 1, \cdots, k)$. All of them can be directly established from the training set.

$$P(H_1^l) = \frac{s + \sum_{i=1}^m \vec{y}_{x_i}(l)}{s \times 2 + m}, P(H_0^l) = 1 - P(H_1^l) \tag{13}$$

$$P(E_j^l | H_1^l) = \frac{s + c[j]}{s \times (k+1) + \sum_{p=0}^k c[p]} \tag{14}$$

$$P(E_j^l | H_0^l) = \frac{s + c'[j]}{s \times (k+1) + \sum_{p=0}^k c'[p]} \tag{15}$$

where $c[j]$ counts the number of instances in training set with label l whose the k nearest neighbors has exactly j instances with label l, $c'[j]$ is similar to $c[j]$, it counts the number of training instances unlabeled label l but whose the k nearest neighbors has exactly j instances with label l. And s is a smoothing parameter which is set to be 1.

Finally, for a test instance t, let $N(t)$ represents the k nearest neighbors of t in the training set. The membership counting vector can be calculated using $\mathbf{C}_t(l) = \sum_{a \in N(t)} \mathbf{y}_a(l)$. And the category vector of t is obtained using the following maximum a posterior principle:

$$\begin{aligned} \vec{y}_t(l) &= \arg\max_{b \in \{0,1\}} P(H_b^l | E^l_{\mathbf{C}_t(l)}) \\ &= \arg\max_{b \in \{0,1\}} P(H_b^l) P(E^l_{\mathbf{C}_t(l)} | H_b^l) \end{aligned} \tag{16}$$

In other words, if $b = 1$ makes $P(H_b^l|E^l_{\mathbf{C}_t(l)})$ bigger, then $y_t(l) = 1$, else $y_t(l) = 0$.

E. MODEL VALIDATION AND PERFORMANCE EVALUATION

When considering the model validation method following the 4th rule in Chou's 5-steps rule [44], in this paper, the jackknife cross-validation method which is considered to be the most reasonable and objective is used [64]. For a given dataset with N instances, the principle of the jackknife test is to select one individual in the data set as an independent test sample and the remaining $N-1$ individuals as the training set until all the individuals in the dataset have been tested. A definite result is obtained after jackknife test.

In multi-label classification problems, some evaluation metrics are used for performance measurement to better evaluate the capabilities of the multi-label classifiers. Two evaluation metrics named overall locative accuracy (OLA) and overall actual accuracy (OAA) which are often used in multi-label subcellular location prediction are used in this paper. Here we denote that $\mathcal{M}(Q_i)$ is the predicted label of the i-th sample and $\mathcal{L}(Q_i)$ is the true label of the i-th sample. The overall locative accuracy (OLA) is:

$$OLA = \frac{1}{\sum_{i=1}^{N} |\mathcal{L}(Q_i)|} \sum_{i=1}^{N} |\mathcal{M}(Q_i) \cap \mathcal{L}(Q_i)| \tag{17}$$

and the overall actual accuracy (OAA) is:

$$OAA = \frac{1}{N} \sum_{i=1}^{N} \bigtriangleup[\mathcal{M}(Q_i), \mathcal{L}(Q_i)] \tag{18}$$

where

$$\bigtriangleup[\mathcal{M}(Q_i), \mathcal{L}(Q_i)] = \begin{cases} 1, & if\ \mathcal{M}(Q_i) = \mathcal{L}(Q_i) \\ 0, & otherwise \end{cases} \tag{19}$$

The flowchart of the proposed subcellular location prediction model is detailed in Figure 2. Using flowchart or graphic approaches to study biological and medical systems can provide an intuitive vision and useful insights for helping analyze complicated relations therein as shown by the pioneering papers from the then Chairman of Nobel Prize Committee Sture Forsen [65].

Step 1: Input the protein samples in Gram-positive dataset and Gram-negative dataset, respectively. Using the proposed feature expression methods AECA-PSSM and PSSM-DWT to calculate features of proteins.

Step 2: Using MLDA method to reduce the dimension of the feature vector and remove redundant information.

Step 3: Employing ML-KNN to predict the subcellular locations of the protein samples.

