Chemometrics and Intelligent Laboratory Systems 208 (2021) 104216

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

MpsLDA-ProSVM: Predicting multi-label protein subcellular localization by wMLDAe dimensionality reduction and ProSVM classifier

Qi Zhang [a,b,1], Shan Li [c,1], Qingmei Zhang [a,b], Yandan Zhang [a,b], Yu Han [a,b], Ruixin Chen [a,b], Bin Yu [a,b,d,e,*]

[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] Key Laboratory of Computational Science and Application of Hainan Province, Haikou, 571158, China

ARTICLE INFO

Keywords:
Multi-label protein subcellular localization
Multi-information fusion
wMLDAe dimensionality reduction
ProSVM classifier

ABSTRACT

Multi-label proteins play a significant role in life processes such as cell growth, development, and reproduction. Exploring protein subcellular localization (SCL) is a direct way to better understand the function of multi-label proteins in cells. This paper firstly presents a new prediction model named MpsLDA-ProSVM which predicts the SCL of multi-label proteins. Firstly, we utilize four coding algorithms including pseudo position-specific scoring matrix (PsePSSM), gene ontology (GO), conjoint triad (CT) and pseudo amino acid composition (PseAAC) to draw the feature information from protein sequences. Then, for the first time, we use a weighted multi-label linear discriminant analysis framework based on entropy weight form (wMLDAe) to refine and purify features. Finally, we input the optimal feature subset into the multi-label learning with label-specific features (LIFT) and multi-label k-nearest neighbor (ML-KNN) algorithms to obtain a synthetic ranking of relevant labels, and then use Prediction and Relevance Ordering based SVM (ProSVM) classifier to predict the SCLs. Tested by leave-one-out cross-validation (LOOCV), the overall actual accuracy on virus, plant, Gram-positive bacteria and Gram-negative bacteria datasets are 98.06%, 98.97%, 99.81% and 98.49%, which are 0.56%–9.16%, 1.07%–30.87%, 0.21%–6.91% and 3.99%–8.59% higher than other advanced methods respectively. By comparison, the model MpsLDA-ProSVM can effectively predict the specific location of multi-label proteins in cells.

1. Introduction

The nature of proteome research refers to study the characteristics of proteins in a large range, including protein-protein interactions [1,2], post-translational modifications [3], protein expression levels [4] and etc. From this, we can get a comprehensive understanding of disease occurrence, drug development and other processes at the protein level [5,6]. However, proteins exert their characteristics only in the corresponding subcellular location. So determining the protein subcellular location is of great significance for understanding the protein characteristics and functional diversity [7], which is also considered to lay the foundation for protein design and synthesis. In recent years, researchers have put forward a number of forecasting methods to predict SCLs, but most of the predictors can only process unit proteins [8,9]. However, experiments have found that some proteins can appear in two or more different subcellular positions, which are called multi-label proteins [10–12]. For example, acyl-CoA synthetase is found in both the endoplasmic reticulum and mitochondrial outer membrane [13]; Antioxidant defense proteins are presented in mitochondria, peroxisome and cytosol [14]. These multi-label proteins play a critical role in physical transportation, information transferring, energy conversion and other life activities of cells [15]. Therefore, it is urgent to explore the SCL of multi-label proteins.

The first task of SCL prediction of multi-label proteins is feature extraction. At present, the commonly used features include evolutionary information, physical and chemical properties, sequence information, structure information, and etc. For example, in order to explore the function of viral proteins, Thakur et al. [16] used three different feature

* 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.2020.104216
Received 16 October 2020; Received in revised form 19 November 2020; Accepted 27 November 2020
Available online 30 November 2020
0169-7439/© 2020 Elsevier B.V. All rights reserved.

extraction algorithms including dipeptide composition (DC), amino acid composition (AAC), physicochemical properties (PHY) and their hybrids to develop a server " MSLVP ″ for predicting SCLs. Zhou et al. [17] use AAC, conserved domain database (CDD), position specific scoring matrix (PSSM) and GO to extract and fuse the features of human proteins and proposed a new method Hum-mPLoc 3.0 based on amino acid sequence. Cheng et al. [18] fused GO information and PseAAC to extract features and predict the multiple subcellular locations from animal dataset.

However, single feature extraction method cannot fully characterize the feature information in the sequences, so researchers take the way of feature fusion to characterize the proteins multidimensionally. But the high-dimensional data after fusion contains a lot of irrelevant redundancy, which will seriously mislead the prediction. Dimensionality reduction can solve such problems, remove irrelevant information, and improve the running speed of the model. At present, Li et al. [19] constructed a robust multi-label kernelized fuzzy rough set model named RMFRS. The model combined the kernel information of sample space and label space, and combined multi-core learning with a fuzzy rough set to complete feature selection. According to information theory, Zhang et al. [20] put forward a feature selection method named multi-label feature selection based on label redundancy (LRFS), which used conditional mutual information between each label and candidate features, and combined the feature redundant terms to determine the final feature vector. Chen et al. [21] put forward a new method of extended adaptive least absolute shrinkage and selection operator (EALasso). This method preserves the oracle properties of identifying the correct subset model and has optimal estimation accuracy. The convergence of the algorithm was achieved by using the iterative optimization algorithm.

Machine learning algorithm acts as the last step for SCL prediction. Recently, Wang et al. [22] used the algorithm of integrating multiple classifier chains for prediction on the multi-label protein datasets of Gram-negative bacteria and Gram-positive bacteria. The overall locative accuracy of the two datasets is 94.1% and 94.4%, and the overall absolute accuracy is 92.4% and 94.0%. Wan et al. [23] built an integrated multi-label classifier HPSLPred, which was proved to have the highest accuracy and average precision of 75.89% on human datasets. Javed and Hayat [10] used a ranking support vector machine and multi-label k-nearest neighbor classifiers for prediction about virus and Gram-positive bacteria datasets. The overall accuracy on the two datasets were 80.46% and 85.54%, respectively.

Although existing prediction methods have achieved good prediction performance, in order to further improve the prediction performance, it is necessary to build a prediction model by combining multiple information, processing high-dimensional data, and selecting the appropriate classifier. Inspired by this, this paper proposes a new algorithm called MpsLDA-ProSVM to predict the SCL of multi-label proteins. Firstly, we fuse the feature vectors which are extracted by PseAAC, CT, PsePSSM and GO. Secondly, for the first time, wMLDAe is used to remove redundancy and get the best feature vectors. At last, we first combine the ML-KNN and LIFT algorithms to obtain the label score of each instance, so as to get the classification and ranking of the relevant labels, and then predict the multi-label proteins through the ProSVM classifier to determine the final prediction model. In this paper, two training datasets virus and plant datasets and two test datasets Gram-positive bacteria and Gram-negative bacteria datasets are used for evaluation by LOOCV. By experiment, our proposed MpsLDA-ProSVM method can helpfully improve the performance of multi-label protein SCL prediction.

2. Materials and methods

2.1. Datasets

In this study, datasets of virus proteins and plant proteins are selected as training set to construct the model, and datasets of Gram-negative bacteria and Gram-positive bacteria are selected as test set to examine the feasibility of the model. There are 252 protein sequences in the virus dataset [24]. It contains 207 different sequences located in 6 subcellular positions, among which 165 belong to one position, 39 belong to two positions and 3 belong to three positions. The plant dataset [25] has 1055 protein sequences. It contains 978 different sequences located in 12 subcellular positions. Among them, 904 belong to one position, 71 belong to two positions, and 3 belong to three positions. The breakdowns of these two training sets are listed in Fig. 1 and Fig. 2.

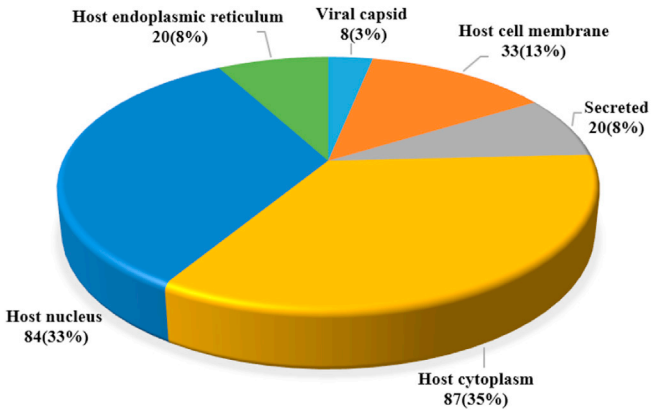

Fig. 1. The distribution of virus dataset in different subcellular locations. The number of proteins represents the number of "location proteins" in the corresponding subcellular location. There are 252 protein sequences in six subcellular locations of the virus dataset, and the homology of any one protein with other protein sequences in the same subcellular location is limited to 25%.

The Gram-positive bacteria [26] has 523 protein sequences. It contains 519 different sequences located in 4 subcellular positions, among which 515 belong to one position and 4 belong to two positions. The Gram-negative bacteria [27] has 1456 protein sequences. It contains 1392 distinct sequences located in 8 subcellular locations, of which 1328 belong to one location and 64 belong to two locations. Showing the specific classification of protein sequences about two test sets are in Supplementary Table S1 and S2.

