1 **MpsLDA-ProSVM: predicting multi-label protein subcellular** 2 **localization by wMLDAe dimensionality reduction and ProSVM** 3 **classifier**

4 Qi Zhang ^{a,b,1}, Shan Li ^{c,1}, Bin Yu ^{a,b,d,*}, Yang Li ^e, Yandan Zhang ^a, Qin Ma ^e, Yusen Zhang ^f

5 ^a *College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao*
6 *266061, China*

7 ^b *Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science*
8 *and Technology, Qingdao 266061, China*

9 ^c *School of Mathematics and Statistics, Central South University, Changsha 410083, China*

10 ^d *School of Life Sciences, University of Science and Technology of China, Hefei 230027, China*

11 ^e *Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus,*
12 *Ohio 43210, USA*

13 ^f *School of Mathematics and Statistics, Shandong University, Weihai 264209, China*
14

15 **ABSTRACT**

16 Proteins play a significant part in life processes such as cell growth, development, and reproduction.
17 Exploring protein subcellular localization (SCL) is a direct way to better understand the function of
18 proteins in cells. Studies have found that more and more proteins belong to multiple subcellular
19 locations, and these proteins are called multi-label proteins. They not only play a key role in cell
20 life activities, but also play an indispensable role in medicine and drug development. This article
21 first presents a new prediction model, MpsLDA-ProSVM, to predict the SCL of multi-label proteins.
22 Firstly, the physical and chemical information, evolution information, sequence information and
23 annotation information of protein sequences are fused. Then, for the first time, use a weighted multi-
24 label linear discriminant analysis framework based on entropy weight form (wMLDAe) to refine
25 and purify features, reduce the difficulty of learning. Finally, input the optimal feature subset into
26 the multi-label learning with label-specific features (LIFT) and multi-label k-nearest neighbor (ML-
27 KNN) algorithms to obtain a synthetic ranking of relevant labels, and then use Prediction and
28 Relevance Ordering based SVM (ProSVM) classifier to predict the SCLs. This method can rank
29 and classify related tags at the same time, which greatly improves the efficiency of the model. Tested
30 by jackknife method, the overall actual accuracy (OAA) on virus, plant, Gram-positive bacteria and
31 Gram-negative bacteria datasets are 98.06%, 98.97%, 99.81% and 98.49%, which are 0.56%-9.16%,
32 5.37%-30.87%, 3.51%-6.91% and 3.99%-8.59% higher than other advanced methods respectively.
33 The source codes and datasets are available at [https://github.com/QUST-AIBBDRC/MpsLDA-](https://github.com/QUST-AIBBDRC/MpsLDA-ProSVM/)
34 [ProSVM/](https://github.com/QUST-AIBBDRC/MpsLDA-ProSVM/).

35 **Keywords:** Multi-label protein subcellular localization; Multi-information fusion; wMLDAe
36 dimensionality reduction; ProSVM classifier.

* Corresponding author.

E-mail address: yubin@qust.edu.cn (B. Yu).

¹ These authors contributed equally to this work.

1. Introduction

The nature of proteome research refers to studying the characteristics of proteins in a large range, including protein-protein interactions [1, 2], post-translational modifications [3], protein expression levels [4], etc. From this, we can get a comprehensive understanding of cell metabolism, disease occurrence and other processes at the protein level. However, proteins exert their characteristics only in the corresponding subcellular location [5], so determining the protein subcellular location is of great significance for understanding the protein characteristics and functional diversity [6], which is 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 [7]. However, experiments have found that some proteins can appear in two or more different Subcellular position, which are called multi-label proteins [8]. For example, acyl-CoA synthetase is found in both the endoplasmic reticulum and mitochondrial outer membrane [9]; Antioxidant defense proteins are present in mitochondria, peroxisome and cytosol [10]. These multi-label proteins play a critical role in physical transportation, information transferring, energy conversion and other life activities of cells [11]. 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 and structure information. For example, in order to explore the function of viral proteins, Thakur et al. [12] used three different feature extraction algorithms dipeptide composition (DC), amino acid composition (AAC), physicochemical properties (PHY) and their hybrids to develop a server "MSLVP" for predicting SCLs. Zhou et al. [13] used position-specific scoring matrix (PSSM), amino acid composition (AAC), conserved domain database (CDD) and gene ontology (GO) annotation to extract and fuse the features of human proteins and propose a new method Hum-mPLoc 3.0 based on amino acid sequence. Cheng et al. [14] fused GO information and pseudo amino acid composition (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. [15] 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 fuzzy rough set to complete feature selection. According to information theory, Zhang et al. [16] 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.

