Bioinformatics, 36(4), 2020, 1074–1081
doi: 10.1093/bioinformatics/btz734
Advance Access Publication Date: 11 October 2019
Original Paper

Sequence analysis

SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting

Bin Yu[1,2,3,4,*], Wenying Qiu[1,3], Cheng Chen[1,3], Anjun Ma[5], Jing Jiang[5,6], Hongyan Zhou[1,3] and Qin Ma[5,*]

[1]College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China, [2]School of Life Sciences, University of Science and Technology of China, Hefei 230027, China, [3]Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao 266061, China, [4]School of Mathematics and Statistics, Changsha University of Science and Technology, Changsha 410114, China, [5]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA and [6]School of Aerospace Engineering, Xiamen University, Xiamen 361001, China

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on January 25, 2019; revised on September 4, 2019; editorial decision on September 19, 2019; accepted on September 25, 2019

Abstract

Motivation: Mitochondria are an essential organelle in most eukaryotes. They not only play an important role in energy metabolism but also take part in many critical cytopathological processes. Abnormal mitochondria can trigger a series of human diseases, such as Parkinson's disease, multifactor disorder and Type-II diabetes. Protein submitochondrial localization enables the understanding of protein function in studying disease pathogenesis and drug design.

Results: We proposed a new method, SubMito-XGBoost, for protein submitochondrial localization prediction. Three steps are included: (i) the *g*-gap dipeptide composition (*g*-gap DC), pseudo-amino acid composition (PseAAC), autocorrelation function (ACF) and Bi-gram position-specific scoring matrix (Bi-gram PSSM) are employed to extract protein sequence features, (ii) Synthetic Minority Oversampling Technique (SMOTE) is used to balance samples, and the ReliefF algorithm is applied for feature selection and (iii) the obtained feature vectors are fed into XGBoost to predict protein submitochondrial locations. SubMito-XGBoost has obtained satisfactory prediction results by the leave-one-out-cross-validation (LOOCV) compared with existing methods. The prediction accuracies of the SubMito-XGBoost method on the two training datasets M317 and M983 were 97.7% and 98.9%, which are 2.8–12.5% and 3.8–9.9% higher than other methods, respectively. The prediction accuracy of the independent test set M495 was 94.8%, which is significantly better than the existing studies. The proposed method also achieves satisfactory predictive performance on plant and non-plant protein submitochondrial datasets. SubMito-XGBoost also plays an important role in new drug design for the treatment of related diseases.

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

Contact: yubin@qust.edu.cn or qin.ma@osumc.edu.

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

1 Introduction

Mitochondria are a vital organelle in eukaryotic cells and are involved in critical physiological processes such as cell differentiation, cell information transmission and apoptosis and growth. The bias of proteins submitochondrial localization can lead to a series of damage interactions and cause serious diseases such as Parkinson's disease (Burbulla *et al.*, 2017), multifactor disorder (Shi *et al.*, 2011) and Type-II diabetes (Gerbitz *et al.*, 1996). Therefore, the study of protein submitochondrial localizations is of great significance to provide a theoretical basis for the exploration of the pathogenesis, diagnosis and development of new drugs at the molecular level for

human disease. In recent research at the sub-cellular level, some significant results have been achieved in the prediction of the location of subcellular structures (Chou and Shen, 2006; Yu *et al.*, 2018). Traditional biological methods such as cell separation, electron microscopy and fluorescence microscopy (Li *et al.*, 2015; Mei, 2012) are not very feasible for analyzing large-scale biological data. However, computational models can not only analyze a large amount of biological data, but can also make preliminary predictions of unknown located protein sequences, which is an excellent auxiliary for traditional biological experimental methods.

Recently, researchers have developed some protein submitochondrial localization prediction models using feature extraction, feature selection and machine learning algorithms, such as SubMito (Du and Li, 2006), GPLoc (Nanni and Lumini, 2008), Predict_SubMito (Zeng *et al.*, 2009), SubIdent (Shi *et al.*, 2011), MitoLoc (Zakeri *et al.*, 2011), SubMito-PSPCP (Du and Yu, 2013) and TetraMito (Lin *et al.*, 2013). The first submitochondrial protein localization study was done by Du and Li (2006) to develop SubMito, which first fuses amino acid component information and dipeptide component information, and then combines nine physical and chemical properties of the protein sequence into the support vector machine (SVM) classifier. Lin *et al.* (2013) developed TetraMito, which used the binomial distribution to select the tetrapeptide components, and generated the dataset M317 with prediction accuracy (ACC) of 94%, however, not considering the physicochemical properties. Nanni and Lumini (2008) proposed the GPLoc algorithm, which fused PseAAC and the physicochemical properties of the Amino Acid index (AAIndex) database to extract the feature of protein sequences from the dataset M317. Shi *et al.* (2011) constructed the SubIdent to locate the submitochondria on dataset M317 using discrete wavelet transform and combining with protein physicochemical properties such as the hydrophilic, hydrophobic and polarity values for feature extraction. Du and Yu (2013) constructed the SubMito-PSPCP to predict the protein submitochondria location. Qiu *et al.* (2018) proposed PseAAC-PsePSSM-WD method to predict the protein submitochondria locations. Mei (2012) proposed the multi-kernel transfer learning model for submitochondrial protein localization, which uses the transfer learning model to measure the individual contribution of GO's molecular function, cellular composition, and biological process. Since each feature extraction method only extracts part of the information of the protein sequence, research has been done to integrate multiple methods for protein subcellular information extraction. Jiao *et al.* (Jiao and Du, 2017) proposed a method integrating functional domain enrichment score, position-specific physicochemical properties (PSPCP) and PseAAC obtaining a prediction ACC of 94%. Li *et al.* (2015) fused position specific scoring matrix and gene ontology to extract feature sequences and achieved ideal prediction ACC.