2.2. Feature extraction

In our study, we use pseudo position-specific scoring matrix, gene ontology, pseudo amino acid composition and conjoint triad to draw the information from sequence and construct the initial feature vector.

2.2.1. Pseudo amino acid composition (PseAAC)

PseAAC [28] model put forward by Chou [29] can not only extract the location information of protein sequence, but also reflect the essential

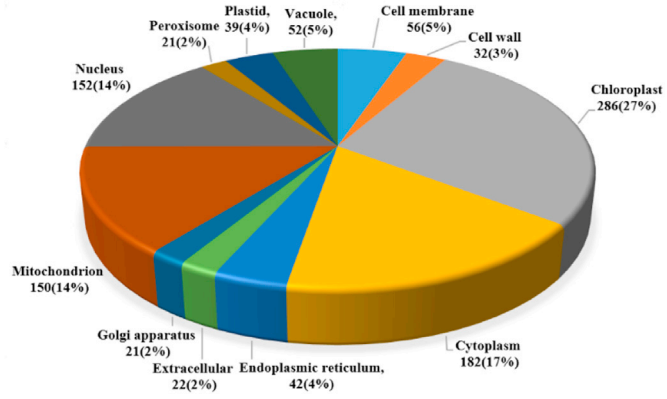

Fig. 2. The distribution of plant dataset in different subcellular locations. The number of proteins represents the number of "location proteins" in the corresponding subcellular location. There are 1055 protein sequences in twelve subcellular locations of the plant dataset, and the homology of any one protein with other protein sequences in the same subcellular location is limited to 25%.

composition information. According to the characteristics of 20 amino acids, each amino acid sequence is normalized to $(20 + \lambda)$-dimensional vector, in which the first 20 dimensions represent amino acid composition information, while the last λ dimensions represent pseudo amino acid composition information. At this point, the sequence is converted to the following feature vector:

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

among them,

$$p_u = \begin{cases} \dfrac{f_\mu}{\sum_{i=1}^{20} f_\mu + \omega \sum_{k=1}^{\lambda} \tau_k} & 1 \le \mu \le 20 \\ \dfrac{\omega \tau_{\mu-20}}{\sum_{i=1}^{20} f_\mu + \omega \sum_{k=1}^{\lambda} \tau_k} & 21 \le \mu \le 20 + \lambda \end{cases} \tag{2}$$

where ω represents the weighting factor, which refers to the order information of amino acids, and the value is 0.05 [30]. In this study, the value of λ is determined by the prediction accuracy, which is the optimal parameter value.

2.2.2. Pseudo position-specific scoring matrix (PsePSSM)

PsePSSM [31,32] draws evolutionary information from sequences with this method following two steps. Firstly, the PSI-BLAST [33] program is used to align the protein sequences to obtain a position-specific scoring matrix (PSSM) [34]. For protein sequences of length L, the PSSM algorithm obtains $L \times$ 20-dimensional feature vectors, and the results are as follows:

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

where the row indicates the corresponding amino acid position in the sequence, and the 20 amino acid types that can be mutated are listed.

Secondly, the PSSM matrix is standardized, and then use PsePSSM algorithm to generate $(20 + 20 \times \xi)$-dimensional eigenvectors for each protein sequence in order to unify the dimensions of the matrix. The feature extraction results are as follows:

$$P_{P_{sePSSM}} = \left(\overline{P_1}, \overline{P_2}, \cdots, \overline{P_{20}}, \theta_1^1, \theta_2^1, \cdots, \theta_{20}^1, \cdots, \theta_1^\xi, \theta_2^\xi, \cdots, \theta_{20}^\xi\right)^{'} \tag{4}$$

where $\overline{P_j}$ represents the component of amino acid j in PSSM. θ_j^ξ is the ξ-order related factor for the j-type amino acid. Similarly, the value of ξ depends on the prediction accuracy, which is the optimal feature extraction parameter.

2.2.3. Conjoint triad (CT)

In order to draw sequence information from proteins, Shen et al. [35] proposed CT [36] algorithm. Firstly, according to the dipole and volume of the amino acid side chain, the 20 amino acids are divided into seven categories. Secondly, every three of the protein sequences are grouped to form a triplet. By analogy, each sequence can generate $7\times 7\times 7 = 343$ triplets. The frequency of each triplet is calculated to form a 343-dimensional feature vector. The formula is as follows:

$$d_i = \frac{f_i - \min\{f_1, f_2, \cdots, f_{343}\}}{\max\{f_1, f_2, \cdots, f_{343}\}} \quad i = 1, 2, \cdots, 343 \tag{5}$$

2.2.4. Gene ontology (GO)

GO is a new expression method based on GO database, which is used to draw annotation information from protein sequence. The GO model uses GO ID to annotate the characteristics of protein sequences. It annotates genes and proteins based on three aspects including biological processes, molecular functions, and cell composition. The BLASTP is used to search for homologous proteins in the SWISS-PROT database and the sequence with similarities $\ge 60\%$ is deleted. For the remaining protein sequences S_i, the GO [37,38] of the corresponding protein is determined in the GO database according to the index number in the sequence. Its manifestation is:

$$p_i = \left[G_1, G_2, \cdots, G_j, \cdots, G_{|S_j|}\right]^{'} \quad i = 1, 2, \cdots, |S_i| \tag{6}$$

where $|S_i|$ is the size of the S_i. If the GO number exists, $G_j = 1$; otherwise, $G_j = 0$.

2.3. Weighted multi-label linear discriminant analysis framework based on entropy weight form (wMLDAe)

Linear discriminant analysis is one of the representative methods of supervised dimensionality reduction. For multi-label data, Xu [39] proposed a weighted multi-label linear discriminant analysis framework based on binary and related weight forms. Assuming n multi-label training instances and T class label sets are given, and W is defined as $n \times T$-dimensional non-negative weight matrix, that is:

$$W = [w_1, w_2, \cdots, w_n]^{'} \tag{7}$$

where $w_i = [w_{i1}, \cdots, w_{ik}, \cdots, w_{iT}]^{'}, i < n$ represents the $i-th$ instance. And the between-class, within-class and total scatter matrices are defined as:

$$\begin{aligned} S_b^M &= \sum_{k=1}^{T} \left(\sum_{i=1}^{n} w_{ik} \right) \left(m_k^M - m^M\right) \left(m_k^M - m^M\right)^{'} \\ S_w^M &= \sum_{k=1}^{T} \sum_{i=1}^{n} w_{ik} \left(x_i - m_k^M\right) \left(x_i - m_k^M\right)^{'} \\ S_t^M &= \sum_{k=1}^{T} \sum_{i=1}^{n} w_{ik} (x_i - m^M)(x_i - m^M)^{'} \end{aligned} \tag{8}$$

where m^M is the total mean vector and m_k^M is the mean vector of the $k-th$ class. Chen [40] put forward an entropy-based weight form and is introduced into the feature selection process of this section. Entropy is a metric to measure the uncertainty of random variables [41]. To the $i-th$ instance, the entropy can be defined as:

$$h_i = -\sum_{k=1}^{m_i} P_{ik} \ln(P_{ik}) = -\ln\left(\frac{1}{\|y_i\|_1}\right) \tag{9}$$

Let w_{ik} indicate the probability whose the $i-th$ instance belongs to the label of the $k-th$ class, which can be estimated by formula (10):

$$w_{ik} = e^{-h_i} = \frac{1}{m_i} = \frac{1}{\|y_i\|_1} \tag{10}$$

This shows that the probability of the relevant labels is equal to the inverse of the number of labels. At this time, the sum of the weights of each instance is 1, and the total scatter matrix is:

$$S_t^M = \sum_{k=1}^{T} \sum_{i=1}^{n} w_{ik} (x_i - m^M)(x_i - m^M)^{'} = \sum_{i=1}^{n} (x_i - m^M)(x_i - m^M)^{'} \tag{11}$$

The results of formula (11) is defined in the same way as those of the single-label LDA algorithm [42]. In the wMLDA algorithm, the projection matrix P is obtained by solving the generalized eigenvalue problem, that is:

$$S_b^M P = \Lambda S_w^M P \tag{12}$$

where Λ is composed of C non-negative eigenvalues, that is

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_C \geq 0$. A wMLDA having the form of the above entropy weight is added with a suffix 'e', which is called wMLDAe. In this paper, through taking full account of the label information, we use wMLDAe to process the fused vector and avoid over-fitting of the model.