III. EXPERIMENTAL RESULTS

In this paper, we test our method on two bacterial protein datasets and analyze it from different aspects. First, we analyze the performance of AECA-PSSM and compare it with traditional distance-based methods. Then, for the feature PSSM-DWT, we use jackknife test to measure the overall locative accuracy and overall actual accuracy corresponding to different wavelet functions and different decomposition scales. We also discuss the selection of number of nearest neighbors (k) for the ML-KNN classifier and the dimension of reduction (p) for the fusion feature. It is worth noting that, the parameters of this model are optimized based on the Gram-positive dataset. After that, we analyze the contribution of different feature extraction methods. Finally, we compare our method with other existing methods.

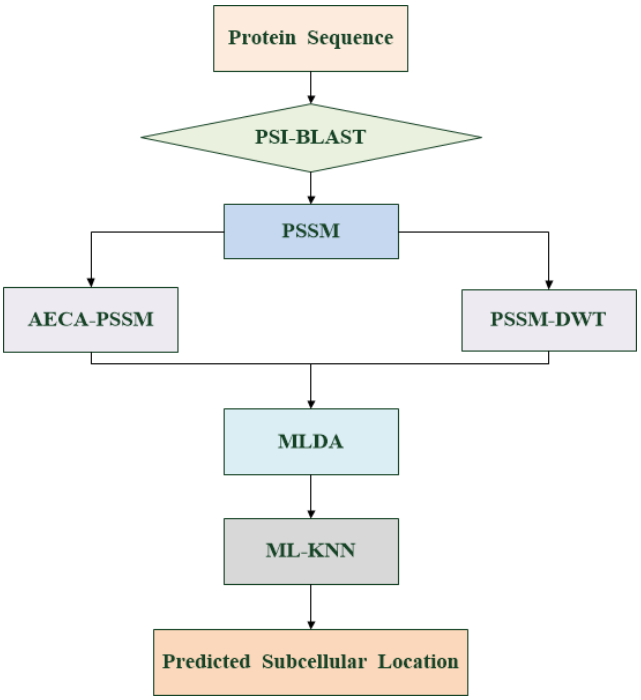


FIGURE 2. Pipeline of the proposed method.

TABLE 3. Comparison with traditional distance-based method.

Feature	Jackknife test (%)			
	Gram-positive		Gram-negative	
	OLA	OAA	OLA	OAA
CD-PSSM	62.52	62.62	57.55	55.60
ED-PSSM	65.58	64.55	58.65	57.47
AECA-PSSM	65.39	65.51	62.36	59.70

A. PERFORMANCE OF ABSOLUTE ENTROPY CORRELATION ANALYSIS

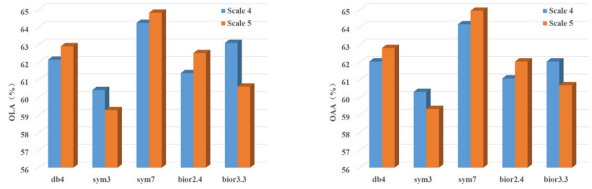
In this paper, we use absolute entropy correlation analysis to obtain the difference between each two columns in PSSM. Table 3 shows the comparison between our method and two traditional distance-based methods: cosine distance (CD-PSSM) and euclidean distance (ED-PSSM). It can be seen that while using AECA-PSSM to obtain the information in PSSM, better results are obtained compared with CD-PSSM and ED-PSSM.

B. SELECTION OF WAVELET FUNCTION AND DECOMPOSITION SCALE

Since different wavelet functions have different processing power for different signals, choosing appropriate wavelet function to process signals will obtain better feature information from the PSSM of the protein sequence.

TABLE 4. Prediction results with different wavelet functions and different decomposition scales on Gram-positive dataset.

Evaluation		Jackknife test(%)				
		db4	sym3	sym7	bior2.4	bior3.3
OLA	4	62.14	60.42	64.24	61.38	63.10
	5	62.91	59.27	64.82	62.52	60.61
OAA	4	62.04	60.31	64.16	61.08	62.04
	5	62.81	59.34	64.93	62.04	60.69



(a) Overall locative accuracy (b) Overall actual accuracy

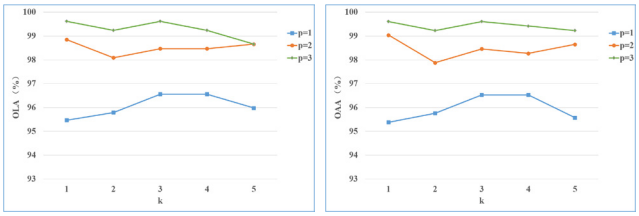
FIGURE 3. Prediction results of Gram-positive dataset under different wavelet functions and decomposition scales.

