Chemometrics and Intelligent Laboratory Systems 206 (2020) 104148

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

DMLDA-LocLIFT: Identification of multi-label protein subcellular localization using DMLDA dimensionality reduction and LIFT classifier

Qi Zhang [a,b,1], Shan Li [c,1], Bin Yu [a,b,d,*], Qingmei Zhang [a,b], Yu Han [a,b], Yan Zhang [e], Qin Ma [f]

[a] *College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao, 266061, China*
[b] *Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao, 266061, China*
[c] *School of Mathematics and Statistics, Central South University, Changsha, 410083, China*
[d] *School of Life Sciences, University of Science and Technology of China, Hefei, 230027, China*
[e] *College of Electromechanical Engineering, Qingdao University of Science and Technology, Qingdao, 266061, China*
[f] *Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, 43210, USA*

ARTICLE INFO

Keywords:
Multi-label protein subcellular localization
Multi-information fusion
DMLDA dimensionality reduction
LIFT classifier

ABSTRACT

Multi-label proteins occur in two or more subcellular locations, which play a vital role in cell development and metabolism. Prediction and analysis of multi-label subcellular localization (SCL) can present new perspective with drug target identification and new drug design. However, the prediction of multi-label protein SCL using biological experiments is expensive and labor-intensive. Therefore, predicting large-scale SCL with machine learning methods has turned into a popular study topic in bioinformatics. In this study, a novel multi-label learning methods for protein SCL prediction, called DMLDA-LocLIFT, is proposed. Firstly, the dipeptide composition (DC), encoding based on grouped weight (EBGW), pseudo amino acid composition (PseAAC), gene ontology (GO) and pseudo-position specific scoring matrix (PsePSSM) are employed to encode subcellular protein sequences. Then, using direct multi-label linear discriminant analysis (DMLDA) to get rid of noise information of the fused feature vector. Lastly, the first-best feature vectors are input into the multi-label learning with Label-specIfic FeaTures (LIFT) classifier to predict. The leave-one-out cross validation (LOOCV) shows that the overall actual accuracy on Gram-negative bacteria, Gram-positive bacteria, plant datasets, virus dataset and human dataset are 98.6%, 99.6%, 97.9%, 94.7% and 96.1% respectively, which are obviously better than other state-of-the-art prediction methods. The proposed model can effectively predict SCL of multi-label proteins and provide references for experimental identification of SCL. The source codes and datasets are available at https://github.com/QUST-AIBBDRC/DMLDA-LocLIFT/.

1. Introduction

Cell growth, development, reproduction, and other vital activities are inextricable with the action of proteins. In the level of cell, proteins only fulfill a function in particular subcellular locations [1]. These positions offer a certain chemical environment and a series of interacting pairs, enabling proteins to perform their functions correctly [2]. SCL is critical for protein function. Abnormal proteins are tightly related to human diseases, such as breast cancer [3], primary human liver tumors [4], preeclampsia [5], brucellosis [6] and etc. Moreover, understanding the location of proteins can present new perspective with drug design and drug target identification [7].

As we can know, with the initiation of many large-scale protein projects, the quantity of proteins in the database has increased exponentially. It is far from satisfying the research which needs to search for the subcellular position of proteins through using experimental methods. So predicting protein SCL with machine learning becomes a problem that should be solved quickly. In the past two decades, several prediction methods [8,9] had been proposed to solve this problem, but these prediction approaches were developed based on a single labeling system in which every proteins only had one subcellular position [10]. However, the experiment shows that more and more proteins can simultaneously occur in two or more different subcellular locations [11]. These multi-label proteins play a more important role in cell vital movements, so the learning of SCL for multi-label protein is particularly important.

Predicting multi-label SCL through machine learning can not only

[*] Corresponding author. College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao, 266061, China.
E-mail address: yubin@qust.edu.cn (B. Yu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.chemolab.2020.104148
Received 10 April 2020; Received in revised form 3 July 2020; Accepted 31 August 2020
Available online 4 September 2020

improves efficiency but also provides some references for experimental methods. To build a SCL model by the method of calculation, the primary task is to draw the feature information from protein sequences. The commonly used feature extraction methods are mainly derived from structural information, physicochemical properties, evolutionary information, and sequence information. Cheng et al. [12] used key gene ontology information into pseudo amino acid composition to draw the useful information from amino acid residues of animal protein. Huang et al. [13] encoded protein sequences by integrating protein sequences information, physicochemical information and evolutionary information. Wang et al. [14] improved the position specific scoring matrix with segmented amino acid composition to extract characteristic message from Gram-positive bacteria and Gram-negative bacteria. Shen and Chou [15] proposed a new method Gneg-mPLoc for identifying gram-negative bacteria, which used three coding methods gene ontology, function domain, and sequence evolution to extract features from Gram-negative bacteria.

Feature fusion can generate high-dimensional matrix, which includes lots of superfluous information, and may influence the performance of the classifier seriously. Dimensional reduction can assist us to avoid these problems, and it is widely applied in pattern recognition and classification. Today, researchers have introduced multiple dimensionality reduction methods in the study of SCL prediction, including principal component analysis (PCA) [16], multi-label informed latent semantic indexing (MLSI) [17], multi-label dimensionality reduction via dependency maximization (MDDM) [18], multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence (MVMD) [19] and etc. Zhang et al. [20] proposed manifold regularized discriminant feature selection (MDFS), which utilized label correlation to establish an optimization framework and used iterative optimization algorithm to solve convexity optimization problems, thus greatly improving the performance of prediction. Zhang and Zhou [18] proposed a dimensionality reduction method MDDM, which used projection strategy to project raw data into lower-dimensional feature space and can make full use of the reliance between related class labels and primitive features. Lin et al. [21] proposed a max-dependency and min-redundancy (MDMR) method, which combined mutual information with maximum correlation and minimum redundancy to solve the multi-label dimension explosion issue and select the best subset for multi-label learning.

Apart from the effect of dimensionality reduction methods, the selection of the classifier is also critical to the prediction. The commonly used multi-label classification methods include multi-label k-nearest neighbor (ML-KNN) [22], multi-label neural network (BP-MLL) [23], multi-label learning by instance differentiation (InsDif) [24], multi-label learning by exploiting label correlations local (ML-LOC) [25] and multi-label learning with label-specific features (LIFT) [26] etc. Wang et al. [27] used the ensemble multiple classifier chain (ECC) classifier set to predict the multi-label proteins of Gram-negative bacteria and Gram-positive bacteria. After testing, the overall absolute accuracy and overall location accuracy of the two datasets were 92.4%, 94.1% and 94.0%, 94.4%, respectively. Wan et al. [28] used a multi-label support vector machine classifier through selecting virus proteins and two plant proteins as the benchmark datasets. LOOCV showed the overall actual accuracy (OAA) and overall location accuracy (OLA) of the three datasets were 93.7% and 97.2%, 93.6% and 95.6%, 72.6% and 78.4%, respectively. Wan et al. [29] used Multi-label LASSO (mLASSO) as a classifier to predict the human proteins dataset and used the LOOCV to assess the performance of the model. The overall actual accuracy (OAA) and overall location accuracy (OLA) were 74.8% and 84.6%.

Although some achievements have been made in protein SCL prediction using machine learning methods, there are still many aspects for melioration. Firstly, in the process of multi-label learning, the characteristic information of protein sequences characterizing SCL has not been fully elucidated. Then, high dimensional matrix after fusion will produce many redundant features. High-dimensional data will not only reduce the speed of operation but also affect the results of the model prediction. It is an urgent problem to be solved how to select a suitable dimensionality reduction method to eliminate noise information, and then reduce the running time and improve the prediction performance. Finally, the experimental data of the multi-label protein subcellular shows geometric growth. It is essential to design a precise, fast and effective prediction tool for SCL.

Enlightened by this, we put forward a new protein SCL prediction method called DMLDA-LocLIFT based on multi-label learning. Firstly, extracting information from protein sequences by dipeptide composition (DC), encoding based on grouped weight (EBGW), gene ontology (GO), pseudo-position specific scoring matrix (PsePSSM) and pseudo amino acid composition (PseAAC), and the best parameter λ values, ξ values, and L values of the model are determined by LOOCV. Secondly, the feature extraction information are combined as the initial features. Compared with other dimensionality reduction methods, DMLDA can capture the effective feature representation by removing the redundancy from raw feature space. At last, the first-best feature subsets are input into the LIFT classifier to predict the location of multi-label protein SCL, which achieves better prediction accuracy than ML-LOC, ML-KNN, mLASSO and INSDIF classifiers. The results indicate that DMLDA-LocLIFT method can significantly increase the prediction accuracy of protein SCL.

2. Materials and methods

2.1. Datasets

Benchmark datasets are generally composed of training dataset and test dataset. The training dataset is to train the model, and test dataset is to assess the model. In this paper, five different multi-label protein subcellular datasets are selected, namely, Gram-negative bacteria dataset [30], Gram-positive bacteria dataset [30], plant dataset [28], virus dataset [31] and human dataset [32]. The Gram-negative bacteria dataset contains 1456 locative proteins. It has 1392 distinct protein sequences located in 8 subcellular positions, among them 1328 belong to one subcellular position and 64 belong to two positions. The Gram-positive bacteria dataset contains 523 locative proteins. It includes 519 distinct protein sequences located in 4 subcellular positions, among them 515 belong to one subcellular position and 4 belong to two positions. The plant dataset, virus dataset and human dataset are used as independent datasets to verify the validation of the model. There are 1055 locative proteins in the plant dataset. The dataset has 978 distinct protein sequences located in 12 subcellular positions, of which 904 belong to one subcellular position, 71 belong to two locations and 3 belong to three positions. The virus dataset contains 252 locative proteins. It has 207 distinct protein sequences located in 6 subcellular positions, of which 165 belong to one subcellular position, 39 belong to two positions and 3 belong to three positions. The human dataset contains 3681 locative proteins. It includes 3106 distinct protein sequences located in 14 subcellular positions, where 2580 belong to one subcellular position, 480 belong to two positions, 43 belong to three positions and 3 belong to four positions. The subcellular location name and number of five datasets are shown in Supplementary Tables S1–S5. The homology of sequences in the five datasets are limited to below 25%.