2.4. Prediction and relevance ordering based SVM (ProSVM)

Assuming that a set of n training instances $\{x_i\}_{i=1}^n$ and T label $L = \{l_1, l_2, \cdots, l_T\}$are given. In order to consider the comprehensive ranking of relevant labels, we firstly combine the ML-KNN [43] and LIFT [44] algorithms to obtain the label score of each training instance, denoted as $g(x_i) = [g_1(x_i), g_2(x_i), \cdots, g_T(x_i)]$. Each label $l_t, t \in \{1, 2, \cdots, T\}$ is assigned a real value $g_t(x_i)$, and the labels are sorted according to the real values to obtain the corresponding set $R_i \in L$ of the ranking of relevant labels. $\phi_x(a)$ is used to represent the label index set that is less relevant than l_a. Both ML-KNN and LIFT algorithm can be used for nonlinear classification, and have solid theoretical basis, low training time complexity and high stability of prediction algorithm. The purpose of prediction and relevance ordering based SVM (ProSVM) [45] is to optimize the PRO loss function, which is non-convex and difficult to optimize. Therefore, we consider optimizing its equivalent loss function as follows:

$$\min_g \lambda \sum_{i=1}^{n} \widehat{L}(x_i, R_i, \phi, g) + \text{Regularizer(g)} \tag{13}$$

where

$$\begin{aligned} &\widehat{L}(x_i, R_i, \phi, g) = \sum_{l_t \in R_i \cup \{\theta\}} \sum_{s \in \phi_{x_i}(t)} (1 + g_s(x_i) - g_t(x_i))_+ / (4h_{s,t}) \\ &h_{s,t} = |g_s(x)| \times |g_t(x) - \{s\}| / 1 + \delta[g_s(x) = g_t(x)] \end{aligned} \tag{14}$$

Regularizer(g) is a regularizer of g, λ is a parameter measuring functional complexity of g and the equivalent loss function.

Suppose $g's$ are linear models with Regularizer$(g) = \sum_{t \in \{1, \cdots, T\} \cup \{\theta\}} \|w_t\|^2/2$ and $g_t(x) = w_t' x, t \in \{1, 2, \cdots, T\} \cup \{\theta\}$, in which θ is the threshold. Let G be the training set, $\overline{w} \triangleq [w_1; \cdots w_T; w_\theta]$, and (13) can be transformed into SVM-type in general form:

$$\begin{aligned} &\min_{\overline{w}, \xi} \ \frac{1}{2}\|\overline{w}\|^2 + \lambda C' \xi \\ &s.t. \ A\overline{w} \geq 1_p - \xi \ \ \xi \geq 0_p \end{aligned} \tag{15}$$

where 1_p is the $p \times 1$-dimensional one matrix and p is the total number of constraints. The vector C represents the weight of the loss, and the matrix A represents the constraint value between samples. ξ is determined by the value of $\overline{w}$.

The alternating direction method of multipliers (ADMM) [46] is a widely used optimization method for constraint problems in machine learning. The idea of decomposition and coordination is used to find the answer of the primitive question by solving its sub problems. Its basic methodology is to use the augmented LaGrange method to solve the optimization problem with equality constraints. Decompose the training set G into N disjoint subsets. Fusion the objective function and constraints through a dual variable to form an augmented LaGrange function. According to the ADMM solution algorithm, the specific process is as follows:

$$\begin{aligned} &\overline{w}_k^0 = \text{argmin}_{\overline{w}^0} L\left(\overline{w}^0, \left\{\overline{w}_k^n, \alpha_{k-1}^n\right\}_{n=1}^N, \eta\right) \\ &\alpha_k^n = \alpha_{k-1}^n + \eta\left(\overline{w}_k^n - \overline{w}_k^0\right)' \ \ \forall n = 1, 2, \cdots, N \end{aligned} \tag{16}$$

Through the above method, it is equivalent to solve the following problem:

$$\min_{\overline{w}^n} \ F\left(\overline{w}^n, G^n\right) + \left(\alpha_{k-1}^n\right)' \overline{w}^n + \frac{\eta}{2}\sum_{n=1}^{N}\left\|\overline{w}^n - \overline{w}_{k-1}^0\right\|^2 \tag{17}$$

This is a quadratic programming (QP) issue and Eq. (17) is similar to the standard SVM problem. At this time, the multi-label problem is converted into multiple binary classification problems (each label corresponds to a binary classification), which can be solved by the latest SVM solver LIBLINEAR [47] to obtain the score of each instance, and classify the instances by the threshold.

2.5. Performance evaluation

Choosing appropriate indicators to judge the quality of a prediction system is an important step to build a prediction model. At present, the commonly used inspection methods include re-substitution, independent test, k-fold cross-validation and leave-one-out cross-validation (LOOCV). Among them, LOOCV used in this paper is regarded as one of the most rigorous inspection methods [48,49]. Specifically, for N protein sequence, one of them is selected as testing, and the remaining N-1 are used as training. We will repeat N times until all sequences are tested.

So as to more intuitively assess the prediction performance about the model, we use eight measures including hamming loss (HL), average precision (AP), ranking loss (RL), F1, PRO loss (PL), coverage (CV), overall location accuracy (OLA) and overall actual accuracy (OAA) [45, 50,51]. Suppose a given multi-label training set $D = \{x_i, y_i\}_{i=1}^n$ and T learning functions $F = \{f_1 f_2, \cdots, f_T\}$, where $y_i \in \{1, -1\}^T$ is the real label of example x_i, and $H = \{h_1, h_2, \cdots, h_T\}$ represents T multi-label classifiers. Y_i^+ (Y_i^-) represents the true relevant (irrelevant) labels set of instance x_i, and P_i^+ (P_i^-) represents the predicted relevant (irrelevant) labels set of instance x_i. $rank_F(x_i, j)$ represents the ranking of j label in the label ranking list. The specific definitions of the indicators are defined as follows:

$$HL = \frac{1}{nT}\sum_{i=1}^{n}\sum_{j=1}^{T}\left[\left[h_i(x_i) \neq y_{ij}\right]\right] \tag{18}$$

$$F1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2\left|P_i^+ \cap Y_i^+\right|}{\left|P_i^+\right| + \left|Y_i^+\right|} \tag{19}$$

$$CV = \frac{1}{T}\left(\frac{1}{n}\sum_{i=1}^{n}\max_{j \in Y_i^+} rank_F(x_i, j) - 1\right) \tag{20}$$

$$RL = \frac{1}{n}\sum_{i=1}^{n}\frac{|SetR_i|}{\left|Y_i^+\right|\left|Y_i^-\right|} \tag{21}$$

where $SetR_i = \{(p, q) \big| f_p(x_i) \leq f_q(x_i), (p, q) \in Y_i^+ \times Y_i^-\}$.

$$AP = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left|Y_i^+\right|}\sum_{j \in Y_i^+}\frac{\left|SetP_{ij}\right|}{rank_F(x_i, j)} \tag{22}$$

where $SetP_{ij} = \{k \in Y_i^+ | rank_F(x_i, k) \leq rank_F(x_i, j)\}$.

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

$$OLA = \frac{1}{\sum_{i=1}^{N}|L(Q_i)|}\sum_{i=1}^{N}|M(Q_i) \cap L(Q_i)| \tag{24}$$

where $M(Q_i), L(Q_i)$ represents the predicted label subset and the real label subset respectively, and