In addition, different decomposition scales also affect the analysis of protein sequences. In order to obtain a better performance of the extracted feature to represent the protein sequences, we test different wavelet functions and decomposition scales on Gram-positive dataset to extract PSSM-DWT feature.

It can be seen from Table 4 that different wavelet functions and different decomposition scales affect the predicted performance of protein subcellular localization. For the Gram-positive bacteria proteins, when the decomposition scale is 5 and wavelet function is sym7, the highest overall locative accuracy 64.82% and overall actual accuracy 64.93% are obtained. In order to analyze the results of different wavelet functions and decomposition scales more intuitively, as is shown in Figure 3, we draw histograms of the overall locative accuracy and overall actual accuracy under different decomposition scales and wavelet functions. It is worth noting that, sym7 wavelet achieves the highest overall locative accuracy and overall actual accuracy at different decomposition scales. In this paper, sym7 wavelet and 5 decomposition scale are chosen.

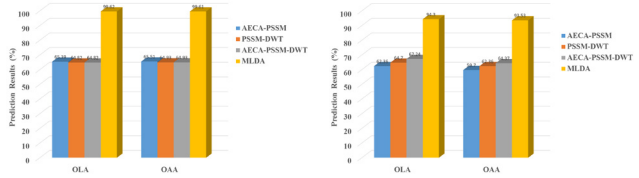
C. SELECTION OF THE NUMBER OF NEAREST NEIGHBORS (K) AND THE DIMENSION OF REDUCTION (P)

The prediction results of the subcellular location are affected by the number of nearest neighbors (k) in the ML-KNN classifier and the dimension of reduction (p) for the fusion feature. We test the parameter k with different values from 1 to 5 and the parameter p with different values from 1 to $L - 1 = 3$ on Gram-positive dataset, and the results are shown in Figure 4. From Figure 4, we can see that on the Gram-positive dataset, when k changes from 1 to 5, the two evaluation measurements (OLA and OAA) are highest when the parameter p equals to 3. When the dimension of reduction is 2 or 3, the OLA and OAA can be more than 97% in all five cases of different nearest neighbors. When $k = 1$ and $p = 3(L - 1)$ is selected, the highest OLA and OAA achieved.



(a) Overall locative accuracy (b) Overall actual accuracy

FIGURE 4. The jackknife test performance changes as k increases with fixed value of p.



(a) Gram-positive (b) Gram-negative

FIGURE 5. Classification results of different feature extraction methods.

D. EFFECT OF DIFFERENT FEATURE EXTRACTION METHODS

In this paper, protein sequence information is extracted using two feature expression methods based on evolutionary information reflected in PSSM. To investigate the performance of the two different features in predicting bacterial proteins, we test each feature using ML-KNN ($k = 1$) algorithm on Gram-positive dataset and Gram-negative dataset. Then, we combined the two features AECA-PSSM and PSSM-DWT to form a fusion feature vector: AECA-PSSM-DWT. After that, MLDA is employed to reduce the dimension of the feature vector as well as eliminate redundant and irrelevant information for multi-label system. According to the experiment above, the extracted fusion feature vector is transformed into a $p = L - 1$ dimensional projection subspace. The results are listed in Figure 5.

In Figure 5, we find that the AECA-PSSM performs better than PSSM-DWT on Gram-positive dataset, and the performance of PSSM-DWT feature is better than AECA-PSSM on Gram-negative dataset. On Gram-negative dataset, while using the fusion feature AECA-PSSM-DWT, we can get a better prediction result which reaches overall locative accuracy of 67.24% and overall actual accuracy of 64.37%. On Gram-positive dataset, while using the fusion feature, the overall locative accuracy and overall actual accuracy have no improvement. The reason is that using the two feature extraction methods, we can only get more information of the proteins, but the redundant and irrelevant information in feature vectors can not be removed which may limits the classifier performance. And after dimensaionality reduction using MLDA, the prediction results are significantly improved on the Gram-positive bacteria dataset and Gram-negative bacteria dataset. On Gram-positive dataset, the overall locative accuracy achieves 99.62% and the overall actual accuracy achievs 99.61% by MLDA algorithm. On Gram-negative dataset, the overall locative accuracy achieves 94.30% and the overall actual accuracy achieves 93.53% by MLDA

TABLE 5. Comparison from different methods on gram-positive dataset by jackknife test.