2.2. Feature extraction

In order to predict the location of multi-label protein subcellular, the first task is to draw the feature from protein sequence and transform the protein sequence information into digital information. In this article, we use five feature extractions to draw the information from protein sequence, namely PseAAC, PsePSSM, EBGW, DC and GO.

2.2.1. Pseudo amino acid composition (PseAAC)

PseAAC [33–35] mainly uses sequence information and physical and chemical properties for feature extraction. Currently, the researchers have widely used the method of PseAAC in proteomics. The method maps the protein sequence into the following feature vectors:

$$P = [p_1, p_2, \cdots, p_{20}, p_{20+1}, \cdots, p_{20+\lambda}]^T \tag{1}$$

Every component is defined as:

$$p_u = \begin{cases} \dfrac{f_u}{\sum\limits_{i=1}^{20} f_u + \omega \sum\limits_{k=1}^{\lambda} \tau_k} & 1 \le \mathrm{u} \le 20 \\ \dfrac{\omega \tau_{u-20}}{\sum\limits_{i=1}^{20} f_u + \omega \sum\limits_{k=1}^{\lambda} \tau_k} & 20 + 1 \le \mathrm{u} \le 20 + \lambda \end{cases} \tag{2}$$

among them, ω is the weight factor and the value is 0.05 in the literature [36]. τ_k is the k nearest correlation factor, which represents sequence information between each amino acid. f_u represents the frequency of the u-th amino acid. Therefore, the preceding 20-dimensions of the eigenvector represent the composition information of amino acids, and the other λ-dimensions are sequence correlation factors which reflect different levels of amino acid sequence information. In this paper, since the five standard datasets selected require that the length of protein sequence should not be too short, the smallest sequence length is 50, so the selection of λ is $\lambda \le 50$. By choosing different parameter λ, the first-best λ can be determined by the precision.

2.2.2. Pseudo-position specific scoring matrix (PsePSSM)

In this paper, PSI-BLAST served to obtain position specific score matrix (PSSM) [37,38] of protein sequence. First, download a Swiss-Prot database provided by NCBI. Secondly, all protein sequences are compared with this Swiss-Prot database by using the PSI-BLAST program. In this course, set the threshold E to 0.001, and the maximum number of iterations is 3. BLOSUM62 is selected as the contrast matrix. The other parameters take the default value, and the PSSM of every sequence is obtained. we can obtain the $L \times 20$ PSSM matrix from protein sequence, among them L refers to the length of each sequence. Such as formula (3).

$$P_{PSSM} = \begin{pmatrix} P_{1,1} P_{1,2} \cdots P_{1,20} \\ \cdots\cdots\cdots \\ P_{i,1} P_{i,2} \cdots P_{i,20} \\ \cdots\cdots\cdots \\ P_{L,1}, P_{L,2} \cdots P_{L,20} \end{pmatrix} \tag{3}$$

where $P_{i,j}$ is the score of amino acid residues in a sequence that are mutated to j-type amino acids during evolution.

Firstly, the above matrix is standardized, and the value of PSSM matrix is transformed into 0 to 1 by formula (4):

$$f(x) = 1/(1 + e^{-x}) \tag{4}$$

In this article, PsePSSM algorithm [39,40] is applied to draw the features from protein sequences. And process of feature extraction is as follows:

$$P_{PsePSSM} = \left(\overline{P_1}, \overline{P_2}, \cdots, \overline{P_{20}}, \theta_1^1, \theta_2^1, \cdots, \theta_{20}^1, \cdots, \theta_1^\xi, \theta_2^\xi, \cdots, \theta_{20}^\xi\right)^T \tag{5}$$

$$\theta_j^\xi = \frac{1}{L - \xi} \sum_{i=1}^{L-\xi} \left(P_{i,j} - P_{(i+\xi),j}\right)^2 \; (\mathrm{j} = 1, \cdots, 20; \xi < \mathrm{L}, \xi \ne 0) \tag{6}$$

among them, $\overline{P_j}$ denotes the component of amino acid j in PSSM. θ_j^ξ represents the ξ-order relevant factors about amino acid type j. The order ξ correlation factor can be understood as the sequence order effect. θ_j^1 is a related factor by connecting the most consecutive PSSM scores along the protein chain of j-type amino acids; θ_j^2 that by connecting the second most consecutive PSSM scores; and so forth. As can be seen from formula (5), a sequence can be produced a $20 +\ 20 \times\ \xi$-dimensional feature vector.

2.2.3. Encoding based on grouped weight (EBGW)

According to the physicochemical properties about amino acids, Zhang [41] et al. used the idea of “coarse-grained” and “grouping” to reduce protein sequences into binary feature sequences, and they introduced the normal weight function based on the feature sequences to realize the EBGW [42,43]. Based on physical and chemical properties (hydrophobicity and charge) of the 20 basic amino acids which constitute the protein sequence, they were divided into four categories:

neutral and hydrophobic amino acids. $C1 = \{G, A, V, L, I, M, P, F, W\}$

neutral and polarity amino acids. $C2 = \{Q, N, S, T, Y, C\}$

acidic amino acids. $C3 = \{D, E\}$

alkaline amino acids. $C4 = \{H, K, R\}$

Combining $C1,\ C2,\ C3,\ C4$ in pairs and obtaining $\{C1, C2\}$ vs $\{C3, C4\}$, $\{C1, C3\}$ vs $\{C2, C4\}$ and $\{C1, C4\}$ vs $\{C2, C3\}$ three new divisions. Let $P = p_1 p_2 \cdots p_L$ be a protein sequence, we can use three homomorphic maps $f_1(P) f_2(P)$, $f_3(P)$ to transform P into three binary sequences. The definition is as follows:

$$f_1(p_i) = \begin{cases} 1 \; p_i \in \{C1, C2\} \\ 0 \; p_i \in \{C3, C4\} \end{cases} \tag{7}$$

$$f_2(p_i) = \begin{cases} 1 \; p_i \in \{C1, C3\} \\ 0 \; p_i \in \{C2, C4\} \end{cases} \tag{8}$$

$$f_3(p_i) = \begin{cases} 1 \; p_i \in \{C1, C4\} \\ 0 \; p_i \in \{C2, C3\} \end{cases} \tag{9}$$

$f_1(P), f_2(P), f_3(P)$ are binary sequences with three lengths of L. These sequences are divided into several subsequences with increasing lengths in turn. Each $f_i(P)$ generates an L-dimensional feature vector. Therefore, for a sequence of length L, it can be produced a $3L$-dimension vector.

2.2.4. Dipeptide composition (DC)

DC [44] is to calculate the frequency of occurrence about amino acid pairs. The vector elements extracted is shown in equation (10):

$$\begin{aligned} & V_{DC}(P) = (f_1, f_2, \cdots, f_{400}) \\ & f_i = \frac{R_i}{N - 1} \quad i = 1, 2, \cdots, 400 \end{aligned} \tag{10}$$

where R_i denotes the frequency of amino acid pairs and N represents the length of sequence. There are $20 \times 20 = 400$ ways to combine 20 amino acid forms into amino acid pairs.

2.2.5. Gene ontology (GO)

For a protein sequence P_i in the dataset, using BLASTP to search for P_i homologous proteins in SWISS-PROT database. The parameter E is set to 0.001. In the obtained homologous sequence, protein sequences of similarity $\ge 60\%$ is screened out and formed a protein set with P_i, which is recorded as S_i. For each protein sequence in set S_i, the GO [45,46] of the corresponding protein is searched in the GO database by using the index number of the sequence.

Equation (11) is used to construct a GO binary vector. The elements in the vector indicate whether or not to exist a GO number. Due to many GO numbers, delete those GO numbers that do not exist in the corresponding dataset to reduce redundant information.

$$p_i = \left[G_1, G_2, \cdots G_j, \cdots G_{|S_j|}\right]^T i = 1, 2, \cdots |S_i| \tag{11}$$

where $|S_i|$ is the size of the S_i set, $G_j = \begin{cases} 1 \;, & \text{if } \; \mathrm{p}_i \in G_{|S_j|} \\ 0 \;, & \text{otherwise} \end{cases}$.

2.3. *Direct multi-label discriminant analysis (DMLDA)*

A multi-label dataset containing n samples and K labels are defined as $\{x_i, y_i\}_{i=1}^{n}$, where $x_i \in R^d$,$y_i \in \{0,1\}^K$. If x_i relates to the $k-th$ label, then $y_i(k) = 1$ and 0 otherwise. Each label is treated as a category, and the samples are divided into K groups. The number of samples in the category k is represented by n_k. We denote the sample set is $X = [x_1, \cdots x_n]^T$ and the label indicator set is $Y = [Y_1, \cdots Y_n]^T = [y_{(1)}, \cdots y_{(K)}]$, where $y_i \in \{0,1\}^n$ is the label indicator vector of category k.