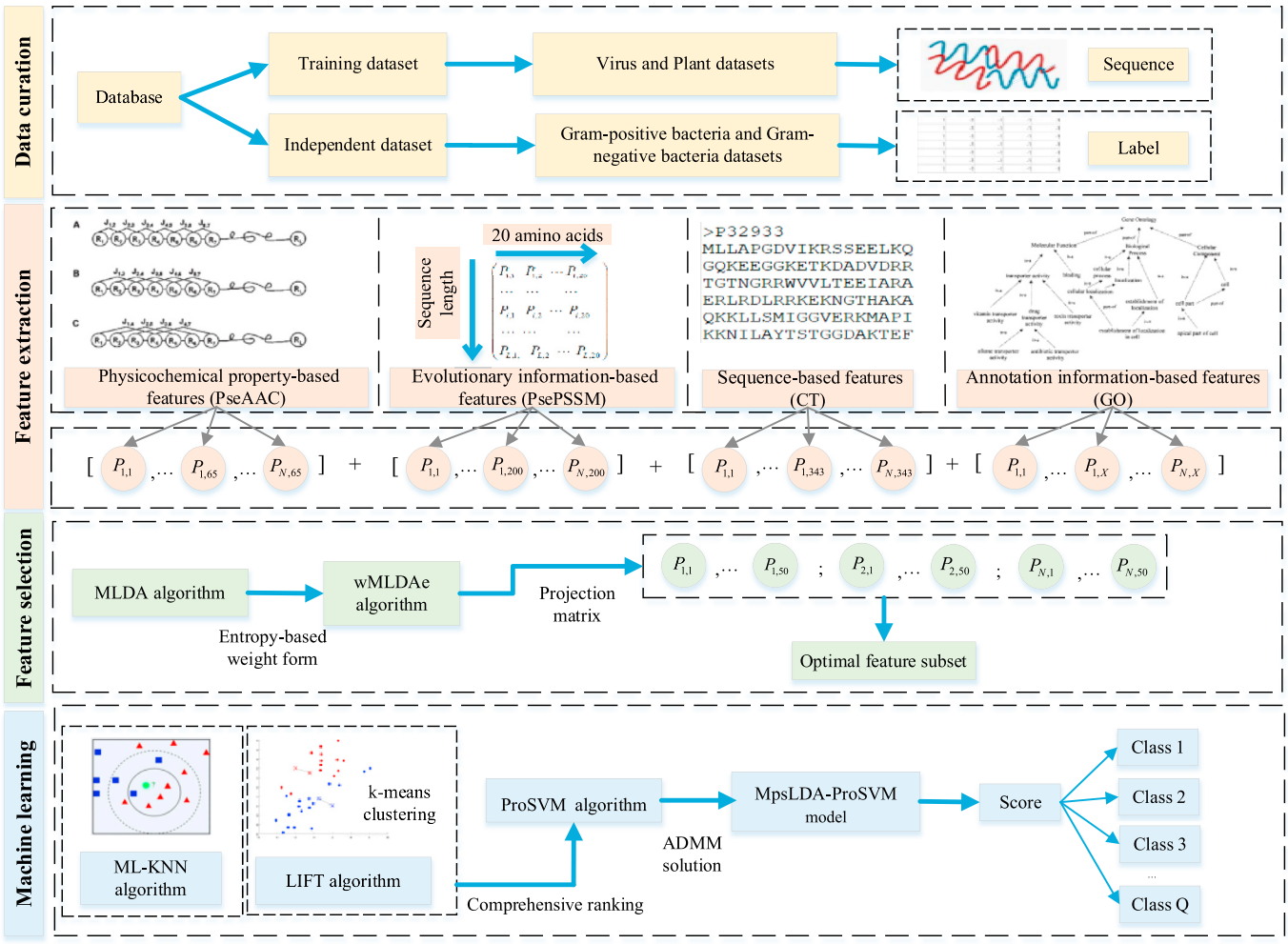

Fig. 3. The prediction pipeline of MpsLDA-ProSVM.

$$|M(Q_i),\quad L(Q_i)| = \begin{cases} 1 & |M(Q_i) = L(Q_i)| \\ 0 & otherwise \end{cases} \tag{25}$$

According to the definition in 2.4, this paper uses the PRO loss evaluation index, which refers to the ranking of relevant labels and the predicted value of labels. Specifically defined as:

$$PL = \sum_{l_t \in R_i \cup \{\theta\}} \sum_{s \in \phi_x(t)} \frac{1 + \delta[g_t(x) = g_s(x)]}{4|g_t(x)| \times |g_s(x) - \{t\}|} \gamma_{t,s} \tag{26}$$

where $\gamma_{t,s}$ is the modified 0–1 error. Specifically, $\gamma_{t,s} = 1$ if $g_t(x) < g_s(x)$, $\gamma_{t,s} = 1/2$ if $g_t(x) = g_s(x)$, otherwise the value is 0.

For PL, HL, CV and RL, the smaller the predicted value is, the better the classification performance. For AP, F1, OAA, and OLA, the larger the predicted value is, the better the classification performance.

2.6. Illustration of the MpsLDA-ProSVM workflow

For convenience, in this study, we put forward a multi-label protein SCL prediction method, MpsLDA-ProSVM, and the pipeline is shown in Fig. 3. The source codes and datasets are available at https://github.com/QUST-AIBBDRC/MpsLDA-ProSVM/.

The specific steps of the MpsLDA-ProSVM method are:

Step 1: Obtain the protein sequences of virus and plant datasets and make the true class labels of the corresponding sequences according to the classification.

Step 2: Extract the protein sequences to get the feature information by using PsePSSM, CT, PseAAC and GO. The parameters of feature extraction are determined by inputting feature information into ProSVM classifier.

Step 3: Combine the vectors from four feature extraction methods, and use wMLDAe algorithm to select the best feature subset.

Step 4: Input the features selected in ***Step 3*** to the ML-KNN and LIFT classifiers to obtain the classification and ranking of relevant labels, and then use the ProSVM classifier test to perform multi-label protein SCL prediction by LOOCV, leading to the MpsLDA-ProSVM model.

Step 5: Use the Gram-positive bacteria and Gram-negative bacteria datasets as independent test sets, and HL, OAA, OLA and other measures as final evaluation indicators to assess the performance of the MpsLDA-ProSVM model.

3. Results and discussion

3.1. Selection of optimal parameters λ and ξ

In order to build the optimal protein SCL prediction model and extract the optimal feature vectors from the protein sequences, the parameters of the model need to be continuously adjusted. Specially, the settings of the built-in parameters λ and ξ in the algorithm will directly affect the model performance for virus and plant datasets. To determine the values of the parameters λ and ξ, we set the ξ value to 1 to 10 with an interval of 1 . Because the shortest sequence length in the independent test set is 50, so parameter λ ranges from 5 to 49 with an interval of 5. The results of

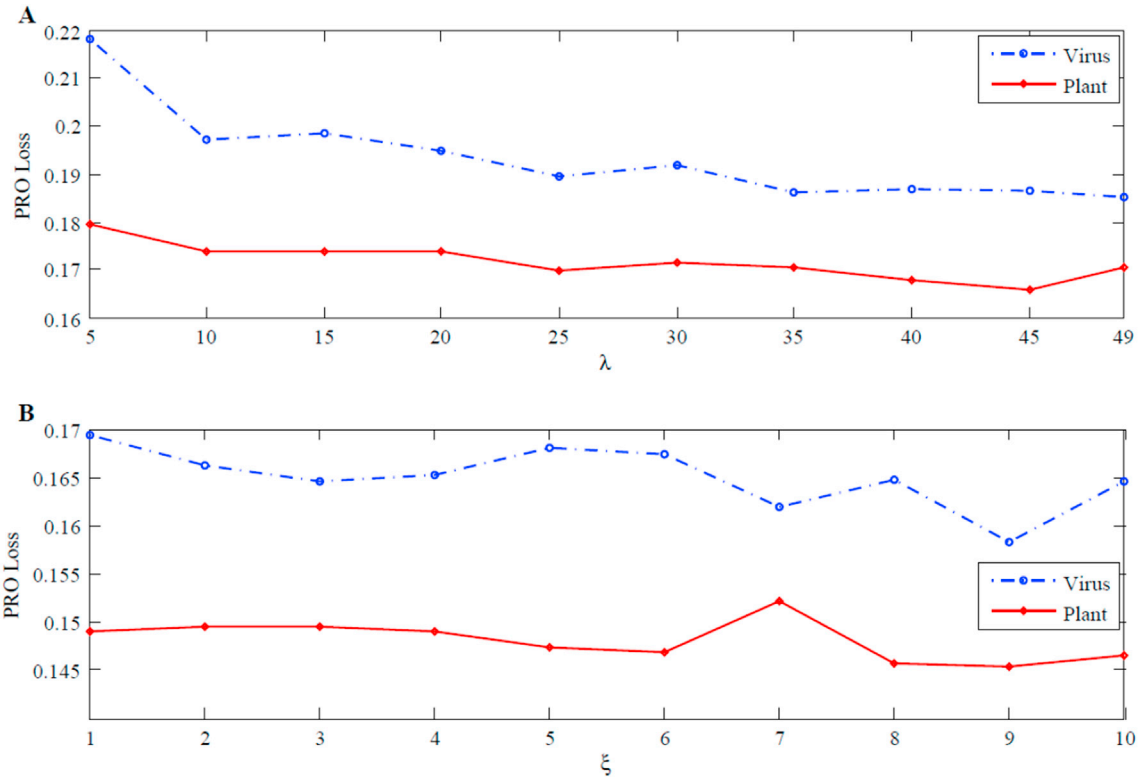

Fig. 4. The effect of different parameters on PL value of virus and plant datasets. (A) The PL value obtained using the PseAAC algorithm. (B) The PL value obtained using the PsePSSM algorithm.

feature extraction under different parameters are input into the ProSVM classifier, tested by LOOCV, and PL is selected as an evaluation index to determine the final parameters. The change of PL value of virus and plant datasets by selecting different parameter values are shown in Fig. 4. The specific results of other evaluation indexes are shown in Supplementary Table S3 and S4.

In Fig. 4, the prediction result of the classifier varies with different parameter values. For the value of parameter λ, $\lambda = 45$ is selected under comprehensive consideration. At this time, when using PseAAC for feature extraction, a total of $(20 + \lambda = 65)$-dimensional feature vectors are generated. For the value of parameter ξ, when using PsePSSM algorithm for feature extraction, select $\xi = 9$. Currently, a protein sequence generates a $(20 + 20 \times \xi = 200)$-dimensional feature vectors.