Location	Jackknife test					
	Gpos-PLoc [29]	Gpos-mPLoc [45]	iLoc-Gpos [28]	Gpos-ECC-mPLoc [31]	Gram-LocEN [30]	This paper
Cell membrane	-	-	96.0%	96.5%	97.7%	98.9%
Cell wall	-	-	66.7%	66.7%	94.4%	100%
Cytoplasm	-	-	95.2%	96.2%	97.1%	100%
Extracellular	-	-	89.4%	92.7%	95.1%	100%
OLA	72.5%	82.2%	93.1%	94.4%	96.8%	99.6%
OAA	-	-	92.9%	94.0%	96.3%	99.6%

TABLE 6. Comparison from different methods on gram-negative dataset by jackknife test.

Location	Jackknife test					
	Gneg-PLoc [32]	Gneg-mPLoc [27]	iLoc-Gneg [66]	Gneg-ECC-mPLoc [31]	Gram-LocEN [30]	This paper
Cell inner membrane	81.5%	94.3%	96.8%	95.5%	97.1%	96.8%
Cell outer membrane	54.8%	84.7%	83.1%	94.4%	89.5%	90.3%
Cytoplasm	88.3%	87.1%	89.5%	92.2%	92.4%	95.4%
Extracellular	44.4%	59.4%	86.5%	93.2%	97.0%	88.7%
Fimbrium	34.4%	87.5%	93.8%	93.8%	100%	96.9%
Flagellum	0%	0%	100%	100%	100%	100%
Nucleoid	0%	0%	50.0%	87.5%	87.5%	87.5%
Periplasm	48.3%	85.6%	89.4%	94.4%	89.4%	90.6%
OLA	71.5%	85.7%	91.4%	94.1%	95.3%	94.3%
OAA	-	-	89.9%	92.4%	94.5%	93.5%

dimensionality reduction. It indicates that MLDA efficiently remove the redundant and irrelevant information and obtain discriminative features which improve the performance of predicting the subcellular location of bacterial proteins.

E. COMPARISON WITH OTHER EXITING PREDICTION METHODS

Finally, we compare our proposed method with other exiting GO based methods for predicting the subcellular locations of bacterial proteins to investigate the performance of our method. The prediction results are listed in Table 5 and Table 6.

As shown in Table 5, in terms of predicting the subcellular location of Gram-positive bacterial protein, our proposed method achieves 99.6% overall actual accuracy and 99.6% overall locative accuracy, which is better than other mentioned predictors. Specifically, The OLA of the proposed method for the Gram-positive bacterial dataset is 2.8%–27.1% higher than other methods and the OAA of the proposed method is 6.7%, 5.6%, 3.3% better than iLoc-Gpos, Gpos-ECC-mPLoc and Gram-LocEN, respectively.

As shown in Table 6, for Gram-negative dataset, our method achieves 94.3% overall locative accuracy and 93.5% overall actual accuracy which performs better than the first four predictors. The OLA of Gram-negative dataset is 0.2%–22.8% higher than the first four methods and the OAA is 3.6%, 1.1% higher than iLoc-Gneg and Gneg-ECC-mPLoc, respectively. Compared with Gram-LocEN, the OLA of our method is 1% lower and the OAA is 1% lower. But in this paper, the dimension of the feature vector in Gram-negative dataset is only 7 and the results are acceptable.

We also list three other measurements in Table 7 which are used in [33], [46] to evaluate the performance of our method. The average precision is the high the better while coverage and ranking loss are the lower the better.

In order to further explore the universality of our proposed method in the field of predicting the subcellular location of protein, we test it on the Virus protein dataset. The Virus dataset contains proteins located in six subcellular locations. The details of the Virus protein dataset are shown in [23]. For the Virus dataset, we also directly used optimized parameters obtained from the Gram-positive dataset: sym7 wavelet, 5 scales, $k = 1$ and $p = L - 1$ ($p = 5$). We can find that the overall actual accuracy can reach 97.1% in Virus dataset. The prediction results on the Virus dataset are also satisfactory.

