Analytical Biochemistry 609 (2020) 113903

Contents lists available at ScienceDirect

Analytical Biochemistry

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

Prediction of protein crotonylation sites through LightGBM classifier based on SMOTE and elastic net

Yaning Liu [a,b,1], Zhaomin Yu [a,b,1], Cheng Chen [a,b], Yu Han [a,b], Bin Yu [a,b,c,*]

[a] College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao, 266061, China
[b] Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao, 266061, China
[c] School of Life Sciences, University of Science and Technology of China, Hefei, 230027, China

ARTICLE INFO

Keywords:
Crotonylation sites
Multi-feature fusion
SMOTE
Elastic net
LightGBM

ABSTRACT

Lysine crotonylation is an important protein post-translational modification, which plays an important role in the process of chromosome organization and nucleic acid metabolism. Recognition of crotonylation sites is important to understand the function and mechanism of proteins. Traditional experimental methods are time-consuming and expensive, and can't predict crotonylation sites quickly and accurately. Therefore, this paper proposes a novel crotonylation sites prediction method called LightGBM-CroSite. First, binary encoding (BE), position weight amino acid composition (PWAA), encoding based on grouped weight (EBGW), k nearest neighbors (KNN), pseudo-position specific scoring matrix (PsePSSM) are used to extract features of protein sequences and obtain the original feature space. Second, the elastic net is used to remove redundant information and select the optimal feature subset. Third, the synthetic minority oversampling technique (SMOTE) is used to balance the samples. Finally, the balanced feature vectors are input into LightGBM to predict the crotonylation sites. According to the result of jackknife test, the Accuracy (ACC), Matthew's correlation coefficient (MCC) and area under ROC curve (AUC) are 98.99%, 0.9798 and 0.9996, respectively. Compared with other state-of-the-art methods, the results show that our method has a better model performance on the crotonylation sites prediction. The source code and all datasets are available at https://github.com/QUST-AIBBDRC/LightGBM-CroSite/.

1. Introduction

Protein post-translational modification (PTM) is an important chemical process. Covalent modification to control the function or activity of proteins is a feature of the regulation of most physiological activities such as signal transduction, proliferation, development and differentiation [1]. At present, more than 400 post-translational modifications have been found, the most common of which are phosphorylation [2], methylation [3], acetylation [4] and glycosylation [5]. lysine crotonylation, a newly discovered post-translational modification, plays an important role in many processes such as chromosome organization, nucleic acid metabolism and gene expression [6]. This has become a research focus in the field of post-translational modification. The identification of crotonylation sites is an important prerequisite for the study of crotonylation modification. Methods to predict crotonylation sites require four processes. First, the features of such sites must be extracted from the protein sequences. Second, the most informative subset of features must be selected. Third, the effect of imbalance between the number of positive and negative samples in the dataset must be mitigated. Fourth, a powerful classifier for sites prediction must be selected and trained.

Extracting features of protein sequences is the basic step to construct the crotonylation sites prediction model. Features of protein sequences are mainly composed of sequence feature, physicochemical feature, evolutionary information and structural feature. Multi-feature fusion makes the information more comprehensive and makes the prediction performance more accurate. Liu et al. [7] incorporated four kinds of amino acid pairwise coupling information into the general PseAAC to extract the physicochemical feature of protein sequences in the phosphorylation sites prediction. Wang et al. [8] used amino acid composition, dipeptide composition, encoding based on grouped weight, k nearest neighbors, position-specific amino acid propensity, position-weighted amino acid composition and pseudo-position specific scoring matrix to extract sequence feature, evolutionary information,

* Corresponding author. College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao, 266061, China.
E-mail address: yubin@qust.edu.cn (B. Yu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ab.2020.113903
Received 11 January 2020; Received in revised form 27 July 2020; Accepted 5 August 2020
Available online 15 August 2020
0003-2697/

and physicochemical feature of protein sequences. Xu et al. [9] constructed S-Nitrosylation sites predictor called iSNO-PseAAC, which incorporated position specific amino acid propensity (PSAAP) into pseudo amino acid composition (PseAAC) to extract the evolutionary information and physicochemical feature of proteins sequences. Wang et al. [10] used amino acid composition, binary encoding, encoding based on grouped weight, k nearest neighbors and pseudo-position specific scoring matrix to extract features of protein sequences, and the performance of malonylation sites prediction was improved significantly.