[17] 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 the optimal estimation accuracy. The convergence of the algorithm was achieved by using the iterative optimization algorithm.

Machine learning algorithm such as classifiers acts as the last step for SCL prediction of multi-label proteins. Recently, Wang et al. [18] used the algorithm of integrating multiple classifier chains (ECC) for prediction on the multi-label protein datasets of Gram-negative bacteria and Gram-positive bacteria. The overall prediction accuracy (OLA) and the overall actual accuracy (OAA) about these two datasets were 94.1%, 92.4% and 94.4%, 94.0%, respectively. Wan et al. [19] 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 [8] used 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, four feature coding methods, PseAAC, CT, PsePSSM and GO, are used to draw the feature information from protein sequences, and the information is feature-fused. Secondly, for the first time, weighted multi-label linear discriminant analysis framework based on entropy weight form (wMLDAe) is used to remove redundancy and get the best feature vectors. At last, these feature vectors are first input to 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 jackknife method, and the overall actual accuracy (OAA) on virus, plant, Gram-positive bacteria and Gram-negative bacteria datasets are 98.06%, 98.97%, 99.81% and 98.49%, respectively, and the overall location accuracy (OLA) are 98.41%, 99.33%, 99.81% and 99.04% respectively. 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 207 different protein sequences in the virus dataset [20], which are 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 [21] has 978 different protein sequences, which are 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.

The Gram-positive bacteria [22] contains 519 different protein sequences, which are located in 4 subcellular positions, among which 515 belong to one position and 4 belong to two positions. The Gram-negative bacteria [23] contains 1392 distinct protein sequences, which are 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. The homology of any one protein with other protein sequences in the same subcellular location is limited to below 25%.

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%.

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%.

2.2. Feature extraction

In the article, we use pseudo position-specific scoring matrix, GO, 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

The pseudo amino acid composition (PseAAC) [7] model put forward by Chou [24] can not only extract the location information of protein sequence, but also reflect the essential 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:

$$P = [p_1, p_2, \dots, p_{20}, p_{20+1}, \dots, p_{20+\lambda}]^T \quad (1)$$

Among them,

$$p_u = \begin{cases} \frac{f_\mu}{\sum_{\mu=1}^{20} f_\mu + \omega \sum_{k=1}^{\lambda} \tau_k}, & 1 \leq \mu \leq 20 \\ \frac{\omega \tau_{\mu-20}}{\sum_{\mu=1}^{20} f_\mu + \omega \sum_{k=1}^{\lambda} \tau_k}, & 21 \leq \mu \leq 20 + \lambda \end{cases} \quad (2)$$

where ω represents the weighting factor, the value is 0.05 [25], and refers to the order information between the amino acids in the protein sequence. 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

Pseudo position-specific scoring matrix (PsePSSM) [26] draws evolution information from sequences with this method following two steps. First, in the NR database, align the protein sequences using the PSI-BLAST [27] program to obtain a position-specific scoring matrix (PSSM) [28]. For protein sequences of length L , the PSSM algorithm obtains $L \times 20$ -dimensional feature vectors, and the results are as follows:

$$P_{PSSM} = \begin{pmatrix} P_{1,1} & P_{1,2} & \dots & P_{1,20} \\ \dots & \dots & \dots & \dots \\ P_{i,1} & P_{i,2} & \dots & P_{i,20} \\ \dots & \dots & \dots & \dots \\ P_{L,1} & P_{L,2} & \dots & P_{L,20} \end{pmatrix} \quad (3)$$

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:

$$P_{sePSSM} = (\overline{P_1}, \overline{P_2}, \dots, \overline{P_{20}}, \theta_1^1, \theta_2^1, \dots, \theta_{20}^1, \dots, \theta_1^\xi, \theta_2^\xi, \dots, \theta_{20}^\xi) \quad (4)$$

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

In order to draw sequence information from proteins, Shen et al. [29] proposed conjoint triad (CT) [30, 31] algorithm based on the dipole and volume of amino acid side chains. First, 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. Calculate the frequency of each triplet to form a 343-dimensional feature vector. The formula is as follows:

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

2.2.4. Gene ontology

Gene ontology 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: biological processes, molecular functions, and cell composition. Use BLASTP to search for homologous proteins in the SWISS-PROT database and remove sequences with similarities $\geq 60\%$. For the remaining protein sequences S_i , the GO information [32, 33] of the corresponding protein is determined in the GO database according to the index number in the sequence. Its manifestation is:

$$p_i = [G_1, G_2, \dots, G_j, \dots, G_{|S_i|}]^T, \quad i = 1, 2, \dots, |S_i| \quad (6)$$

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

Linear discriminant analysis is one of the most widely used feature selection methods. For multi-label data, Xu [34] proposed a weighted multi-label linear discriminant analysis framework based on binary and related weight forms. Assuming l multi-label training instances and q class label sets are given, define W as $l \times q$ -dimensional non-negative weight matrix, that is:

$$W = [w_1, w_2, \dots, w_l]^T \quad (7)$$

where $w_i = [w_{i1}, \dots, w_{ik}, \dots, w_{iq}]^T, i < l$ represents the i -th instance, define the between-class, within-class and total scatter matrices as:

$$S_b^M = \sum_{k=1}^q \left(\sum_{i=1}^l w_{ik} \right) (m_k^M - m^M)(m_k^M - m^M)',$$

$$S_w^M = \sum_{k=1}^q \sum_{i=1}^l w_{ik} (x_i - m_k^M)(x_i - m_k^M)', \quad (8)$$

$$S_t^M = \sum_{k=1}^q \sum_{i=1}^l w_{ik} (x_i - m^M)(x_i - m^M)',$$

where m^M is the total mean vector and m_k^M is the mean vector of the k -th class. Chen [35] 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 [36]. To the i -th instance, the entropy can be defined as:

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

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

$$w_{ik} = e^{-h_i} = \frac{1}{m_i} = \frac{1}{\|y_i\|_1}, \quad (10)$$

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:

$$S_t^M = \sum_{k=1}^q \sum_{i=1}^l w_{ik} (x_i - m^M)(x_i - m^M)' = \sum_{i=1}^l (x_i - m^M)(x_i - m^M)', \quad (11)$$

The results of the formula (11) are defined in the same way as those of the single-label LDA algorithm [37]. In the wMLDA algorithm, the projection matrix P is obtained by solving the generalized eigenvalue problem, that is:

$$S_b^M P = \Lambda S_w^M P. \quad (12)$$

where Λ is composed of C non-negative eigenvalues, that is $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_C \geq 0$. A wMLDA having the form of the above entropy weight is added with a suffix 'e', which is called wMLDAe. In this paper, the algorithm is used to process the fused vector, taking full account of the label information, avoiding over-fitting of the model, and improving the running speed.

2.4. Prediction and Relevance Ordering based SVM

Assume that a set of n training instances $\{x_i\}_{i=1}^n$ and T label $L = \{l_1, l_2, \dots, l_T\}$ are given. In order to consider the comprehensive ranking of relevant labels, we first use the ML-KNN [38] and LIFT [39] algorithms to obtain the label score of each training instance, denoted as $g(x_i) = [g_1(x_i), g_2(x_i), \dots, g_T(x_i)]$. Each label $l_t, t \in \{1, 2, \dots, T\}$ 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) [40] 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:

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

where

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

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

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

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 $\bar{w}$.

The Alternating Direction Method of Multipliers (ADMM) [41] 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:

$$\begin{aligned} \bar{w}_k^0 &= \arg \min_{\bar{w}^0} L\left(\bar{w}^0, \left\{ \bar{w}_k^n, \alpha_{k-1}^n \right\}_{n=1}^N, \eta\right), \\ \alpha_k^n &= \alpha_{k-1}^n + \eta \left(\bar{w}_k^n - \bar{w}_k^0 \right)', \quad \forall n = 1, 2, \dots, N \end{aligned} \quad (16)$$

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

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

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 [42] to obtain the score of each instance, and classify the instances by the threshold.