TABLE 7. Results of average precision, coverage and ranking loss.

Dataset	Jackknife test		
	average precision↑	coverage ↓	ranking loss ↓
Gram-positive	0.9990	0.0154	0.0019
Gram-negative	0.9705	0.1695	0.0159

TABLE 8. Results on virus dataset.

	OLA	OAA	average precision	coverage	ranking loss
AD-SVM [67]	96.0%	93.2%	-	-	-
Javed et al. [46]	-	94.7%	0.88	0.514	0.065
This paper	98.8%	97.1%	0.9964	0.2319	0.0024

IV. CONCLUSION

In this paper, we proposed a novel method based on evolutionary information and multi-label linear discriminant analysis to predict the subcellular location of bacterial proteins via Chou's 5-steps rule. We extract two novel features based on the evolutionary information embedded in PSSM, namely AECA-PSSM and PSSM-DWT. Through the two feature extraction methods, effective evolutionary information about the protein sequence is obtained from PSSM. After fusing the two features, MLDA is applied to reduce the dimension and complexity of the fusion feature by mapping the fusion feature vector into a lower feature space. And ML-KNN is used to solve the multi-label problem for predicting the subcellular location of multi-site bacterial proteins. It can be concluded that the proposed method is rational and feasible to predict the subcellular location of multi-site Gram-positive bacterial proteins and Gram-negative bacterial proteins. We also test our method on Virus dataset and obtain satisfactory results.

According to rule 5 of Chou's 5-steps rule [44] and a series of recent publications [68]–[70] in demonstrating new findings or approaches, user-friendly and publicly accessible web-servers will significantly enhance their impacts on medical science [42] and driving medicinal chemistry into an unprecedented revolution [47]. We shall make efforts in our future work to provide a web-server for our proposed method.

