Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Original Paper

Subject Section

Predicting the multi-label protein subcellular localization through multi-information fusion and MLSI dimensionality reduction based on MLFE classifier

Yushuang Liu[1,2], Shuping Jin[1,2], Hongli Gao[1,2], Xue Wang[1,2], Congjing Wang[1,2], Weifeng Zhou[1,2] and Bin Yu[1,2,*]

[1]College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China

[2]Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao 266061, China

Associate Editor: XXXXXXX
Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Multi-label protein subcellular localization (SCL) is an indispensable way to study protein function. It can locate a certain protein (such as the human transmembrane protein that promotes the invasion of the SARS-CoV-2) or expression product at a specific location in a cell, which can provide a reference for clinical treatment of diseases such as COVID-19.

Results: The paper proposes a novel method named ML-locMLFE. First of all, six feature extraction methods are adopted to obtain protein effective information. These methods include pseudo amino acid composition (PseAAC), encoding based on grouped weight (EBGW), gene ontology (GO), multi-scale continuous and discontinuous (MCD), residue probing transformation (RPT) and evolutionary distance transformation (EDT). In the next part, we utilize the multi-label information latent semantic index (MLSI) method to avoid the interference of redundant information. In the end, multi-label learning with feature induced labeling information enrichment (MLFE) is adopted to predict the multi-label protein SCL. The Gram-positive bacteria dataset is chosen as a training set, while the Gram-negative bacteria dataset, virus dataset, newPlant dataset and SARS-CoV-2 dataset as the test sets. The overall actual accuracy (OAA) of the first four datasets is 99.23%, 93.82%, 93.24%, and 96.72% by the leave-one-out cross validation (LOOCV). It is worth mentioning that the OAA prediction result of our predictor on the SARS-CoV-2 dataset is 72.73%. The results indicate that the ML-locMLFE method has obvious advantages in predicting the SCL of multi-label protein, which provides new ideas for further research on the SCL of multi-label protein.

Availability and implementation: The source codes and data are publicly available at https://github.com/QUST-AIBBDRC/ML-locMLFE/.

Contact: yubin@qust.edu.cn.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The structure and function of protein are various, but they can only play a role in the right SCL (Costa, et al., 2018; Chu, et al., 2020). When protein structure changes, it will cause diseases, such as kidney disease (Ivanova, et al., 2008), myocarditis (Jang, et al., 1988), diabetes (Brownlee, 1995), dermatomyositis (Brownlee, 1995), muscle atrophy (Sneddon, et al., 2000). With the continuous increase of data and the expansion of research directions (Wan, et al., 2015; Wan, et al., 2017), the traditional machine learning methods cannot achieve good prediction results (Zhang, et al., 2020; Marilyn, et al., 2020). Firstly, the traditional machine learning methods are time-consuming and labor-intensive. Secondly, the protein not only exists in one SCL, but also may exist in two or multiple SCL. The prediction method for a single protein site ignores the situation of two or more subcellular locations (Du, et al., 2020; Faisal, et al., 2020). Finally, the high-dimensional space formed by multi-information fusion increases the interference of redundant information on the prediction results (Yu, et al., 2018). Therefore,

this paper mainly optimizes the feature extraction, feature selection and classifier to improve the prediction accuracy.

Since protein sequences cannot be directly used for calculation, it must be transformed into digital information for further study (Yu, et al., 2020a). Zhang et al. (Zhang, et al., 2020) utilized dipeptide composition (DC), position specific scoring matrix (PsePSSM), PseAAC, GO, and EBGW to extract protein information from relevant datasets. Wu et al. (Wu, et al., 2012) adopted GO and evolutionary information to develop a new predictor iLoc-Gpos in Gram-positive bacteria dataset. Wan et al. (Wan, et al., 2014) employed the relationship between GO terms to predict the SCL of plant dataset. Zhang et al. (Zhang, et al., 2021b) used position specific scoring matrix-transition probability composition (PSSM-TPC), DC, GO, PseAAC, PsePSSM and differential evolution (DE) algorithm to assign five single feature weight vector.

The feature fusion method can combine multiple information of protein sequences (Yu, et al., 2021). But the interference of redundant information on the prediction results will gradually increase (Shi, et al., 2019; Fan, et al., 2021) with the increase of dimension. In order to eliminate the useless features in the original space, researchers have proposed a variety of dimensionality reduction methods. Zhang et al. (Zhang, et al., 2019) put forward a manifold regularized discriminant feature selection (MDFS) algorithm to improve performance by optimizing feature selection framework and considering label correlation. Zhang and Zhou (Zhang and Zhou, 2010) suggested a multi-label dimension reduction method based on dependency maximization (MDDM) to maximize the dependency of original features and related category labels to make the dimension reduction process more efficient. Xu et al. (Xu, et al., 2016) came up with the multi-label feature extraction algorithm via feature variance and feature-label dependence (MVMD) method, which integrated two least squares formulas and used the maximum feature variance and the correlation of feature label to select the best feature vector. Zhang et al. (Zhang, et al., 2020) presented global relevance and redundancy optimization (GRRO) method composed of feature relevance, label relevance and feature redundancy, which greatly improved computing efficiency.

Choosing a suitable classifier is crucial for predicting the SCL of proteins. Wan et al. (Wan, et al., 2018; Wang, et al., 2021) proposed an adaptive decision-making scheme for support vector machines (AD-SVM) to obtain the OAA on virus dataset was 93.24%, and the OLA was 96.03%. Wang et al. (Wang, et al., 2015) used the ensemble multiple classifier chain (ECC) to predict the protein SCL of Gram-negative bacteria dataset, and the OAA was 94.03%, the OLA was 94.46%. Shen et al. (Shen, et al., 2019) used the multi-kernel support vector machine classifier to predict the two human datasets multi-label protein SCL, and the average precision reached 70.65% and 68.89%, respectively, compared with the results of other methods, the result was the best.

In order to improve the accuracy of prediction, we propose a new model called ML-locMLFE to predict the SCL of multi-label protein. Six feature extraction methods are used to transform protein sequences into digital information. Therefore this paper needs to fuse six types of feature information .Then we use the MLSI to classify and recognize the most effective information from many features. Finally, the MLFE is utilized to predict the SCL of multi-label protein. Compared with other methods, ML-locMLFE is more superior in predicting the SCL of multi-label protein.

