Briefings in Bioinformatics, 22(5), 2021, 1–11

https://doi.org/10.1093/bib/bbab012
Problem solving protocol

Accurate prediction of multi-label protein subcellular localization through multi-view feature learning with RBRL classifier

Qi Zhang†, Yandan Zhang†, Shan Li, Yu Han, Shuping Jin, Haiming Gu and Bin Yu

Corresponding author: Bin Yu, College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China; School of Life Sciences, University of Science and Technology of China, Hefei 230027, China; Key Laboratory of Computational Science and Application of Hainan Province, Haikou 571158, China. E-mail: yubin@qust.edu.cn
†These authors contributed equally to this work.

Abstract

Multi-label proteins can participate in carrier transportation, enzyme catalysis, hormone regulation and other life activities. Meanwhile, they play a key role in the fields of biopharmaceuticals, gene and cell therapy. This article proposes a prediction method called Mps-mvRBRL to predict the subcellular localization (SCL) of multi-label protein. Firstly, pseudo position-specific scoring matrix, dipeptide composition, position specific scoring matrix-transition probability composition, gene ontology and pseudo amino acid composition algorithms are used to obtain numerical information from different views. Based on the contribution of five individual feature extraction methods, differential evolution is used for the first time to learn the weight of single feature, and then these original features use a weighted combination method to fuse multi-view information. Secondly, the fused high-dimensional features use a weighted linear discriminant analysis framework based on binary weight form to eliminate irrelevant information. Finally, the best feature vector is input into the joint ranking support vector machine and binary relevance with robust low-rank learning classifier to predict the SCL. After applying leave-one-out cross-validation, the overall actual accuracy (OAA) and overall location accuracy (OLA) of Mps-mvRBRL on the training set of Gram-positive bacteria are both 99.81%. The OAA on the test sets of plant, virus and

Qi Zhang is a master student at the College of Mathematics and Physics, Qingdao University of Science and Technology, China. Her research interests are bioinformatics and machine learning.
Yandan Zhang is currently an associate professor at the College of Mathematics and Physics, Qingdao University of Science and Technology, China. Her research interests are bioinformatics, machine learning and algorithms.
Shan Li is a PhD student at the School of Mathematics and Statistics, Central South University, China. Her research interests are bioinformatics and machine learning.
Yu Han is a lecturer at the College of Mathematics and Physics, Qingdao University of Science and Technology, China. Her research interests include bioinformatics and machine learning.
Shuping Jin is a master student at the College of Mathematics and Physics, Qingdao University of Science and Technology, China. Her research interests are bioinformatics and machine learning.
Haiming Gu is a professor at the College of Mathematics and Physics, Qingdao University of Science and Technology, China. His research interests include artificial intelligence, machine learning and algorithms.
Bin Yu is an associate professor at the College of Mathematics and Physics, Qingdao University of Science and Technology, China. His research interests include bioinformatics, artificial intelligence and biomedical image processing.
Submitted: 20 October 2020; **Received (in revised form):** 12 December 2020

Gram-negative bacteria datasets are 97.24%, 98.55% and 98.20%, respectively, and the OLA are 97.16%, 97.62% and 98.28%, respectively. The results show that the model achieves good prediction performance for predicting the SCL of multi-label protein.

Key words: multi-view information fusion; wMLDAb dimension reduction; RBRL classifier; multi-label protein subcellular localization

Introduction

Protein is the material basis of life and it is a major ingredient of cell and tissue that make up the human body. The protein synthesized in the cell can only perform its function in the correct subcellular localization (SCL). The abnormal protein SCL can affect the function of the protein and it is closely related to human diseases, such as minor salivary gland tumors [1], breast cancer [2], kidney stone [3], liver tumors [4], etc. Therefore, the study of protein SCL prediction has very important significance [5]. However, most prediction methods are based on single-location protein [6], but studies find that more and more protein sequences can exist in multiple locations at the same time [7]. These multi-label proteins have special biological functions. Therefore, studying the SCL of multi-label protein is the current focus of bioinformatics.

The feature extraction is the basis of various prediction algorithms, which helps to understand the relationship between protein sequences and functions. Javed and Hayat [8] used pseudo-amino acid composition using physicochemical properties and split amino acid composition to obtain the information from Gram-positive bacteria and virus datasets. Wan and Mak [9] proposed a novel adaptive-decision based multi-label support vector machine classifier (AD-SVM) by retrieving gene ontology (GO) items and constructing GO vectors to extract information from virus and plant datasets. Zhang et al. [10] improved the autocorrelation function algorithm on the basis of position-specific scoring matrix (PSSM) and used the Kullback–Leibler divergence to draw the information.

Currently, researchers adopt feature fusion method to comprehensively consider the information in the sequence. However, feature fusion is easy to cause ‘dimension disaster’. A large amount of redundant information can interfere with the prediction performance of classifier. Lin et al. [11] combines mutual information with max-dependency and min-redundancy (MDMR) algorithm by considering feature dependence and feature redundancy of multi-label features to select the optimal feature subset. Xu et al. [12]. put forward a new method multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence (MVMD) to reduce dimension where the least square method is integrated by linear combination. This algorithm can maximize feature variance and feature label dependency, and finally we can obtain the best feature matrix composed of all positive eigenvalues. Zhang et al. [13] proposed MDFS algorithm to build a low-dimensional embedding based on the original feature space to develop local and global label correlation. And the distinguishing feature of multi-label learning is searched by incorporating $l_{2,1}$-norm regularization into the learning framework.

Machine learning algorithm is currently the most widely used method in data analysis, and it is also one of the important supporting technologies of bioinformatics. Wan et al. [14] used two multi-label chloroplast proteins to evaluate the performance of EnTrans-Chlo. This model used a non-transductive algorithm to replace the transductive algorithm, namely multi-label support vector machine (ML-SVM) algorithm. After applying leave-one-out cross-validation (LOOCV), the overall actual accuracy (OAA) of the two proteins were 60.03% and 61.10%, respectively. Hasan et al. [15] introduced a support vector machine algorithm based on multiple kernel learning. Part of the data from Höglund and DBMLoc was selected to form a combined dataset to assess the performance about model, and the accuracy of new data was 77.1%. Wan et al. [16] proposed a multi-label elastic-net (EN) as a classification algorithm to test the performance of Gram-LocEN. After applying LOOCV, the OAA and overall location accuracy (OLA) of the Gram-positive bacteria and Gram-negative bacteria datasets were 96.3% and 96.8%, 94.5% and 95.3%, respectively.

Even though the existing methods can predict the SCL well, they still need to be improved in terms of operating speed and stability. Inspired by this, this article proposes a new model for predicting the SCL of multi-label protein, called Mps-mvRBRL. Firstly, we use pseudo position-specific scoring matrix (PsePSSM), dipeptide composition (DC), position-specific scoring matrix-transition probability composition (PSSM-TPC), GO and pseudo amino acid composition (PseAAC) algorithms for feature extraction. For optimizing the expression ability of multi-view features, differential evolution (DE) algorithm is used for the first time to assign five single feature weight vectors, and the multi-view information is merged in a weighted form. Secondly, the weighted linear discriminant analysis framework based on binary weight form (wMLDAb) is used to eliminate useless information from the fusion data. Finally, we use joint ranking support vector machine and binary relevance with robust low-rank learning (RBRL) classification to predict the SCL. The experimental results show that Mps-mvRBRL has good stability and can effectively predict the SCL of multi-label protein.