2.5. Performance evaluation

According to Chou's five-step rule [43], choosing appropriate indicators to judge the quality of a prediction system is an important steps to build a prediction model. At present, the commonly used inspection methods include re-substitution, independent test, k-fold cross-validation and jackknife. Among them, jackknife method used in this paper is regarded as one of the most rigorous inspection methods [26]. For the jackknife test, specifically, any one of the N protein sequences is selected as testing, and the other N-1 sequences are used as training, which is repeated for N times until all protein 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) [40, 44, 45]. Suppose a given multi-label training set $D = \{x_i, y_i\}_{i=1}^n$ and T learning functions $F = \{f_1, f_2, \dots, f_T\}$, where $y_i \in \{1, -1\}^T$ is the real label of example x_i , and $H = \{h_1, h_2, \dots, h_T\}$ 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 as follows:

$$HL = \frac{1}{nT} \sum_{i=1}^n \sum_{j=1}^T \mathbb{I}[h_j(x_i) \neq y_{ij}] \quad (18)$$

$$F1 = \frac{1}{n} \sum_{i=1}^n \frac{2|P_i^+ \cap Y_i^+|}{|P_i^+| + |Y_i^+|} \quad (19)$$

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

$$RL = \frac{1}{n} \sum_{i=1}^n \frac{|SetR_i|}{|Y_i^+| + |Y_i^-|} \quad (21)$$

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

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

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

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

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

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

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

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:

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

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 article, we put forward a multi-label protein SCL prediction method, MpsLDA-ProSVM, and the pipeline is shown in Fig. 3. The experiment environment is: windows Server 2012R2 Intel (R) Xeon (TM) CPU E5-2650 @ 2.30GHz 2.30GHz with 32.0GB of RAM, MATLAB2014a.

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 the jackknife, 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.

Fig. 3. The prediction pipeline of MpsLDA-ProSVM. (A) The protein sequences of the training and testing sets and their label vectors. (B) PsePSSM, CT, PseAAC, GO are used for feature extraction and fusion. (C) wMLDAe algorithm is used to reduce dimensions and select the optimal feature subset. (D) First, we use ML-KNN and LIFT algorithm to get the comprehensive ranking of relevant labels, then input it into ProSVM classifier for prediction, and construct the MpsLDA-ProSVM model. (E) Tested by the jackknife method and compared with other methods in the paper to assess the performance of the 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. Set the λ value at 5 intervals. Because the shortest protein sequence length in the independent test set is 50, the value of parameter λ ranges from 5 to 49. The results of feature extraction under different parameters are input into the ProSVM classifier, tested by jackknife method, 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.

It can be seen from Fig. 4 that the prediction result of the classifier varies with different parameter values. The smaller the PL value, the better the performance of the model. 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.

334 For the value of parameter ξ , when using PsePSSM algorithm for feature extraction, select $\xi=9$.
 335 At this time, a protein sequence generates a $(20+20 \times \xi=200)$ -dimensional feature vector.

336
 337 **Fig. 4.** The effect of different parameters on PL value of virus and plant datasets. (A) The PL value
 338 obtained using the PseAAC algorithm; (B) The PL value obtained using the PsePSSM algorithm.

339 3.2. Effect of feature extraction algorithm on results

340 The PseAAC and PsePSSM algorithm are used to draw the physical and chemical information
 341 and evolution information from sequence. CT algorithm and GO annotation are used to draw the
 342 sequence information and annotation information. However, using only a single feature coding
 343 method cannot fully extract useful information about protein sequences, so researchers have fused
 344 the feature vectors extracted from various kinds of information together. To deal with the
 345 "dimensional disaster" after fusion, wMLDAe is used to reduce the number of dimensions generated
 346 after fusion. To verify the superiority of feature fusion, we use ProSVM classifier to predict the
 347 information after single-feature and multi-feature fusion. The jackknife method is used for testing,
 348 and PL, HL, RL, and F1 are used as evaluation indicators to assess the performance. The prediction
 349 results of virus and plant datasets using four single feature extraction algorithms (GO, PsePSSM,
 350 CT, PseAAC) and a fusion method using wMLDAe dimensionality reduction
 351 (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)

In Table1 that for the 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 17.31%, 14.50% and 23.71% 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%. Similarly, for the 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.

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

3.3. Effect of dimensional reduction algorithm on results

In this study, drawing the feature information by four feature extraction algorithms are fused, and many redundant features are also brought. This kind of noise information has some influence on protein SCL. To obtain the optimal feature vector, PCA [46], MVMD [47], MLSI [48], MDDM

[49], DMLDA [50] 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 the wMLDAe algorithm, in order to more clearly analyze the results when the virus and plant datasets take different dimensions, the specific results are shown in Supplementary Table S5 and S6.

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.