2 Materials and methods

2.1 Datasets

Five datasets are used to verify the effectiveness of the model. The Gram-positive bacteria dataset (Dehzangi, et al., 2015) is the training set, while the Gram-negative bacteria dataset (Dehzangi, et al., 2015), the virus dataset (Shen, et al., 2010), the SARS-CoV-2 dataset (Zhang, et al., 2020) and the newPlant dataset (Wan, et al., 2012) are the test sets together. The Gram-positive bacteria dataset, Gram-negative bacteria dataset, virus dataset come from the Swiss-Prot database, and the breakdown of each dataset is shown in Supplementary Tables S1-S3. We have obtained data from the UniProt database of the past three years to construct a new plant dataset (named as the newPlant dataset). The detailed breakdown is given in Supplementary Table S4. As a newly mutated coronavirus, the SARS-CoV-2 can cause great harm to human health. Therefore, the accurate identification of the SCL of the SARS-CoV-2 protein is helpful to analyze the pathogenic mechanism of the virus. The SARS-CoV-2 dataset is constructed from the UniProt database, and the detailed breakdown is shown in Supplementary Table S5. The homology of the five datasets is less than 25%.

2.2 Feature encoding

The quality of features has a crucial impact on the predictive ability of the model. Therefore, a suitable feature extraction method is an extremely critical step in predicting the SCL of multi-label protein. Six methods, namely PseAAC, EBGW, GO, RPT, EDT and MCD, are adopted here.

PseAAC: PseAAC is a commonly used feature extraction method to predict SCL. According to Chou et al. (Chou, 2010), PseAAC mainly reflects the protein sequence information (Bahar, et al., 1997; Sahu, et al., 2020; Zhang, et al., 2021a). The algorithm can be expressed by

$$L=\left[l_1, l_2, \cdots, l_{20}, l_{20+1}, \cdots, l_{20+\beta}\right]^T \tag{1}$$

$$l_v=\begin{cases}\dfrac{f_v}{\sum_{j=1}^{20} f_v+\omega\sum_{k=1}^{\beta}\eta_k} & 1\le v\le 20\\ \dfrac{\omega\eta_{v-20}}{\sum_{j=1}^{20} f_v+\omega\sum_{k=1}^{\beta}\eta_k} & 20+1\le v\le 20+\beta\end{cases} \tag{2}$$

where η_k represents the level sequence correlation factor, f_v represents the frequency of the v -th amino acid in the protein, ω is the weighting factor, and the value selected in this paper is 0.05 (Chou, 2010). Because β is the characteristic parameter, a $20+\beta$ dimensional feature vector will be formed finally.

EBGW: The physical and chemical properties are one of the important properties of protein. Zhang et al. (Zhang, et al., 2006) proposed EBGW, which divided amino acids into four categories, as shown in Table 1.

Table 1. The 20 amino acids are divided into four groups (K1-K4)

Group	Amino acids
neutral and hydrophobic amino acids (K1)	A, F, G, I, L, M, P, V, W
neutral and polarity amino acids (K2)	C, N, Q, S, T, Y
acidic amino acids (K3)	K, H, R
alkaline amino acids (K4)	D, E

Three disjoint combinations can be obtained from Table 1. According to formulas (3), (4) and (5), the protein sequences are converted into 3 binary sequences.

$$f_1(p_i)=\begin{cases}1 & p_i\in\{K1,K2\}\\ 0 & p_i\in\{K3,K4\}\end{cases} \tag{3}$$

$$f_2(p_i)=\begin{cases}1 & p_i\in\{K1,K3\}\\ 0 & p_i\in\{K2,K4\}\end{cases} \tag{4}$$

$$f_3(p_i)=\begin{cases}1 & p_i\in\{K1,K4\}\\ 0 & p_i\in\{K2,K3\}\end{cases} \tag{5}$$

The length of three binary sequences is L . These sequences are divided into multiple subsequences and the subsequence length is progressive

increase. Each will form L dimensional feature vector, so three binary sequences form $3*L$ dimensional vector.

GO: When using GO model to extract GO information of each protein sequence, it is usually divided into two steps (Huang, et al., 2008; Shen, et al., 2020). One is GO terms, and another is GO vector. The BLASTP is used to search from the Swiss-Prot database and retain homologous proteins (denoted as Y_i) with a similarity greater than or equal to 60% with protein P_i. Parameter E is set to 0.001 (Zhang, et al., 2018) in the above steps. In the GOA database, we searched for accession number (ACs) of each protein in Y_i, which are obtained from the Swiss-Prot database. Then, the corresponding GO terms were obtained (Xiao, et al., 2011). Then, GO feature vector is constructed as:

$$p_i = \left[G_1, G_2, \cdots, G_{|Y_i|}\right]^T \ i = 1, 2, \cdots, |Y_i| \tag{6}$$

here $|Y_i|$ is the size of Y_i dataset, $G_{|Y_i|} = \begin{cases} 1, p_i \in G_{|Y_i|} \\ 0, otherwise \end{cases}$.

RPT: RPT is a feature extraction method that reflects the evolutionary information of protein sequences (Jeong, et al., 2011). In the PSSM, domains with similar conservations are grouped according to conservation scores (Wang, et al., 2019; Zhang, et al., 2021). Here, each particular columns corresponding are standard amino acids in the PSSM. The 20 amino acids are separated 20 groups as rows in the PSSM. Then, calculate the sum of PSSM values for each element in each column, and form a 20×20 dimensional matrix, which is the RPT matrix. The matrix is expressed as follow:

$$RPT = \begin{pmatrix} H_{1,1} & H_{1,2} & \cdots & H_{1,20} \\ H_{2,1} & H_{2,2} & \cdots & H_{2,20} \\ \cdots & \cdots & \cdots & \cdots \\ H_{20,1} & H_{20,2} & \cdots & H_{20,20} \end{pmatrix} \tag{7}$$

Therefore, the matrix can be transformed into a 400 dimensional feature vector $V = [b_{S_{1,1}}, b_{S_{2,1}}, \cdots, b_{S_{i,j}}, \cdots, b_{S_{20,20}}]$, where $b_{S_{i,j}}$ is obtained by equation (8).

$$b_{S_{i,j}} = \frac{H_{n,m}}{L}(n, m = 1, 2, \cdots, 20) \tag{8}$$

EDT: EDT is an effective method to calculate the non-occurrence information possibility of two amino acids (Jeong, et al., 2011). The two amino acids with interval of $d(1, 2, \cdots, L_{\min} - 1)$, where $L_{\min}$ is the shortest sequence length in the dataset. The feature vector of EDT is denoted as:

$$P = (f(A_1, A_1), f(A_1, A_2), \cdots, f(A_1, A_{20}), \cdots, f(A_x, A_y), \cdots, f(A_{20}, A_{20})) \tag{9}$$

The $f(A_x, A_y)$ is non-occurrence possibility of two amino acids with interval d. It is calculated by Formula (10):

$$f(A_x, A_y) = \sum_{d=1}^{D} \frac{1}{L-d} \sum_{i=1}^{L-d} \left(A_{i,x} - A_{i+d,y}\right)^2 \tag{10}$$

where $A_{i,x}, A_{i+d,y}$ are the element in the PSSM, A_x, A_y are any two of the 20 amino acids, D is the maximum value in d.