Feature extraction of protein sequences is based on a variety of feature extraction methods in which dimension is often relatively high and contains a lot of redundant information, reducing the predictive performance of the model. Dimension reduction methods filter the feature vectors in high-dimensional data to eliminate the unnecessary features. Commonly used methods are principal component analysis (PCA) (Ahmad *et al.*, 2016), information gain (Wen *et al.*, 2016), maximum relevance and minimum redundancy (mRMR) (Khan *et al.*, 2017), maximum relevance maximum distance (MRMD) (Zou *et al.*, 2016), singular value decomposition (SVD) (Silvério-Machado *et al.*, 2015) and local linear discriminant analysis (Yu *et al.*, 2018). Ahmad *et al.* (Ahmad *et al.*, 2016) fused dipeptide composition (DPC), split amino acid composition (SAAC) and composition and translation methods for protein sequence feature extraction, then used PCA to select optimal feature vectors, and the satisfactory prediction ACC was obtained. Therefore, it is essential to select features for the sequence information after feature fusion.

Machine learning methodologies are commonly used in the research field of protein localization prediction, such as hidden Markov model (HMM) (Fariselli *et al.*, 2005), support vector machine (SVM) (Li *et al.*, 2015), K-nearest neighbor (KNN) (Chou and Shen, 2006), increment of diversity (ID) (Chen, 2012) and ensemble learning classifier (Wang *et al.*, 2018). Ensemble learning has become a hotspot in the field of bioinformatics (Wang *et al.*, 2018) to predict submitochondrial proteins locations. Zakeri *et al.* (2011) proposed MitoLoc applying ordered weighted averaging (OWA) to ten commonly used feature extraction methods to construct the better SVM classifier. But this method may have a lower ACC rate for a few classes due to data imbalance. Therefore, the classification problem of imbalanced datasets needs to be further studied, and there is still much room for improvement based on the integrated learning prediction model.

Here we propose SubMito-XGBoost for protein submitochondrial localization prediction using a three-step pipeline: (i) methods of *g*-gap dipeptide composition (*g*-Gap DC), PseAAC, autocorrelation function (ACF) and Bi-gram position specific scoring matrix (Bi-gram PSSM) are integrated to extract protein sequences of submitochondria; (ii) the feature-extracted dataset is processed using the synthetic minority oversampling technique (SMOTE) method (Chawla *et al.*, 2002) to ensure a balance among submitochondrial protein classes, and the ReliefF algorithm (Kira and Rendell, 1992) is used to select the optimal feature subsets; (iii) the optimal feature vectors are input into the XGBoost classifier to predict the locations of the protein submitochondria. To evaluate the performance of the SubMito-XGBoost model, the leave-one-out-cross-validation (LOOCV) was applied to datasets M317 and M983. As a result, the prediction ACC of the two datasets were 97.65% and 98.94%, respectively, and the optimal parameter combination was determined. An additional dataset M495 was used as an independent test data for model prediction. The overall prediction ACC of SubMito-XGBoost was 94.83% showing a significant improvement of protein submitochondrial localization prediction compared with other existing tools.

2 Materials and methods

2.1 Datasets

Three submitochondrial protein datasets, M317, M495 and M983, were selected as objective representative baselines to construct the computational model for precise localization analysis. The datasets M317 and M983, constructed by Du *et al.* (Du and Li, 2006; Du and Yu, 2013) with sequence similarities not exceeding 40%, were extracted from the Uniprot database (Release 2016_04, https://www.uniprot.org/) (UniProt, 2015) and used as test sets for parameter selection. The dataset M495 (Lin *et al.*, 2013) with similarity not exceeding 25% was used for independent test and has been experimentally annotated as the subcellular location 'Mitochondrion'. The M495 protein data was extracted from the UniProt database (http://lin.uestc.edu.cn/-server/subMito/data). Proteins in each of the three datasets were grouped into three sub-regions: Intermembrane, Matrix and Outer membrane. The number of sub-region proteins in the three submitochondrial datasets are shown in Table 1.

2.2 Feature extraction

The *g*-gap dipeptide composition (*g*-Gap DC) introduces important intrinsic correlation information of protein sequences in a multi-dimensional space, which has been successfully applied in protein structure predictions (Ding and Li, 2015). Among them, the parameter *g* selection plays an important role in the *g*-Gap DC feature

Table 1. Structure distribution of submitochondrial datasets

Datasets	Number of proteins			
	Inner membrane	Matrix	Outer membrane	Total
M317	131	145	41	317
M495	254	132	109	495
M983	661	177	145	983

method. The PseAAC (Chou, 2005) combines the sequence information of amino acid sequences with the physicochemical properties of amino acid residues. In this paper, we used the network server constructed by Shen and Chou (2008), selecting the Type1 type in the pseudo amino acid model and two amino acid attribute features, Hydrophobicity and Hydrophilicity, with the weight parameter ω set to 0.05. The selection of the parameter λ cannot exceed the length of the shortest sequence in the submitochondrial protein dataset, and the shortest sequence lengths of the datasets M317, M983 and M495 are 54, 102 and 49, respectively, so the range of λ is 0 to 48. Thus, each time a protein sequence is entered, the server returns a $20+\lambda$-dimensional feature vector. The auto-correlation function (ACF) was proposed by Bu *et al.* (Bu *et al.*, 2010) in 1999 to extract features from protein sequences by amino acid index. The ACF algorithm converts the protein sequence into a numerical sequence based on the indices of the 20 amino acids and then defines the autocorrelation function of the numerical sequence. We used the ACF feature extraction process with parameter m, where m is an integer and less than L. We transformed a protein sequence into a self-correlation function eigenvector of fixed length $10 \times m$. In the end, a feature extraction method based on Bi-gram position-specific scoring matrix (Bi-gram PSSM) (Khan *et al.*, 2017) was adopted. This solves the shortcomings of calculating the feature vector of the original protein sequence when the feature vector is very sparse. Moreover, it converts protein sequences of different lengths into fixed-length feature vectors, retaining a large amount of evolutionary information. We used the PSI-BLAST program to compare the imported submitochondrial protein sequences in FASTA format with the non-redundant protein database nr, where the parameter is set to 3 iterations, the value of e is equal to 0.001, and the other parameters adopt the default settings to obtain the PSSM of each protein sequence. The Bi-gram PSSM method generated a 400-dimensional feature vectors for the protein sequence. The specific theoretical process of the four feature extraction methods is found in Supplementary Si1–Si4.