It can be known from Fig. 5 that when different dimensionality reduction methods and different dimensions are selected, the prediction results obtained are also different. Compared with other algorithm, wMLDAe algorithm has the best prediction performance. For the 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 the 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 can reduce noise and redundant information by introducing entropy weight. The wMLDAe algorithm is significantly better than the PCA, MVMD, MLSI, MDDM and DMLDA algorithms. It shows that the algorithm can not only remove the irrelevant information in the feature vector, but also retain the effective features in the sequence and form the optimal feature subset, so as to promote the running speed and increase the prediction accuracy about the model. Therefore, we choose wMLDAe algorithm for dimension reduction and select the dimension of 50. The performance of this model is optimal.

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 [38], LIFT [39], Rank-SVM [51], ML-RBF [52], ML-GKR [14] and ProSVM classification algorithms to perform protein SCL prediction on virus and plant datasets, respectively. The jackknife method 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.

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.

From Fig. 6, when we use the ProSVM algorithm, the prediction performance is the best. For the 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. The prediction results obtained by using ProSVM algorithm in other indicators are also significantly better than other algorithms. Similarly, for the 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 use kernel functions for mapping to high-dimensional space to optimize nonlinear classification problems. Both are susceptible to the choice of kernel functions and parameters, and for Rank-SVM algorithm, it is difficult to run when the amount of data is large. For the ML-RBF algorithm, it is generally combined with K-means

clustering, which has the problems of low cluster quality and unstable performance. As the increase of training samples, the number of hidden neurons in RBF network also increases, which increases the complexity of the network and the amount of calculation. From the virus and plant datasets, it can also be concluded that the prediction effect of ML-RBF is not ideal. For ML-GKR algorithm, the benchmark data is decomposed by binary correlation, so that the multi-label problem can be converted into a single-label problem. However, 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.

3.5. Comparison with other methods

In the paper, to clarify the prediction performance about this model, we chose to use the same evaluation index for comparison based on the jackknife method. In this paper, the virus and plant datasets are predicted. Firstly, features are extracted and fused by using PseAAC, PsePSSM, GO information 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. The comparative results of virus and plant datasets are shown in Table 3 and 4.

Table 3

Comparison of prediction results of different methods in virus dataset.

Locations		Methods				
		mGOASVM [53]	HybridGO- Loc [54]	mPLR- Loc [55]	MSLVP [12]	ProSVM
1	Viral capsid	1.0000	1.0000	1.0000	0.9920	1.0000
2	Host cell membrane	0.9700	0.9700	0.9090	0.9920	0.9697
3	Host endoplasmic reticulum	0.8500	0.9000	0.8500	0.9960	1.0000
4	Host cytoplasm	0.9770	0.9660	0.9890	0.9880	0.9885
5	Host nucleus	0.9760	0.9880	0.964	0.9960	0.9762
6	Secreted	1.0000	1.0000	0.8500	0.9960	1.0000
PL		-	-	-	-	0.0133
HL		0.0260	0.0160	0.0230	-	0.0064
RL		-	-	-	-	0.0144
AP		0.9390	0.9650	0.957	-	0.9961
CV		-	-	-	-	0.0466
F1		0.9500	0.9680	0.9550	-	0.9885
OAA		0.8890	0.9370	0.9030	0.9750	0.9806
OLA		0.9680	0.9720	0.9480	0.9930	0.9841

Note: '-' indicates that the corresponding references do not provide the results of the corresponding indicators.

Table 4

Comparison of prediction results of different methods in plant dataset.

	Locations	Methods				
		iLoc-Plant [56]	mGOASVM [53]	HybridGO-Loc [54]	mPLR-Loc [55]	ProSVM
1	Cell membrane	0.6960	0.9460	0.9110	0.8930	1.0000
2	Cell wall	0.5940	0.8440	0.8750	0.7810	0.9688
3	Chloroplast	0.8810	0.9510	0.9720	0.9830	1.0000
4	Cytoplasm	0.6260	0.9560	0.9230	0.9010	0.9890
5	Endoplasmic reticulum	0.5000	0.9050	0.9050	0.8330	1.0000
6	Extracellular	0.0910	1.0000	0.9550	0.8640	1.0000
7	Golgi apparatus	0.7620	0.9050	0.9050	0.8570	0.9524
8	Mitochondrion	0.7470	1.0000	0.9930	0.9930	1.0000
9	Nucleus	0.9210	0.9930	0.9870	0.9610	0.9934
10	Peroxisome	0.2860	1.0000	1.0000	1.0000	1.0000
11	Plastid	0.1790	1.0000	0.9740	0.9230	0.9744
12	Vacuole	0.5380	0.9420	0.9230	0.9420	0.9808
PL		-	-	-	-	0.0021
HL		-	0.0130	0.0070	0.0100	0.0015
RL		-	-	-	-	0.0047
AP		-	0.9330	0.9720	0.9560	0.9978
CV		-	-	-	-	0.0114
F1		-	0.9420	0.9660	0.9490	0.9945
OAA		0.6810	0.8740	0.9360	0.9080	0.9897
OLA		0.7170	0.9620	0.9560	0.9370	0.9933

Note: ‘-’ indicates that the corresponding references do not provide the results of the corresponding indicators.