MCD: Due to the influence factors of continuous and discontinuous fragments in protein sequences, You et al. (You, et al., 2014) proposed the MCD feature extraction method. The method converts protein sequence into digital information by binary method. For example, a protein sequence 'AVDCALSK' is randomly selected and transformed into a digital model '11321476' via MCD calculation. Then, the sequence is divided into 10 regions, thus composition (C), transition (T) and distribution (D) are employed to represent protein characteristics and each descriptor can be calculated. Finally, a 630 dimensional feature vector is formed by all descriptors from 10 regions.

2.3 Multi-label informed latent semantic indexing

Assuming the feature space contains N samples, and each sample size is M dimensional feature vector, but we will reduce to L dimension. MLSI (Yu, et al., 2005) defines the input matrix $X = [x_1, x_2, \cdots, x_i, \cdots, x_N] \in R^{N\times M}$, where x_i is the M dimensional feature vector. The output matrix $Y = [y_1, y_2, \cdots, y_i, \cdots, y_N] \in R^{N\times L}$ and y_i is the L dimensional feature vector. Kernel function $k_x(\cdot, \cdot)$ represents inner product as:

$$k_x(x_i, x_j) = \langle x_i, x_j \rangle \tag{11}$$

Similar kernel function $k_y(\cdot, \cdot)$ is expressed as Equation (12), and the kernel matrix $K_y = YY^T$ is obtained.

$$k_y(y_i, y_j) = \langle y_i, y_j \rangle \tag{12}$$

The kernel calculation matrix C is as (13):

$$C = (1-\beta)K_x + \beta K_y \tag{13}$$

Then, for generalized eigenvalue problems,

$$\left[K_x C^{-1} K_x + \gamma K_x\right]\alpha = \tilde{\lambda} K_x^2 \alpha \tag{14}$$

where coefficient α requires $\alpha^T K_x^2 \alpha = 1$, $\tilde{\lambda}_1 \le \cdots \le \tilde{\lambda}_N$. By formula (14), the generalized eigenvalues $\alpha_1, \cdots, \alpha_N$ are calculated and the first K eigenvalues are used as mappings. The i-th mapping function can be obtained by scaling the eigenvalues:

$$\varphi_j(x) = \frac{1}{\sqrt{\tilde{\lambda}_j}} w_j^T x = \frac{1}{\sqrt{\tilde{\lambda}_j}} \sum_{i=1}^{N} (\alpha_j)_i k_x(x_i, x) \tag{15}$$

Then, $\lambda = 1/\tilde{\lambda}$ and formula (15) is rewritten as:

$$K_x^2 \alpha = \lambda \left[K_x C^{-1} K_x + \gamma K_x\right]\alpha \tag{16}$$

Finally, K dimensional vector with the largest eigenvalue is selected.

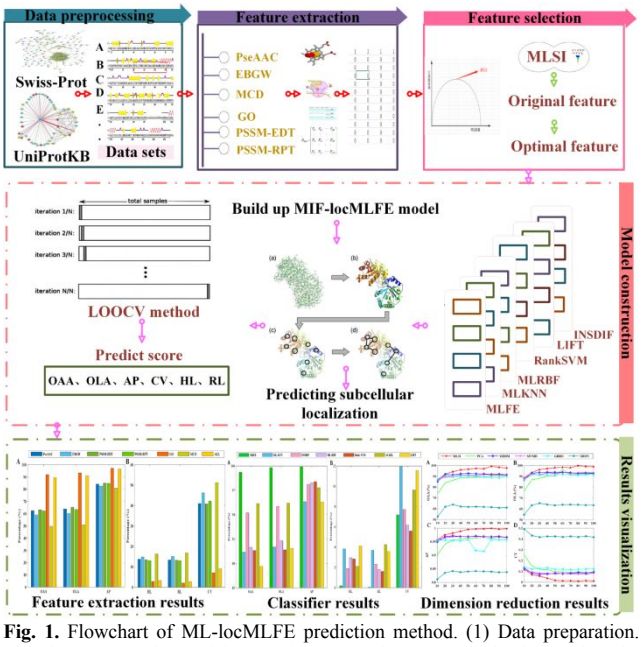

Fig. 1. Flowchart of ML-locMLFE prediction method. (1) Data preparation. Five datasets are obtained from Swiss-Prot and UniProtKB databases, then the corresponding protein sequences and real label are also achieved. (2) Feature extraction. PseAAC, EBGW, MCD, RPT, EDT and GO are used to convert protein sequence information into digital information, then the six features vector are fused. (3) Feature selection. The MLSI method identifies the most effective information to form the optimal feature subset. (4) Model construction. Combining step (3), the optimal feature subset is integrated into the classifier MLFE, and the ML-locMLFE model is constructed based on LOOCV. (5) Model evaluation. Gram-positive bacteria dataset is used to evaluate the effectiveness of ML-locMLFE, and Gram-negative bacteria dataset, virus dataset, SARS-CoV-2 dataset, newPlant dataset are used to verify the performance of ML-locMLFE. Both the training set and the test sets will choose OAA, OLA, HL, RL, AP and CV as evaluation indicators

2.4 Multi-label learning with feature induced labeling information enrichment

If training sample is denoted as (x_i, Y_i), p is the number of training sample, and given the enriched labeling information U, the original training sample can be transformed into $D=\{(x_i,u_i)\mid 1\le i\le p\}$. The response variables u_i can measure the model through the multi-output regression technology. MLFE algorithm (Zhang, et al., 2018) uses minimization to obtain the objective function of the regression model:

$$\begin{aligned}\Omega(\Theta,\ b)&=\beta_1\sum_{i=1}^{p}\Omega_1(u_i)+\beta_2\sum_{i=1}^{p}\sum_{j=1}^{q}\Omega_2(o_{ij})\\&+\beta_3\sum_{i=1}^{p}\sum_{j=1}^{q}\Omega_3(r_{ij})+\frac{1}{2}\sum_{j=1}^{q}\left\|\theta_j\right\|_2^2\end{aligned}\tag{17}$$

here $\Theta=[\theta_1,\theta_2,\cdots,\theta_q]$, $b=[b_1,b_2,\cdots,b_q]^T$ represent weight matrix and deviation vector of regression model respectively and q is the number of class label.