Considering that labels are interrelated, the label relationship between two classes in the DMLDA algorithm [47] is defined as follows.

$$C_{kl} = \cos(y_{(k)}, y_{(l)}) = \frac{\langle y_{(k)}, y_{(l)} \rangle}{\|y_{(k)}\| \|y_{(l)}\|} \tag{12}$$

Compared with the single-label dataset, the samples containing multiple labels are computed repeatedly. The utilization of normalized matrix $Z = [z_1, \cdots z_n]^T \in R^{n \times K}$ corrects the repeated computation problem.

$$z_i = \frac{y_i C}{\|y_i C\|_{\ell_1}} \tag{13}$$

where $\|\cdot\|_{\ell_1}$ represents the ℓ_1-norm of the vector.

Computing between-class scatter matrices:

$$S_b = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} (1 - Z_{ik})(x_i - m_k)(x_i - m_k)^T \right) \tag{14}$$

Computing within-class scatter matrices:

$$S_w = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} Z_{ik}(x_i - m_k)(x_i - m_k)^T \right) \tag{15}$$

where m_k is the average value of the $k-th$ class and m is the global average value of the multi-tagged dataset.

Solving the objective function of MLDA:

$$\max_G \frac{tr(G^T S_b G)}{tr(G^T S_w G)} \tag{16}$$

where G is a transformation matrix projecting dataset into lower q-dimensional space, and the projection vector is the q feature vectors corresponding to the q largest eigenvalues of $S_w^{-1} S_b$. The matrix after dimensionality reduction becomes: $T = G^{'}X$.

2.4. *Multi-label learning with Label-specIfic FeaTures (LIFT)*

Suppose that given m multi-label training sample $D = \{(x_i, Y_i) | 1 \le i \le m\}$, where $x_i \in X$ is the feature vector of d-dimensional and $Y_i \in Y$ is the corresponding category label of x_i. The LIFT classification algorithm [26] completes the construction of tag features and the induction of classification model through the following specific steps.

Firstly, the sample data of each category are clustered, and the clustering centers of positive and negative examples in each category are obtained. Specifically, for the class label $l_k \in Y$, positive training instance set P_k and negative training instance set N_k are represented via equations (17) and (18), respectively.

$$P_k = \{x_i | (x_i, Y_i) \in D, l_k \in Y_i\} \tag{17}$$

$$N_k = \{x_i | (x_i, Y_i) \in D, l_k \notin Y_i\} \tag{18}$$

where, P_k and N_k contain training examples with and without l_k label, respectively.

Secondly, in order to further study the properties of P_k and N_k, LIFT adopts clustering method. In this paper, K-means clustering [48] is applied to positive training instance set P_k and negative training instance set N_k. Assuming that P_k is divided into m_k^+ disjoint clusters and N_k is divided into m_k^- disjoint clusters, and then two groups of clustering centers $\{p_1^k, p_2^k, \cdots p_m^k\}$ and $\{n_1^k, n_2^k, \cdots n_m^k\}$ can be obtained respectively. To avoid the disturbance caused by imbalance, LIFT sets P_k and N_k to the same number of clusters, namely $m_k^+ = m_k^- = m_k$. Specifically, the number of clusters divided by P_k and N_k is set as equation (19):

$$m_k = [r \cdot \min(|P_k|, |N_k|)] \tag{19}$$

where $|\cdot|$ represents the cardinality set back and $r \in [0,1]$ is the ratio parameter of the number of reserved clusters.

By computing the distance among the sample and the clustering center, a mapping φ_k is formed from d-dimensional space to the $2m_k$-dimensional space, and the feature subset of the label in each instance is obtained as a basis for judging whether the label exists or not. The φ_k is described as:

$$\varphi_k(x) = \left[d(x, p_1^k), \cdots, d(x, p_m^k), d(x, n_1^k), \cdots, d(x, n_m^k) \right] \tag{20}$$

Finally, q classification models are constructed and recorded as $\{g_1, g_2, \ldots, g_q\}$. For each category labeled $l_k \in Y$, a binary classification training sample B_k containing m samples are created from the multi-label training sample D and mapping φ_k, and B_k can be formulated as:

$$B_k = \{(\varphi_k(x_i), Y_i(k)) | (x_i, Y_i) \in D\} \tag{21}$$

where $Y_i(k) = \begin{cases} +1 & l_k \in Y_i \\ -1 & otherwise \end{cases}$. Based on the new training set, LIBSVM [49] is selected for binary learner in this study. Thus, the prediction outcomes of each category label are obtained.

2.5. *Performance evaluation and model construction*

In statistical learning, re-substitution, leave-one-out cross validation (LOOCV) [50], k-fold cross-validation [51] and independent test are usually used to check out the effectiveness of the model. We use LOOCV to assess the validation of the model. Specifically, the sample is divided into N small samples, one of which is selected as the testing sample, and others are regarded as training samples. This process repeats N times until all small samples are tested. To analyze the prediction results more intuitively, the overall location accuracy (OLA) and the overall actual accuracy (OAA) are used to assess the prediction model [52]. The equations are as follows:

$$OLA = \frac{1}{\sum_{i=1}^{N} |L(Q_i)|} \sum_{i=1}^{N} |M(Q_i) \cap L(Q_i)| \tag{22}$$

$$OAA = \frac{1}{N} \sum_{i=1}^{N} \Delta |M(Q_i), L(Q_i)| \tag{23}$$

where, $L(Q_i)$ and $M(Q_i)$ represent the real label and predicted label, $|\cdot|$ counts the labels in the collection, $\cap$ is the intersection of the two collections, $|M(Q_i), L(Q_i)| = \begin{cases} 1 \ M(Q_i) = L(Q_i) \\ 0 \ otherwise \end{cases}$.

In addition, some other performance indicators, such as Coverage (CV), Average Precision (AP), Ranking Loss (RL) and Hamming Loss (HL) [53], are also used to assess the model. A multi-label dataset is defined as $\{X_i, Y_i\}_{i=1}^{n}$, where $Y_i \subset Y$ is the correct label set, Y'_i is the predicted label set, and n is the number of sequences. And indicators are defined as follows:

(1) Hamming Loss (HL):

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cup Y'| - |Y_i \cap Y'|}{M} \tag{24}$$

where M refers to the number of labels.

(2) Coverage (CV):

$$\frac{1}{n}\sum_{i=1}^{n}\max_{y\in Y_i} rank(f(X_i, y)) - 1 \tag{25}$$

where $rank(f(X_i, y))$ is the ranking of all tags when sorted in descending order.

(3) Ranking Loss (RL):

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Y_i||\overline{Y}_i|}|RAL(X_i)| \tag{26}$$

where $RAL(X_i) = \{(y_j, y_k)\big| f(X_i, y_j) \le f(X_i, y_k), (y_j, y_k) \in Y_i \times \overline{Y}_i\}$, and $f(X_i, y_j)$ refers to $y_j \subset Y$ is a partial label of X_i and $\overline{Y}_i$ is a complement of Y_i.

(4) Average Precision (AP):

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Y_i|}\sum_{y_j\in Y_i} AVP(X_i) \tag{27}$$

where $AVP(X_i) = \frac{\left|y_k | rank(f(X_i, y_k)) \le rank(f(X_i, y_j)), y_k \in Y_i\right|}{rank(f(X_i, y_j))}$.

For HL, CV and RL, the smaller the value, the better the performance of the model. For AP, OAA and OLA, the larger the value, the better the performance.

In our study, the protein SCL prediction method is called DMLDA-LocLIFT. The calculation process is displayed in Fig. 1. MATLAB2014a are implemented to fulfill the prediction, and the simulation environment is as follows: Windows Server 2012R2 Intel (R) Xeon (TM) CPU E5-2650 @ 2.30 GHz with 32.0 GB of RAM.

The steps of DMLDA-LocLIFT are depicted as follows:

Step 1 Prepare protein sequences of Gram-positive bacteria and Gram-negative bacteria datasets and corresponding class labels of proteins.

Step 2 Use PseAAC, EBGW, PsePSSM, DC and GO to draw the various information from sequence and five feature extraction results are fused.

Step 3 Use DMLDA method to decrease the dimension of the fused feature in **Step 2** and LIFT is selected as the classifier. The optimal feature vectors via feature selection are determined by OAA, OLA and other evaluation indexes obtained using the LOOCV.

Step 4 By the LOOCV, the first-best feature subsets are input into LIFT classifier to predict the protein SCL, and the DMLDA-LocLIFT model is determined.

Step 5 Calculate OAA, OLA and other indicators of the two training sets to assess the validation of the model.

Step 6 The plant dataset, virus and human datasets are used as independent datasets to examine the performance of DMLDA-LocLIFT model.