Materials and Methods

Datasets

In the paper, four benchmark datasets are used to predict the SCL of multi-label protein. Gram-positive bacteria [17] are used as a training set and it contains a total of 523 protein sequences belonging to four subcellular locations. The virus [18], plant [19] and Gram-negative bacteria datasets are used as independent test sets. The virus dataset contains a total of 252 protein sequences belonging to six subcellular locations. The plant dataset contains a total of 1055 protein sequences belonging to twelve subcellular locations. The Gram-negative bacteria [16] contain a total of 1456 protein sequences belonging to eight subcellular locations. The homology of protein sequences in the datasets is restricted to 25%. Supplementary Tables S1–S4 show the specific classifications of dataset.

Feature extraction

We use PsePSSM, DC, PSSM-TPC, GO and PseAAC to obtain the information. The specific introduction of five algorithms is shown in Supplementary Si1–Si5. Among them, PSSM [20] algorithm only considers the evolution information about all amino acid residues, but it ignores the order information. Therefore, the researchers combined the PseAAC and PSSM and proposed the PsePSSM algorithm [21]. The DC [22] algorithm can extract sequence information, which refers to the frequency of two adjacent amino acid residues in the sequence. In the evolution process, for obtaining the sequence information about protein, a method of transition probability composition is used by the PSSM matrix to form the PSSM-TPC algorithm [23]. GO [24] is a standardization project of gene function annotation information. Knowing the GO information of the protein can predict its SCL. And for the application of GO, we have a few instructions. The specific introduction is shown in Supplementary Si6. The PseAAC [25] algorithm was proposed by Chou [26] based on sequence information and physicochemical properties of amino acids.

Learning weights for combining multi-view features

At present, most researchers use multi-view information fusion to extract the effective information in the sequence. They just choose the concatenation method to simply combine the feature extraction results. However, this combination method ignores the relative importance of individual features and it easily causes the 'curse of dimensionality'. Therefore, the nature of individual feature extraction method should be understood, and we use the weighted combination method to fuse multi-view information through given feature weights to achieve the best predictive ability.

In this paper, the DE algorithm [27] is used to solve the global optimization problem. It can retain good individuals through repeated iterations. DE algorithm solves some optimization problems in complex environments by adjusting its search strategy and has strong global convergence and robustness. At present, the DE algorithm is applied in many fields, such as artificial neural networks, bioinformatics and signal processing [28, 29]. In this article, the specific algorithm steps of the DE algorithm are as follows.

1. Initialization

For optimization problems:

$$\begin{aligned} &\max f\left(w_1, w_2, \cdots w_D\right) \\ &s.t\ \ w_j^L \le w_j \le w_j^U, j = 1, 2, \cdots D \end{aligned} \tag{1}$$

where $f(w_i)$ refers to the maximum value of OAA, and D refers to the number of single feature extraction methods. In this article, $D = 5$, w_j^L and w_j^U represent the upper and lower bounds of the value range of the j-th component. In this article, $w_j \in [-2, 2]$. The built-in parameters of the maximum generation $G_{\max}$, population size N, crossover rate CR and scaling factor F are set as 100, 100, 0.5 and 0.5, respectively.

The initial population $P^g = \left\{w_1^g, w_2^g, \cdots w_N^g\right\}$ is randomly generated, where $w_i^g = \left(w_{i,1}^g, w_{i,2}^g, \cdots w_{i,5}^g\right)^T$ represents the i-th individual in the g-th generation population.

2. Mutation

DE algorithm implements individual mutation by a differential strategy. Different individuals in P^g are randomly selected, and their vectors are scaled to perform vector synthesis with the individuals to be mutated. That is:

$$m_i^{g+1} = w_{r_1}^g + F \cdot \left(w_{r_2}^g - w_{r_3}^g\right) \tag{2}$$

where $w_{r_1}^g$, $w_{r_2}^g$ and $w_{r_3}^g$ are three randomly different individuals in subset $P^g - \left\{w_i^g\right\}$.

3. Crossover

In order to increase the diversity of the next-generation solutions, this paper performs crossover operations between individuals on the g-generation population w_i^g and its variant intermediate m_i^{g+1}. That is:

$$t_{i,k}^{g+1} = \begin{cases} m_{i,k}^{g+1}, if \text{ rand}\,(0,1) < CR \text{ } or \text{ } k = k_{rand} \\ w_{i,k}^g, \qquad\qquad otherwise \end{cases} \tag{3}$$

where k_{rand} is a random integer of $[1, 2, \cdots D]$.

4. Selection

According to the greedy strategy, the DE algorithm uses the target individual w_i^g and the $t_{i,k}^{g+1}$ obtained in the previous step to select the individuals of the next-generation population through the evaluation function f. That is:

$$w_i^{g+1} = \begin{cases} t_i^{g+1}, & if\ f\left(t_i^{g+1}\right) \le f\left(w_i^g\right) \\ w_i^g, & \text{otherwise} \end{cases} \tag{4}$$

When $g > G_{\max}$, the algorithm should be terminated and output w_{best}. Otherwise repeat steps 2–4. In this article, we use DE algorithm to find the weights of PsePSSM, DC, PSSM-TPC, GO and PseAAC algorithms as $W_{best} = (0.8609, 0.4200, 0.5426, 1.6844, 0.1858)^T$.

Weighted linear discriminant analysis framework based on binary weight form

Linear discriminant analysis (LDA) is a classic linear learning method in statistics, pattern recognition and machine learning methods, and it is usually used for dimension reduction. The traditional LDA method can only be used to deal with the binary classification problem. When there are more than two categories, it needs to find a subspace that retains the variability of all categories. Wang et al. [30] proposed a multi-label LDA (MLDA). They defined a label correlation matrix and converted a binary vector to a real vector to summarize the contribution of each instance of the label. On this basis, Xu [31] added a binary weight framework to MLDA (wMLDAb) to extract the optimal features of multi-label data. The specific introduction is shown in Supplementary Si7.

Joint ranking support vector machine and binary relevance with robust low-rank learning

Ranking support vector machine (Rank-SVM) [32] and binary correlation (BR) [33] are two independent and representative multi-label classification prediction models. Specifically, Rank-SVM algorithm can optimize ranking loss (RL) and solve the problem of imbalanced class. BR algorithm converts multi-label classification task to multiple binary classification problems and optimizes hamming loss (HL). In this paper, we use Rank-SVM and BR algorithm combined with robust RBRL [34] to predict the SCL of multi-label protein.

For a training example $D = \{X, Y\}$, $X = [x_1; \cdots ; x_n] \in \mathbb{R}^{n \times m}$ is the input vector and n means the number of samples. m represents the sample dimension. $Y = [y_1; \cdots ; y_n] \in \{-1, 1\}^{n \times K}$ is the label vector, and K is the number of classes. For each instance x_i,

$f = xW + b$ provides the predicted value, where $b = [b_1, \cdots, b_K]$ is the deviation vector and $W = [w^1, \cdots, w^K]$ is the parameter matrix.