2.3 Synthetic minority oversampling technique

SMOTE proposed by Chawla *et al.* (2002) is a commonly used oversampling method. It uses a small number of samples to control the generation and distribution of artificial samples to achieve dataset class balance, thus effectively solving the problem of classification over-fitting. The specific theoretical process of the SMOTE method is located in Supplementary Si5.

The submitochondrial datasets M317, M983 and M495 used in this paper all have sample imbalance problems. Among them, the matrix proteins in the dataset M317 is about three times than that of the outer membrane proteins, the intermembrane proteins in the dataset M495 are nearly twice as large as the outer membrane protein sequences, and the intermembrane proteins in the dataset M983 is about five times than that of the outer membrane proteins. When the sample unbalanced dataset predicts the model, it often causes the prediction results that are biased toward the majority. To deal with the problem of dataset category imbalance, this paper uses SMOTE algorithm to select a few types of outer membrane protein samples and create new synthetic samples along the line segments connecting k nearest neighbors by $x_{new} = x_i + rand(0,1) \times (y_i - x_i), i = 1, 2, \ldots, N$, so that the three types of protein sequences are approximately equal, and finally, the dataset reaches equilibrium.

2.4 eXtreme gradient boosting

Extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016) is an ensemble learning algorithm based on gradient boosting, which is applied by researchers to bioactive molecular prediction (Babajide and Saeed, 2016), miRNA-disease association prediction (Chen *et al.*, 2018) and post-translation modified locus prediction (Zhao *et al.*, 2018). XGBoost is an optimization model that combines a linear model with a boosting tree model. It uses not only the first derivative but also the second derivative of the loss function for second-order derivation. This allows the algorithm to converge to the global optimality faster and improve the efficiency of the optimal solution of the model.

For a given n sample and m feature datasets $D = \{(x_i, y_i) (|D| = n, x_i \in R^m, y_i \in R)\}$, the XGBoost algorithm objective function is defined as

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \tag{1}$$

where l is the loss function, so the smaller l is, the better the performance of the algorithm. f_t is the t-th tree. Then, according to Taylor expansion optimization of the objective function, the second-order Taylor of the loss function after the t iteration is obtained, and the approximate loss function is obtained as the formula (2):

$$L^{(t)} \simeq \sum_{i=1}^{k} [l(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{2}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ represents the first derivative of each sample and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ represents the second derivative of each sample, and the loss function depends only on the first and second derivatives of each data point.

When using the XGBoost integrated learning algorithm to predict protein submitochondria, the essential steps are to optimize the XGBoost algorithm's general parameters, booster parameters and learning parameters.

2.5 Performance evaluation and model construction

At present, the test methods for evaluating the performance of the model based on statistical theory include independent test, k-fold cross-validation test and LOOCV. Among them, LOOCV can reflect the ACC and generalization ability of the algorithm objectively and rigorously, so it is widely used in protein submitochondrial localization prediction (Du and Yu, 2013; Lin *et al.*, 2013; Zakeri *et al.*, 2011).

Due to the imbalance of the data used in this paper, six standard statistical indicators are used to measure the statistical prediction information obtained after the prediction: sensitivity is also the recall rate (Sn), specificity (Sp), ACC, Matthew's correlation coefficient (MCC), precision (Pre) and F-measure ($F1$). They are defined as follows:

$$Sn = \frac{TP}{TP + FN} \tag{3}$$

$$Sp = \frac{TN}{TN + FP} \tag{4}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{6}$$

$$Pre = \frac{TP}{TP + FP} \tag{7}$$

$$F1 = 2 \times \frac{Pre \times Sn}{Pre + Sn} \tag{8}$$

$F1$ is a statistic used to evaluate the generalization performance of the model, which is a weighted average of the precision and the recall. The performance of the model can be well evaluated in the case of unbalanced datasets. The higher the value of $F1$, the better the performance of the classifier. The MCC can measure the specificity and sensitivity of the model at the same time. When MCC = 1, the prediction is entirely correct. In this paper, by using the sensitivity, specificity, overall ACC and Matthew's correlation coefficient of Supplementary Si6 to evaluate model performance, it is not only

easy for researchers to understand, but also plays an essential role in the development of sequence analysis research.

2.6 Illustration of the SubMito-XGBoost workflow

For convenience, the protein submitochondrial localization prediction method proposed in this paper is called SubMito-XGBoost, with a workflow is shown in Figure 1. The detailed steps are listed as following:

Step 1: *Data preparation and feature extraction.* The protein sequences of the submitochondrial datasets M317, M983 and M495 were obtained from UniProt database. Then, *g*-Gap DC, PseAAC, ACF and Bi-gram PSSM were used to extract the initial feature set. We acquired $400 + (20 + \lambda) + (10 \times m) + 400$ dimensional feature vectors by fusing the four feature extraction methods.

Step 2: *Sampling and dimensional reduction.* The SMOTE algorithm was employed to balance the dataset. Thus, the number of protein sequences of the inner membrane, the outer membrane and the matrix are balanced. To eliminate redundant information, we selected the optimal feature subset as the input of the next step via the ReliefF algorithm (Supplementary Si7).

Step 3: *Model construction and model evaluation.* The balanced samples were integrated into XGBoost classifier, and the SubMito-XGBoost model was built up. Then we used the LOOCV on M317 and M983. To further evaluate the prediction performance, M495 was employed as an independent test set to verify the validity of the model.

3 Results and discussion

The experimental environment of all the following results was Intel (R) Core (TM) i5-3210M CPU @ 2.50 GHz 2.50 GHz 12.0GB of memory, MATLAB2014a and RStudio programming implementation.

3.1 Feature extraction method selection of optimal parameters

The submitochondrial protein sequences were extracted by the integration of *g*-Gap DC, PseAAC, ACF and Bi-gram PSSM methods. In *g*-Gap DC, the choice of the parameter *g* has a significant impact on the model prediction performance, which represents the correlation between any amino acid and *g* residue intervals. To find the highest prediction performance, the *g* parameter in *g*-Gap DC was set in the range from 0 to 10, the λ parameter was chosen from 0 to 45 for PseAAC and the*m* parameter of ACF was chosen from 1 to 45. Then, the feature extracted vectors were classified by XGBoost classifier, the prediction result was evaluated by LOOCV, and the optimal value was determined by the overall ACC on the datasets M317 and M983. The prediction ACC of the three different feature extraction methods on datasets M317 and M983 are shown in Tables 2–4.