3.2. Effect of feature extraction algorithm on results

PseAAC can not only avoid losing the sequence information of the protein sequence, but also use the properties of amino acid residues to dig out the physical and chemical information of the sequence. PsePSSM not only considers the evolutionary information of amino acid residues, but also takes into account the sequence information. CT extracts sequence information from proteins through the dipole and volume properties of amino acid side chains. GO is composed of three independent ontologies of biological process, molecular function and cellular component, and provides annotations for gene products. GO is a feature extraction algorithm that has been used frequently in recent years, which can effectively extract annotation information in protein sequences. Therefore, this paper uses these four algorithms to convert protein sequences into numerical features. However, we cannot fully extract useful information of protein sequences by using a single feature coding method. So, researchers fuse the feature vectors extracted from various kinds of information together. To deal with the "dimensional disaster" after fusion, wMLDAe is used to reduce the number of dimensions generated after fusion. To verify the superiority of feature fusion, we use ProSVM classifier to predict the information after single-feature and multi-feature fusion. LOOCV is used for testing, and such as PL, HL, RL, and F1 are used as evaluation indicators to assess the performance. The prediction results of virus and plant datasets using four single feature extraction algorithms (GO, PsePSSM, CT, PseAAC) and a fusion method using wMLDAe dimensionality reduction (PseAAC + PsePSSM + CT + GO (wMLDAe)) are shown in Table 1.

Table 1
Comparison of the results of different feature extraction algorithms on virus and plant datasets.

Datasets	Algorithms	PL	HL	RL	AP	CV	F1
Virus	PseAAC	0.1864	0.2326	0.1741	0.7270	0.2842	0.5146
	PsePSSM	0.1583	0.2512	0.1630	0.7574	0.2439	0.5607
	CT	0.2504	0.2271	0.5750	0.6235	0.3712	0.4166
	GO	0.0413	0.0435	0.0976	0.9246	0.4734	0.8969
	ALL	**0.0133**	0.0064	0.0144	0.9961	0.0466	0.9885
Plant	PseAAC	0.1660	0.2292	0.1717	0.5815	0.0841	0.3861
	PsePSSM	0.1454	0.2162	0.1559	0.6139	0.0859	0.4159
	CT	0.2225	0.2049	0.5426	0.4302	0.2821	0.3235
	GO	0.0522	0.0389	0.1613	0.8354	0.0843	0.7941
	ALL	**0.0021**	0.0015	0.0047	0.9978	0.0114	0.9945

ALL: PseAAC + PsePSSM + CT + GO(wMLDAe). *Note*: The best results of PL have been bolded.

Table 2
Dimensions generated by the four feature extraction methods.

	PseAAC	PsePSSM	CT	GO	Total
Virus	65D	200D	343D	749D	1357D
Plant	65D	200D	343D	3091D	3699D

In Table 1, for virus dataset, when only a single feature extraction method is used, the performance of GO information is the best. The PL index of GO information is 14.51%, 11.70% and 20.91% lower than that of PseAAC, PsePSSM and CT respectively. The values in other indicators in GO information are also significantly better than other single feature extraction algorithms. However, the various indexes of feature fusion data based on wMLDAe dimension reduction show superior characteristics. Its PL, HL, RL, AP, CV and F1 values are 1.33%, 0.64%, 1.44%, 99.61%, 4.66% and 98.85%, respectively. Similarly, for plant dataset, the performance of GO information in the single feature extraction method is the best, which is significantly better than the index values of other feature extraction methods. Its corresponding PL index is 11.38%, 9.32%, and 17.03% lower than PseAAC, PsePSSM, and CT, respectively. And the various indexes of feature fusion data based on wMLDAe dimension reduction show superior characteristics. Its PL, HL, RL, AP, CV, and F1 values are 0.21%, 0.15%, 0.47%, 99.78%, 1.14%, and 99.45%.

Through comparison, this paper chooses the fusion method based on wMLDAe dimension reduction. At this time, the virus dataset obtains a total of 1357-dimensional feature vectors and the plant dataset obtains a total of 3699-dimensional feature vectors. And the dimensions generated by the four feature extraction methods are shown in Table 2.

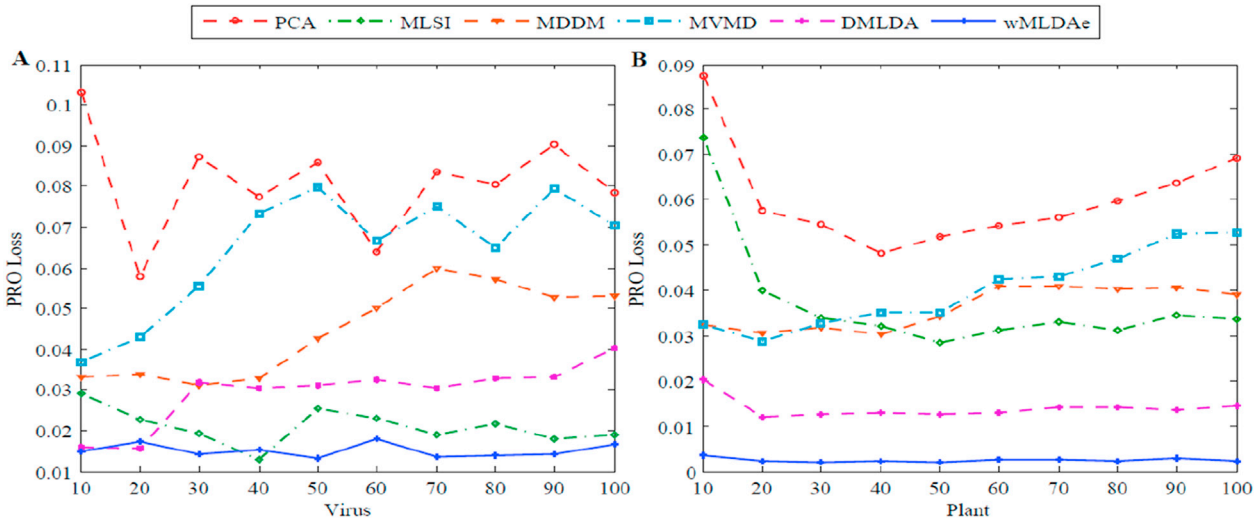

Fig. 5. The effect of different dimensionality reduction methods and different dimensions for virus and plant datasets. (A) PL values of virus dataset with different methods and different dimensions. (B) PL values of plant dataset with different methods and different dimensions.

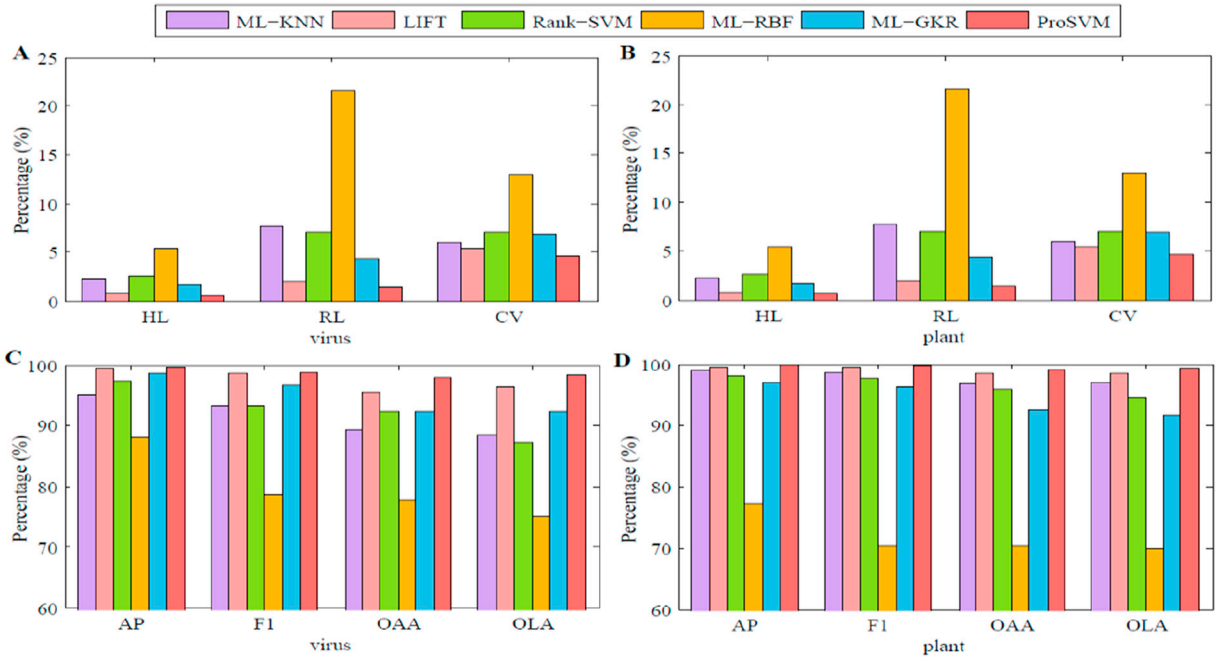

Fig. 6. Prediction results for virus and plant datasets when selecting different classifiers. (A) and (C) The results of HL, RL, CV, AP, F1, OAA, OLA values obtained by different classifiers in virus dataset. (B) and (D) The results of HL, RL, CV, AP, F1, OAA, OLA values obtained by different classifiers in plant dataset.