The purpose of Rank-SVM is to minimize RL, and its target optimization problem is

$$\begin{array}{l} \min\limits_{w} \frac{1}{2} \sum\limits_{j=1}^{l} \left\| w^j \right\|^2 + \lambda_2 \sum\limits_{i=1}^{n} \frac{1}{|Y_i^+||Y_i^-|} \sum\limits_{p \in Y_i^+} \sum\limits_{q \in Y_i^-} \xi_{pq}^i \\ s.t. \langle w^p, x_i \rangle - \langle w^q, x_i \rangle \geq 1 - \xi_{pq}^i, (p, q) \in Y_i^+ \times Y_i^- \\ \xi_{pq}^i \geq 0, i = 1, \cdots, n \end{array} , \qquad (5)$$

where $Y_i^+(Y_i^-)$ represents the relevant (irrelevant) label index set of the instance x_i. $|\cdot|$ represents the cardinality of the set, and λ_2 is a parameter that controls the complexity of the model.

The thresholding step is applied to Rank-SVM through the BR algorithm. At this time, it can be regarded as an extension of BR. The target optimization problem is

$$\begin{array}{l} \min\limits_{w} \frac{1}{2} \left\| \left(\left| E - Y \cdot (XW) \right|_+ \right)^2 \right\|_1 + \frac{\lambda_1}{2} \|W\|_F^2 + \\ \frac{\lambda_2}{2} \sum\limits_{i=1}^{n} \frac{1}{|Y_i^+||Y_i^-|} \sum\limits_{p \in Y_i^+} \sum\limits_{q \in Y_i^-} \max\left(0, 2 - \langle w^p - w^q, x_i \rangle\right)^2 \end{array} , \qquad (6)$$

among them, E is a matrix whose elements are all 1, and $\circ$ represents the Hadamard product.

In order to take advantage of high-order label correlation, Wu et al. [34] added a low-rank constraint to W, and the objective function is as follows:

$$\begin{array}{l} \min\limits_{w} \frac{1}{2} \left\| \left(\left| E - Y \cdot (XW) \right|_+ \right)^2 \right\|_1 + \frac{\lambda_1}{2} \|W\|_F^2 + \lambda_3 \|W\|_* \\ \frac{\lambda_2}{2} \sum\limits_{i=1}^{n} \frac{1}{|Y_i^+||Y_i^-|} \sum\limits_{p \in Y_i^+} \sum\limits_{q \in Y_i^-} \max\left(0, 2 - \langle w^p - w^q, x_i \rangle\right)^2 \end{array} \qquad (7)$$

In order to solve the objective function of the above formula, Wu et al. used the accelerated proximal gradient (APG) method [35] to analyze the problem. In general, APG is used to solve non-smooth convex problems. That is, $\min\limits_{W \in H} \Gamma(W) = f(W) + g(W)$. For the above formula, $f_r(W)$ is used to represent the last term in the above formula and decompose its objective function into

$$\begin{array}{l} f(W) = \frac{1}{2} \left\| \left(\left| E - Y \cdot (XW) \right|_+ \right)^2 \right\|_1 + \frac{\lambda_1}{2} \|W\|_F^2 + \lambda_2 f_r(W) \\ g(W) = \lambda_3 \|W\|_* \end{array} . \qquad (8)$$

Therefore, the optimization problem is solved by solving the gradient function of $f(W)$ and calculating its Lipschitz constant. The gradient of $f(W)$ is

$$\begin{array}{l} \nabla_w f(W) = X^T \left(\left| E - Y \cdot (XW) \right|_+ \cdot (-Y) \right) + \lambda_1 W + \lambda_2 \nabla_w f_r(W) \\ where\ \nabla_w f_r(W) = \left[\frac{\partial f_r}{\partial w^1}, \cdots, \frac{\partial f_r}{\partial w^l} \right] \end{array} \qquad (9)$$

The Lipschitz constant of $\nabla_w f(W)$ is

$$l_f = \sqrt{3\left(\|X\|_F^2\right)^2 + 3\lambda_1^2 + 3\left(\lambda_2 L_{f_r}\right)^2} \qquad (10)$$

In summary, this algorithm integrates the threshold to the ranking process of Rank-SVM through the BR algorithm, which not only inherits the virtue from Rank-SVM and BR algorithms but also avoids the defect of two classification algorithms.

Performance evaluation

In the learning, we use LOOCV [36, 37] for testing models. Specifically, LOOCV can obtain as much effective information as possible from limited data and is also regarded as the most rigorous evaluation method. This article uses six multi-label evaluation indicators [13, 38], including RL, coverage (CV), HL, average precision (AP), OAA and OLA.

Assume that $D = \{(x_i, Y_i) | 1 \leq i \leq n\}$ is the test set, and $f(x_i)$ is predicted label set. L is the true label of instance x_i and it denotes as $L = \{l_1, l_2, \cdots, l_q\}$. q is the number of category labels. Its specific definition is as follows:

(1) RL

$$RL = \frac{1}{t} \sum_{i=1}^{n} \frac{\left| \left\{ (l_j, l_k) \middle| f_j(x_i) \leq f_k(x_i), (l_j, l_k) \in Y_i \times \overline{Y_i} \right\} \right|}{|Y_i| \left| \overline{Y_i} \right|}, \qquad (11)$$

where $\overline{Y_i}$ represents the complement of Y_i in L.

(2) CV

$$CV = \frac{1}{q} \left(\frac{1}{n} \sum_{i=1}^{n} \underset{l_k \in Y_i}{\max} \text{rank}(x_i, l_k) - 1 \right), \qquad (12)$$

where $\text{rank}(x_i, l_k)$ is the ranking of the predicted label in the label ranking list.

(3) HL

$$HL = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{q} \left| f(x_i) \oplus Y_i \right|_1, \qquad (13)$$

where $\oplus$ is the symmetric difference between examples, and $|\cdot|_1$ is the 1-norm.

(4) AP

$$AP = \frac{1}{t} \sum_{i=1}^{n} \frac{1}{|Y_i|} \sum_{l_j, l_k \in Y_i} \frac{\left| L_i = \left\{ l_j \middle| \text{rank}(x_i, l_j) \leq \text{rank}(x_i, l_k) \right\} \right|}{\text{rank}(x_i, l_k)} \qquad (14)$$

(5) OAA

$$OAA = \frac{1}{N} \sum_{i=1}^{N} \Delta \left| M(Q_j), L(Q_j) \right| \qquad (15)$$

(6) OLA

$$OLA = \frac{1}{\sum_{i=1}^{N} |L(Q_j)|} \sum_{i=1}^{N} \left| M(Q_j) \cap L(Q_j) \right|, \qquad (16)$$

where $M(Q_j)$ and $L(Q_j)$ represent predicted label set and true label set, respectively, and $|M(Q_j), L(Q_j)| = \begin{cases} 1 & M(Q_j) = L(Q_j) \\ 0 & otherwise \end{cases}$.

Illustration of the Mps-mvRBRL workflow

In the paper, we put forward a novel method Mps-mvRBRL for predicting the SCL, and the flowchart is shown in Figure 1. The source codes and datasets are available at https://github.com/QUST-AIBBDRC/Mps-mvRBRL/.