In order to obtain the optimal objective function, newton weighted least squares iterative method (IRWLS) (Sanchez-Fernandez, et al., 2004; Tsoumakas, et al., 2001) is used. In the iterative process, the descent direction of model optimization is determined by solving the linear solution of the equation. Let $\{\Theta^{(k)}, b^{(k)}\}$ represents the current model after the k-th iteration, and Equation (18) is obtained based on the first order Taylor expansion.

$$\begin{aligned}\Omega'(\Theta,\ b)&=\beta_1\left(\sum_{i=1}^{p}\Omega_1\left(u_i^{(k)}\right)+\frac{d\Omega_1(u)}{du}\bigg|_{u_i^{(k)}}\frac{\left(e_i^{(k)}\right)^T}{u_i^{(k)}}\left(e_i-e_i^{(k)}\right)\right)\\&+\beta_2\sum_{i=1}^{p}\sum_{j=1}^{q}\Omega_2(o_{ij})+\beta_3\sum_{i=1}^{p}\sum_{j=1}^{q}\Omega_3(r_{ij})+\frac{1}{2}\sum_{j=1}^{q}\left\|\theta_j\right\|_2^2\end{aligned}\tag{18}$$

where $e_i^{(k)}$ and $u_i^{(k)}$ can be calculated under the current model $\{\Theta^{(k)}, b^{(k)}\}$. In order to identify the analytical solution of the descent direction, it is necessary to construct the quadratic approximation value of $d\Omega_1(u)/u$:

$$\begin{aligned}\Omega''(\Theta,\ b)&=\beta_1\left(\sum_{i=1}^{p}\Omega_1\left(u_i^{(k)}\right)+\frac{d\Omega_1(u)}{du}\bigg|_{u_i^{(k)}}\frac{u_i^2-\left(u_i^{(k)}\right)^2}{2u_i^{(k)}}\right)\\&+\beta_2\sum_{i=1}^{p}\sum_{j=1}^{q}\Omega_2(o_{ij})+\beta_3\sum_{i=1}^{p}\sum_{j=1}^{q}\Omega_3(r_{ij})+\frac{1}{2}\sum_{j=1}^{q}\left\|\theta_j\right\|_2^2\\&=\frac{1}{2}\beta_1\sum_{i=1}^{p}\sum_{j=1}^{q}\alpha_i u_i^2+\beta_2\sum_{i=1}^{p}\sum_{j=1}^{q}\Omega_2(o_{ij})\\&+\beta_3\sum_{i=1}^{p}\sum_{j=1}^{q}\Omega_3(r_{ij})+\frac{1}{2}\sum_{j=1}^{q}\left\|\theta_j\right\|_2^2+\Upsilon\end{aligned}\tag{19}$$

where Υ is a constant term.

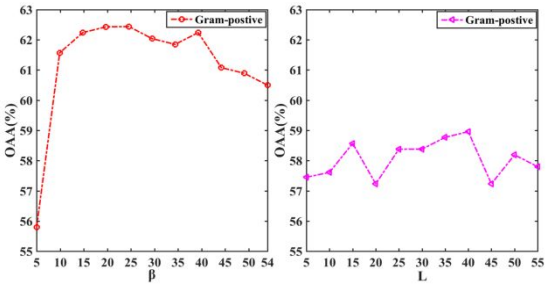

Fig. 2. The OAA obtained by Gram-positive bacteria dataset under different parameters. (A) The OAA reaches its maximum at $\beta=25$, therefore, $\beta=25$ is selected as the optimal parameter of PseAAC and forms a $20+\beta=45$ dimensional vector. (B) The OAA reaches its maximum at $L=40$, therefore, we select $L=40$ as the best parameter of EBGW and form a $3*L=120$ dimensional vector

2.5 Performance evaluation

The cross-validation method can avoid over-fitting to some extent. The commonly methods include K-fold cross-validation (Jia, et al., 2018), LOOCV (Yu, et al., 2020b), self-compatibility method (Bringi, et al., 2001) and independent sample test (Timothy, et al., 1987). Compared with other cross validation methods, LOOCV is deterministic and has high sample utilization (Cheng, et al., 2017). Therefore, the LOOCV test is introduced in this paper to evaluate the effectiveness of the model with the OAA, OLA, Hamming Loss (HL), Coverage (CV), Ranking Loss (RL), Average precision (AP) as indicators. Six evaluation indicators are defined as the following.

$$OAA=\frac{1}{W}\sum_{i=1}^{W}\Delta\left|Y_i(U_i),Y'(U_i)\right|\tag{20}$$

$$OLA=\frac{1}{\sum_{i=1}^{W}\left|Y_i(U_i)\right|}\sum_{i=1}^{W}\left|Y_i(U_i)\cap Y'(U_i)\right|\tag{21}$$

where W is the number of training sample. $Y_i(U_i)$ and $Y'(U_i)$ represent prediction label and real label, $\Delta\left|Y_i(U_i),Y'(U_i)\right|=\begin{cases}1, & Y_i(U_i)=Y'(U_i)\\ 0, & otherwise\end{cases}$.

$$HL=\frac{1}{W}\sum_{i=1}^{W}\frac{\left|Y_i(U_i)\cup Y'(U_i)\right|-\left|Y_i(U_i)\cap Y'(U_i)\right|}{G}\tag{22}$$

where G is the number of labels.

$$CV=\frac{1}{W}\sum_{i=1}^{W}\underset{y\in Y_i}{macrank}\left(f(X_i,y)\right)-1\tag{23}$$

where $rank(f(X_i,y))-1$ makes all labels rank down and get the corresponding ranking.

$$RL=\frac{1}{W}\sum_{i=1}^{W}\frac{1}{\left|Y_i\right|\left|\overline{Y}\right|}\left|RAL(X_i)\right|\tag{24}$$

where $RAL(X_i)=\{(y_j,y_k)\mid f(X_i,y_j)\le f(X_i,y_k),(y_j,y_k)\in Y_i\times\overline{Y_i}\}$ and $y_i\in Y$. $f(X_i,y_j)$ is part of the label of X_i, $\overline{Y}$ is a supplement to Y_i.

$$AP=\frac{1}{W}\sum_{i=1}^{W}\frac{1}{\left|Y_i\right|}\sum_{y_j\in Y_i}AVP(T_i)\tag{25}$$

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

This study proposes a new method ML-locMLFE for predicting the SCL of multi-label proteins and the detailed process is displayed in Fig. 1.