3. Results

3.1. Selection of optimal parameters λ, ξ and L

In this study, the selection of the parameter λ, ξ and L value of PseAAC, PsePSSM and EBGW algorithms respectively play a vital part in the model. To get the best feature information, it is necessary to continuously debug the parameters. For the datasets of Gram-negative bacteria and Gram-positive bacteria, seek out the best parameters for feature extraction, the ξ values are set to 0 to 10 in turn; the L values are set from 5 to 45 with an interval of 5. Since the minimum length of the

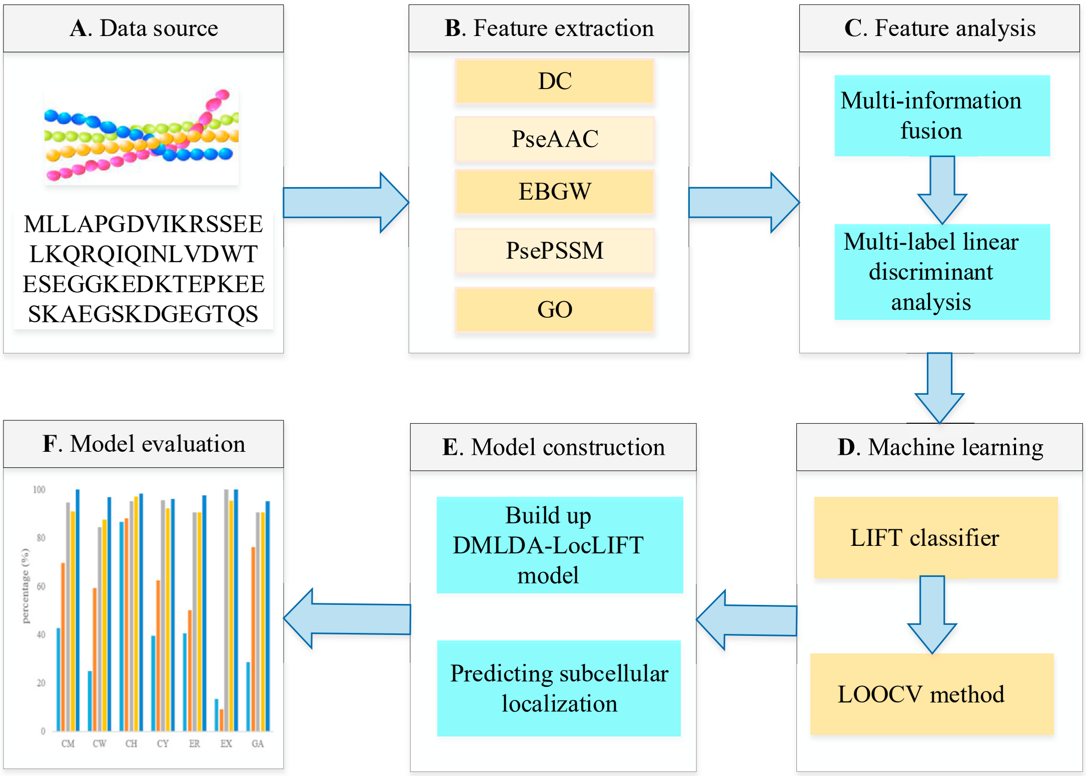

Fig. 1. The prediction pipeline of DMLDA-LocLIFT.

selected sequence is 50 when constructing protein datasets, the λ values are set from 0 to 49 in turn. Tested by the LOOCV, two training datasets are predicted by the LIFT classifier, and all prediction results are shown in Supplementary Tables S6–S11. Fig. 2 shows the change in OAA values corresponding to different parameters selected under Gram-negative bacteria and Gram-positive bacteria datasets.

In Fig. 2, when using PseAAC to feature extraction, the OAA of Gram-positive bacteria reach the highest if the λ value is 45. Gram-negative bacteria achieve the highest accuracy if the λ value is 49. Considering comprehensively $\lambda = 49$ is taken as the optimal parameter of PseAAC and $20 + \lambda = 69$-dimensional feature vectors are generated. When using PsePSSM to extract feature, $\xi = 3$ for Gram-positive bacteria and $\xi = 9$ for Gram-negative bacteria, the OAA value reaches the best. Considering comprehensively $\xi = 3$ is taken as the optimal parameter of PsePSSM. At this time, $20 + 20 \times \xi = 80$-dimensional feature vectors are generated. When using EBGW for feature extraction, Gram-negative bacteria and Gram-positive bacteria reach the highest accuracy if the L value is 40, so $3 \times L = 120$-dimensional feature vectors are generated.

3.2. Effect of feature extraction algorithm on results

The method of feature extraction is to transform the protein sequence information into digital information. Here, we take a fusion of five feature extraction results to obtain more feature information, but many redundant information will inevitably be produced. Therefore, the DMLDA dimensionality reduction method is used to eliminate noise and preserve effective feature information in protein sequences. In this section, using five separate feature extraction methods to compare the results of feature fusion based on DMLDA dimensionality reduction. We adopt the LIFT classifier and use LOOCV to evaluate. The specific prediction results about six different methods are obtained as shown in Supplementary Table S12. In order to analyze the prediction results more intuitively, Fig. 3 shows the changes of OAA and OLA values of Gram-negative bacteria and Gram-positive bacteria with six different methods.

As can be seen from Fig. 3, for single feature extraction methods, GO information has the greatest impact on multi-label protein SCL prediction. About the Gram-negative bacteria dataset, the feature fusion method based on DMLDA dimensionality reduction, OAA and OLA reach 98.6% and 98.6%, which are 7.7% and 5.8% higher than GO information. About the Gram-positive bacteria dataset, the feature fusion method based on DMLDA dimensionality reduction, OAA and OLA reach 99.6% and 99.4%, which are 10.0% and 7.0% higher than GO information. Moreover, the prediction results of every class of proteins is better than other single feature extraction methods. Therefore, we choose to combine the five feature extraction algorithms, and use DMLDA to get rid of noise information caused by fusion.

Among them, when making use of PseAAC algorithm for feature extraction, 69-dimensional feature vectors can be produced. When employing PsePSSM algorithm, 80-dimensional feature vectors can be produced. When using EBGW algorithm, 120-dimensional feature vectors can be produced. A 400-dimensional feature vector is generated when protein sequences are extracted using a DC algorithm. When GO information is used for feature extraction, 1717-dimensions feature vectors are generated for Gram-negative bacteria and 912-dimensions feature vectors are generated for Gram-positive bacteria. Therefore, when using five single feature extraction algorithms to fuse, a total of 2386-dimensions feature vectors are generated for Gram-negative bacterial dataset, and 1581-dimensions feature vectors are generated for Gram-positive bacterial dataset.

3.3. Effect of dimensional reduction algorithm on results

For the Gram-negative bacteria and Gram-positive bacteria datasets, fusing five feature extraction results to draw the features from protein sequences, but it also brings a lot of redundant features. To reach the desired prediction accuracy of protein SCL, PCA, MLSI, MDDM, MVMD, MDFS and DMLDA are used to reduce the dimension. Through LIFT multi-label classification, LOOCV is applied to test the results, and the OAA value of two datasets are obtained by selecting six different dimensionality reduction methods, as shown in Supplementary Table S13. To analyze the predicted performance of two datasets more intuitively under different dimensionality reduction methods, Fig. 4 and Fig. 5 show the comparison results of OAA, HL, CV and AP values.

In Figs. 4 and 5, we compare the six dimensionality reduction methods of PCA, MLSI, MDDM, MVMD, MDFS and DMLDA. For the Gram-negative dataset, DMLDA is used to reduce dimension and 10 features are selected as the feature subset. The OAA reaches 98.7%, which is 36.9%, 31.4%, 15.9%, 22.6% and 45.8% higher than PCA, MLSI, MDDM, MVMD and MDFS methods, respectively. Similarly, for the Gram-positive dataset, DMLDA is used to reduce dimension and 60

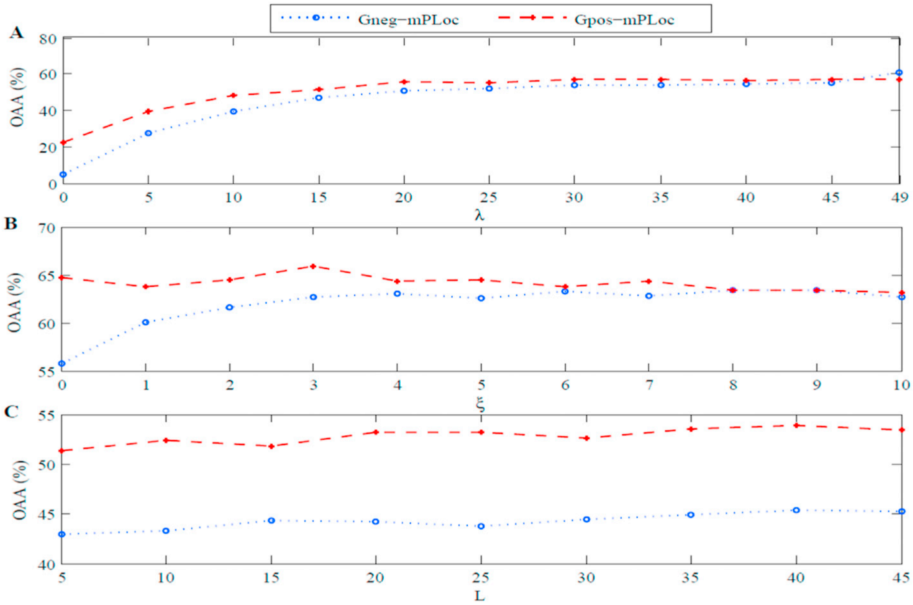

Fig. 2. The influence of different parameters on the OAA values of Gram-negative bacteria and Gram-positive bacteria datasets. (A) OAA values of PseAAC algorithm. (B) OAA values of PsePSSM algorithm. (C) OAA value of EBGW algorithm.