The steps of Mps-mvRBRL method are as follows:

(1) The protein sequences of the training set Gram-positive bacteria dataset and the independent test sets virus, plant and Gram-negative bacteria datasets were obtained from the SWISS-PROT database and got the true label set according to the sequence category.
(2) The PsePSSM, DC, PSSM-TPC, GO and PseAAC algorithms were used to extract various information in the sequence and we used DE algorithm to study the weights. Multi-view information is fused by the weighted form to constitute the initial features of the model.
(3) The features fused in (2) were passed through wMLDAb to clear away noise information and choose the best feature subset.
(4) After applying LOOCV test, the best subset in (3) was input into RBRL to predict multi-label protein SCL and constructed the Mps-mvRBRL model.
(5) Virus, plant and Gram-negative bacteria datasets were considered as independent test sets and six measurement methods were used to assess Mps-mvRBRL model.

Results and discussion

Selection of optimal parameters λ and ξ

For feature extraction methods containing parameters, we need to separately analyze various prediction results under different parameters. The parameter ξ ranges from 0 to 10, and the interval is 1. The parameter λ is from 5 to 49, and the interval is 5. After employing LOOCV, the feature vectors under different parameters are input into the RBRL classifier to obtain the six evaluation index values, as shown in Supplementary Table S5 and Table S6. And the changes in the OAA values of the PseAAC and PsePSSM algorithms are shown in Supplementary Figure S1. Through the results, when using PseAAC and PsePSSM algorithms in other processes, $\lambda = 20$ and $\xi = 10$ are used.

Effect of feature extraction algorithm on results

Feature extraction contains a variety of feature coding methods, but it is difficult to fully describe each feature of the sequence by only using an individual coding method. Therefore, this paper adopts the method of multi-view feature fusion and uses the feature vectors extracted by PsePSSM, DC, PSSM-TPC, GO and PseAAC algorithms to fuse. We use two multi-view feature fusion methods. One is to use the concatenation method to directly fuse the five feature extraction results, namely PsePSSM+DC+PSSM-TPC+GO+PseAAC. The other is to use the DE algorithm to assign weights according to the contribution value of each single feature and then perform feature fusion, namely w_1 PsePSSM + w_2 DC + w_3 PSSM-TPC + w_4 GO + w_5 PseAAC. So as to highlight the advantages of multi-view feature fusion, we input the feature vectors of five individual feature encoding methods and two feature fusions into the RBRL classifier. After employing LOOCV, the specific results are shown in Supplementary Table S7. At this time, for Gram-positive bacteria dataset, this paper uses five feature coding methods to generate a total of 1972-dimension feature vectors. In order to evaluate the robustness of the five feature extraction methods and two fusion methods, the comparison of ROC and PR curves are shown in Figure 2.

It can be seen from Figure 2 that for the five single feature extraction algorithms, when the GO algorithm is used, the area under the curve (AUC) value is 0.9622, which is 13.21%, 12.13%, 13.69%, and 11.84% higher than those of PseAAC, PsePSSM, DC and PSSM-TPC, respectively. Similarly, when we use GO algorithm to extract information, the AUPR value is 0.9385, which is 27.72%, 24.74%, 27.02% and 26.54% higher than those of five feature extraction algorithms, respectively. Through analysis, among the five single feature extraction methods, the GO algorithm has the best robustness. In the two fusion methods, when the fusion method of DE algorithm is used, the AUC value is 0.9643, which is 3.75% higher than those of ordinary fusion. Similarly, when the fusion method of DE algorithm is used, the AUPR value is 0.9428, which is 10.57% higher than those of ordinary fusion. Through analysis, among the two fusion methods, the fusion method of DE algorithm has the best robustness.

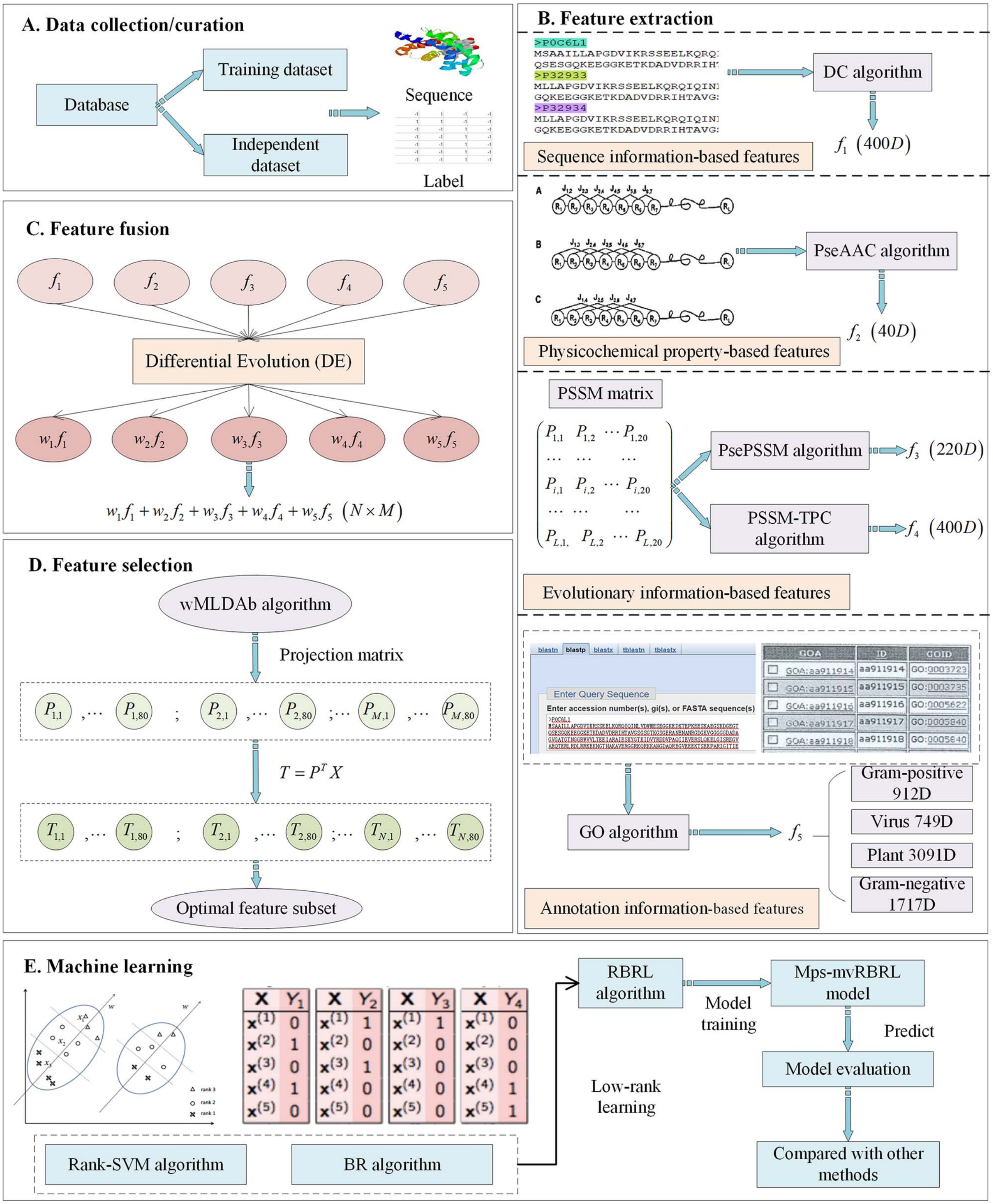
Effect of dimensional reduction algorithm on results

Feature fusion can produce high-dimensional data and often accompany many redundant features, which seriously affects the performance of the classifier. Therefore, dimension reduction is the main problem that we need to solve at present. In order to select the best feature subset, this paper selects Gram-positive bacteria as the research object. We use principal component analysis (PCA) [39], multi-label informed latent semantic indexing (MLSI) [40], multi-label dimensionality reduction via dependency maximization (MDDM) [41], MVMD [12] and wMLDAb algorithms to reduce the dimension, respectively. The dimension of five algorithms is taken 10 to 100 with an interval of 10. After employing LOOCV, the changes of the six evaluation indicators by using five dimensionality reduction algorithms are shown in Figure 3. In order to evaluate the robustness of the five dimensionality reduction methods, the comparison and analysis of ROC and PR curves are shown in Supplementary Figure S2.