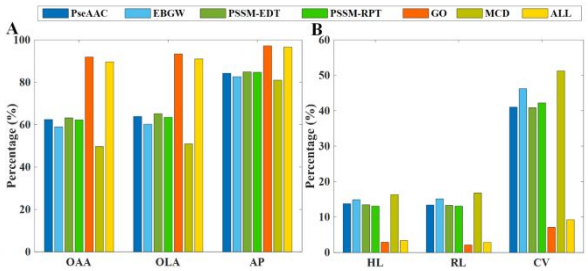

Fig. 3. Comparison of results based on seven different methods for Gram-positive bacteria. ALL: PseAAC+EBGW+EDT+RPT+GO+MCD. The six single feature extraction methods, the GO method has greatest contribution rate to the model. For the fusion features, the OAA and OLA are lower than GO due to the increase of redundant information in the fusion feature space. But compared with the other five single characteristics, the OAA is 26.40-39.89% higher than other methods, and the OLA is 25.81%-40.15% higher than other methods. Therefore, the fusion features can represent the overall characteristics of the protein and improve the accuracy of the model prediction

3 Results

3.1 The result analysis of feature encoding parameters β and L

PseAAC and EBGW have different characteristic information by setting different parameters. Since the minimum length of all protein sequences of the Gram-positive bacteria dataset is 55, the parameter of PseAAC is set from 5 to 54 and the parameter of EBGW is set from 5 to 55. Through the LOOCV test, the characteristic information obtained from each parameter is put into the classifier MLFE, and the specific evaluation index values of the different parameter results are listed in the Supplementary Table S6-S7. The optimal OAA obtained from PseAAC and EBGW are 62.44% and 58.96%. The comparison results under different parameters are shown in Fig. 2.

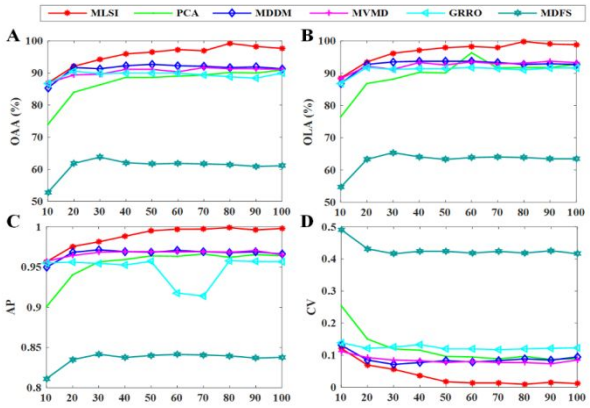

Fig. 4. Comparison results based on different dimension reduction methods. When the 80-dimensional feature subset is obtained by MLSI, the results of OAA value and OLA value both reach the highest. This method uses the linear correlation of input information and output information to select the feature subset, which greatly improves the ability of prediction. At the same time, MLSI has increased the OAA by 7.52%-37.77%, and the OLA by 6.69%-35.95% compared with other methods. Therefore, MLSI is chosen as the feature selection method

3.2 The influence of feature extraction methods on results

This article uses a total of six feature extraction methods. Among them, the PseAAC method not only considers the sequence information of the protein but also includes the position information of the amino acids in the sequence. The EBGW method is based on the physical and chemical properties of amino acids to effectively extract the physical and chemical information of proteins. The MCD method uses multiple regions as features to extract the physical and chemical information of protein sequences. The GO method extracts the annotation information of the protein, which can essentially analyze the properties of genes and gene products. Because the EDT method considers the evolutionary information of the protein, it can reflect the probability of two different amino acids. The RPT method obtains the evolutionary information of the protein by grouping the evolution scores in PSSM. Therefore, the six feature extraction methods obtain effective information from the different characteristics of the protein, which greatly improves the prediction performance of the model. Through the LOOCV test, six single feature vectors are put into MLFE, and GO has the largest contribution rate among all single features, and its OAA and OLA reach 91.91% and 93.31%, respectively. However, single feature information cannot represent all important information. Therefore, the six feature extraction results need to be fused. We extract 912 dimensional feature vectors from GO, 45 dimensional feature vectors from PseAAC, 120 dimensional feature vectors from EBGW, 400 dimensional feature vectors from RPT and EDT, respectively, and 630 dimensional feature vectors from MCD. After the final fusion, 2507 dimensional feature vectors are obtained. Through the LOOCV test, the comparison results of single and fusion features are given in Fig. 3.

3.3 Analysis of feature selection results

Feature selection method can reduce spatial dimensions and decrease model training time. Therefore, this paper uses principal component analysis (PCA) (Abdi, et al., 1987), GRRO (Zhang, et al., 2020), MDFS (Zhang, et al., 2019), MDDM (Zhang and Zhou, 2010), MVMD (Xu, et al., 2016), MLSI (Yu, et al., 2005) to eliminate irrelevant features. Through LOOCV test, the feature subset obtained by each method is put into MLFE. Then, the OAA of MLSI reaches 99.23%, and the OLA reaches 99.81%, which are both optimal. The algorithm not only retains the original input features, but also captures the correlation of output dimensions, which greatly improves the performance of model prediction. On the Gram-positive bacteria dataset, the MLSI method selects different dimensions to obtain the prediction results which are shown in Supplementary Table S8, and the comparison results of different methods can be found in Fig. 4.