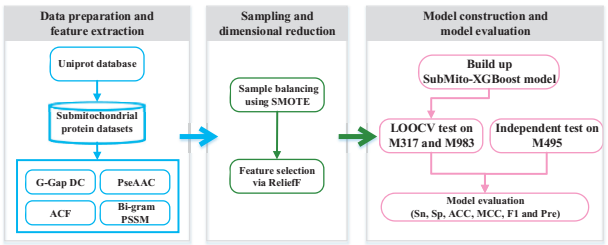

Fig. 1. Flowchart of protein submitochondrial localization prediction based on SubMito-XGBoost method. G-Gap DC represents the *g*-gap dipeptide composition, PseAAC represents the pseudo amino acid composition, ACF represents autocorrelation function, and Bi-gram PSSM represents Bi-gram position-specific scoring matrix. *g*-Gap DC, PseAAC, ACF and Bi-gram PSSM were used to extract feature information. Then the SMOTE was employed to balance samples, and the optimal feature subset was determined via ReliefF. LOOCV represents the leave-one-out-cross-validation. Finally, we used an independent test set to verify the validation of the model

It can be seen from Table 2 that when using *g*-Gap DC to extract submitochondrial protein features, when $g = 4$, the highest ACC (ACC) of dataset M317 is 75.39%. And when the *g* is 4 and 5, the highest ACC of the dataset M983 is 68.97%. To more intuitively compare the performance of the model under different *g* values, Supplementary Figure S1 shows the trend of the ACC obtained by the datasets M317 and M983. Therefore, 4-gap DC is chosen as the primary feature of the *g*-Gap DC feature extraction method. To unify the model parameters in datasets M317 and M983, the optimal λ value was 45 for PseAAC (Supplementary Figure S2), and the corresponding dimension is $20 + \lambda = 65$; For ACF, we selected $m = 10$, the trend of the ACC obtained by the datasets M317 and M983 is shown in Supplementary Figure S3, and each protein generates a $10 \times m = 100$ dimension feature vector. The specific discussion results are as Supplementary Sr1, Sr2.

We also used Bi-gram PSSM to extract the evolutionary information of submitochondria protein sequences. The sensitivity, specificity, MCC and ACC of the submitochondrial datasets M317 and M983 are shown in Supplementary Table S1. The datasets M317 and M983 have higher overall prediction ACC of 86.12 and 83.42%, respectively, which can effectively extract protein sequence information.

To extract the feature information of protein sequences more comprehensively and realize the complementarity between different feature information, this paper draws on the method of feature fusion. On the M317 and M983 datasets, the four protein-sequence feature extraction methods of *g*-Gap DC, PseAAC, ACF and Bi-gram PSSM were combined, and the XGBoost was used for classification. For comparison, Table 5 lists the predicted results of the four single feature information predictions with the predicted results obtained after the fusion. It can be seen from Table 5 that for dataset M317, using the four feature extraction methods after fusion gives an ACC of 87.38%, which is higher than the*g*-Gap DC, PseAAC, ACF and Bi-gram PSSM methods, demonstrating that the fusion of

Table 2. Prediction results of submitochondrial proteins obtained by *g*-Gap DC method

Data	*g* parameter									
	0	1	2	3	4	5	6	7	8	9
M317	73.19	71.92	71.29	72.24	75.39	71.29	69.09	69.72	70.98	70.66
M983	68.06	67.55	67.96	68.67	68.97	68.97	67.24	67.34	66.84	67.75

Table 3. Prediction results of submitochondrial proteins obtained by PseAAC method

Data	λ parameter									
	0	5	10	15	20	25	30	35	40	45
M317	76.97	80.13	82.02	80.76	82.02	81.70	82.02	84.54	83.28	82.97
M983	71.41	73.75	73.65	73.04	73.04	72.23	72.84	72.23	74.36	75.18

Table 4. Prediction results of submitochondrial proteins obtained by ACF method

Data	*m* parameter									
	1	5	10	15	20	25	30	35	40	45
M317	75.39	79.50	80.13	79.50	77.29	78.86	78.86	79.18	79.81	79.18
M983	68.57	71.01	71.41	70.30	70.50	70.50	71.11	71.01	69.18	71.52

Table 5. Prediction results of datasets M317 and M983 by four feature extraction methods

Data	Methods	Inner membrane	Matrix	Outer membrane	ACC	Feature numbers
M317	*g*-Gap DC	80.15	89.66	9.76	75.39	400
	PseAAC	84.73	95.17	34.15	82.97	65
	ACF	82.44	92.41	29.27	80.13	100
	Bi-gram PSSM	86.26	94. 48	56.10	86.12	400
	ALL	90.84	93.79	53.66	87.38	965
M983	*g*-Gap DC	96.07	14.69	11.72	68.97	400
	PseAAC	94.86	32.77	37.24	75.18	65
	ACF	90.32	37.85	26.21	71.41	100
	Bi-gram PSSM	93.80	65.54	57.93	83.42	400
	ALL	93.95	55.37	44.14	79.65	965

Note: ALL stands for fusion g-Gap DC, PseAAC, ACF and Bi-gram PSSM four feature extraction methods.

these methods can obtain more useful biometric information from the protein sequence. Supplementary Sr3 lists the overall prediction ACC of the four single feature extraction methods for the dataset M983. The fusion of four extraction methods to extract the protein sequences can obtain more useful biometric information, which is effective for improving the performance of the protein submitochondrial localization model.

3.2 Influence of SMOTE algorithm on prediction results

The submitochondrial protein datasets M317 and M983 used in this paper are highly unbalanced. Among them, the ratio of inner membrane to outer membrane in dataset M983 and M317 are 1: 5 and 1: 4, respectively. To balance the samples, this paper first used SMOTE re-sampling for the vector obtained after feature extraction, and then input it into the XGBoost for classification and used the LOOCV to test the result. For the datasets M317 and M983, the model prediction results based on the SMOTE algorithm training are shown in Supplementary Table S2. Figure 2 displays the predicted ACC rate of the three classes of proteins in the datasets M317 and M983 before and after using the SMOTE algorithm. The prediction ACC was significantly improved in both datasets and subgroups; the outer membrane protein had the most significant change. The predicted changes to matrix proteins are also significant. Therefore, the SMOTE algorithm was used in the following analysis to balance samples.