In Figure 3, compared with PCA, MLSI, MDDM and MVMD, it can be seen from the comparison of six evaluation indicators that the prediction effect of wMLDAb algorithm is optimized. For the wMLDAb algorithm, when the dimension is 80, the OAA and OLA values both are 99.81%, which are 3.86% and 3.83% higher than those of initial 10 dimensions. At this time, the performance index reaches the best. The wMLDAb algorithm removes redundant features in high-dimensional vectors by considering the relevant information between labels and finally selects the best feature subset, thereby improving the prediction accuracy. Therefore, this article chooses wMLDAb as dimensionality reduction algorithm.

Selection of classification algorithms

The classifier occupies a dominant position in the model and appropriate classifier can improve the accuracy about the model. This paper uses Gram-positive bacteria as the research object and selects multi-label K-nearest neighbor [42], multi-label learning with label-specific features [43], Rank-SVM [32], multi-instance multi-label learning radial basis function [44], multi-instance multi-label learning K-nearest neighbor [45], multi-label Gaussian kernel regression [46] and RBRL algorithms to predict the SCL. After applying LOOCV, the prediction results of six evaluation indicators are shown in Figure 4.



X	Y_1	X	Y_2	X	Y_3	X	Y_4
$x^{(1)}$	0	$x^{(1)}$	1	$x^{(1)}$	1	$x^{(1)}$	0
$x^{(2)}$	1	$x^{(2)}$	0	$x^{(2)}$	0	$x^{(2)}$	0
$x^{(3)}$	0	$x^{(3)}$	1	$x^{(3)}$	0	$x^{(3)}$	0
$x^{(4)}$	1	$x^{(4)}$	0	$x^{(4)}$	0	$x^{(4)}$	1
$x^{(5)}$	0	$x^{(5)}$	0	$x^{(5)}$	0	$x^{(5)}$	1

GOA	ID	GOID
□ GOA:aa911914	aa911914	GO:0003723
□ GOA:aa911915	aa911915	GO:0003735
□ GOA:aa911916	aa911916	GO:0005622
□ GOA:aa911917	aa911917	GO:0005840
□ GOA:aa911918	aa911918	GO:0005840

Figure 1. The overall framework of Mps-mvRBRL.

In Figure 4, compared with other classification algorithms, the RBRL has the highest OAA and OLA of both 99.81%, which are 0.20%–3.86% and 0.57%–4.21% higher than those of other classifiers. And by comparing other indicators, the prediction results that we get by using RBRL are also better than other classifiers. The RBRL algorithm can not only solve the optimization problem of HL and RL but also consider the correlation of labels. Therefore, this article chooses RBRL as classification algorithm. Supplementary Table S8 and Table S9 show the specific results and parameter selection of the seven classification algorithms.

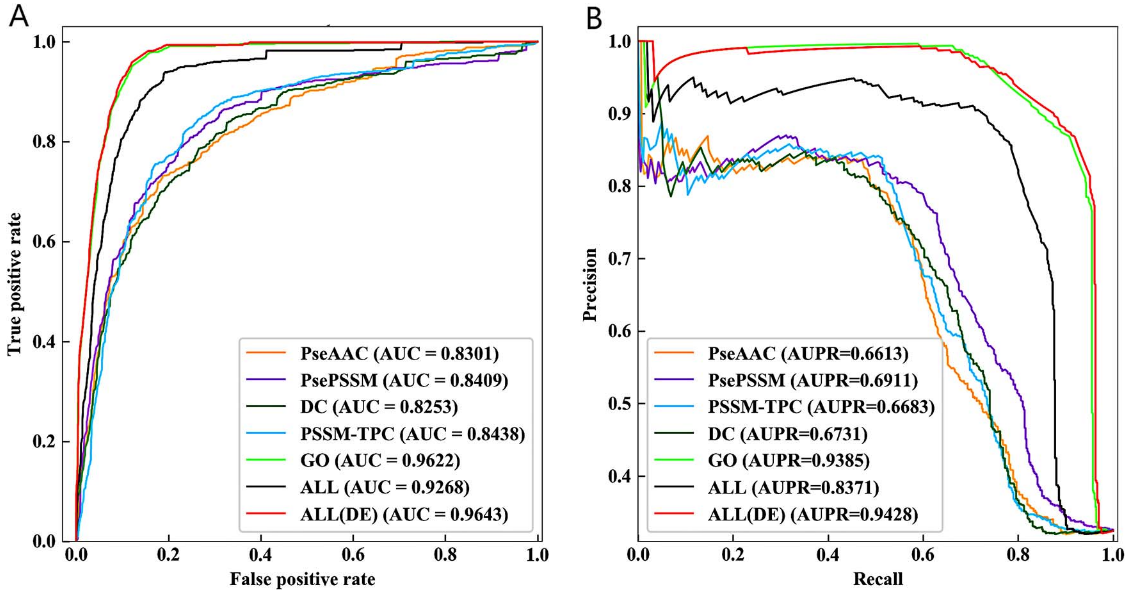


Figure 2. Comparison of ROC and PR curves for five feature extraction methods and two fusion methods. All:PseAAC+PsePSSM+DC+GO+PSSM-TPC. (A) The ROC curves of seven methods on Gram-positive bacteria dataset. (B) The PR curves of seven methods on Gram-positive bacteria dataset.