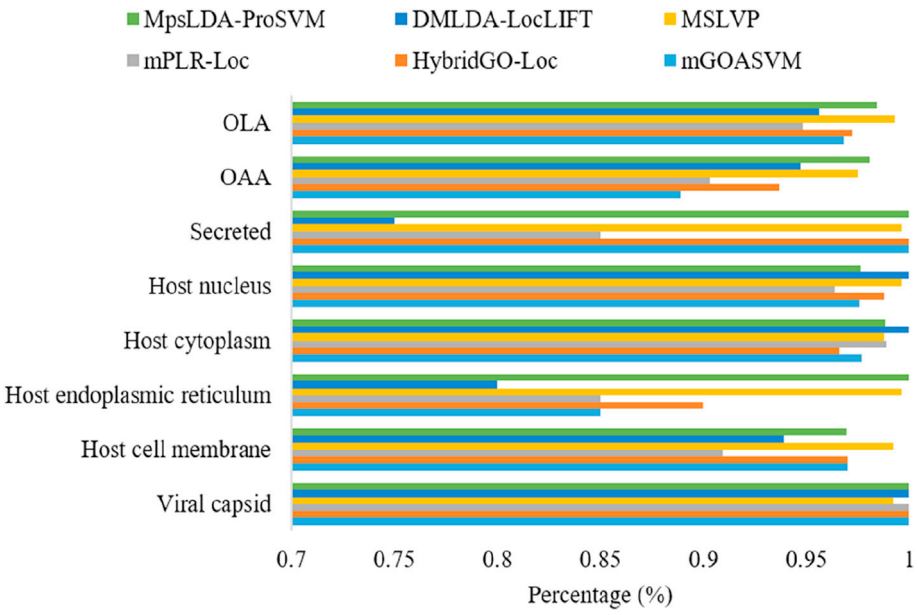

Fig. 7. Prediction results of MpsLDA-ProSVM and other models for virus dataset. *Note*: mGOASVM reported by Ref. [59]; HybridGO-Loc reported by Ref. [60]; mPLR-Loc reported by Ref. [61]; MSLVP reported by Ref. [16]; DMLDA-LocLIFT reported by Ref. [12].

3.3. Effect of dimensional reduction algorithm on results

To obtain the optimal feature vector, principal component analysis (PCA) [52], multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence (MVMD) [53], multi-label informed latent semantic indexing (MLSI) [54], multi-label dimensionality reduction via dependency maximization (MDDM) [55], direct multi-label linear discriminant analysis (DMLDA) [56] and wMLDAe algorithm are used to reduce the dimensionality. Different dimensionality reduction methods use different dimension combination to obtain different prediction results, so set the dimension to 10 to 100 and the interval to 10. The ProSVM classifier is selected to predict the feature subsets after dimensionality reduction, and the PL is used as the evaluation index. Fig. 5 shows the changes in the PL value of the virus and plant datasets when six different dimensionality reduction methods are selected. When using wMLDAe algorithm, the specific results of virus and plant datasets are shown in Supplementary Table S5 and S6.

In Fig. 5, for virus dataset, when using wMLDAe algorithm to reduce dimensions and selecting the dimension of 50, the lowest PL value is 1.33%, which is 7.27%, 6.63%, 1.21%, 2.93% and 1.79% lower than that of PCA, MVMD, MLSI, MDDM and DMLDA, respectively. For plant dataset, when using wMLDAe algorithm and selecting 50-dimensions, the PL value is 0.21%, which is 4.97%, 3.29%, 2.63%, 3.21% and 1.05% lower than that of PCA, MVMD, MLSI, MDDM and DMLDA, respectively. The results show that wMLDAe algorithm is significantly better than other algorithms. Therefore, we choose wMLDAe algorithm for dimension reduction and select the dimension of 50. The performance of this model is optimal.

For the virus and plant datasets, the features obtained by wMLDAe for virus and plant datasets are different because of the different classes of proteins and the number of protein sequences in both datasets. The matrix after dimensionality reduction of virus is 207×50 dimensions, and the matrix after dimensionality reduction of plant is 978×50 dimensions. We get projection matrix by the wMLDAe algorithm, so it does not affect the function of the single feature extraction method. Therefore, the biological significance of protein sequences in virus and plant datasets is still the physicochemical information, evolutionary information, sequence information and annotation information.

3.4. Selection of classification algorithms

The choice of classifier plays a vital role in construction of model. Selecting an appropriate classification algorithm can optimize the complexity of the model and improve the prediction accuracy. For the feature vectors based on wMLDAe dimensionality reduction, this paper uses ML-KNN [43], LIFT [44], ranking support vector machine (Rank-SVM) [57], multi-label radial basis function neural network (ML-RBF) [58], multi-label Gaussian kernel regression (ML-GKR) [18] and ProSVM classification algorithms to perform protein SCL prediction on virus and plant datasets, respectively. LOOCV is used to test, and seven measurement methods such as HL, RL, OAA and OLA are used as evaluation indexes. The prediction results obtained by selecting different classification algorithms for the two training sets are shown in Fig. 6. The specific experimental results for the six classifiers are shown in Supplementary Table S7. For the comparison algorithm, the parameter configuration proposed in this paper is shown in Supplementary Table S8.

From Fig. 6, when we use the ProSVM algorithm, the prediction performance is the best. For virus dataset, the OAA and OLA values obtained by using ProSVM algorithm are 98.06% and 98.41% respectively, which are higher than those obtained by using ML-KNN, LIFT, Rank-SVM, ML-RBF and ML-GKR algorithms. Similarly, for plant dataset, the OAA and OLA values obtained by ProSVM are 98.97% and 99.33% respectively, which are higher than other algorithms. For other indicators, the prediction results obtained using the ProSVM algorithm are also better than other algorithms.

In Fig. 6, by comparing the prediction results of the virus and plant datasets respectively, we find that the model has the best stability and prediction performance when the ProSVM algorithm is selected. However, the ML-KNN algorithm has the problems of large computational cost and slow prediction speed. For LIFT and Rank-SVM algorithms, both are susceptible to the choice of kernel functions and parameters. For the ML-RBF algorithm, with the increase of training samples, the number of hidden neurons in RBF network also rises, and the complexity and computation of the model also gradually increase. For ML-GKR algorithm, it is necessary to use complete feature information for prediction, and it is easy to lose effectiveness in high-dimensional data. In summary, we select ProSVM algorithm as the classifier to predict the SCL of multi-label proteins.

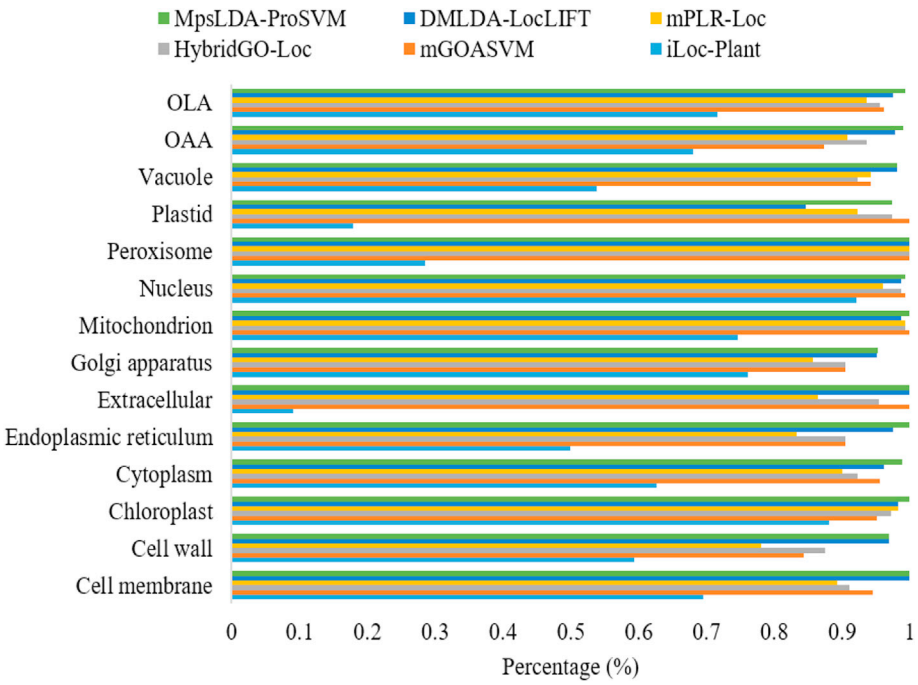

Fig. 8. Prediction results of MpsLDA-ProSVM and other models for plant dataset. *Note*: iLoc-Plant reported by Ref. [62]; mGOASVM reported by Ref. [59]; HybridGO-Loc reported by Refs. [60]; mPLR-Loc reported by Ref. [61]; DMLDA-LocLIFT reported by Ref. [12].