Multi-feature fusion will produce redundant and noise information and increase the computational expense of the model. Feature selection [11] is an indispensable step in the model prediction process, and it can improve the model performance through removing redundant information and retaining important features. Cui et al. [12] used Lasso method to remove redundant information and retain effective features in the protein ubiquitination sites prediction. Wang et al. [13] proposed an evolutionary screening algorithm (ESA) in the protein phosphorylation sites prediction, which simulated natural selection to screen the features. Cao et al. [14] proposed the two-step feature optimization method to predict the phosphorylation sites. By selecting prominent features, the prediction result of seven fungal phosphorylation sites was improved significantly.

In the sites prediction, the imbalance of the dataset will lead to the deviation of prediction results. It is necessary to choose a suitable algorithm to deal with the imbalance problem of dataset. He et al. [15] proposed a novel adaptive synthetic (ADASYN) sampling approach, which can adaptively generated a small amount of data based on the weighted distribution. Hong et al. [16] proposed a kernel classifier construction algorithm, which used orthogonal forward selection (OFS) to balance the datasets. Wang et al. [17] used synthetic minority oversampling technique (SMOTE) to solve the imbalance in the Protein-protein interaction sites prediction and the accuracy of training dataset and independent test dataset were improved effectively.

The machine learning methods have good prediction performances and they are widely used in PTM field, the most common of which are random forest (RF) [18,19], support vector machine (SVM) [20,21], logistic regression (LR) [22,23] and Naive Bayes (NB) [24,25]. RF is the ensemble classifier based on decision tree. SVM is an algorithm based on statistical learning theory and LR is generalized linear regression analysis model. NB is a method based on Bayes' theorem and independent hypothesis of characteristic conditions. Jia et al. [26] used random forest classifier to construct the predictor pSuc-Lys, and the accuracy, specificity, and AUC of the succinylation sites prediction reached 90.83%, 95.97%, and 0.9325, respectively. Ju et al. [27] used the biased support vector machine algorithm to construct the glutaryldiacylation sites predictor called GlutPred. Yu et al. [5] proposed the predictor PredGly based on support vector machine, and the ACC, AUC, Sn, Sp and MCC on the dataset reached 89.25%, 83.16%, 88.67%, 89.85% and 78.51%, respectively. Li et al. [23] proposed the predictor Quokka employing logistic regression, and the prediction results of phosphorylation sites was improved.

At present, there are several methods of the crotonylation sites prediction. Huang et al. [28] proposed the predictor CrotPred, which used a discrete hidden markov model to predict the histone crotonylation sites. Qiu et al. [29] used position weight amino acid composition and support vector machine to predict crotonylation sites. In 2016, Qiu et al. [30] incorporated the sequence-coupled effects into the general PseAAC and used ensemble random forest algorithm to identify crotonylation sites. In 2018, Qiu et al. [31] incorporated five tiers of amino acid pairwise couplings into the general PsePAAC and used ensemble random forest algorithm to predict crotonylation sites. Ju et al. [32]used composition of k-spaced amino acid pairs feature and support vector machine to improve the results of crotonylation sites prediction. Although the existing methods have achieved good results, there are still some shortcomings. First, the extracted protein features are not comprehensive enough. Second, the bad impact on the algorithm of the data imbalance problem is not considered. Third, there is still room for improvement in the accuracy of crotonylation sites prediction.

Based on the three issues mentioned above, this paper proposes a novel method to predict crotonylation sites called LightGBM-CroSite. Considering the sequence-based feature, physicochemical property-based feature and evolutionary-derived feature of protein sequences, we use feature extraction methods of BE, PWAA, EBGW, KNN and PsePSSM to extract features of protein sequences and fuse the five features to obtain the original feature space. In order to eliminate the redundant and noise information in original feature space, we choose elastic net to reduce the dimension of original feature space. Then, SMOTE is used to reduce the influence of the data imbalance problem. At last, we choose the LightGBM classifier with good performance to predict the crotonylation sites. The ACC, MCC and AUC of LightGBM-CroSite are 98.99%, 0.9798 and 0.9996 respectively. Compared with other crotonylation sites prediction methods, LightGBM-CroSite method significantly improves the prediction performance of the crotonylation sites.