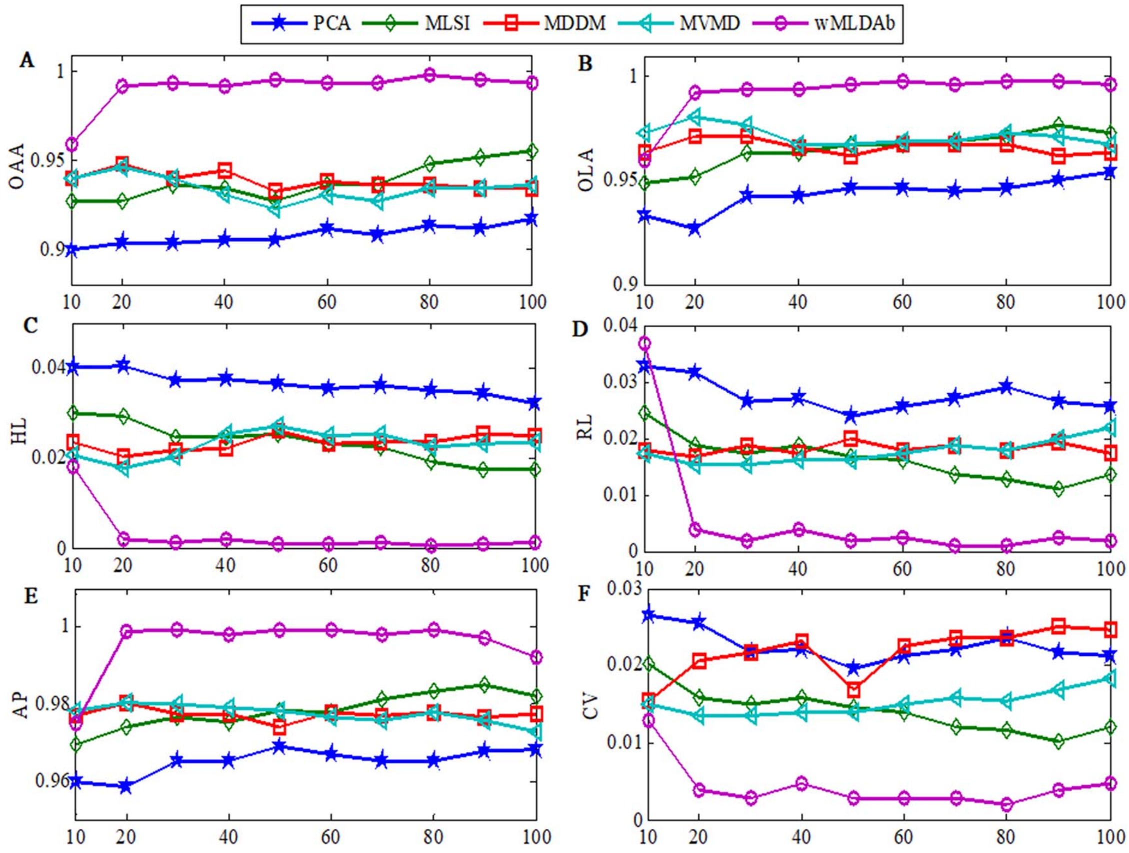


Figure 3. The prediction results of Gram-positive bacteria with five-dimensionality reduction algorithms. (A) overall actual accuracy, (B) overall location accuracy, (C) hamming loss, (D) ranking loss, (E) average precision, (F) coverage.

Comparison with other methods

We select the training set of Gram-positive bacteria dataset to verify the prediction capability of Mps-mvRBRL. The results of Mps-mvRBRL are compared with iLoc-Gpos [47], Gpos-ECC-mPLoc [48] and Gram-LocEN [16] based on the LOOCV. And the datasets used in each model are the same. The specific comparison results of Mps-mvRBRL and other models are shown in Supplementary Table S10.

This article uses the virus, plant and Gram-negative bacteria datasets to test the performance of Mps-mvRBRL and compares

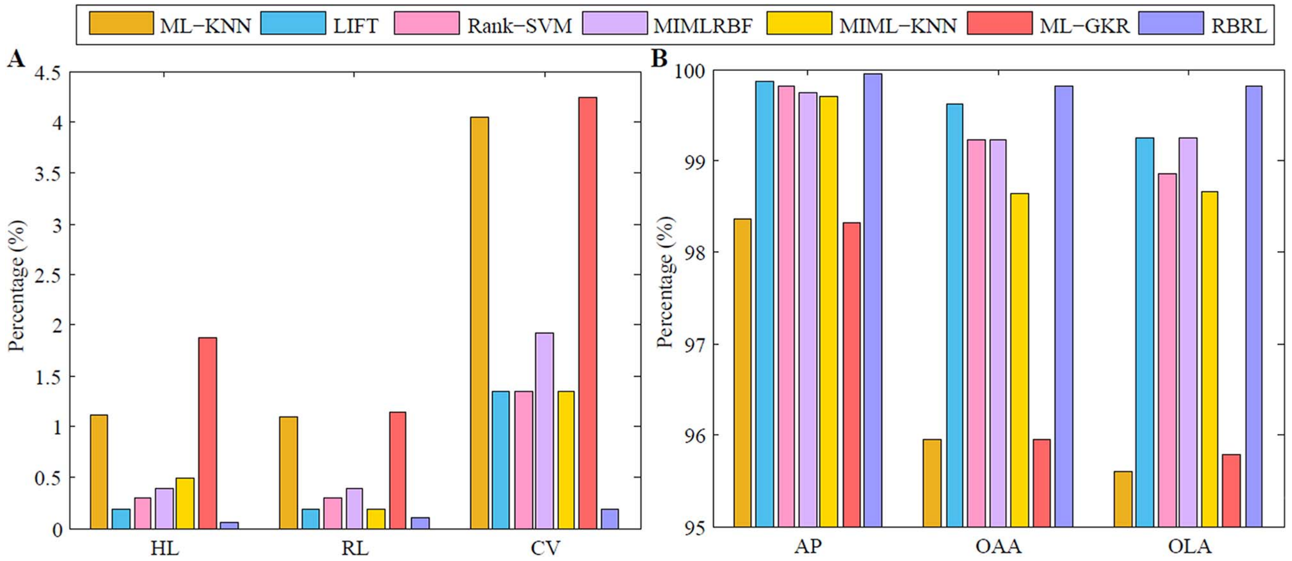


Figure 4. The prediction results of Gram-positive bacteria with seven classification algorithms. (A) The results of HL, RL and CV values obtained by seven classifiers in Gram-positive bacteria dataset. (B) The results of OAA, OLA and AP values obtained by seven classifiers in Gram-positive bacteria dataset.

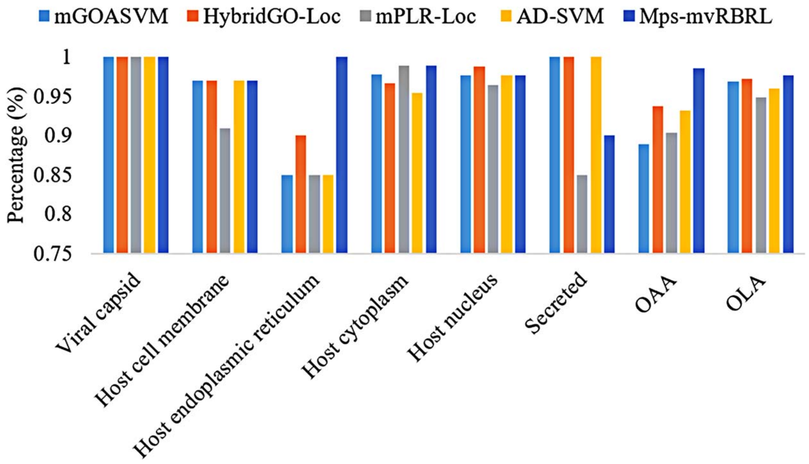


Figure 5. Comparison of different prediction methods on virus dataset.