3.3 Effect of feature selection methods on predicted results

Information redundancy and unrelated features will reduce the prediction performance of the model. Six feature selection methods, i.e. Kernel principal component analysis (KPCA) (Xu *et al.*, 2010), multi-dimensional scaling (MDS) (Gorman and Sawatari, 1985), mutual information (MI) (Li *et al.*, 2017), maximum relevance and minimum redundancy (mRMR) (Khan *et al.*, 2017), maximum relevance maximum distance (MRMD) (Zou *et al.*, 2016) and ReliefF algorithm (Kira and Rendell, 1992), were compared for prediction ACC changes of different dimensions on protein submitochondrial localization (Fig. 3). The ACC of the datasets M317 and M983 were obtained under different dimensionality reduction algorithms and different dimensions, as shown in Supplementary Tables S3 and S4.

For the M317 dataset, the highest ACC of 97.65% was obtained by using the ReliefF algorithm to reduce the protein submitochondria at the 350-dimensions. Compared with the prediction results of the four feature extraction methods without dimensionality reduction, it is 10.27% higher. For dataset M983, the highest ACC of 98.94% was obtained by using the ReliefF algorithm to reduce the protein submitochondria at the 350-dimensions. Meanwhile, compared with the prediction results of the four feature extraction methods without dimensionality reduction, it is 19.29% higher. By comparing the ACC of different dimensionality reduction

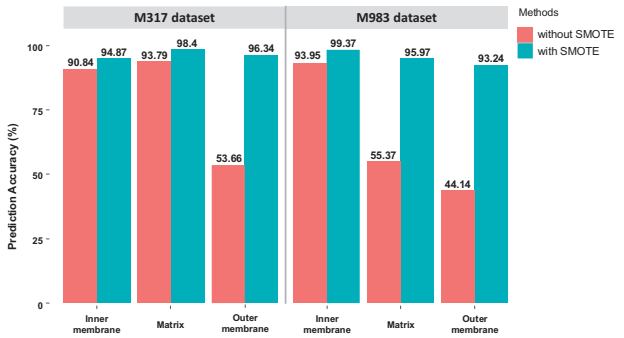

Fig. 2. Submitochondrial datasets M317 and M983 use the SMOTE algorithm to predict trends

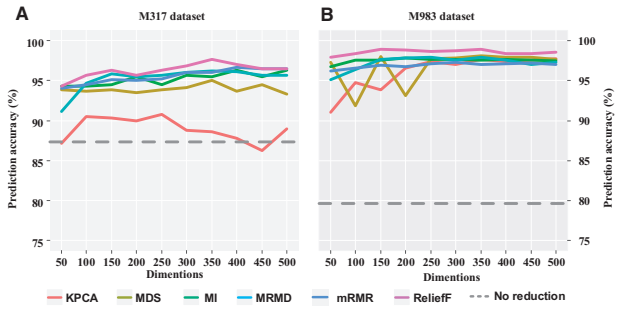

Fig. 3. The effect of six different dimensionality reduction algorithms on the prediction accuracy of datasets. (**A**) The prediction accuracy of M317 under different dimensions. (**B**) The prediction accuracy of M983 under different dimensions. The dashed lines indicate the accuracy score without dimensionality reductions

algorithms in different dimensions, we can find the optimal feature dimension location prediction using the ReliefF algorithm. It is possible that the ReliefF algorithm can weigh the features, effectively eliminating redundant variables and improving prediction ACC. Meanwhile, it also considers the correlation between samples and categories and effectively selects the optimal feature subset. Considering this, this paper uses the ReliefF algorithm to reduce dimensionality, and the optimal dimension selects 350 dimensions.

3.4 Selection of parameters of XGBoost algorithm

The XGBoost algorithm is an improvement based on the gradient boosting tree (Sheridan *et al.*, 2016), which can further improve the classification ACC of the model. In this paper, we found that tuning the hyperparameter combination through grid search on some global optimization problems became a better choice for the parameter optimization method of XGBoost algorithm. We selected variables from the influential parameters nrounds, max_depth, mid_child_weight, col_sample_bytree, gamma, eta, alpha, traversed all the parameter combinations, and determined the prediction result by the LOOCV. The variable selection takes the ACC as the model fitness function, makes the ACC value the largest, and obtains the optimal parameter combination corresponding to each dataset. The optimal combination of parameters and the results for datasets M317 and M983 are shown in Supplementary Tables S5 and S6. The XGBoost algorithm based on the optimal parameters predicted the datasets M317 and M983, which have ACC of 97.65 and 98.94%, respectively. The higher F1 values were obtained, which were 96.62 and 97.27%, respectively. This indicates that the XGBoost algorithm is ideal for positioning prediction, which can significantly improve the prediction of various locations of protein submitochondria.

3.5 Selection of classification algorithms

On the same dataset, the efficiency and ACC of the prediction method are closely related to the selection of the algorithm, and the

prediction results obtained by different classifiers are entirely different. This paper choose XGBoost to compare with other algorithms, including decision tree (DT) (Hostettler *et al.*, 2018), random forest (RF) (Taherzadeh *et al.*, 2018), support vector machine (SVM) (Li *et al.*, 2015), naïve Bayes (NB) (He *et al.*, 2017), LibD3C (Lin *et al.*, 2014) and K-nearest neighbor (KNN) (Shen *et al.*, 2006). The KNN uses the Euclidean distance with the nearest neighbor number of 40. The SVM uses radial basis kernel function, and the LibD3C, NB, RF and DT use default parameters. The prediction results obtained by the datasets M317 and M983 under different classifiers are shown in Table 6.