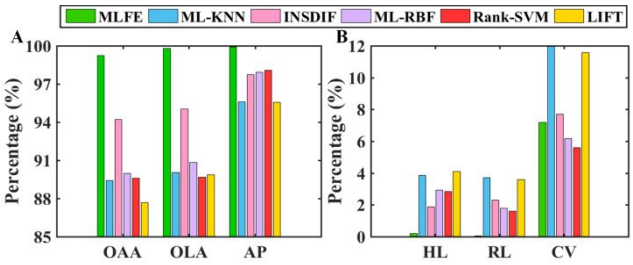

Fig. 5. Prediction results of six classifiers on Gram-positive bacteria dataset. The MLFE classifier is used to predict the multi-label protein subcellular localization, the OAA and OLA are both highest and above 99%. The algorithm uses sparse reconstruction of the training samples to represent the bottom layer of the feature space. At the same time, the OAA of MLFE is 5.01%-11.56% higher than the other five classifiers, and the OLA is 4.78%-10.14% higher than the other five classifiers. In summary, MLFE can effectively link feature information with label information, which improves the prediction performance of the model

3.4 Comparative Analysis of Classification Algorithms

In order to verify the effectiveness of MLFE, we take five classifiers as comparison. That are multi-label k nearest neighbor (ML-KNN) (Gonzalez-Lopez, et al., 2018), multi-label radial basis function (ML-RBF) (Zhang, 2009), multi-label learning with label-specific features (LIFT) (Zhang, et al., 2015), ranking support vector machine (Rank-SVM) (Tayal, et al., 2018), multi-label learning by instance differentiation (INSDIF) (Zhang, et al., 2007). The optimal feature subset obtained by the MLSI is put into the six classifiers. Through the LOOCV test, the results of OAA and OLA obtained from thte MLFE classifier are 99.23% and 99.81%, respectively. The algorithm uses the sparse reconstruction information between the training samples as features, and the reconstruction information is passed into the label space to enrich the original labels as numerical labels, thereby enhancing the effectiveness of the label information. The comparison results of different methods are shown in Fig. 5, and the corresponding ROC and PR curves are shown in Fig. 6. The specific parameter values obtained through different algorithms are shown in Supplementary Table S9.

3.5 Comparison with other methods

With the continuous development of multi-label protein SCL research, many researchers use machine learning methods to predict. To prove the superiority of the ML-locMLFE, we compare the results of the four datasets with other methods. On the Gram-positive dataset, the results of this article are compared with the results of iLoc-pos (Wu, et al., 2012), Gpos-ECC-mPLoc (Wang, et al., 2015) and Gram-LocEN (Wan, et al., 2017). The results of different methods are listed in Fig. 7. On the Gram-negative dataset, the results of this article are compared with the results of iLoc-Gneg (Chou and Shen, 2006), Gneg-ECC-mPLoc (Wang, et al., 2015) and Gram-LocEN (Wan, et al., 2017). On the virus dataset, the results of this article are compared with the results of mGOASVM (Wan, et al., 2012), AD-SVM (Wan, et al., 2018)， mPLR-Loc (Wan, et al., 2015). On the newPlant dataset, the results of this article are compared with the results of Plant-mPLoc (Chou and Shen, 2010), mPLR-Loc (Wan, et al., 2015)，HybridGO-Loc (Wan, et al., 2014). The result of different comparison on the newPlant dataset is shown in Supplementary Table S10, and the results of different methods on other datasets are listed in Supplementary Fig.S1-Fig.S2.

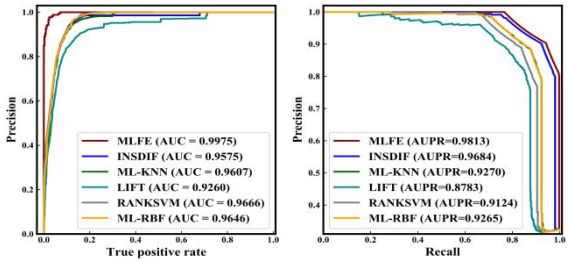

Fig. 6. Comparison of ROC and PR curves of six classifiers. (A) The ROC curve of the Gram-positive bacteria dataset corresponding to the six classifiers. (B) The PR curves of the Gram-positive bacteria dataset corresponding to the six classifiers. The ROC and PR curves are usually used to evaluate the quality of the model. The closer the ROC curve is to the upper left corner, the higher the accuracy of the model. Conversely, the closer the PR curve is to the upper right corner, the higher the accuracy of the model The AUC value and AUPR value of MLFE are both optimal. The AUC value of MLFE is 99.75%, which is 3.09%-7.15% higher than the AUC values of the other five classifiers, and the AUPR value is 98.13%, which is 1.29%-10.30% higher than the other five classifiers

3.6 Prediction multi-label protein SCL of SARS-CoV-2

Since the SARS-CoV-2 has brought us great influence, it is important to locate the subcellular location of SARS-CoV-2 protein accurately and quickly. Many researchers have found a way to treat COVID-19 by analyzing the pathogenesis of SARS-CoV-2. German scientist (Hoffmann, et al., 2020) found that SARS-CoV-2 transmission depended on transmembrane protease serine 2 (TMPRSS2), the protease inhibitors of TMPRSS2 can block SARS-CoV-2 into cells. Xu et al. (Xu, et al., 2020) predicted that the TMPRSS2 can bind to some monomeric compounds independently by studying the protein properties and three dimensional structure of TMPRSS2. Therefore, it is shown that the TMPRSS2 is a serine protease anchored on the cell membrane at the amino-terminal transmembrane region, and its inhibitor can be used as the treatment of COVID-19. The SARS-CoV-2 protein information are obtained by PseAAC, EBGW, GO, RPT, EDT and MCD to form the original feature space. With the continuous increase of the dimensionality, the interference of redundant information on the result is gradually significant. Thus, we use MLSI to obtain the optimal feature subset. Using the MLFE algorithm, the OAA is 72.73%, and the OLA is 69.23%. The specific result comparison is shown in Table 2.

Table 2 shows that the SARS-CoV-2 protein mainly exists in Plasma membrane, nucleus, Golgi apparatus and other subcellular. We obtained 26 proteins from the UniProt database and found that the ninth protein is TMPRSS2, which is located on the plasma membrane and accurately predicted by ML-locMLFE. The SARS-CoV-2 dataset is too small to optimize the model, the stability of the model is relatively low. Therefore, the prediction results of Cytoskeleton, Endoplasmic reticulum, Endosome, Golgi apparatus and Lysosome are not ideal, but the OAA and OLA are 72.73 % and 69.23% by the ML-locMLFE method. The model presented in this paper can not only predict the SCL of important protein in the SARS-CoV-2 quickly and accurately, but also provide a theoretical basis for the treatment of SARS-CoV-2 pneumonia and drug research.