results with other methods. The other methods used in the model are unchanged. After employing LOOCV, the prediction results by using the Mps-mvRBRL method on the three datasets are compared with the same dataset in the model mGOASVM [49], HybridGO-Loc [50], mPLR-Loc [51], AD-SVM [9], iLoc-Gneg [52], Gneg-ECC-mPLoc [48] and Gram-LocEN [16]. The prediction results of each site are shown in Figure 5–7. The OAA and OLA values of Mps-mvRBRL are 98.55% and 97.62% for virus dataset, respectively, which are 4.85%–9.65% and 0.42%–2.82% higher than the accuracy of mGOASVM, HybridGO-Loc, mPLR-Loc and AD-SVM. The OAA and OLA values of Mps-mvRBRL are 97.24% and 97.16% for the plant dataset, respectively, which are 3.64%–9.84% and 0.96%–3.46% higher than the accuracy of mGOASVM, HybridGO-Loc, mPLR-Loc and AD-SVM. The OAA and OLA values of Mps-mvRBRL are 98.20% and 98.28% for Gram-negative bacteria dataset, respectively, which are 3.70%–8.30% and 2.98%–6.88% higher than the accuracy of iLoc-Gneg, Gneg-ECC-mPLoc and Gram-LocEN. The specific comparison results of Mps-mvRBRL and other models are shown in Supplementary Table S11–S13. Therefore, this model can provide a strong guarantee for studying the function of protein in each cell compartment.

Conclusion

This paper puts forward a new method Mps-mvRBRL to predict SCL. Firstly, we use PsePSSM, DC, PSSM-TPC, GO and PseAAC methods to extract numerical information from multiple views. According to the nature of the single-feature coding algorithm, this paper uses the DE algorithm for the first time to learn the weight vector, and the multi-view information is fused by the weighted form. As an efficient evolutionary algorithm for global optimization, DE algorithm has a strong global convergence ability and robustness. Secondly, wMLDAb is applied for the first time to process the fused high-dimensional data. The algorithm makes full use of the label correlation and assigns binary weights on the basis of MLDA. It also can remove irrelevant information in the sequence while retaining the effective features. Finally, we input the best feature subset into RBRL to predict. This algorithm not only takes advantage of the Rank-SVM and BR algorithms but also solves the optimization problems of RL and HL. It also uses an APG method to further utilize the correlation of labels. Based on the above methods, LOOCV test shows that the model Mps-mvRBRL has achieved good prediction results. We also expect

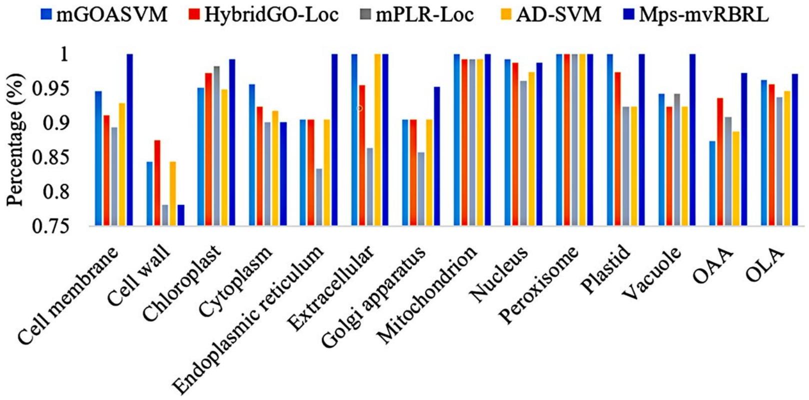


Figure 6. Comparison of different prediction methods on plant dataset.

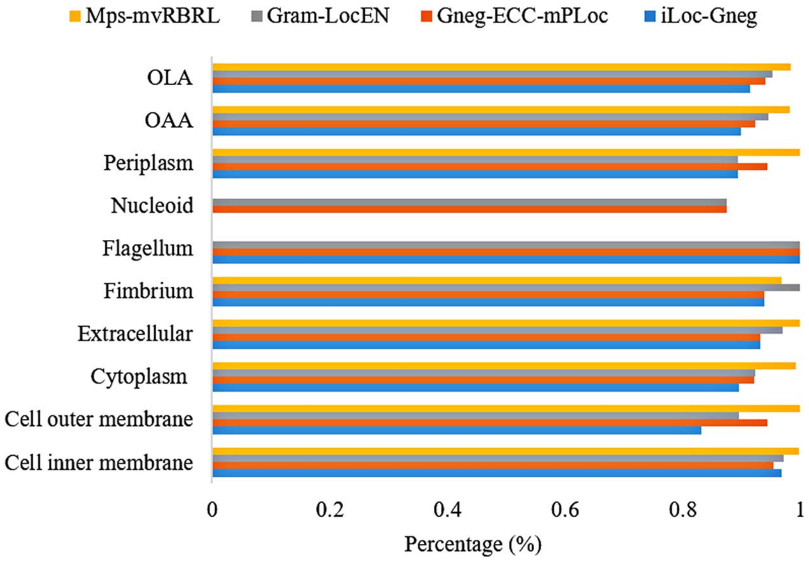


Figure 7. Comparison of different prediction methods on Gram-negative bacteria dataset.

that Mps-mvRBRL can be applied to more multi-label protein attribute predictions. Further research shows that some multi-label proteins are not fixed in organelles, but they exist in the free state. This type of protein usually has a special physiological function in the living body, so we will focus on this type of protein and predict its SCL by studying its location rules in the future.

Key Points

- We propose a new predictor, Mps-mvRBRL, to predict SCL of multi-label protein.
- This article adopts DE algorithm to learn the weight vector of five single-features and fuses them based on the weighted form. The wMLDAb algorithm is first used to remove irrelevant information. And we use Rank-SVM and BR algorithm combined with robust low-rank learning to predict the subcellular location.
- The results of this paper on four benchmark datasets show that Mps-mvRBRL achieves good prediction performance for predicting the SCL of multi-label protein.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Funding

National Natural Science Foundation of China (61863010); Key Research and Development Program of Shandong Province of China (2019GGX101001); Natural Science Foundation of Shandong Province of China (ZR2018MC007); Key Laboratory Open Foundation of Hainan Province (JSKX202001).

References

1. Campbell JB, Crocker J, Shenoi PM. S-100 protein localization in minor salivary gland tumours: an aid to diagnosis. *J Laryngol Otol* 1988;**102**:905–8.
2. Chen Y, Chen CF, Riley DJ, *et al*. Aberrant subcellular localization of BRCA1 in breast cancer. *Science* 1995;**270**:789–91.
3. Hung MC, Link W. Protein localization in disease and therapy. *J Cell Sci* 2011;**124**:3381–92.