It can be seen from Table 3 and 4, for virus dataset, when using MpsLDA-ProSVM method, the OAA and OLA are 98.06% and 98.41% respectively, which are 4.36%~9.16% and 1.21%~3.61% higher than those using mGOASVM, HybridGO-Loc and mPLR-Loc methods. Although the OLA value obtained by using the MSLVP method is 0.89% higher than that under the MpsLDA-ProSVM method, the OAA value is 0.56% lower than that. For plant dataset, when using MpsLDA-ProSVM method, the OAA and OLA are 98.97% and 99.33% respectively, which are 5.37%~30.87% and 3.13%~27.63% higher than those using iLoc-Plant, mGOASVM, HybridGO-Loc and mPLR-Loc methods. Moreover, MpsLDA-ProSVM is superior to other methods in the prediction of each location of the two datasets.

So as to further verify the rationality and validity about this model, the datasets of Gram-positive bacteria and Gram-negative bacteria are used as test sets, and by the jackknife method, we obtain the specific prediction results about two test sets. Selecting the same dataset for comparison, the comparison results are shown in Table 5 and Table 6. For Gram-positive bacteria dataset, when using MpsLDA-ProSVM method, the OAA and OLA are all 99.81%, which are 3.51%~6.91% and 3.01%~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 3.74%~7.64% higher than those using other methods. And in each protein, 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.

Table 5

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

Locations		Methods			
		iLoc-Gpos [57]	Gpos- ECC- mPLoc [18]	Gram- LocEN [58]	MpsLDA- ProSVM
1	Cell membrane	0.9600	0.9650	0.9770	1.0000
2	Cell wall	0.6670	0.6670	0.9440	1.0000
3	Cytoplasm	0.9520	0.9620	0.9710	1.0000
4	Extracellular	0.8940	0.9270	0.9510	0.9918
PL		-	-	-	2.4085E-04
HL		-	-	0.0160	4.8170E-04
RL		-	-	-	9.6339E-04
AP		-	-	0.7100	1.0000
CV		-	-	-	0.0029
F1		-	-	0.9700	0.9994
OAA		0.9290	0.9400	0.9630	0.9981
OLA		0.9310	0.9440	0.9680	0.9981

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

Table 6

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

Locations		Methods			
		iLoc- Gneg [59]	Gneg-ECC- mPLoc [18]	Gram- LocEN [58]	MpsLDA- ProSVM
1	Cell inner membrane	0.9680	0.9550	0.9710	1.0000
2	Cell outer membrane	0.8310	0.9440	0.8950	1.0000
3	Cytoplasm	0.8950	0.9220	0.9240	0.9659
4	Extracellular	0.8650	0.9320	0.9700	1.0000
5	Fimbrium	0.9380	0.9380	1.0000	1.0000
6	Flagellum	1.0000	1.0000	1.0000	1.0000
7	Nucleoid	0.5000	0.8750	0.8750	1.0000
8	Periplasm	0.8940	0.9440	0.8940	1.0000
PL		-	-	-	0.0031
HL		-	-	0.0270	0.0019
RL		-	-	-	0.0057
AP		-	-	0.9520	0.9962
CV		-	-	-	0.0081
F1		-	-	0.9420	0.9950
OAA		0.8990	0.9240	0.9450	0.9849
OLA		0.9140	0.9410	0.9530	0.9904

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

4. Conclusion

The precise location of the SCL of 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, MpsLDA-ProSVM, to predict the SCL of multi-label proteins. First, the fusion of the feature vectors extracted by PseAAC, PsePSSM, GO and CT can fully mine the annotation information, sequence information, evolution information and physicochemical information in protein sequences. 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, fully considers the label information, and improves the speed of 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 use ML-KNN and LIFT algorithm to get the predicted label scores of each instance, summarize the scores and sort them to get the ranking, and then the objective function is optimized according to the ADMM algorithm to obtain the final prediction result. This algorithm further improves the efficiency of the model. 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. 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.