2. Materials and methods

2.1. Datasets

The objective and representative dataset is an important prerequisite to construct the crotonylation sites prediction model. In this paper, Qiu et al. [29] established the crotonylation sites dataset, which was collected from the UniProt database, consisting of 343 lysine crotonylation sites from 101 protein sequences. The protein sequence with crotonylation sites verified by experiments is taken as positive sample, and the protein sequence without crotonylation sites is taken as negative sample. In order to avoid the negative impact of sequence homology and redundancy on the model performance, the homologous and redundant sequences in the dataset are screened out. Finally, the dataset consists of 159 positive samples and 847 negative samples, which can be downloaded from http://lin.uestc.edu.cn/database/Kcr/data.docx.

2.2. Feature extraction

The feature extraction of protein sequences is the basic step in the crotonylation sites prediction, which will affect the prediction model effectiveness. In this paper, binary encoding (BE) [21], position weight amino acid composition (PWAA) [33], encoding based on grouped weight (EBGW) [34], k nearest neighbors (KNN) [35] and pseudo-position specific scoring matrix (PsePSSM) [36] are selected to extract protein sequence information for crotonylation sites prediction.

2.2.1. Binary encoding

In order to reflect the amino acid residues position and composition information of the sequence, binary encoding (BE) [21] is used to encode amino acids. 20 common amino acids can be converted into 21-dimensional vectors consisting of 0 and 1 element according to the order of ACDEFGHIKLMNPQRSTVWYX. For example, alanine A is converted into the vector (10000000000000000000), cysteine C is converted into the vector (01000000000000000000), tyrosine Y is converted into the vector (00000000000000000100), and other unusual amino acids are uniformly converted into the vector (00000000000000000001). A protein sequence of length N is encoded as a 21 × N-dimensional feature vector.

2.2.2. Position weight amino acid composition

Position weight amino acid composition (PWAA) [33] is a common feature extraction method, which mainly extracts the position information of amino acid residues in protein sequences. For the protein sequence with the length $2n+1$, a_i $(i = 1, 2, \ldots, 20)$ is the amino acid residues and the position information of a_i can be expressed as follows:

$$C_i = \frac{1}{n(n+1)} \sum_{j=1}^{n} x_{i,j} \left(j + \frac{j}{n} \right) \quad j = -n, \ldots, n \tag{1}$$

where n represents the number of upstream or downstream residues of the protein sequence sample. If a_i is the j-th residue of the protein sequence sample, $x_{i,j} = 1$, otherwise $x_{i,j} = 0$.

2.2.3. Encoding based on grouped weight

The encoding based on grouped weight (EBGW) [34] is employed to extract protein sequences features. Since the side chain groups have different characteristics, amino acids have different physical and chemical properties such as hydrophobicity and charged character. According to these properties, the 20 common amino acids are divided into the following four different types:

Neutral and non-polarity amino acids. $C_1 = \{G, A, V, L, I, M, P, F, W\}$

Neutral and polarity amino acids. $C_2 = \{Q, N, S, T, Y, C\}$

Acidic amino acids. $C_3 = \{D, E\}$

Basic amino acids. $C_4 = \{H, K, R\}$

Four types of amino acids can be combined into three groups of $D_1 = \{C_1, C_2\} vs \ D_2 = \{C_3, C_4\}$, $D_1 = \{C_1, C_3\} \ vs \ D_2 = \{C_2, C_4\}$ and $D_1 = \{C_1, C_4\} \ vs \ D_2 = \{C_2, C_3\}$.

Given a protein sequence $P = P_1 P_2 P_3 \ldots P_N$, the amino acid $P_i (i = 1, 2, \ldots, N)$ is encoded according to the following rules:

$$f(P_i) = \begin{cases} 1 & P_i \in D_1 \\ 0 & P_i \in D_2 \end{cases} \tag{2}$$

Three binary sequences $H_1(p)$, $H_2(p)$, $H_3(p)$ of length N are obtained. For each binary sequence $H_i(p)$, it is divided into L subsequences whose length increases in turn. By calculating the frequency of 1 in each subsequence, we can get an L-dimensional vector. The formula is as follows:

$$Length_k = \left\lceil \frac{kN}{L} \right\rceil (k = 1, 2, \ldots, L) \tag{3}$$

$$X_i(k) = \frac{sum(k)}{Length_k} \quad (k = 1, 2, \ldots, L) \tag{4}$$

where $sum(k)$ represents the number of 1 in the k-th subsequences. The protein sequence is finally transformed into $(X_1(1), X_1(2), \ldots, X_1(L), X_2(1), X_2(2), \ldots, X_2(L), X_3(1), X_3(2), \ldots, X_3(L))$.