It can be seen from Table 6 that when using the XGBoost to predict the dataset M317, the ACC is 97.65%, which is 2.96–26.78% higher than the KNN, NB, SVM, DT, LibD3C and RF classifiers. From the perspective of the running time of the algorithm, the XGBoost is significantly shorter than the DT, RF and LibD3C classifiers. In dataset M983, the ACC of 98.94% was 2.13–26.83% higher than other classifiers. The specific prediction results of dataset M317 and M983 are shown in Supplementary Sr4, Tables S7 and S8. To more intuitively compare the predictive performance of the seven classifiers, the variation of the overall prediction ACC in the datasets M317 and M983 under different classification algorithms is shown in Supplementary Figure S4.

In summary, in the protein submitochondrial localization prediction, the two datasets used the XGBoost classifier to obtain the highest overall ACC, which is significantly higher than the other six classifiers. This is because the XGBoost introduces regularization terms, making the model simple with good generalization performance.

3.6 Performance of the prediction model

In this paper, the datasets M317 and M983 were extracted by the Bi-gram PSSM, PseAAC, *g*-Gap DC and ACF methods to generate the 965-dimensional feature vector, and then the sample data was processed by SMOTE to make the data class balance. ReliefF was used to select the optimal feature vector. Finally, the XGBoost was input to the submitochondrial protein localization, and the prediction result was tested by the LOOCV method. To evaluate the generalization performance of the model, the independent dataset M495 was used to evaluate the predictive ability of the SubMito-XGBoost model, with the optimal parameters of $g = 4, \lambda = 45, m = 10$ and the optimal dimension is 350. The specific protein submitochondrial prediction results are shown in Table 7.

According to the result in Table 7, the ACC of dataset M317, M983 and M495 are 97.65, 98.94 and 94.83%, respectively. The experimental results show that the SubMito-XGBoost model can accurately predict the location of the inner membrane, matrix and outer membrane proteins in the two datasets. The sensitivity of the intermembrane protein, matrix and outer membrane proteins were 96.67, 93.80 and 94.54%, respectively. The above results indicate that the SubMito-XGBoost method proposed in this paper has excellent performance and can effectively predict the locations of protein submitochondria.

Table 6. Prediction results of protein submitochondrial dataset M317 and M983 under different algorithms

Datasets	Classifiers	LOOCV				
		Sn (%)	Sp (%)	MCC	ACC (%)	Running time (Min)
M317	KNN	79.20	91.27	0.6844	79.53	10'16
	NB	70.75	83.64	0.5637	70.87	18'53
	SVM	83.89	91.27	0.7517	83.66	140'12
	DT	91.49	95.55	0.8705	91.54	28'37
	LibD3C	93.45	93.28	0.9020	93.50	1003'17
	RF	94.65	97.21	0.9205	94.69	185'03
	XGBoost	96.54	98.27	0.9492	97.65	19'42
M983	KNN	77.97	80.11	0.6789	78.49	30'56
	NB	73.31	77.53	0.6605	72.11	58'11
	SVM	82.76	89.89	0.7989	87.26	210'29
	DT	93.00	96.38	0.8986	94.30	91'07
	LibD3C	94.12	95.24	0.9433	96.81	2811'48
	RF	95.15	96.28	0.9506	96.68	385'52
	XGBoost	98.54	97.27	0.9492	98.94	78'08

3.7 Comparison with other methods

In order to objectively evaluate the predictions of the SubMito-XGBoost method, Figure 4 compares the results of this method with other prediction methods on datasets M317, M983 and M495. The specific prediction results of different methods can be seen in Supplementary Table S9.

It can be seen from Figure 4 that for the dataset M317, the ACC of SubMito-XGBoost is 97.7%, which is higher than that obtained by other prediction methods. The SubMito proposed by Jiao *et al.* (Jiao and Du, 2017) is 12.5% higher. Compared with Nanni *et al.* (Nanni and Lumini, 2008), the ACC of the GP-Loc is 8.7% higher. The SubMito is only 51.2%, which is far lower than the SubMito-XGBoost. The results show that the SubMito-XGBoost can reduce the prediction error caused by sample imbalance, the ACC is significantly higher than other existing models, and better prediction results are obtained on the dataset M317. The prediction of the dataset M983 showed ACC of 98.9%. The SubMito-XGBoost is 9.9% higher the SubMito-PSPCP proposed by Du *et al.* (Du and Yu, 2013). After the model was processed by SMOTE method, the sensitivity was increased by 17.9% and 21.6%. Compared with the method proposed by Ahmad *et al.* (Ahmad *et al.*, 2016), the prediction of intermembrane proteins was increased by 12.0%.

For the independent test set M495, the SubMito-XGBoost yielded ACC of 94.8%, which is higher than the TetraMito constructed by Lin *et al.* (2013). In summary, the SubMito-XGBoost method has good prediction results on all three submitochondrial datasets, which sufficiently indicates that the prediction method constructed in this paper can become a dominant protein submitochondrial localization prediction tool.

3.8 Performance of SubMito-XGBoost based plant and no-plant datasets

Chloroplast proteins are likely to use targeting mechanisms with similar mitochondrial protein features because mitochondria, chloroplasts and bacteria have many similarities in DNA molecular structure and protein synthesis properties. Therefore, when predicting protein submitochondrial localization, it is important to study chloroplasts and perform separate studies of organisms without chloroplasts. In order to test the generalization performance of the SubMito-XGBoost model, the original protein sub-mitochondrial dataset was divided into plant protein submitochondrial dataset and non-plant protein submitochondrial dataset for localization prediction based on species attribute. Among them, there are 172 submitochondrial samples of plant proteins and 1, 623 non-plant samples, with specific data structure shown in Supplementary Table S10.