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

This work was supported by the National Nature Science Foundation of China (61863010, 61877064, U1806202), the Key Research and Development Program of Shandong Province of China (2019GGX101001), and the Natural Science Foundation of Shandong Province of China (ZR2018MC007).

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] Z.R. Wu, Q. Liao, B. Liu, idenPC-MIIP: identify protein complexes from weighted PPI networks using mutual important interacting partner relation, *Brief. Bioinform.* 2020 doi: 10.1093/bib/bbaa016.
- [3] L. Zhang, G.X. Yu, D.W. Xia, J. Wang, Protein–protein interactions prediction based on ensemble deep neural networks, *Neurocomputing* 324 (2019) 10-19.
- [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 (4) (2019) 570-583.
- [5] S. Cockcroft, P. Raghu, Phospholipid transport protein function at organelle contact sites, *Curr. Opin. Cell Biol.* 53 (2018) 52-60.
- [6] 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.
- [7] 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.
- [8] F. Javed, M. Hayat, Predicting subcellular localization of multi-label proteins by incorporating the sequence features into Chou's PseAAC, *Genomics* 111 (6) (2019) 1325-1332.
- [9] M.A. Block, J. Jouhet, Lipid trafficking at endoplasmic reticulum–chloroplast membrane contact sites, *Curr. Opin. Cell Biol.* 35 (2015) 21-29.
- [10] J.C. Mueller, C. Andreoli, H. Prokisch, T. Meitinger, Mechanisms for multiple intracellular localization of human mitochondrial proteins, *Mitochondrion* 3 (6) (2004) 315–325.
- [11] L. Liu, L. Tang, X. Jin, W. Zhou, A multi-label supervised topic model conditioned on arbitrary features for gene function prediction, *Genes* 10 (1) (2019) 57.
- [12] A. Thakur, A. Rajput, M. Kumar, MSLVP: prediction of multiple subcellular localization of viral proteins using a support vector machine, *Mol. BioSyst.* 12 (8) (2016) 2572.
- [13] 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 (6) (2016) 843-853.
- [14] 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 (3) (2018) 398-406.
- [15] 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.
- [16] P. Zhang, G.X. Liu, W.F. Gao, Distinguishing two types of labels for multi-label feature selection, *Pattern Recogn* 95 (2019) 72-82.

569 [17] S.B. Chen, Y.M. Zhang, C.H.Q. Ding, J. Zhang, B. Luo, Extended adaptive Lasso for multi-
570 class and multi-label feature selection, *Knowl-Based Syst.* 173 (2019) 28-36.

571 [18] X. Wang, J. Zhang, G.Z. Li, Multi-location gram-positive and gram-negative bacterial protein
572 subcellular localization using gene ontology and multi-label classifier ensemble, *BMC*
573 *Bioinformatics* 16 (S12) (2015) S1.

574 [19] S.B. Wan, Y.C. Duan, Q. Zou, HPSLPred: an ensemble multi-label classifier for human protein
575 subcellular location prediction with imbalanced source, *Proteomics* 17 (17-18) (2017) 17-18.

576 [20] H.B. Shen, K.C. Chou, Virus-mPLoc: a fusion classifier for viral protein subcellular location
577 prediction by incorporating multiple sites, *J Biomol Struct Dyn*, 26 (2010) 175–186.

578 [21] K.C. Chou, H.B. Shen, Plant-mPLoc: A top-down strategy to augment the power for predicting
579 plant protein subcellular localization, *Plos One* 5 (2010) e11335.

580 [22] H.B. Shen, K.C. Chou, Gpos-mPLoc: a top-down approach to improve the quality of predicting
581 subcellular localization of Gram-positive bacterial proteins, *Protein Pept. Lett.* 16 (12) (2009)
582 1478–1484.

583 [23] H.B. Shen, K.C. Chou, Gneg-mPLoc: A top-down strategy to enhance the quality of predicting
584 subcellular localization of Gram-negative bacterial proteins, *J. Theor. Biol.* 264 (2010) 326–
585 333.

586 [24] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily
587 classes, *Bioinformatics* 21 (1) (2005) 10-19.

588 [25] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition,
589 *Proteins* 43 (3) (2010) 246-255.

590 [26] B. Yu, S. Li, W.Y. Qiu, M.H. Wang, J.W. Du, Y.S. Zhang, X.Chen, Prediction of subcellular
591 location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on
592 LFDA dimensionality reduction, *BMC Genomics* 19 (1) (2018) 478.