2.2.4. K nearest neighbors

The K nearest neighbor (KNN) method [35] extracts features of protein sequences based on the similarity of local sequences. Firstly, we calculate the distance between the test site and all known sites in the dataset. For two local protein sequences $S_1 = (S_1(1), S_1(2), \ldots, S_1(N))$ and $S_2 = (S_2(1), S_2(2), \ldots, S_2(N))$, the distance between two local protein sequences is defined as

$$Dist(S_1, S_2) = 1 - \frac{\sum_{j=1}^{L} sim(S_1(j), S_2(j))}{N} (j = 1, 2, \ldots, N) \tag{5}$$

where N represents the number of residues in the protein sequence fragment. The amino acid similarity score matrix sim is defined as

$$sim(S_1(j), S_2(j)) = \frac{Matrix(S_1(j), S_2(j)) - \min\{Matrix\}}{\max\{Matrix\} - \min\{Matrix\}} (j = 1, 2, \ldots, N) \tag{6}$$

where $S_i(j)$ $(i = 1, 2)$ represents the amino acid residue of the protein sequence segment, $Matrix$ is the BLOSUM62 substitution matrix, and $\max/\min\{Matrix\}$ represents the maximum/minimum value in the BLOSUM62 substitution matrix. The extraction process of KNN is as follows: (a) sort all the calculated distance values in ascending order; (b) select the k nearest neighbors of the test site; and (c) calculate the percentage of positive neighbors in k nearest neighbors. In this paper, we select k as 2, 4, 8, 16, 32, 64, 128, 256 to extract multiple features for sites prediction.

2.2.5. Pseudo-position specific scoring matrix

Pseudo-position-specific scoring matrix (PsePSSM) [36] is an important method proposed by Jones et al., which is used to extract the evolutionary information of protein sequences. Given a protein sequence of length N, we compare it with Swiss-prot non-redundant database by iterative PSI-BLAST search method [37], and get its corresponding PSSM matrix with dimension $N \times 20$.Then the $_{\text{PSSM matrix}}$ is processed as follows:

(a) Filter the row vectors of the matrix. The maximum value MAX_{P_i} of each row is compared with the fixed value T, and the row vector whose maximum value less than T is deleted.

(b) Standardize the resulting matrix, and each element in the matrix is scaled into a number that is between 0 and 1 according to the following formula:

$$f(x) = \frac{1}{(1 + e^{-x})} \tag{7}$$

(c) The column vectors of the normalized matrix are averaged, and the PSSM-AAC model is obtained as follows:

$$P_{PSSM-AAC} = [\overline{x_1}, \overline{x_2}, \ldots, \overline{x_{20}}] \tag{8}$$

where, $\overline{x_j} (j = 1, 2, \ldots, 20)$ represents the component of amino acid type j in PSSM.

(d) Calculate the ξth-order correlation factor of amino acid type j.

$$\theta_j^{\xi} = \frac{1}{L - \xi} \sum_{i=1}^{L-\xi} \left(P_{i,j} - P_{i+\xi,j} \right)^2 (j = 1, 2, \ldots, 20; \xi \neq 0; \xi < L) \tag{9}$$

Finally, The protein sequence is transformed into a feature vector $P_{PsePSSM} = (x_1, x_2, \ldots, x_{20}, \theta_1^1, \theta_2^1, \ldots, \theta_{20}^1, \ldots, \theta_1^{\xi}, \theta_2^{\xi}, \ldots, \theta_{20}^{\xi})$.

2.3. Elastic net

Lasso is an effective feature selection method proposed by Tibshirani [38]. By introducing ℓ_1-norm into simple linear equation, Lasso method can continuously shrink and automatically select variables at the same time, but it can't deal with the prediction of correlation variables. To solve this problem, Zou and Hastie [39] proposed elastic net method based on Lasso. By introducing both ℓ_1-norm and ℓ_2-norm into the simple linear regression model, elastic net method [40] not only retains the advantages $_{\text{of Lasso, but also}}$ predicts the relevant variables. Given a dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, the general linear regression model can be as follows:

$$\min_{\omega} \sum_{i=1}^{m} \left(y_i - \omega^T x_i \right)^2 \tag{10}$$

The elastic net method which contains both ℓ_1-norm and ℓ_2-norm is defined as follows:

$$\min \frac{1}{2 \times n} \|y - X\omega\|_2^2 - \alpha \times \beta \|\omega\|_1 + \frac{1}{2} \alpha \times (1 - \beta) \|\omega\|_2^2 \tag{11}$$

where $\|\omega\|_1$ is ℓ_1-norm, $\|\omega\|_2^2$ is ℓ_2-norm.