REFERENCES

[1] 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, Jan. 2018.
[2] S. Wang, W. Li, Y. Fei, Z. Cao, D. Xu, and H. Guo, "An improved process for generating uniform PSSMs and its application in protein subcellular localization via various global dimension reduction techniques," *IEEE Access*, vol. 7, pp. 42384–42395, 2019.
[3] K.-C. Chou, "Structural bioinformatics and its impact to biomedical science," *Current Medicinal Chem.*, vol. 11, no. 16, pp. 2105–2134, Aug. 2004.
[4] Z. Ju and S.-Y. Wang, "Prediction of lysine formylation sites using the composition of k-spaced amino acid pairs via Chou's 5-steps rule and general pseudo components," *Genomics*, vol. 112, no. 1, pp. 859–866, Jan. 2020.
[5] K. Chou, "Progresses in predicting post-translational modification," *Int. J. Peptide Res. Therapeutics*, 2019, doi: 10.1007/s10989-019-09893-5.
[6] K.-C. Chou, "Impacts of bioinformatics to medicinal chemistry," *Medicinal Chem.*, vol. 11, no. 3, pp. 218–234, Mar. 2015.
[7] S. Hua and Z. Sun, "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics*, vol. 17, no. 8, pp. 721–728, Aug. 2001.
[8] H. Lin, H. Ding, F.-B. Guo, A.-Y. Zhang, and J. Huang, "Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition," *Protein Peptide Lett.*, vol. 15, no. 7, pp. 739–744, 2008.
[9] L. Zhang, B. Liao, D. Li, and W. Zhu, "A novel representation for apoptosis protein subcellular localization prediction using support vector machine," *J. Theor. Biol.*, vol. 259, no. 2, pp. 361–365, Jul. 2009.
[10] H. Saini, G. Raicar, A. Dehzangi, S. Lal, and A. Sharma, "Subcellular localization for gram positive and gram negative bacterial proteins using linear interpolation smoothing model," *J. Theor. Biol.*, vol. 386, pp. 25–33, Dec. 2015.
[11] C. Song and F. Shi, "Subcellular location of apoptosis proteins based on chaos game representation," in *Proc. Int. Conf. Future Biomed. Inf. Eng. (FBIE)*, Dec. 2009, pp. 194–196.
[12] B. Yu, S. Li, C. Chen, J. Xu, W. Qiu, X. Wu, and R. Chen, "Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition," *Chemometric Intell. Lab. Syst.*, vol. 167, pp. 102–112, Aug. 2017.
[13] Y. Liang, S. Liu, and S. Zhang, "Geary autocorrelation and DCCA coefficient: Application to predict apoptosis protein subcellular localization via PSSM," *Phys. A, Stat. Mech. Appl.*, vol. 467, pp. 296–306, Feb. 2017.
[14] T. Habib, C. Zhang, J. Y. Yang, M. Q. Yang, and Y. Deng, "Supervised learning method for the prediction of subcellular localization of proteins using amino acid and amino acid pair composition," *BMC Genomics*, vol. 9, no. 1, p. S16, 2008.
[15] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Struct., Function, Bioinf.*, vol. 43, no. 3, pp. 246–255, 2001.
[16] T. Liu, X. Zheng, C. Wang, and J. Wang, "Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: An approach from auto covariance transformation," *Protein Peptide Lett.*, vol. 17, no. 10, pp. 1263–1269, Oct. 2010.
[17] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, and A. Sattar, "Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC," *J. Theor. Biol.*, vol. 364, pp. 284–294, Jan. 2015.
[18] Q. Xiang, B. Liao, X. Li, H. Xu, J. Chen, Z. Shi, Q. Dai, and Y. Yao, "Subcellular localization prediction of apoptosis proteins based on evolutionary information and support vector machine," *Artif. Intell. Med.*, vol. 78, pp. 41–46, May 2017.
[19] S. Zhang and Y. Liang, "Predicting apoptosis protein subcellular localization by integrating auto-cross correlation and PSSM into Chou's PseAAC," *J. Theor. Biol.*, vol. 457, pp. 163–169, Nov. 2018.
[20] P. Mundra, M. Kumar, K. K. Kumar, V. K. Jayaraman, and B. D. Kulkarni, "Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM," *Pattern Recognit. Lett.*, vol. 28, no. 13, pp. 1610–1615, Oct. 2007.