Table 7. Prediction results for submitochondrial of datasets M317, M983 and M495

Datasets	Structure class	Sn (%)	Sp (%)	MCC	ACC (%)
M317	Inner membrane	95.36	99.14	0.9524	97.65
	Matrix	97.93	97.73	0.9539	
	Outer membrane	96.34	97.94	0.9414	
M983	Inner membrane	99.06	96.16	0.9559	98.94
	Matrix	97.57	99.52	0.9595	
	Outer membrane	98.99	99.50	0.9604	
M495	Inner membrane	96.67	91.22	0.9026	94.83
	Matrix	93.80	98.63	0.9318	
	Outer membrane	94.54	97.02	0.9191	

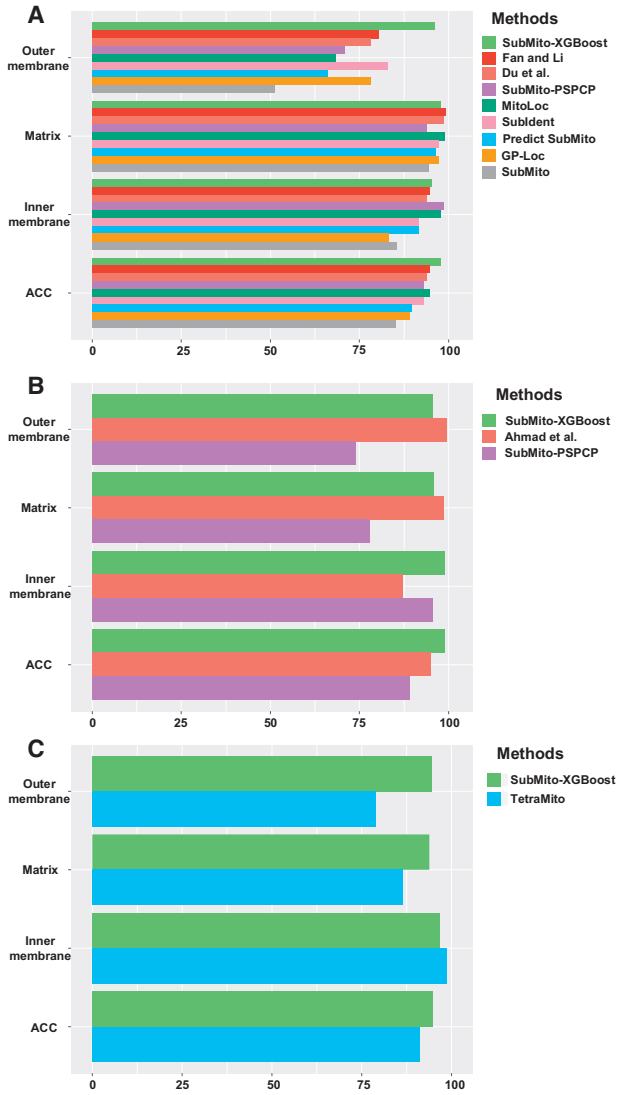

Fig. 4. Comparison of prediction accuracy among different methods in submitochondrial datasets. (**A**) Prediction results of different methods on M317. (**B**) Prediction results of different methods on M983. (**C**) Prediction results of different methods on M495

In the plant protein submitochondrial dataset, the accuracy of the SubMito-XGBoost method was 96.09%, and the accuracy of the intermembrane, matrix and outer membrane proteins were 98.89, 94.29 and 94.35%, respectively. In the non-plant protein submitochondrial dataset, the accuracy of SubMito-XGBoost was 90.94%, and the accuracy of the intermembrane, matrix and outer membrane proteins were 97.21, 83.57 and 87.88%, respectively.

This paper chose XGBoost algorithm for the comparison with other machine learning algorithms, namely decision tree algorithm, random forest, support vector machine, naive Bayes, LibD3C and K-nearest neighbor. The overall prediction accuracy and prediction accuracy of each classifier were obtained, and all the predicted results for plant and non-plant protein submitochondria under different algorithms are shown in Supplementary Tables S11 and S12. In the plant and non-plant protein submitochondrial datasets, the SubMito-XGBoost method showed advanced performance. Specifically, the overall prediction accuracy of plant protein submitochondria was 1.06–12.24% higher than that of KNN, NB, SVM, DT, LibD3C and RF algorithms. The overall prediction accuracy of non-plant protein submitochondria was 1.31–18.92% higher than that predicted by KNN, NB, SVM, DT, LibD3C and RF. In summary, the method of this paper can effectively predict plant and non-plant protein submitochondria.

4 Conclusion

In the big data era, the sequence data accumulated in protein databases has grown exponentially. How to use machine learning methods to accurately predict the locations of protein submitochondria has become a challenging task in bioinformatics and proteomics research. This paper proposes a new protein submitochondrial localization prediction method, SubMito-XGBoost, to fuse the composition information, physicochemical properties, and evolutionary information of protein sequences. It also avoids the model overfitting problem and improves the model generalization ability using the SMOTE algorithm. Meanwhile, the ReliefF algorithm takes into account the correlation between classes and features, and weights the protein sequence features and puts them into the model, making SubMito-XGBoost very effective to process multi-class high-dimensional data. When positioning multiple classes of protein submitochondria, the ACC using XGBoost is improved by combining various weak classifiers into a strong classifier, and multi-threading is supported to improve operational efficiency. The experimental results showed that SubMito-XGBoost can not only effectively improve the prediction ACC of protein submitochondrial localization, but also accurately predict the location of the intermembrane, matrix and outer membrane proteins. We believe that SubMito-XGBoost can be used to predict protein submitochondrial localization and other substructure localization. Although SubMito-XGBoost improves the ACC of protein submitochondrial prediction to a certain extent, there is still much room for improvement in prediction ACC and algorithm efficiency. Additionally, the proposed method is only applicable to the prediction of protein submitochondrial localization and cannot accurately identify non-mitochondrial proteins. In the future, we will try more feature extraction and feature selection methods to improve the performance of SubMito-XGBoost, build large-scale benchmarking datasets, and implement deep learning method to predict protein submitochondrial localization.

Acknowledgements

We thank anonymous reviewers for valuable suggestions and comments. We also thank Ms Marlena Merling for her help in polishing the language of this manuscript.

Funding

This work was supported by the National Nature Science Foundation of China (No. 61863010), the Key Research and Development Program of Shandong Province of China (2019GGX101001), the Natural Science Foundation of Shandong Province of China (Nos. ZR2017MA014 and ZR2018MC007), the Project of Shandong Province Higher Educational Science and Technology Program (No. J17KA159) and the Scientific Research Fund of Hunan Provincial Key Laboratory of Mathematical Modelling and Analysis in Engineering (No. 2018MMAEZD10). This work used the Extreme Science and Engineering Discovery Environment, which is supported by the National Science Foundation (No. ACI-1548562).

Conflict of Interest: none declared.