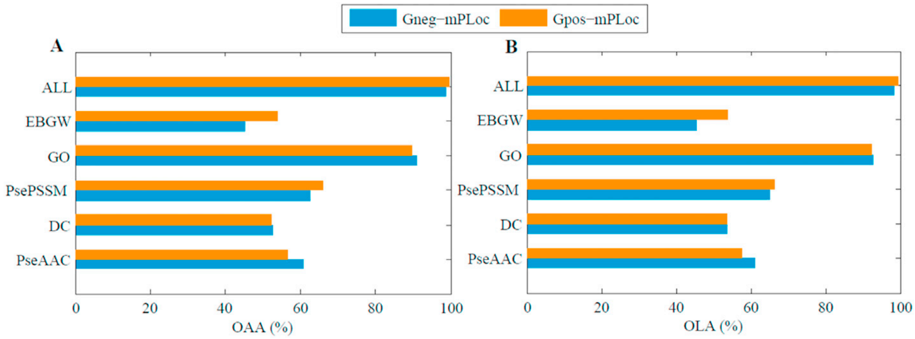

Fig. 3. The influence about six different methods on the prediction accuracy of Gram-positive bacteria and Gram-negative bacteria. ALL: PseAAC + PsePSSM + EBGW + DC + GO (DMLDA) (A) Overall Actual Accuracy, (B) Overall Location Accuracy.

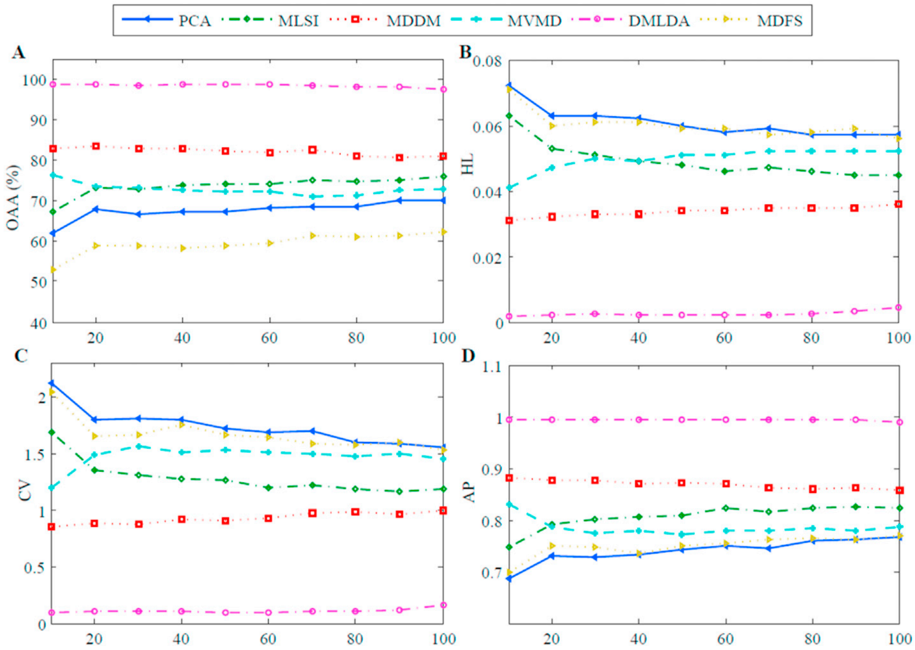

Fig. 4. The effect of selecting different dimensionality reduction methods on the predictive results of Gram-negative bacteria. (A) Overall Actual Accuracy, (B) Hamming Loss, (C) Coverage, (D) Average Precision.

features are selected as the feature subset. The OAA reaches 99.6%, which is 33.3%, 13.3%, 17.7%, 28.5% and 29.9% higher than PCA, MLSI, MDDM, MVMD and MDFS methods, respectively. Moreover, for two bacterial datasets, the DMLDA method achieves the best performance under hamming loss, coverage, and other indicators. By comprehensively considering the comparison results, we use the DMLDA method to reduce dimension and 60 features are selected as the feature subset. The accuracy of the model is the highest.

3.4. Selection of classification algorithms

To construct a model for multi-label protein SCL, the appropriate classifier algorithm has a great influence on the performance about the model. In this study, we compare LIFT with four classification algorithms such as ML-KNN, ML-LOC, INSDIF, and mLASSO. Among them, the ML-KNN classification algorithm chooses $k = 1$. Tested by the LOOCV, the specific outcomes of two datasets are shown in Supplementary Tables S14 and S15. Fig. 6 shows the changes in six performance indexes of Gram-negative bacteria and Gram-positive bacteria with different classifier algorithms.

In Fig. 6, for the Gram-negative bacteria dataset, the LIFT algorithm is used to predict the dataset. The OAA values is 98.6%, which is 0.3%–2.0% higher than the ML-KNN, INSDIF, ML-LOC and mLASSO classifiers. For the Gram-positive bacteria dataset, we use the LIFT algorithm as the classifier of the model. The OLA and OAA of the model reach 99.4% and 99.6%, which are 0.2%–0.4% higher than the OAA values of ML-KNN, INSDIF, ML-LOC and mLASSO, and 0.2%–1.0% higher than the OLA values of other classification algorithms. By comparing the results of HL, AP and other indicators, the performance of the LIFT is superior to other classifiers for datasets of Gram-negative bacteria and Gram-positive bacteria.

Through the analysis of OLA, OAA and other performance indicators of two datasets on five classifiers, a prediction model with high stability is selected. For the two training sets in this paper, the ML-KNN algorithm needs a lot of space to store instances, which has high complexity and unstable prediction performance. Although ML-LOC, mLASSO and INSDIF have considerable prediction results, they are running slowly and still have rising space. Since LIFT changes the multi-label question into the single-label question, and label information in the dataset is fully utilized. Through the clustering analysis of each label, the characteristics of the label are constructed, and the instances are trained and tested by the clustering results. All things considered, we choose the LIFT algorithm as the model classifier.

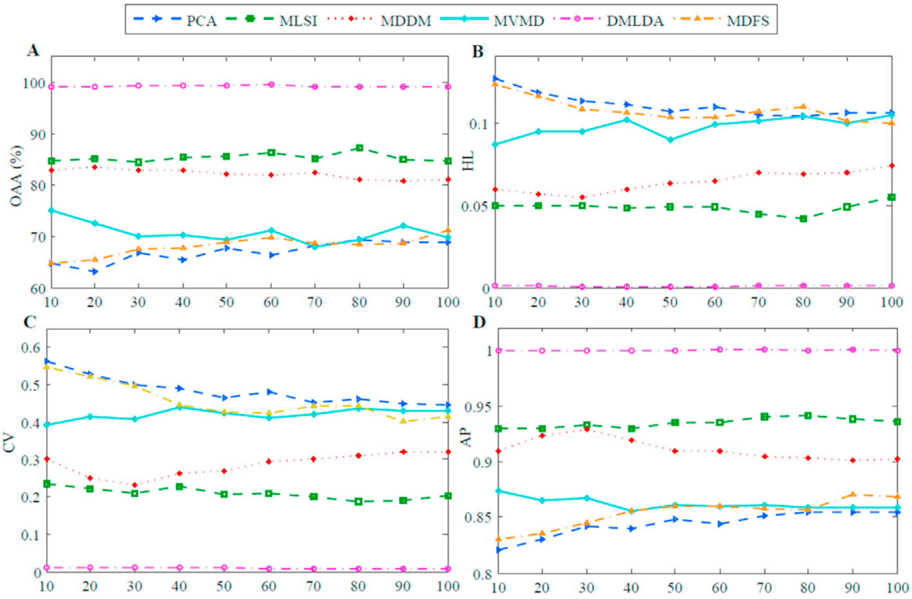

Fig. 5. The effect of selecting different dimensionality reduction methods on the predictive results of Gram-positive bacteria. (A) Overall Actual Accuracy, (B) Hamming Loss, (C) Coverage, (D) Average Precision.

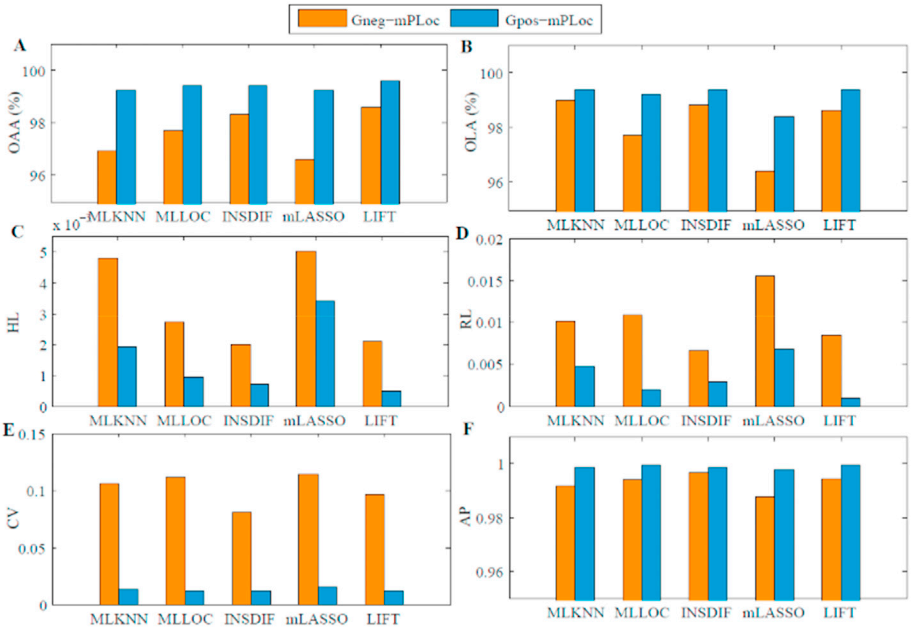

Fig. 6. The effect of different classifiers on the prediction results of Gram-negative bacteria and Gram-positive bacteria datasets. (A) Overall Actual Accuracy, (B) Overall Location Accuracy, (C) Hamming Loss, (D) Ranking Loss, (E) Coverage, (F) Average Precision.