[21] S. Wan, M.-W. Mak, and S.-Y. Kung, "Ensemble linear neighborhood propagation for predicting subchloroplast localization of multi-location proteins," *J. Proteome Res.*, vol. 15, no. 12, pp. 4755–4762, Dec. 2016.
[22] H.-C. Yi, Z.-H. You, D.-S. Huang, X. Li, T.-H. Jiang, and L.-P. Li, "A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information," *Mol. Therapy-Nucleic Acids*, vol. 11, pp. 337–344, Jun. 2018.
[23] H.-B. Shen and K.-C. Chou, "Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells," *Biopolymers*, vol. 85, no. 3, pp. 233–240, Feb. 2007.
[24] 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, 2014, Art. no. e89545.
[25] X. Xiao, Z.-C. Wu, and K.-C. Chou, "ILoc-virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites," *J. Theor. Biol.*, vol. 284, no. 1, pp. 42–51, Sep. 2011.
[26] S. Wan, M.-W. Mak, and S.-Y. Kung, "MGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines," *BMC Bioinf.*, vol. 13, no. 1, p. 290, 2012.
[27] H.-B. Shen and K.-C. Chou, "Gneg-mPLoc: A top-down strategy to enhance the quality of predicting subcellular localization of gram-negative bacterial proteins," *J. Theor. Biol.*, vol. 264, no. 2, pp. 326–333, May 2010.
[28] Z.-C. Wu, X. Xiao, and K.-C. Chou, "iLoc-Gpos: A multi-layer classifier for predicting the subcellular localization of singleplex and multiplex Gram-positive bacterial proteins," *Protein Peptide Lett.*, vol. 19, no. 1, pp. 4–14, Jan. 2012.
[29] H.-B. Shen and K.-C. Chou, "Gpos-PLoc: An ensemble classifier for predicting subcellular localization of gram-positive bacterial proteins," *Protein Eng., Des. Selection*, vol. 20, no. 1, pp. 39–46, Jan. 2007.
[30] 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," *Chemometric Intell. Lab. Syst.*, vol. 162, pp. 1–9, Mar. 2017.
[31] X. Wang, J. Zhang, and G.-Z. Li, "Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble," *BMC Bioinf.*, vol. 16, no. S12, p. S1, 2015.
[32] K.-C. Chou and H.-B. Shen, "Large-scale predictions of Gram-negative bacterial protein subcellular locations," *J. Proteome Res.*, vol. 5, no. 12, pp. 3420–3428, Dec. 2006.
[33] L. Wei, M. Liao, X. Gao, J. Wang, and W. Lin, "MGOF-loc: A novel ensemble learning method for human protein subcellular localization prediction," *Neurocomputing*, vol. 217, pp. 73–82, Dec. 2016.
[34] M. Bhasin and G. P. S. Raghava, "ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST," *Nucleic Acids Res.*, vol. 32, pp. W414–W419, Jul. 2004.
[35] K.-C. Chou, Z.-C. Wu, and X. Xiao, "iLoc-Hum: Using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Mol. BioSyst.*, vol. 8, no. 2, pp. 629–641, 2012.
[36] D. Nogami, Y. Nakano, and Y. Taguchi, "Protein subcellular location prediction using principal component analysis," IPSJ SIG Tech. Rep. 2013-BIO-34, Jun. 2013, pp. 1–2.
[37] Y. Tang, L. Xie, and L. Chen, "IAPSL-IF: Identification of apoptosis protein subcellular location using integrative features captured from amino acid sequences," *Int. J. Mol. Sci.*, vol. 19, no. 4, p. 1190, 2018.
[38] B. Yu, S. Li, W. Qiu, M. Wang, J. Du, Y. Zhang, and X. Chen, "Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction," *BMC Genomics*, vol. 19, no. 1, p. 478, Dec. 2018.
[39] M. Kabir, S. Ahmad, M. Iqbal, and M. Hayat, "INR-2L: A two-level sequence-based predictor developed via Chou's 5-steps rule and general PseAAC for identifying nuclear receptors and their families," *Genomics*, vol. 112, no. 1, pp. 276–285, Jan. 2020.
[40] Y. Shen, J. Tang, and F. Guo, "Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC," *J. Theor. Biol.*, vol. 462, pp. 230–239, Feb. 2019.