References

Ahmad,K. *et al.* (2016) Prediction of protein submitochondrial locations by incorporating dipeptide composition into chou's general pseudo amino acid composition. *J. Membr. Biol.*, **249**, 1–12.

Babajide,M.I. and Saeed,F. (2016) Bioactive molecule prediction using extreme gradient boosting. *Molecules*, **21**, 983.

Bu,W.S. *et al.* (2010) Prediction of protein (domain) structural classes based on amino-acid index. *FEBS J.*, **266**, 1043–1049.

Burbulla,L.F. *et al.* (2017) Dopamine oxidation mediates mitochondrial and lysosomal dysfunction in Parkinson's disease. *Science*, **357**, 1255–1261.

Chawla,N.V. *et al.* (2002) SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, **16**, 321–357.

Chen,T. and Guestrin,C. (2016) XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Chen,X. *et al*. (2018) EGBMMDA: extreme gradient boosting machine for miRNA-disease association prediction. *Cell Death Dis*., **9**, 3.
Chen,Y.-L. *et al*. (2012) Using increment of diversity to predict mitochondrial proteins of malaria parasite: integrating pseudo-amino acid composition and structural alphabet. *Amino Acids*, **42**, 1309–1316.
Chou,K.C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **21**, 10–19.
Chou,K.C. and Shen,H.B. (2006) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. *J. Proteome Res*., **5**, 1888–1897.
Ding,H. and Li,D. (2015) Identification of mitochondrial proteins of malaria parasite using analysis of variance. *Amino Acids*, **47**, 329–333.
Du,P. and Li,Y. (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. *BMC Bioinformatics*, **7**, 518-518.
Du,P. and Yu,Y. (2013) SubMito-PSPCP: predicting protein submitochondrial locations by hybridizing positional specific physicochemical properties with pseudoamino acid compositions. *Biomed Res. Int*., **2013**, 1.
Fariselli,P. *et al*. (2005) A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinformatics*, **6**, S12.
Gerbitz,K.D. *et al*. (1996) Mitochondria and diabetes. Genetic, biochemical, and clinical implications of the cellular energy circuit. *Diabetes*, **45**, 113.
Gorman,R. and Sawatari,T. (1985) The use of multidimensional perceptual models in the selection of sonar echo features. *J. Acoust. Soc. Am*., **77**, 1178–1184.
He,B. *et al*. (2017) NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers. *Bioinformatics*, **33**, 2296–2306.
Hostettler,I.C. *et al*. (2018) Decision tree analysis in subarachnoid hemorrhage: prediction of outcome parameters during the course of aneurysmal subarachnoid hemorrhage using decision tree analysis. *J. Neurosurg*., **1**, 1–12.
Jiao,Y.S. and Du,P.F. (2017) Predicting protein submitochondrial locations by incorporating the positional-specific physicochemical properties into Chou's general pseudo-amino acid compositions. *J. Theor. Biol*., **416**, 81–87.
Khan,M. *et al*. (2017) Bi-PSSM: position specific scoring matrix based intelligent Computational model for identification of mycobacterial membrane proteins. *J. Theor. Biol*., **435**, 116–124.
Kira,K. and Rendell,L.A. (1992) The feature selection problem: traditional methods and a new algorithm. In: *Tenth National Conference on Artificial Intelligence*. p. 129–134.
Li,F. *et al*. (2017) Granular multi-label feature selection based on mutual information. *Pattern Recogn*., **67**, 410–423.
Li,L. *et al*. (2015) Protein submitochondrial localization from integrated sequence representation and SVM-based backward feature extraction. *Mol. Biosyst*., **11**, 170–177.
Lin,C. *et al*. (2014) LibD3C: ensemble classifiers with a clustering and dynamic selection strategy. *Neurocomputing*, **123**, 424–435.
Lin,H. *et al*. (2013) Using over-represented tetrapeptides to predict protein submitochondria locations. *Acta Biotheor*., **61**, 259–268.
Mei,S. (2012) Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization. *J. Theor. Biol*., **293**, 121–130.
Nanni,L. and Lumini,A. (2008) Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids*, **34**, 653–660.
Qiu,W.Y. *et al*. (2018) Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition. *J. Theor. Biol*., **450**, 86–103.
Shen,H. *et al*. (2006) Fuzzy KNN for predicting membrane protein types from pseudo-amino acid composition. *J. Theor. Biol*., **240**, 9–13.
Shen,H.B. and Chou,K.C. (2008) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem*., **373**, 386–388.
Sheridan,R.P. *et al*. (2016) Extreme gradient boosting as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model*., **56**, 2353–2360.
Shi,S. *et al*. (2011) Identify submitochondria and subchloroplast locations with pseudo amino acid composition: approach from the strategy of discrete wavelet transform feature extraction. *BBA Mol. Cell Res*., **1813**, 424–430.
Silvério-Machado,R. *et al*. (2015) Retrieval of Enterobacteriaceae drug targets using singular value decomposition. *Bioinformatics*, **31**, 1267–1273.
Taherzadeh,G. *et al*. (2018) Structure-based prediction of protein-peptide binding regions using Random Forest. *Bioinformatics*, **34**, 477.
UniProt,C. (2015) UniProt: a hub for protein information. *Nucleic Acids Res*., **43**, 204–212.
Wang,X.Y. *et al*. (2018) Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics*, doi:10.1093/bioinformatics/bty995.
Wen,P.P. *et al*. (2016) Accurate in silico prediction of species-specific methylation sites based on information gain feature optimization. *Bioinformatics*, **32**, 3107–3115.
Xu,Y. *et al*. (2010) Producing computationally efficient KPCA-based feature extraction for classification problems. *Electr. Lett*., **46**, 452–453.
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.
Zakeri,P. *et al*. (2011) Prediction of protein submitochondria locations based on data fusion of various features of sequences. *J. Theor. Biol*., **269**, 208–216.
Zeng,Y.H. *et al*. (2009) Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J. Theor. Biol*., **259**, 366–372.
Zhao,Z. *et al*. (2018) Imbalance learning for the prediction of N6-Methylation sites in mRNAs. *BMC Genomics*, **19**, 574.
Zou,Q. *et al*. (2016) A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing*, **173**, 346–354.