Table 3
Comparison of different prediction methods for protein SCL on Gram-positive bacteria dataset.

	Locations	Methods				
		iLoc-Gpos [63]	Gpos-ECC-mPLoc [22]	Gram-LocEN [64]	DMLDA-LocLIFT [12]	MpsLDA-ProSVM
1	Cell membrane	0.9600	0.9650	0.9770	0.9940	1.0000
2	Cell wall	0.6670	0.6670	0.9440	0.9440	1.0000
3	Cytoplasm	0.9520	0.9620	0.9710	1.0000	1.0000
4	Extracellular	0.8940	0.9270	0.9510	0.9920	0.9918
PL		–	–	–	–	2.4085E-04
HL		–	–	0.0160	4.81E-04	4.8170E-04
RL		–	–	–	9.63E-04	9.6339E-04
AP		–	–	0.7100	0.9990	1.0000
CV		–	–	–	0.0120	0.0029
F1		–	–	0.9700	–	0.9994
OAA		0.9290	0.9400	0.9630	0.9960	**0.9981**
OLA		0.9310	0.9440	0.9680	0.9940	**0.9981**

Note: '-' indicates that the relevant references do not offer the results of the relevant indicators. The best results of OAA and OLA have been bolded.

Table 4
Comparison of different prediction methods for protein SCL on Gram-negative bacteria dataset.

	Locations	Methods				
		iLoc-Gneg [65]	Gneg-ECC-mPLoc [22]	Gram-LocEN [64]	DMLDA-LocLIFT [12]	MpsLDA-ProSVM
1	Cell inner membrane	0.9680	0.9550	0.9710	0.9950	1.0000
2	Cell outer membrane	0.8310	0.9440	0.8950	0.9680	1.0000
3	Cytoplasm	0.8950	0.9220	0.9240	0.9830	0.9659
4	Extracellular	0.8650	0.9320	0.9700	0.9850	1.0000
5	Fimbrium	0.9380	0.9380	1.0000	0.9690	1.0000
6	Flagellum	1.0000	1.0000	1.0000	0.7500	1.0000
7	Nucleoid	0.5000	0.8750	0.8750	1.0000	1.0000
8	Periplasm	0.8940	0.9440	0.8940	0.9940	1.0000
PL		–	–	–	–	0.0031
HL		–	–	0.0270	0.0020	0.0019
RL		–	–	–	0.0080	0.0057
AP		–	–	0.9520	0.9950	0.9962
CV		–	–	–	0.0960	0.0081
F1		–	–	0.9420	–	0.9950
OAA		0.8990	0.9240	0.9450	**0.9860**	0.9849
OLA		0.9140	0.9410	0.9530	0.9860	**0.9904**

Note: '-' indicates that the relevant references do not offer the results of the relevant indicators. The best results of OAA and OLA have been bolded.

3.5. Comparison with other methods

In the paper, to clarify the prediction performance about this model, we choose to use the same evaluation index for comparison based on LOOCV. In this study, the virus and plant datasets are predicted. Firstly, features are extracted and fused by using PseAAC, PsePSSM, GO and CT, then the fused high-dimensional vector is reduced by using wMLDAe algorithm to get the best feature vector, and finally it is entered into ProSVM to get the prediction results. Fig. 7 and Fig. 8 show comparison diagrams of MpsLDA-ProSVM and other models in the two training sets. The concrete comparative results of virus and plant datasets are shown in Supplementary Table S9 and Table S10.

In Figs. 7 and 8, for virus dataset, when using MpsLDA-ProSVM method, the OAA and OLA are 98.06% and 98.41% respectively, which are 0.56%~9.16% and 1.21%~3.61% higher than those using other methods. For plant dataset, when using MpsLDA-ProSVM method, the OAA and OLA are 98.97% and 99.33% respectively, which are 1.07% ~30.87% and 1.83%~27.63% higher than those using other methods.

For further verifying the rationality and validity about this model, the datasets of Gram-positive bacteria and Gram-negative bacteria are used as test sets. Tested by LOOCV, we obtain the specific prediction results about two test sets. Selecting the same dataset for comparison, the results

are shown in Table 3 and Table 4. For Gram-positive bacteria dataset, when using MpsLDA-ProSVM method, the OAA and OLA are all 99.81%, which are 0.21%~6.91% and 0.41%~6.71% higher than those using other methods, respectively. For Gram-negative bacteria dataset, when using MpsLDA-ProSVM method, the OAA and OLA are 98.49% and 99.04% respectively, which are 3.99%~8.59% and 0.44%~7.64% higher than those using other methods. And in other indicators, the prediction accuracy obtained by using MpsLDA-ProSVM method is obviously better than other methods. Therefore, the prediction model proposed in this paper has good generalization ability and achieves satisfactory results.

For these comparison methods, MpsLDA-ProSVM uses the same PseAAC, GO and ML-KNN algorithm as iLoc-Plant [62], iLoc-Gpos [63] and iLoc-Gneg [65], the same GO algorithm as Gpos-ECC-mPLoc [22] and Gneg-ECC-mPLoc [22] and the same PseAAC, PsePSSM, GO and LIFT algorithm as DMLDA-LocLIFT [12]. The other methods are different.

4. Conclusion

The precise location of the SCL for multi-label proteins can not only clarify the various functions of proteins, but also promote the knowledge of the pathogenesis with some diseases and the research of new drugs. This paper proposes a new method named MpsLDA-ProSVM. Firstly, the annotation information, sequence information, evolutionary information and physicochemical information of protein sequence are considered in feature extraction. Multi-information fusion can achieve the effect of complementary advantages between different features, thereby enhancing the description ability of features. It can avoid the phenomenon of large single feature differences, and provide a basic guarantee for building efficient models. Secondly, in order to avoid dimension disaster, wMLDAe is used for the first time to remove irrelevant information. It applies the entropy weight form to the MLDA method, and fully considers the label information. The wMLDAe algorithm can not only remove the irrelevant information in the feature vector, but also retain the effective features in the sequence and form the best feature subset, so as to promote the running speed and increase the prediction accuracy about the model. Finally, the feature vector is input to the ProSVM classifier to predict multi-label proteins. For getting the comprehensive ranking of relevant labels, we first combine the ML-KNN and LIFT algorithms to get the predicted label scores of each instance, summarize the scores and sort them to get the ranking, and then combined with the Pro Loss function, the objective function is optimized according to the ADMM algorithm to obtain the final prediction result. Solve the multi-label classification problem through the problem transformation method, and finally convert the multi-label problem into multiple binary classification problems. This reduces the computational complexity of the model, and the processed labels are relatively independent, and the overall prediction performance of the model will not be easily affected by the changes of individual labels. This algorithm further improves the efficiency of the model. Experiments prove that the model MpsLDA-ProSVM also has the advantages of high stability and strong generalization ability. Combined with the above methods, the model has achieved satisfactory prediction results and is obviously superior to other prediction methods.

With the continuous collection and identification of multi-label proteins, we expect that MpsLDA-ProSVM can be used on datasets of more species. We also hope that the MpsLDA-ProSVM model can be used to solve other multi-label problems, such as identifying drug-target interactions [66], anatomical therapeutic chemical classification [67], predicting biological activity of antimicrobial peptides [68], and etc. In the dimensionality reduction method used in this study, the algorithm emphasizes the importance of the objective function by increasing the weight form. It can not only be used for multi-label protein subcellular location prediction, but also in other bioinformatics fields, such as protein function prediction, biopharmaceutical classification, antimicrobial peptides and their functional types prediction, and etc. In recent years, deep learning has successfully predicted the sites of multi-label data by virtue of its powerful generalization ability. In the future, we will try to use the deep learning to predict multi-label protein subcellular location.

Author 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. Qingmei Zhang: Formal analysis, Investigation, Methodology, Validation, Visualization. Yandan Zhang: Formal analysis, Investigation, Methodology, Validation. Yu Han: Writing- Reviewing and Editing. Ruixin Chen: Writing- Reviewing and Editing. Bin Yu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Validation, Writing- Reviewing and 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

The authors sincerely thank the anonymous reviewers for their valuable 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), the Natural Science Foundation of Shandong Province of China (No. ZR2018MC007), and the Key Laboratory Open Foundation of Hainan Province (No. JSKX202001).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2020.104216.