[41] K.-C. Chou, "Impacts of pseudo amino acid components and 5-steps rule to proteomics and proteome analysis," *Current Topics Medicinal Chem.*, vol. 19, no. 25, pp. 2283–2300, Nov. 2019.
[42] K.-C. Chou, "Advances in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs," *Current Medicinal Chem.*, vol. 26, no. 26, pp. 4918–4943, Oct. 2019.
[43] K.-C. Chou, "An insightful 10-year recollection since the emergence of the 5-steps rule," *Current Pharmaceutical Des.*, vol. 25, no. 40, pp. 4223–4234, Jan. 2020.
[44] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, Mar. 2011.
[45] H.-B. Shen and K.-C. Chou, "Gpos-mPLoc: A top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins," *Protein Peptide Lett.*, vol. 16, no. 12, pp. 1478–1484, Dec. 2009.
[46] F. Javed and M. Hayat, "Predicting subcellular localization of multi-label proteins by incorporating the sequence features into Chou's PseAAC," *Genomics*, vol. 111, no. 6, pp. 1325–1332, Dec. 2019.
[47] K.-C. Chou, "An unprecedented revolution in medicinal chemistry driven by the progress of biological science," *Current Topics Medicinal Chem.*, vol. 17, no. 21, pp. 2337–2358, Jul. 2017.
[48] H.-B. Shen and K.-C. Chou, "PseAAC: A flexible Web server for generating various kinds of protein pseudo amino acid composition," *Anal. Biochem.*, vol. 373, no. 2, pp. 386–388, Feb. 2008.
[49] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions," *Anal. Biochem.*, vol. 425, no. 2, pp. 117–119, 2012.
[50] D.-S. Cao, Q.-S. Xu, and Y.-Z. Liang, "propy: A tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, no. 7, pp. 960–962, 2013.
[51] P. Du, S. Gu, and Y. Jiao, "PseAAC-General: Fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets," *Int. J. Mol. Sci.*, vol. 15, no. 3, pp. 3495–3506, 2014.
[52] W. Chen, T.-Y. Lei, D.-C. Jin, H. Lin, and K.-C. Chou, "PseKNC: A flexible Web server for generating pseudo K-tuple nucleotide composition," *Anal. Biochem.*, vol. 456, pp. 53–60, Jul. 2014.
[53] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou, "Pse-in-one: A Web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W65–W71, Jul. 2015.
[54] B. Liu, H. Wu, and K.-C. Chou, "Pse-in-One 2.0: An improved package of Web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Natural Sci.*, vol. 09, no. 04, pp. 67–91, 2017.
[55] S. Altschul, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.
[56] J. Wang, B. Yang, J. Revote, A. Leier, T. T. Marquez-Lago, G. Webb, J. Song, K.-C. Chou, and T. Lithgow, "POSSUM: A bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles," *Bioinformatics*, vol. 33, no. 17, pp. 2756–2758, Sep. 2017.
[57] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.
[58] P. M. Shemi and E. M. Shareena, "Analysis of ECG signal denoising using discrete wavelet transform," in *Proc. IEEE Int. Conf. Eng. Technol. (ICETECH)*, Mar. 2016, pp. 713–718.
[59] L. Nanni, S. Brahnam, and A. Lumini, "Wavelet images and Chou's pseudo amino acid composition for protein classification," *Amino Acids*, vol. 43, no. 2, pp. 657–665, 2012.
[60] L. Nanni, A. Lumini, and S. Brahnam, "An empirical study of different approaches for protein classification," *Sci. World J.*, vol. 2014, Jun. 2014, Art. no. 236717.
[61] T. Wang and J. Yang, "Predicting subcellular localization of Gram-negative bacterial proteins by linear dimensionality reduction method," *Protein Peptide Lett.*, vol. 17, no. 1, pp. 32–37, Jan. 2010.
[62] H. Wang, C. Ding, and H. Huang, "Multi-label linear discriminant analysis," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer-Verlag, 2010, pp. 126–139.
[63] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.
[64] M. R. Uddin, A. Sharma, D. M. Farid, M. M. Rahman, A. Dehzangi, and S. Shatabda, "EvoStruct-sub: An accurate gram-positive protein subcellular localization predictor using evolutionary and structural features," *J. Theor. Biol.*, vol. 443, pp. 138–146, Apr. 2018.
[65] C. Kuochen, R. Carter, and S. Forsen, "A new graphical-method for deriving rate-equations for complicated mechanisms," *Chem. Scripta*, vol. 18, no. 2, pp. 82–86, 1981.
[66] X. Xiao, Z.-C. Wu, and K.-C. Chou, "A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites," *PLoS ONE*, vol. 6, no. 6, 2011, Art. no. e20592.
[67] S. Wan and M.-W. Mak, "Predicting subcellular localization of multi-location proteins by improving support vector machines with an adaptive-decision scheme," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 3, pp. 399–411, Mar. 2018.
[68] 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.
[69] X. Cheng, X. Xiao, and K.-C. Chou, "PLoc-mVirus: Predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC," *Gene*, vol. 628, pp. 315–321, Sep. 2017.
[70] X. Xiao, X. Cheng, G. Chen, Q. Mao, and K.-C. Chou, "pLoc_bal-mVirus: Predict subcellular localization of multi-label virus proteins by Chou's general PseAAC and IHTS treatment to balance training dataset," *Medicinal Chem.*, vol. 15, no. 5, pp. 496–509, 2019.

LEI DU received the B.S. degree in electronic information science and technology from the School of Information Science and Engineering, University of Jinan, Jinan, China, in 2018, where she is currently pursuing the M.S. degree in signal and information processing. Her current research interests include machine learning and bioinformatics.

QINGFANG MENG was born in 1979. She received the master's and Ph.D. degrees from the School of Information Science and Engineering, Shandong University, in 2005 and 2008, respectively. She was a Professor and the Master's Tutor with the School of Information Science and Engineering, University of Jinan, China. She has published more than 40 research articles, where 14 are indexed by SCI and more than 30 are indexed by EI. Her research interests include bioinformatics, computational intelligence, biomedical signal processing, and nonlinear time series analysis.

HUI JIANG received the B.S. degree in electronic information engineering from the Department of Information Engineering, College of Mechanical and Electrical Engineering, Anhui Polytechnic University, in 2017. She is currently pursuing the M.S. degree in signal and information processing with the University of Jinan. Her research interests include bioinformatics, complex networks, and nonlinear time series analysis.

YANG LI received the B.S. degree in communication engineering from the School of Information Science and Engineering, University of Jinan, in 2017, where she is currently pursuing the M.S. degree in signal and information processing. Her current research interests include machine learning, complex networks, and nonlinear time series analysis.