3.5. Comparison with other methods

At present, researchers have put forward many methods to predict protein SCL of Gram-negative bacteria and Gram-positive bacteria datasets. To prove the validity of the model proposed in this paper, we contrast its prediction results to the identical dataset based on the LOOCV. Firstly, PseAAC, EBGW, PsePSSM, GO and DC are used to extract and fuse the features, and DMLDA is applied to select the optimal feature subset, and then using LIFT classifier to obtain the prediction results. The comparative results of Gram-negative bacteria and Gram-positive bacteria are shown in Table 1 and Table 2.

Table 1 shows that the OLA of DMLDA-LocLIFT for the Gram-negative bacteria is 98.6%, which is 3.3%–27.1% better than other methods. The OAA is 98.6%, which is 4.1%–8.7% better than other prediction methods. Table 2 shows that the OLA of DMLDA-LocLIFT for the Gram-positive bacteria is 99.4%, which is 2.6%–6.3% better than other methods. The OAA is 99.6%, which is 3.3%–6.7% better than other prediction methods. Moreover, for each site, the accuracy of DMLDA-LocLIFT method is superior to other prediction methods. By comparison, it is shown that the DMLDA-LocLIFT method proposed in this study has achieved desirable effect.

To verify the predictive capability about the model, we use plant dataset, virus dataset and human dataset as the independent datasets, and its results are compared with other methods in Tables 3–5. For the plant dataset, the OLA of the proposed model DMLDA-LocLIFT is 97.5%, which is 1.3%–33.8% higher than other prediction algorithms. The OAA

Table 1
Comparison of different prediction methods for protein subcellular localization on Gram-negative bacteria.

Locations	LOOCV				
	Methods				
	Gneg-PLoc [63]	iLoc-Gneg [54]	Gneg-ECC-mPLoc [27]	Gram-LocEN [55]	DMLDA-LocLIFT
Cell inner membrane	454/557 = 0.815	539/557 = 0.968	532/557 = 0.955	541/557 = 0.971	554/557 = 0.995
Cell outer membrane	68/124 = 0.548	103/124 = 0.831	105/124 = 0.944	111/124 = 0.895	120/124 = 0.968
Cytoplasm	362/410 = 0.883	367/410 = 0.895	378/410 = 0.922	379/410 = 0.924	403/410 = 0.983
Extracellular	59/133 = 0.444	115/133=0.865	124/133 = 0.932	129/133 = 0.970	131/133 = 0.985
Fimbrium	11/32 = 0.344	30/32 = 0.938	30/32 = 0.938	32/32 = 1.000	31/32 = 0.969
Flagellum	0/12 = 0.000	12/12 = 1.000	12/12 = 1.000	12/12 = 1.000	9/12 = 0.750
Nucleoid	0/8 = 0.000	4/8=0.500	7/8 = 0.875	7/8 = 0.875	8/8 = 1.000
Periplasm	87/180 = 0.483	161/180 = 0.894	170/180 = 0.944	161/180 = 0.894	179/180 = 0.994
OLA	1041/1456 = 0.715	1331/1456 = 0.914	1370/1456 = 0.941	1387/1456 = 0.953	1435/1456 = 0.986
OAA	–	1252/1392 = 0.899	1286/1392 = 0.924	1315/1392 = 0.945	1372/1392 = 0.986
HL	–	–	–	0.027	0.002
AP	–	–	–	0.952	0.995
RL	–	–	–	–	0.008
CV	–	–	–	–	0.096

Note: "-" indicates that the relevant references do not offer the results of the corresponding indicators.

Table 2
Comparison of different prediction methods for protein subcellular localization on Gram-positive bacteria.

Locations	LOOCV			
	Methods			
	iLoc-Gpos [56]	Gpos-ECC-mPLoc [27]	Gram-LocEN [55]	DMLDA-LocLIFT
Cell membrane	167/174 = 0.960	168/174 = 0.965	170/174 = 0.977	173/174 = 0.994
Cell wall	12/18 = 0.667	12/18 = 0.667	17/18 = 0.944	17/18 = 0.944
Cytoplasm	198/208 = 0.952	200/208 = 0.962	202/208 = 0.971	208/208 = 1.000
Extracellular	110/123 = 0.894	114/123 = 0.927	117/123 = 0.951	122/123 = 0.992
OLA	487/523 = 0.931	494/523 = 0.944	506/523 = 0.968	520/523 = 0.994
OAA	482/519 = 0.929	488/519 = 0.940	500/519 = 0.963	517/519 = 0.996
HL	–	–	0.016	4.81E-04
AP	–	–	0.971	0.999
RL	–	–	–	9.63E-04
CV	–	–	–	0.012

Note: "-" indicates that the relevant references do not offer the results of the corresponding indicators.

is 97.9%, which is 4.3%–29.8% better than other methods, and the comparison chart between the methods is shown in Supplementary Fig. S1. For virus dataset, the performance of DMLDA-LocLIFT is significantly better than other methods. Although the OLA value is slightly lower than mGOASVM and AD-SVM, the OAA obtained by the model in this paper is 94.7%, which is 1.5%–19.9% better than other methods. For human dataset, the OLA of DMLDA-LocLIFT method is 97.3%, which is 11.2%–21% better than other methods. The OAA is 96.1%, which is

Table 3
Comparison of different prediction methods for protein subcellular localization on plant dataset.

Locations	LOOCV				
	Methods				
	Plant-mPLoc [57]	iLoc-Plant [58]	mGOASVM [52]	HybridGO-Loc [28]	DMLDA-LocLIFT
Cell membrane	24/56 = 0.429	39/56 = 0.696	53/56 = 0.946	51/56 = 0.911	56/56 = 1.000
Cell wall	8/32 = 0.250	19/32 = 0.594	27/32 = 0.844	28/32 = 0.875	31/32 = 0.969
Chloroplast	248/286 = 0.867	252/286 = 0.881	272/286 = 0.951	278/286 = 0.972	281/286 = 0.983
Cytoplasm	72/182 = 0.396	114/182 = 0.626	174/182 = 0.956	168/182 = 0.923	175/182 = 0.962
Endoplasmic reticulum	17/42 = 0.405	21/42 = 0.500	38/42 = 0.905	38/42 = 0.905	41/42 = 0.976
Extracellular	3/22 = 0.136	2/22 = 0.091	22/22 = 1.000	21/22 = 0.955	22/22 = 1.000
Golgi apparatus	6/21 = 0.286	16/21 = 0.762	19/21 = 0.905	19/21 = 0.905	20/21 = 0.952
Mitochondrion	114/150 = 0.760	112/150 = 0.747	150/150=1.000	149/150 = 0.993	148/150 = 0.987
Nucleus	136/152 = 0.895	140/152 = 0.921	151/152 = 0.993	150/152 = 0.987	150/152 = 0.987
Peroxisome	14/21 = 0.667	6/21 = 0.286	21/21 = 1.000	21/21 = 1.000	21/21 = 1.000
Plastid	4/39 = 0.103	7/39 = 0.179	39/39 = 1.000	38/39 = 0.974	33/39 = 0.846
Vacuole	26/52 = 0.500	28/52 = 0.538	49/52 = 0.942	48/52 = 0.923	51/52 = 0.981
OLA	672/1055 = 0.637	756/1055 = 0.717	1015/1055 = 0.962	1009/1055 = 0.956	1029/1055 = 0.975
OAA	–	666/978 = 0.681	855/978 = 0.874	915/978 = 0.936	957/978 = 0.979
HL	–	–	0.013	0.007	0.002
AP	–	–	0.933	0.972	0.989
RL	–	–	–	–	0.016
CV	–	–	–	–	0.017

Note: "-" indicates that the relevant references do not offer the results of the corresponding indicators.

Table 4
Comparison of different prediction methods for protein subcellular localization on virus dataset.

Locations	LOOCV				
	Methods				
	iLoc-Virus [59]	mGOASVM [52]	AD-SVM [60]	mPLR-Loc [46]	DMLDA-LocLIFT
Viral capsid	8/8 = 1.000	8/8 = 1.000	8/8 = 1.000	8/8 = 1.000	8/8 = 1.000
Host cell membrane	25/33 = 0.758	32/33 = 0.970	32/33 = 0.970	30/33 = 0.909	31/33 =0.939
Host endoplasmic reticulum	15/20 = 0.750	17/20 = 0.850	17/20 = 0.850	17/20 = 0.850	16/20 = 0.800
Host cytoplasm	64/87 = 0.736	85/87 = 0.977	83/87 = 0.954	86/87 = 0.989	87/87 = 1.000
Host nucleus	70/84 = 0.833	82/84 = 0.976	82/84 = 0.976	81/84 = 0.964	84/84 = 1.000
Secreted	15/20 = 0.750	20/20 = 1.000	20/20 = 1.000	17/20 = 0.850	15/20 = 0.750
OLA	197/252 = 0.782	244/252 = 0.968	242/252 = 0.960	239/252 = 0.948	241/252 = 0.956
OAA	155/207 = 0.748	184/207 = 0.889	193/207 = 0.932	187/207 = 0.903	196/207 = 0.947
HL	–	0.026	0.019	0.023	0.010
AP	–	0.939	0.960	0.957	0.987
RL	–		–	–	0.035
CV	–		–	–	0.217

Note: "-" indicates that the relevant references do not offer the results of the corresponding indicators.