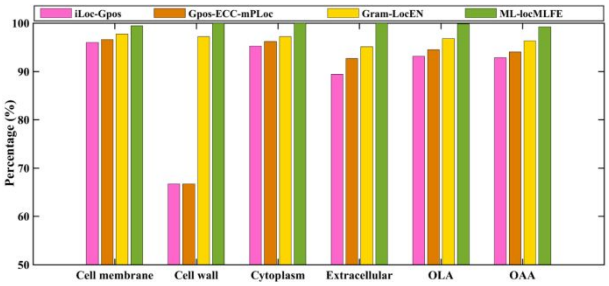

Fig. 7. On the Gram-positive bacteria dataset, the ML-locMLFE is compared with other methods by LOOCV test. The OAA and the OLA are 99.23% and 99.81% by MLFE, which is 2.93%-6.33%, 3.01%-6.71% higher than other methods. In addition, the prediction results of the four types of subcellular locations by this method are 99.42%, 100.00%, 100.00%, and 100.00%, which are 1.72%-3.42%, 5.60%-33.33%, 2.84%-4.80, 4.88%-10.57% higher than other methods, respectively. Therefore, the ML-locMLFE is superior to other methods using the same dataset

Table 2. Prediction of protein SCL using ML-locMLFE on SARS-CoV-2 dataset

Locations	ML-locMLFE
Plasma membrane	10/10=1.0000
Cytoskeleton	0/1=0.0000
Endoplasmic reticulum	0/1=0.0000
Endosome	0/2=0.0000
Golgi apparatus	0/1=0.0000
Lysosome	0/2=0.0000
Mitochondrion	0/2=0.0000
Nucleus	7/7=1.0000

4 Conclusion

It is significant to understand the structure and function of protein by using machine learning methods to predict the SCL of multi-label protein. Firstly, PseAAC, EBGW, GO, RPT, EDT, MCD are used to extract important information about the various properties of proteins. Among them, the GO method has the highest prediction accuracy and the largest contribution rate compared with the other five methods. The annotation information of genes and gene products extracted using the GO method can provide important evidence for the study of protein functions. Secondly, it is the first time to use MLSI as feature selection in the prediction of multi-label protein SCL.

This method can map input features to a new feature space, which not only ensures the existence of input information, but also captures the correlation between multiple output information, so that it can more effectively select the best feature subset. Finally, we integrate the optimal feature subset into MLFE. For the first time, MLFE algorithm is used to enrich the original labels of training samples into numerical labels to enhance the effectiveness of multi-label information, which further improves the performance of the model. Through the LOOCV test, the OAA of the Gram-positive bacteria dataset, the Gram-negative bacteria dataset, the virus dataset, the SARS-CoV-2 dataset, and the newPlant dataset are 99.23%, 93.82%, 93.24%, 72.73%, 96.72%, and the OLA are 99.81%, 96.50%, 99.21%, 69.23%, 96.25% respectively. Therefore, the ML-locMLFE proposed in this paper can predict the multi-label protein SCL more accurately. In addition, the ML-locMLFE model can spread to other research fields such as multi-label protein post-translational modification, multi-label mRNA subcellular localization and identification of drug-target interactions. More importantly, the model can accurately predict the SCL of the SARS-CoV-2 protein, and then clarify the pathogenic mechanism of the virus. We hope that our method can provide some insights and help in the clinical treatment of various diseases, including COVID-19. In the next step, we will construct larger scale and diverse data sets to study the SCL of multi-label protein.

Acknowledgements

We thank anonymous reviewers for valuable suggestions and comments.

Funding

This work was supported by the National Natural Science Foundation of China (No. 62172248).

Conflict of Interest: none declared.

References

Abdi,H. and Williams,L.J. (2010) Principal component analysis, *Comput. Stat*., **2**, 433-459.

Bahar,I. *et al*. (1997) Understanding the recognition of protein structural classes by amino acid composition. *Proteins*, **29**, 172-185.

Bringi,V.N. *et al*. (2001) Correcting C-band radar reflectivity and differential reflectivity data for rain attenuation: a self-consistent method with constraints. *IEEE T. Geo. Remote Sens*., **39**, 1906-1915.

Brownlee,M.D. and Michael. (1995) Advanced protein glycosylation in diabetes and aging. *Ann Rev Med*., **46**, 223-234.

Cheng,X. *et al*. (2017) iATC-mISF:a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics*, **33**, 341-346.

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

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

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

Chu,Y.Y. *et al*. (2020) DTI-MLCD: predicting drug-target interactions using multi-label learning with community detection method. *Brief. Bioinform*., **22**, 1-15.

Costa,E.A. *et al.* (2018) Defining the physiological role of SRP in protein-targeting efficiency and specificity. *Science*, **359**, 689-692.

Dehzangi,A. *et al*. (2015) Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionarybased descriptors into Chou's general PseAAC. *J. Theor. Biol*., **364**, 284-294.

Du,L. *et al*. (2020) Using Evolutionary information and multi-label linear discriminant analysis to predict the subcellular location of multi-site bacterial proteins via Chou's 5-steps rule. *IEEE Access*, **8**, 56452-56461.

Fan,Y,L. *et al*. (2021) Multilabel feature selection: a local causal structure learning approach. *IEEE Transactions on Neural Networks and Learning Systems*, **120**, 108169.

Gattani,S. *et al*. (2019) StackCBPred: a stacking based prediction of protein-carbohydrate binding sites from sequence. *Carbohydr. Res.*, **486**, 107857.

Gonzalez-Lopez,J. *et al*. (2018) Distributed nearest neighbor classification for large-scale multi-label data on spark. *Future Generat. Comput. Syst.*, **87**, 66-82.

Hoffmann,M. *et al*. (2020) SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*, **181**, 271-280.

Huang,W.L. *et al*. (2008) ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC Bioinformatics*, **9**, 1-16.

Ivanova,L. *et al*. (2008) Mesenchymal transition in kidney collecting duct epithelial cells. *Am J Physiol Renal Physiol.*, **294**, 1238-1248.

Jang,S.K. *et al*. (1988) A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *J. Virol*., **62**, 2636-2643.

Javed,F. *et al*. (2020) ML-RBF: Predict protein subcellular locations in a multi-label system using evolutionary features. *Chemometr. Intell. Lab. Syst*., **203**, 104055.

Jeong,J.C. *et al*. (2010) On position-specific scoring matrix for protein function prediction. *IEEE ACM T. Comput. Bi*., **8**, 308-315.

Jia,C.Z. *et al*. (2018) O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. *Bioinformatic*s, **12**, 2029-2036.

Marilyn,B. *et al*. (2020) Deep neural network to extract high-level features and labels in multi-label classification problems. *Neurocomputing*, **413**, 259-270.

Sahu,S. *et al*. (2020) Plant-mSubP: a computational framework for the prediction of single- and multi-target protein subcellular localization using integrated machine-learning approaches. *AoB Plants*, **12**, plz068.

Sanchez-Fernández,M. *et al*. (2004) SVM multiregression for nonlinear channel estimation in multiple-input multiple-output systems. *IEEE T. Knowl Data En.*, **52**, 2298-2307.