4. Krutovskikh V, Mazzoleni G, Mironov N, *et al.* Altered homologous and heterologous gap-junctional intercellular communication in primary human liver tumors associated with aberrant protein localization but not gene mutation of connexin 32. *Int J Cancer* 1994;**56**:87–94.
5. Tahir M, Khan A. Protein subcellular localization of fluorescence microscopy images: employing new statistical and Texton based image features and SVM based ensemble classification. *Inform Sciences* 2016;**345**:65–80.
6. Wei LY, Ding YJ, Su R, *et al.* Prediction of human protein subcellular localization using deep learning. *J Parallel Distr Com* 2018;**117**:212–7.
7. Shen YN, Ding YJ, Tang JJ, *et al.* Critical evaluation of web-based prediction tools for human protein subcellular localization. *Brief Bioinform* 2019;**21**:1628–40.
8. Javed F, Hayat M. Predicting subcellular localization of multi-label proteins by incorporating the sequence features into Chou's PseAAC. *Genomics* 2019;**111**:1325–32.
9. Wan SB, Mak MW. Predicting subcellular localization of multi-location proteins by improving support vector machines with an adaptive-decision scheme. *Int J Mach Learn Cyb* 2018;**9**:399–411.
10. Zhang S, Zhang T, Liu C. Prediction of apoptosis protein subcellular localization via heterogeneous features and hierarchical extreme learning machine. *SAR QSAR Environ Res* 2019;**30**:209–28.
11. Lin YJ, Hu QH, Liu JH, *et al.* Multi-label feature selection based on max-dependency and min-redundancy. *Neurocomputing* 2015;**168**:92–103.
12. Xu JH, Liu JL, Yin J, *et al.* A multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence simultaneously. *Knowl-Based Syst* 2016;**98**:172–84.
13. Zhang J, Luo ZM, Li CD, *et al.* Manifold regularized discriminative feature selection for multi-label learning. *Pattern Recogn* 2019;**95**:136–50.
14. Wan SB, Mak MW, Kung SY. Transductive learning for multi-label protein subchloroplast localization prediction. *IEEE ACM T Comput Bi* 2016;**14**:212–24.
15. Hasan MAM, Ahmad S, Molla MKI. Protein subcellular localization prediction using multiple kernel learning based support vector machine. *Mol Biosyst* 2017;**13**:785.
16. Wan SB, Mak MW, Kung SY. Gram-LocEN: interpretable prediction of subcellular multi localization of Gram-positive and Gram-negative bacterial proteins. *Chemometr Intell Lab Syst* 2017;**162**:1–9.
17. Shen HB, Chou KC. Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins. *Protein Pept Lett* 2009;**16**:1478–84.
18. Shen HB, Chou KC. Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. *J Biomol Struct Dyn* 2010;**26**:175–86.
19. Chou KC, Shen HB. Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PLoS One* 2010;**5**: e11335.
20. Jones D. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999; **292**:195–202.
21. Yu B, Yu ZM, Chen C, *et al.* DNNAce: prediction of prokaryote lysine acetylation sites through deep neural networks with multi-information fusion. *Chemometr Intell Lab Syst* 2020;**200**: 103999.
22. Manavalan B, Basith S, Shin TH, *et al.* mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 2018;**35**:2757–65.
23. Ali F, Ahmed S, Swati ZNK, *et al.* DP-BINDER: machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information. *J Comput Aid Mol Des* 2019;**33**:645–58.
24. Zhang CX, Zheng W, Freddolino PL, *et al.* MetaGO: predicting gene ontology of non-homologous proteins through low-resolution protein structure prediction and protein–protein network mapping. *J Mol Biol* 2018;**430**:2256–65.
25. Yu B, Qiu WY, Chen C, *et al.* SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics* 2020;**36**:1074–81.
26. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 2005;**21**:10–9.
27. Hu J, Zhou XG, Zhu YH, *et al.* TargetDBP: accurate DNA-binding protein prediction via sequence-based multi-view feature learning. *IEEE ACM T Comput Bi* 2020;**17**:1419–29.
28. Fister I, Suganthan PN, Fister I, Jr, *et al.* Artificial neural network regression as a local search heuristic for ensemble strategies in differential evolution. *Nonlinear Dynam* 2016;**84**:895–914.
29. Chen N, Chen WN, Zhang J. Fast detection of human using differential evolution. *Signal Process* 2015;**110**:155–63.
30. Wang H, Ding C, Huang H. Multi-label linear discriminant analysis. *Lect Notes Comput Sci* 2010;**6316**:126–39.
31. Xu JH. A weighted linear discriminant analysis framework for multi-label feature extraction. *Neurocomputing* 2018;**275**:107–20.
32. Elisseeff A, Weston J. A kernel method for multi-labelled classification. *Adv Neural Inf Process Syst* 2001;**4**:681–7.
33. Boutell MR, Luo J, Shen X, *et al.* Learning multi-label scene classification. *Pattern Recogn* 2004;**37**:1757–71.
34. Wu GQ, Zheng RB, Tian YJ, *et al.* Joint ranking SVM and binary relevance with robust low-rank learning for multi-label classification. *Neural Netw* 2020;**122**:24–39.
35. Nesterov Y. Smooth minimization of non-smooth functions. *Math Program* 2005;**103**:127–52.
36. Wang XY, Yu B, Ma AJ, *et al.* Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics* 2019;**35**:2395–402.
37. Chen X, Li TH, Zhao Y, *et al.* Deep-belief network for predicting potential miRNA-disease associations. *Brief Bioinform* 2020. doi: https://doi.org/10.1093/bib/bbaa186.
38. Zhang Q, Li S, Yu B, *et al.* DMLDA-LocLIFT: identification of multi-label protein subcellular localization using DMLDA dimensionality reduction and LIFT classifier. *Chemometr Intell Lab Syst* 2020;**206**:104148.
39. Abdi H, Williams LJ. Principal component analysis. *Comput Stat* 2010;**2**:433–59.
40. Yu K, Yu SP, Tresp V. Multi-label informed latent semantic indexing. *Int ACM SIGIR Conf Res Dev Inf Retriev* 2005; 258–65.
41. Zhang Y, Zhou ZH. Multilabel dimensionality reduction via dependency maximization. *ACM Trans Knowl Discov* 2010;**4**:14.
42. Zhang ML, Zhou ZH. ML-KNN: a lazy learning approach to multi label learning. *Pattern Recogn* 2007;**40**:2038–48.

43. Zhang ML, Wu L. LIFT: multi-label learning with label-specific features. *IEEE Trans Pattern Anal Mach Intell* 2015;**37**:107–20.
44. Zhang ML, Wang ZJ. MIMLRBF: RBF neural networks for multi-instance multi-label learning. *Neurocomputing* 2009;**72**:3951–6.
45. Zhang ML. A k-nearest neighbor based multi-instance multi-label learning algorithm In: *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*. 2010.
46. Cheng X, Lin WZ, Xiao X, *et al.* pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC. *Bioinformatics* 2019;**35**:398–406.
47. Wu ZC, Xiao X, Chou KC. iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. *Protein Pept Lett* 2012;**19**:4–14.
48. Wang X, Zhang J, Li GZ. Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble. *BMC Bioinform* 2015;**16**:S1.
49. Wan SB, Mak MW, Kung SY. mGOASVM: multi-label protein subcellular localization based on gene ontology and support vector machines. *BMC Bioinform* 2012;**13**: 290.
50. Wan SB, Mak MW, Kung SY. HybridGO-Loc: mining hybrid features on gene ontology for predicting subcellular localization of multi-location proteins. *PLoS One* 2014;**9**: e89545.
51. Wan SB, Mak MW, Kung SY. mPLR-Loc: an adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction. *Anal Biochem* 2015;**473**:14–27.
52. Xiao X, Wu ZC, Chou KC. A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. *PLoS One* 2011;**6**: e20592.