Table 5
Comparison of different prediction methods for protein subcellular localization on human dataset.

Locations	LOOCV				
	Methods				
	iLoc-Hum [61]	mGOASVM [52]	mLASSO [29]	mEN [62]	DMLDA-LocLIFT
Centrosome	56/77 = 0.727	64/77 = 0.831	42/77 = 0.546	60/77 = 0.779	74/77 = 0.961
Cytoplasm	561/817 = 0.687	683/817 = 0.836	699/817 = 0.856	683/817 = 0.836	799/817 = 0.978
Cytoskeleton	27/79 = 0.342	44/79 = 0.557	29/79 = 0.367	32/79 = 0.405	71/79 = 0.899
Endoplasmic reticulum	166/229 = 0.725	193/229 = 0.843	194/229 = 0.847	190/229 = 0.830	222/229 = 0.969
Endosome	1/24 = 0.042	9/24 = 0.375	1/24 = 0.042	5/24 = 0.208	22/24 = 0.916
Extracellular	325/385 = 0.844	344/385 = 0.894	311/385 = 0.808	314/385 = 0.816	375/385 = 0.974
Golgi apparatus	99/161 = 0.615	131/161 = 0.814	118/161 = 0.733	128/161 = 0.795	156/161 = 0.968
Lysosome	56/77 = 0.727	71/77 = 0.922	62/77 = 0.805	74/77 = 0.961	75/77 = 0.974
Microsome	7/24 = 0.292	18/24 = 0.750	1/24 = 0.042	14/24 = 0.583	22/24 = 0.916
Mitochondrion	284/364 = 0.780	339/364 = 0.931	336/364 = 0.923	336/364 = 0.923	355/364 = 0.975
Nucleus	918/1021 = 0.899	931/1021 = 0.912	922/1021 = 0.903	923/1021 = 0.904	999/1021 = 0.978
Peroxisome	20/47 = 0.426	43/47 = 0.915	34/47 = 0.723	39/47 = 0.830	47/47 = 1.000
Plasma membrane	277/354 = 0.783	288/354 = 0.814	267/354 = 0.754	266/354 = 0.751	347/354 = 0.980
Synapse	12/22 = 0.546	12/22 = 0.546	3/22 = 0.136	13/22 = 0.591	20/22 = 0.909
OLA	2809/3681 = 0.763	3170/3681 = 0.861	3019/3681 = 0.820	3077/3681 = 0.836	3584/3681 = 0.973
OAA	2118/3106 = 0.682	2251/3106 = 0.725	2265/3106 = 0.729	2307/3106 = 0.743	2987/3106 = 0.961
HL	–	0.029	0.029	0.028	0.004
AP	–	0.851	0.859	0.869	0.985
RL	–	–	–	–	0.016
CV	–	–	–	–	0.210

Note: "-" indicates that the relevant references do not offer the results of the corresponding indicators.

21.8%–27.9% better than other methods. Moreover, in these three datasets, for all subcellular locations, the single locative accuracies of DMLDA-LocLIFT are significantly better than other methods. Considering comprehensively, the method of DMLDA-LocLIFT has good prediction performance on three independent datasets.

4. Discussion

In this study, we put forward a multi-label protein prediction method called DMLDA-LocLIFT. A total of five feature extraction algorithms are adopted, namely GO, EBGW, DC, PsePSSM and PseAAC. The utilization of DMLDA algorithm makes the feature subset optimal, and the dimension is 60-dimensions. The LIFT classifier is used to predict the position of the multi-label proteins. Through the LOOCV, the OLA of Gram-negative bacteria, Gram-positive bacteria, plant dataset, virus dataset and human dataset are 98.6%, 99.4%, 97.5%, 95.6% and 97.3%, which are 3.3%–27.1%, 2.6%–6.3%, 1.3%–33.8%, 0.8%–17.4%, 11.2%–21% superior to other advanced methods, respectively. And OAA of five datasets are 98.6%, 99.6%, 97.9%, 94.7% and 96.1%, which are 4.1%–8.7%, 3.3%–6.7%, 4.3%–29.8%, 1.5%–19.9%, 21.8%–27.9% better than other advanced methods, respectively.

The DMLDA-LocLIFT model can effectively predict SCL of multi-label proteins, and its precision is obviously superior to other advanced methods. For the following reasons:

1. The physical and chemical information, evolutionary information, sequence information and annotation information in the fusion protein sequence are extracted by feature extraction. GO information has the highest prediction accuracy and plays a vital role in the multi-label prediction model.
2. DMLDA dimensionality reduction method can eliminate noise information in high-dimensional data via using label correlation, and effectively retain the important feature vectors for recognizing protein subcellular locations.
3. Through introducing the idea of clustering, LIFT transforms multi-label classification question into multiple binary classification questions, that greatly improves the prediction accuracy of multi-label learning and has good model generalization ability.

To explore the SCL of multi-label proteins can help people understand the function of these proteins more accurately, and provide theoretical basis for the pathogenesis of diseases. And understanding the location of proteins can furnish new opinions with drug design and drug target identification, so as to defeat the disease.

5. Conclusion

Studies have found that multi-label proteins carry more cellular metabolic functions, and normal life activities within cells are more dependent on them. Through the study of multi-label protein SCL can help people find out the pathogenesis of more diseases and reveal the special functions of multi-label proteins in drug compounds. With the continuous growth of multi-label protein data, traditional research methods are costly and time-consuming. As such, predicting protein SCL with machine learning becomes a problem that should be solved quickly. A new method of protein SCL prediction based on multi-label learning called DMLDA-LocLIFT is proposed in this study. Firstly, we use PseAAC, PsePSSM, EBGW, DC and GO algorithms to draw the valid information from sequences and fuse them. Then the feature fusion information is processed by DMLDA dimensionality reduction. Finally, the best-first feature subsets are input into LIFT to predict the location on Gram-negative bacteria, Gram-positive bacteria, plant dataset, virus dataset and human dataset. Through the LOOCV, the OLA of five datasets are 98.6%, 99.4%, 97.5%, 95.6% and 97.3%, and OAA are 98.6%, 99.6%, 97.9%, 94.7% and 96.1%, respectively. By comparison, we propose the DMLDA-LocLIFT method which obviously increases the prediction precision of multi-label protein SCL, but the construction of the model has to be improved in many aspects. Deep learning is a powerful tool for finding complex structures in high-dimensional data, and it has the advantage of strong adaptability, easy conversion, and good model generalization ability. So, predicting multi-label SCL using deep learning is our next research direction.

CRediT authorship contribution statement

Qi Zhang: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Validation, Visualization. **Shan Li:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Validation, Visualization. **Bin Yu:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Validation, Writing - review & editing. **Qingmei Zhang:** Formal analysis, Investigation, Methodology, Validation, Visualization. **Yu Han:** Writing - review & editing. **Yan Zhang:** Formal analysis, Investigation, Methodology, Writing - original draft. **Qin Ma:** Formal analysis, Writing - original draft, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank anonymous reviewers for valuable suggestions and comments. This work was supported by the National Natural Science Foundation of China (No. 61863010), the Key Research and Development Program of Shandong Province of China (No. 2019GGX101001), and the Natural Science Foundation of Shandong Province of China (Nos. ZR2018MC007, ZR2019MEE066).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2020.104148.