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

Shen,Y.D. *et al*. (2020) Critical evaluation of web-based prediction tools for human protein subcellular localization. *Brief. Bioinform*., **21**, 1628-1640.

Shen,Y.N. *et al*. (2019) Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. *J. Theor. Biol*., **462**, 230-239.

Shi,H. *et al*. (2019) Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics*, **111**, 1839-1852.

Sneddon,A.A. *et al*. (2000) Amelioration of denervation-induced atrophy by clenbuterol is associated with increased PKC-α activity. *Am J Physiol Endocrinol Metab*., **279**, E188.

Tayal,A. *et al*. (2018) Bounding the difference between RankRC and RankSVM andapplication to multi-level rare class kernel ranking. *Data Min. Knowl. Disc*., **32**, 417-452.

Timothy, *et al*. (1987) Robustness of the two independent samples t-test when applied to ordinal scaled data. *Stat Med*., **6**, 79-90.

Tsoumakas,G. *et al*. (2001) Random k-labelsets for multi-label classification. *IEEE T. Knowl Data En*., **23**, 1079-1089.

Wan,S.B. and Mak,M.W. (2018) Predicting subcellular localization of multi-location proteins by improving support vector machines with an adaptive-decision scheme. *Int. J. Mach. Learn. Cybern*., **9**, 399-411.

Wan,S.B. *et al*. (2012) mGOASVM: multi-label protein subcellular localization based on gene ontology and support vector machines. *BMC Bioinformatics*, **13**, 290.

Wan,S.B. *et al*. (2014) HybridGO-Loc: mining hybrid features on gene ontology for predicting subcellular localization of multi-location proteins. *PLoS One*, **9**, e89545.

Wan,S.B. *et al*. (2015) mPLR-Loc: an adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction. *Anal. Biochem*., **473**, 14-27.

Wan,S.B. *et al*. (2017) Gram-LocEN: interpretable prediction of subcellular multi-localization of Gram-positive and Gram-negative bacterial proteins. *Chemometr. Intell. Lab. Syst.*, **162**, 1-9.

Wang,R. *et al*. (2021) Active k-labelsets ensemble for multilabel classification. *Pattern Recognit*., **109**, 107583.

Wang,X. *et al*. (2015) Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble. *BMC Bioinformatics*, **16**, S1.

Wang,X.Y. *et al*. (2019) Protein-proteininteraction sites prediction by ensemble random forests with synthetic minority oversamplingtechnique. *Bioinformatics*, **35**, 2395-2402.

Wu,Z.C. *et al*. (2012) iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. *Protein. Pept. Lett.*, **19**, 4-14.

Xiao,X. *et al*. (2011) A multi-label classifier for predicting the subcellular localization

of gram-negative bacterial proteins with both single and multiple sites. *PLoS One*, **6**, e20592.

Xu,J.H. *et al*. (2016) A multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence simultaneously. *Knowl Based Syst.*, **98**, 172-184.

Xu,X.Y. *et al*. (2020) Potential monomer compounds for treatment of corona virus disease 2019 (COVID-19) by transmembrane serine proteinase 2 (TMPRSS2). *Drug Evaluation Research*, **43**, 813-821.

You,Z.H. *et al*. (2014) Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. *BMC Bioinformatics*, **15**, 15:S9.

Yu,B. *et al*. (2018) Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction. *BMC Genomics*, **19**, 478.

Yu,B. *et al*. (2020a) Prediction of protein-protein interactions based on L1-regularized logistic regression and gradient tree boosting. *Genom. Proteom. Bioinf.*, **18**, 582-592.

Yu,B. *et al*. (2020b) SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics*, **36**, 1074-1081.

Yu,B. *et al*. (2021) Prediction of protein-protein interactions based on elastic net and deep forest. *Expert Syst. Appl.*, **176**, 114876.

Yu,K. *et al*. (2005) Multi-label informed latent semantic indexing. *in: International ACM SIGIR Conference on Research and Development in Information Retrieval*, **pp**, 258-265.

Zhang,C.X. *et al*. (2018) MetaGO: predicting Gene Ontology of non-homologous proteins through low-resolution protein structure prediction and protein–protein network mapping. *J. Mol. Biol.*, **430**, 2256-2265.

Zhang,J. *et al*. (2019) Manifold regularized discriminative feature selection for multi-label learning. *Pattern Recognit.*, **95**, 136-150.

Zhang,J. *et al*. (2020) Multi-label feature selection via global relevance and redundancy optimization. *in: Proceedings of International Joint Conference on Artificial Intelligence*, **pp**, 2512-2518.

Zhang,J. *et al*. (2020) A systemic and molecular study of subcellular localization of SARS-CoV-2 proteins. *STTT*., **5**, 1-3.

Zhang,M.L. (2009) ML-RBF: RBF neural networks for multi-label learning. *Neural Process Lett*., **29**. 61-74.

Zhang,M.L. and Wu, L. (2015) LIFT: multi-label learning with label-specific features. *IEEE Trans. Pattern Anal*., **37**, 107-120.

Zhang,M.L. *et al*. (2007) Multi-label learning by instance differentiation. *in: National Conference on Artificial Intelligence. AAAI Press*., **7**, 669-674.

Zhang,Q. *et al*. (2020) DMLDA-LocLIFT: Identification of multi-label protein subcellular localization using DMLDA dimensionality reduction and LIFT classifier. *Chemometr. Intell. Lab. Syst.*, **206**, 104148.

Zhang,Q *et al*. (2021a) MpsLDA-ProSVM: Predicting multi-label protein subcellular localization by wMLDAe dimensionality reduction and ProSVM classifier. *Chemometr. Intell. Lab. Syst*., **208**, 104216.

Zhang,Q. *et al*. (2021b) Accurate prediction of multi-label protein subcellular localization through multi-view feature learning with RBRL classifier. *Brief. Bioinform*., **22**, 1-11.

Zhang,Q.M. *et al*. (2021) StackPDB: Predicting DNA-binding proteins based on XGB-RFE feature optimization and stacking ensemble classifier. *Appl. Soft Comput*., **99**, 106921.

Zhang,Q.W. *et al*. (2018) Feature-induced labeling information enrichment for multi-label learning. *in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, **pp**, 4446-4453.

Zhang,Y. and Zhou,Z.H. (2010) Multilabel dimensionality reduction via dependency maximization. *ACM Trans. Knowl. Discov*., **4**, 14.

Zhang,Z.H. *et al*. (2006) A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. *FEBS. Lett*., **580**, 6169-6174.