593 [27] Y. Liang, S. Zhang, S. Ding, Accurate prediction of Gram-negative bacterial secreted protein
594 types by fusing multiple statistical features from PSI-BLAST profile, *Sar Qsar Environ. Res.*
595 29 (6) (2018) 469-481.

596 [28] B. Liu, C.C. Li, K. Yan, DeepSVM-fold: protein fold recognition by combining support vector
597 machines and pairwise sequence similarity scores generated by deep learning networks, *Brief.*
598 *Bioinform.* 2019 doi: 10.1093/bib/bbz098.

599 [29] J. Shen, J. Zhang, X.M. Luo, W.L. Zhu, K.Q. Yu, K.X. Chen, Y.X. Li, H.L. Jiang, Predicting
600 protein-protein interactions based only on sequences information, *Proc. Natl. Acad. Sci.*
601 *USA*.104 (11) (2007) 4337–4341.

602 [30] L.Y. Wei, P. Xing, G. Shi, Z.L. Ji, Q. Zou, Fast prediction of protein methylation sites using a
603 sequence-based feature selection technique, *IEEE/Acm T. Comput. Bi.* 16 (2017) 1264-1273.

604 [31] N. Xiao, D.S. Cao, M.F. Zhu, Q.S. Xu, protr/ProtrWeb: R package and web server for
605 generating various numerical representation schemes of protein sequences, *Bioinformatics* 31
606 (11) (2015) 1857-1859.

- [32] 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 (15) (2018) 2256-2265.
- [33] 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 (D1) (2017) D183-D189.
- [34] J.H. Xu, A weighted linear discriminant analysis framework for multi-label feature extraction, *Neurocomputing* 275 (2018) 107-120.
- [35] W. 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.
- [36] P.M.E. Shulte, H.M. Cheah, Applying Boltzmann's definition of entropy, *Eur. J. Phys.* 19 (4) (1998) 371.
- [37] 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.
- [38] 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, *The 7th International Conference on Biomedical Engineering and Biotechnology* 27 (5) (2019) 185-193.
- [39] M.L. Zhang, L. Wu, LIFT: multi-label learning with label-specific features, *IEEE T. Pattern Anal.* 37 (1) (2014) 107-120.
- [40] M. Xu, Y.F. Li, Z.H. Zhou, Robust multi-label learning with PRO Loss, *IEEE T. Knowl Data En.* (2019) doi: 10.1109/TKDE.2019.2908898.
- [41] G. Banjac, P. Goulart, B. Stellatoet, S. Boyd, Infeasibility detection in the alternating direction method of multipliers for convex optimization, *J. Optimiz. Theory App.* 183 (2019) 490-519.
- [42] 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 (9) (2008) 1871-1874.
- [43] 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.
- [44] 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 Networks* 122 (2020) 24-39.
- [45] 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 (3) (2018) 399-411.
- [46] H. Abdi, L.J. Williams, Principal component analysis, *Wires Comput. Statistics* 2 (4) (2010) 433-459.

- [47] 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.
- [48] K. Yu, S.P. Yu, V. Tresp, Multi-label informed latent semantic indexing, *International Acm Sigir Conference on Research & Development in Information Retrieval.* ACM (2005) 258-265.
- [49] Y. Zhang, Z.H. Zhou, Multilabel dimensionality reduction via dependency maximization, *ACM Trans. Knowl. Discov.* 4 (3) (2010) 14.
- [50] M. Oikonomou, A. Tefas, Direct multi-label linear discriminant analysis, *Communications in Computer & Information Science* 383 (2013) 414-423.
- [51] A. Tayal, T.F. Coleman, Y. Li, Bounding the difference between RankRC and RankSVM and application to multi-level rare class kernel ranking, *Data Min. Knowl. Disc.* 32 (2) (2018) 417-452.
- [52] M.L. Zhang, ML-RBF: RBF neural networks for multi-label learning, *Neural Process Lett.* 29 (2) (2009) 61-74.
- [53] S.B. Wan, M.W. Mak, S.Y. Kung, mGOASVM: multi-label protein subcellular localization based on gene ontology and support vector machines, *BMC Bioinformatics* 13 (1) (2012) 290.
- [54] 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 (3) (2014) e89545.
- [55] 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.
- [56] 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 (12) (2011) 3287-3297.
- [57] 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 (1) (2012) 4-14.
- [58] 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.
- [59] 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 (6) (2011) e20592.