References

[1] E.A. Costa, K. Subramanian, J. Nunnari, J.S. Weissman, Defining the physiological role of SRP in protein-targeting efficiency and specificity, Science 359 (2018) 689–692.
[2] E. Glory, R.F. Murphy, Automated subcellular location determination and high-throughput microscopy, Dev. Cell 12 (2007) 7–16.
[3] D. Schulz, V.R.T. Zanotelli, J.R. Fischer, D. Schapiro, S. Engler, X.K. Lun, H.W. Jackson, B. Bodenmiller, Simultaneous multiplexed imaging of mRNA and proteins with subcellular resolution in breast cancer tissue samples by mass cytometry, Cell Syst. 6 (2018) 25–36.
[4] S. Nuciforo, I. Fofana, M.S. Matter, T. Blumer, D. Calabrese, T. Boldanova, S. Piscuoglio, S. Wieland, F. Ringnalda, G. Schwank, L.M. Terracciano, C.K.Y. Ng, M.H. Heim, Organoid models of human liver cancers derived from tumor needle biopsies, Cell Rep. 24 (2018) 1363–1376.
[5] G.H. Qiao, X.Z. Sun, Increased plasma fatty acid binding protein 4 concentration at the first prenatal visit and its relevance to preeclampsia, Hypertens. Res. 41 (2018) 763–769.
[6] A.M. Rodríguez, M.V. Delpino, M.C. Miraglia, M.M. Costa Franco, P. Barrionuevo, V.A. Dennis, S.C. Oliveira, G.H. Giambartolomei, Brucella abortus-activated microglia induce neuronal death through primary phagocytosis, Glia 65 (2017) 1137–1151.
[7] J.L. He, F. Huang, J. Li, Q.W. Chen, D.L. Chen, J.P. Chen, Bioinformatics analysis of four proteins of Leishmania donovani to guide epitopes vaccine design and drug targets selection, Acta Trop. 191 (2019) 50–59.
[8] Y.Y. Xu, F. Yang, H.B. Shen, Incorporating organelle correlations into semi-supervised learning for protein subcellular localization prediction, Bioinformatics 32 (2016) 2184–2192.
[9] B. Yu, S. Li, C. Chen, J.M. Xu, W.Y. Qiu, X. Wu, R.X. Chen, Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition, Chemometr. Intell. Lab. Syst. 167 (2017) 102–112.
[10] J.J. Almagro Armenteros, C.K. Sønderby, S.K. Sønderby, H. Nielsen, O. Winther, DeepLoc: prediction of protein subcellular localization using deep learning, Bioinformatics 33 (2017) 3387–3395.
[11] H. Zhou, Y. Yang, H.B. Shen, Hum-mPLoc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features, Bioinformatics 33 (2016) 843–853.
[12] X. Cheng, W.Z. Lin, X. Xiao, K.C. Chou, pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC, Bioinformatics 35 (2018) 398–406.
[13] C. Huang, J.Q. Yuan, Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites, Biosystems 113 (2013) 50–57.
[14] S.F. Wang, W.J. Li, Y. Fei, Z.C. Cao, D.S. Xu, H.Y. Guo, An improved process for generating uniform PSSMs and its application in protein subcellular localization via various global dimension reduction techniques, IEEE Access 7 (2019) 42384–42395.
[15] H.B. Shen, 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. 264 (2010) 326–333.
[16] H. Abdi, L.J. Williams, Principal component analysis, Comput. Stat. 2 (2010) 433–459.
[17] K. Yu, S.P. Yu, V. Tresp, Multi-label informed latent semantic indexing, in: International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005, pp. 258–265.
[18] Y. Zhang, Z.H. Zhou, Multilabel dimensionality reduction via dependency maximization, ACM Trans. Knowl. Discov. 4 (2010) 14.
[19] J.H. Xu, J.L. Liu, J. Yin, C.Y. Sun, A multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence simultaneously, Knowl-Based Syst. 98 (2016) 172–184.
[20] J. Zhang, Z.M. Luo, C.D. Li, C.G. Zhou, S.Z. Li, Manifold regularized discriminative feature selection for multi-label learning, Pattern Recogn. 95 (2019) 136–150.
[21] Y.J. Lin, Q.H. Hu, J.H. Liu, J. Duan, Multi-label feature selection based on max-dependency and min-redundancy, Neurocomputing 168 (2015) 92–103.
[22] J. Gonzalez-Lopez, S. Ventura, A. Cano, Distributed nearest neighbor classification for large-scale multi-label data on spark, Future Generat. Comput. Syst. 87 (2018) 66–82.
[23] M.L. Zhang, Z.H. Zhou, Multilabel neural networks with applications to functional genomics and text categorization, in: IEEE T. Knowl. Data. En., 18, 2006, pp. 1338–1351.
[24] M.L. Zhang, Z.H. Zhou, Multi-label learning by instance differentiation, in: National Conference on Artificial Intelligence. AAAI Press., 7, 2007, pp. 669–674.
[25] S.J. Huang, Z.H. Zhou, Multi-label Learning by Exploiting Label Correlations Locally, in: Twenty-sixth AAAI Conference on Artificial Intelligence, 2012, pp. 949–955.
[26] M.L. Zhang, L. Wu, LIFT: multi-label learning with label-specific features, IEEE Trans. Pattern Anal. 37 (2015) 107–120.
[27] X. Wang, J. Zhang, G.Z. Li, Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble, BMC Bioinf. 16 (2015) S1.
[28] S.B. Wan, M.W. Mak, S.Y. Kung, HybridGO-Loc: mining hybrid features on gene ontology for predicting subcellular localization of multi-location proteins, PLoS One 9 (2014), e89545.

[29] S.B. Wan, M.W. Mak, S.Y. Kung, mLASSO-Hum: a LASSO-based interpretable human-protein subcellular localization predictor, J. Theor. Biol. 382 (2015) 223–234.

[30] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC, J. Theor. Biol. 364 (2015) 284–294.

[31] H.B. Shen, K.C. Chou, Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites, J. Biomol. Struct. Dyn. 28 (2010) 175–186.

[32] H.B. Shen, K.C. Chou, A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0, Anal. Biochem. 394 (2009) 269–274.

[33] B. Yu, W.Y. Qiu, C. Chen, A.J. Ma, J. Jiang, H.Y. Zhou, Q. Ma, SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting, Bioinformatics 36 (2020) 1074–1081.

[34] C.R. Xu, L. Ge, Y.S. Zhang, M. Dehmer, I. Gutman, Prediction of therapeutic peptides by incorporating q-Wiener index into Chou's general PseAAC, J. Biomed. Inf. 75 (2017) 63–69.

[35] X.W. Cui, Z.M. Yu, B. Yu, M.H. Wang, B.G. Tian, Q. Ma, UbiSitePred: a novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components, Chemometr. Intell. Lab. Syst. 184 (2019) 28–43.

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

[37] X.J. Zhu, C.Q. Feng, H.Y. Lai, W. Chen, H. Lin, Predicting protein structural classes for low-similarity sequences by evaluating different features, Knowl-Based Syst. 163 (2019) 787–793.

[38] X.Y. Wang, B. Yu, A.J. Ma, C. Chen, B.Q. Liu, Q. Ma, Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique, Bioinformatics 35 (2019) 2395–2402.

[39] H. Shi, S.M. Liu, J.Q. Chen, X. Li, Q. Ma, B. Yu, Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure, Genomics 111 (2019) 1839–1852.

[40] B. Yu, S. Li, W.Y. Qiu, M.H. Wang, J.W. Du, Y.S. Zhang, X. Chen, Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction, BMC Genom. 19 (2018) 478.

[41] Z.H. Zhang, Z.H. Wang, Z.R. Zhang, Y.X. Wang, A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine, FEBS Lett. 580 (2006) 6169–6174.

[42] B. Yu, Z.M. Yu, C. Chen, A.J. Ma, B.Q. Liu, B.G. Tian, Q. Ma, DNNAce: prediction of prokaryote lysine acetylation sites through deep neural networks with multi-information fusion, Chemometr. Intell. Lab. Syst. 200 (2020) 103999.

[43] B.G. Tian, X. Wu, C. Chen, W.Y. Qiu, Q. Ma, B. Yu, Predicting protein-protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach, J. Theor. Biol. 462 (2019) 329–346.

[44] L.Y. Wei, Y.J. Ding, R. Su, J.J. Tang, Q. Zou, Prediction of human protein subcellular localization using deep learning, J. Parallel Distr. Comput. 117 (2018) 212–217.

[45] X.J. Lei, J. Zhao, H. Fujita, A.D. Zhang, Predicting essential proteins based on RNA-Seq, subcellular localization and GO annotation datasets, Knowl-Based Syst. 151 (2018) 136–148.

[46] S.B. Wan, M.W. Mak, S.Y. Kung, mPLR-Loc: an adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction, Anal. Biochem. 473 (2015) 14–27.

[47] M. Oikonomou, A. Tefas, Direct multi-label linear discriminant analysis, Commun. Comput. Inf. Sci. 383 (2013) 414–423.

[48] H.K. Al-Mohair, J.M. Saleh, S.A. Suandi, Hybrid human skin detection using neural network and K-means clustering technique, Appl. Soft Comput. 33 (2015) 337–347.

[49] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–27.

[50] W.Y. Qiu, S. Li, X.W. Cui, Z.M. Yu, M.H. Wang, J.W. Du, Y.J. Peng, B. Yu, Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition, J. Theor. Biol. 450 (2018) 86–103.

[51] C. Chen, Q.M. Zhang, Q. Ma, B. Yu, LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion, Chemometr. Intell. Lab. Syst. 191 (2019) 54–64.

[52] S.B. Wan, M.W. Mak, S.Y. Kung, mGOASVM: multi-label protein subcellular localization based on gene ontology and support vector machines, BMC Bioinf. 13 (2012) 290.

[53] F.F. Luo, W.Z. Guo, Y.L. Yu, G.L. Chen, A multi-label classification algorithm based on kernel extreme learning machine, Neurocomputing 260 (2017) 313–320.

[54] X. Xiao, Z.C. Wu, 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 6 (2011), e20592.

[55] S.B. Wan, M.W. Mak, S.Y. Kung, Gram-LocEN: interpretable prediction of subcellular multi-localization of Gram-positive and Gram-negative bacterial proteins, Chemometr. Intell. Lab. Syst. 162 (2017) 1–9.

[56] Z.C. Wu, X. Xiao, K.C. Chou, iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins, Protein Pept. Lett. 19 (2012) 4–14.

[57] K.C. Chou, H.B. Shen, Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization, PLoS One 5 (2010) e11335.

[58] Z.C. Wu, X. Xiao, K.C. Chou, iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites, Mol. Biosyst. 7 (2011) 3287–3297.

[59] X. Xiao, Z.C. Wu, 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. 284 (2011) 42–51.

[60] S.B. Wan, 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. 9 (2018) 399–411.

[61] K.C. Chou, Z.C. Wu, X. Xiao, iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites, Mol. Biosyst. 8 (2012) 629–641.

[62] S.B. Wan, M.W. Mak, S.Y. Kung, Sparse regressions for predicting and interpreting subcellular localization of multi-label proteins, BMC Bioinf. 17 (2016) 97.

[63] K.C. Chou, H.B. Shen, Large-scale predictions of gram-negative bacterial protein subcellular locations, J. Proteome Res. 5 (2006) 3420–3428.