References

[1] 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.
[2] C. Chen, Q.M. Zhang, B. Yu, Z.M. Yu, P.J. Lawrence, Q. Ma, Y. Zhang, Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier, Comput. Biol. Med. 123 (2020) 103899.
[3] M.H. Wang, X.W. Cui, S. Li, X.H. Yang, A.J. Ma, Y.S. Zhang, B. Yu, DeepMal: accurate prediction of protein malonylation sites by deep neural networks, Chemometr. Intell. Lab. Syst. 207 (2020) 104175.
[4] N. Severe, N.M. Karabacak, K. Gustafsson, N. Baryawno, G. Courties, Y. Kfoury, K.D. Kokkaliaris, C. Rhee, D. Lee, E.W. Scadden, J.E. Garcia-Robledo, T. Brouse, M. Nahrendorf, M. Toner, D.T. Scadden, Stress-induced changes in bone marrow stromal cell populations revealed through single-cell protein expression mapping, Cell Stem. Cell 25 (2019) 570–583.
[5] K.C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (2015) 218–234.
[6] J.Y. Lin, H. Chen, S. Li, Y.S. Liu, X. Li, B. Yu, Accurate prediction of potential druggable proteins based on genetic algorithm and Bagging-SVM ensemble classifier, Artif. Intell. Med. 98 (2019) 35–47.
[7] 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.
[8] 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.
[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 composition, Chemometr. Intell. Lab. Syst. 167 (2017) 102–112.
[10] F. Javed, M. Hayat, Predicting subcellular localization of multi-label proteins by incorporating the sequence features into Chou's PseAAC, Genomics 111 (2019) 1325–1332.
[11] S. Xiang, Q.K. Liang, Y.C. Hu, P. Tang, G. Coppola, D. Zhang, W. Sun, AMC-Net: asymmetric and multi-scale convolutional neural network for multi-label HPA classification, Comput. Methods Progr. Biomed. 178 (2019) 275–287.

[12] Q. Zhang, S. Li, B. Yu, Q.M. Zhang, Y. Han, Y. Zhang, Q. Ma, DMLDA-LocLIFT: identification of multi-label protein subcellular localization using DMLDA dimensionality reduction and LIFT classifier, Chemometr. Intell. Lab. Syst. 206 (2020) 104148.
[13] M.A. Block, J. Jouhet, Lipid trafficking at endoplasmic reticulum-chloroplast membrane contact sites, Curr. Opin. Cell Biol. 35 (2015) 21–29.
[14] J.C. Mueller, C. Andreoli, H. Prokisch, T. Meitinger, Mechanisms for multiple intracellular localization of human mitochondrial proteins, Mitochondrion 3 (2004) 315–325.
[15] L. Liu, L. Tang, X. Jin, W. Zhou, A multi-label supervised topic model conditioned on arbitrary features for gene function prediction, Genes 10 (2019) 57.
[16] A. Thakur, A. Rajput, M. Kumar, MSLVP: prediction of multiple subcellular localization of viral proteins using a support vector machine, Mol. Biosyst. 12 (2016) 2572–2586.
[17] 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.
[18] 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.
[19] Y.W. Li, Y.J. Lin, J.H. Liu, W. Weng, Z.K. Shi, S.X. Wu, Feature selection for multi-label learning based on kernelized fuzzy rough sets, Neurocomputing 318 (2018) 271–286.
[20] P. Zhang, G.X. Liu, W.F. Gao, Distinguishing two types of labels for multi-label feature selection, Pattern Recogn. 95 (2019) 72–82.
[21] S.B. Chen, Y.M. Zhang, C.H.Q. Ding, J. Zhang, B. Luo, Extended adaptive Lasso for multi-class and multi-label feature selection, Knowl-Based Syst. 173 (2019) 28–36.
[22] 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.
[23] S.X. Wan, Y.C. Duan, Q. Zou, HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source, Proteomics 17 (2017) 1700262.
[24] 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.
[25] 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.
[26] H.B. Shen, K.C. Chou, Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins, Protein Pept. Lett. 16 (2009) 1478–1484.
[27] 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.
[28] X.M. Sun, T.Y. Jin, C. Chen, X.W. Cui, Q. Ma, B. Yu, RBPro-RF: use Chou's 5-steps rule to predict RNA-binding proteins via random forest with elastic net, Chemometr. Intell. Lab. Syst. 197 (2020) 103919.
[29] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[30] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins 43 (2010) 246–255.
[31] 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.
[32] 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.
[33] Y. Liang, S. Zhang, S. Ding, Accurate prediction of Gram-negative bacterial secreted protein types by fusing multiple statistical features from PSI-BLAST profile, SAR QSAR Environ. Res. 29 (2018) 469–481.
[34] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292 (1999) 195–202.
[35] J.W. Shen, J. Zhang, X.M. Luo, W.L. Zhu, K.Q. Yu, K.X. Chen, Y.X. Li, H.L. Jiang, Predicting protein-protein interactions based only on sequences information, Proc. Natl. Acad. Sci. U.S.A. 104 (2007) 4337–4341.
[36] 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.
[37] C.X. Zhang, W. Zheng, P.L. Freddolino, Y. Zhang, MetaGO: predicting Gene Ontology of non-homologous proteins through low-resolution protein structure prediction and protein-protein network mapping, J. Mol. Biol. 430 (2018) 2256–2265.
[38] H.Y. Mi, X.S. Huang, A. Muruganujan, H.M. Tang, C. Mills, D. Kang, P.D. Thomas, PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements, Nucleic Acids Res. 45 (2017) D183–D189.
[39] J.H. Xu, A weighted linear discriminant analysis framework for multi-label feature extraction, Neurocomputing 275 (2018) 107–120.
[40] W.Z. Chen, J. Yan, B.Y. Zhang, Z. Chen, Q. Yang, Document transformation for multi-label feature selection in text categorization, Seventh IEEE ICDM (2007) 451–456.
[41] P.M.E. Shutle, H.M. Cheah, Applying Boltzmann's definition of entropy, Eur. J. Phys. 19 (1998) 371.
[42] M.S. Treder, A.K. Porbadnigk, F.S. Avarvand, K.R. Müller, B. Blankertz, The LDA beamformer: optimal estimation of ERP source time series using linear discriminant analysis, Neuroimage 129 (2016) 279–291.
[43] Z.T. Jiang, D. Wang, P. Wu, Y.H. Chen, Predicting subcellular localization of multisite proteins using differently weighted multi-label k-nearest neighbors sets, Technol. Health Care 27 (2019) 185–193.
[44] M.L. Zhang, L. Wu, LIFT: multi-label learning with label-specific features, IEEE Trans. Pattern Anal. 37 (2014) 107–120.
[45] M. Xu, Y.F. Li, Z.H. Zhou, Robust multi-label learning with PRO Loss, IEEE Trans. Knowl. Data Eng. 32 (2019) 1610–1624.
[46] G. Banjac, P. Goulart, B. Stellatoet, S. Boyd, Infeasibility detection in the alternating direction method of multipliers for convex optimization, J. Optim. Theor. Appl. 183 (2019) 490–519.
[47] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, C.J. Lin, LIBLINEAR: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.
[48] 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.
[49] Y.N. Liu, Z.M. Yu, C. Chen, Y. Han, B. Yu, Prediction of protein crotonylation sites through LightGBM classifier based on SMOTE and elastic net, Anal. Biochem. 609 (2020) 113903.
[50] G.Q. Wu, R.B. Zheng, Y.J. Tian, D.L. Liu, Joint Ranking SVM and Binary Relevance with robust low-rank learning for multi-label classification, Neural Network. 122 (2020) 24–39.
[51] 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. Cyb. 9 (2018) 399–411.
[52] H. Abdi, L.J. Williams, Principal component analysis, Comput. Stat. 2 (2010) 433–459.
[53] 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.
[54] K. Yu, S.P. Yu, V. Tresp, Multi-label informed latent semantic indexing, International ACM SIGIR Conference on Research and Development in Information Retrieval (2005) 258–265.
[55] Y. Zhang, Z.H. Zhou, Multilabel dimensionality reduction via dependency maximization, ACM Trans. Knowl. Discov. 4 (2010) 14.
[56] M. Oikonomou, A. Tefas, Direct multi-label linear discriminant analysis, Commun. Comput. Inf. Sci. 383 (2013) 414–423.
[57] A. Tayal, T.F. Coleman, Y.Y. Li, Bounding the difference between RankRC and RankSVM and application to multi-level rare class kernel ranking, Data Min. Knowl. Discov. 32 (2018) 417–452.
[58] M.L. Zhang, ML-RBF: RBF neural networks for multi-label learning, Neural Process. Lett. 29 (2009) 61–74.
[59] 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.
[60] 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.
[61] 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.
[62] 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.
[63] 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.
[64] 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.
[65] X. Xiao, Z.C. Wu, K.C. Chou, A multi-label learning classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites, PloS One 6 (2011), e20592.
[66] Y.Y. Chu, X.Q. Shan, T.H. Chen, M.M. Jiang, Y.J. Wang, Q.K. Wang, D.R. Salahub, Y. Xiong, D.Q. Wei, DTI-MLCD: predicting drug-target interactions using multi-label learning with community detection method, Brief. Bioinformatics (2020). https://doi.org/10.1093/bib/bbaa205.
[67] J.P. Zhou, L. Chen, Z.H. Guo, iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs, Bioinformatics 36 (2020) 1391–1396.
[68] S. Gull, N. Shamim, F. Minhas, AMAP: hierarchical multi-label prediction of biologically active and antimicrobial peptides, Comput. Biol. Med. 107 (2019) 172–181.