2.4. SMOTE

Synthetic minority oversampling technique (SMOTE) [17,41] is a method proposed by Chawla et al. It can reduce the imbalance between different label samples. In the SMOTE, the minority samples are randomly oversampled and the majority samples are randomly undersampled. For each sample x in the minority samples, we firstly select the k nearest neighbors of it. If the sampling rate is n, then we randomly select n samples from the k nearest neighbors. A new sample can be

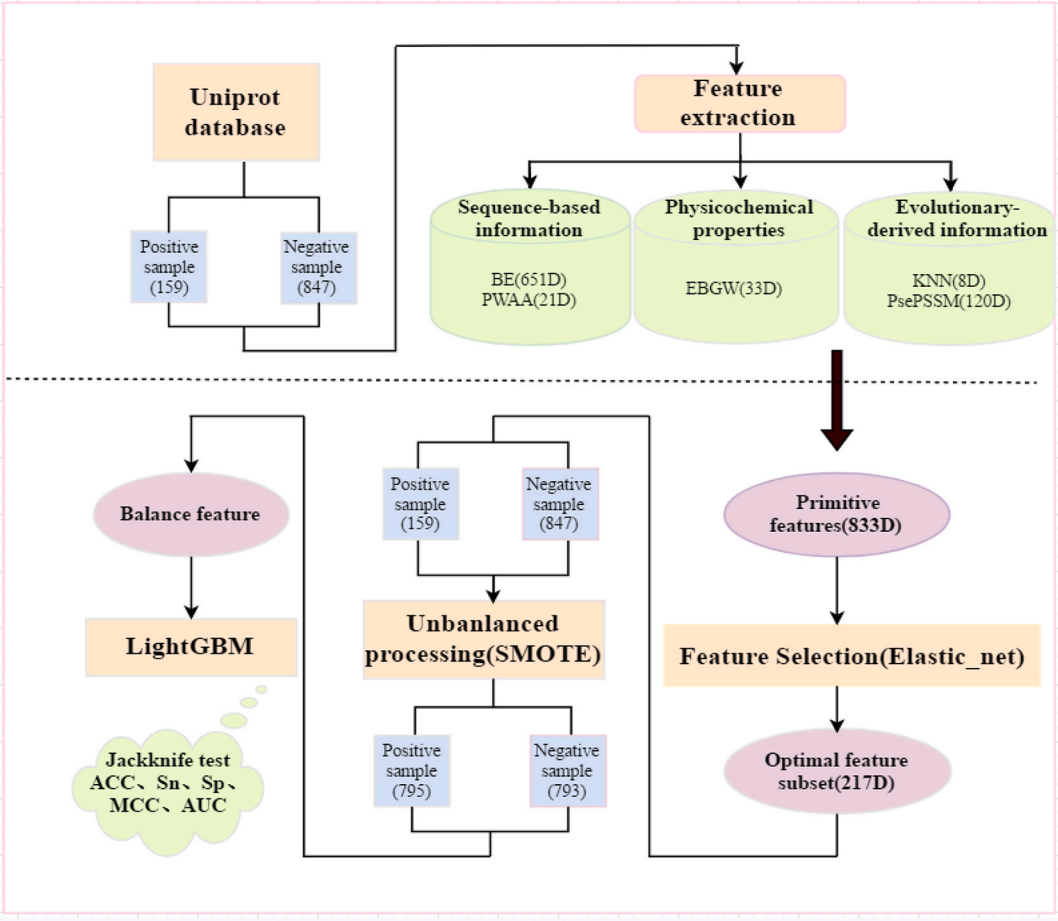

Fig. 1. Pipeline of LightGBM-CroSite for crotonylation sites prediction.

constructed by each sample $x_i(i = 1, 2, \ldots, n)$ and the original sample x as follows:

$$x_{new} = x + rand(0, 1) \times (x_i - x) \tag{12}$$

where $rand(0, 1)$ represents a random number in the interval (0,1). After merging the new samples with the original few samples, a new balanced dataset is generated.

2.5. LightGBM

Gradient boosting decision tree (GBDT) is an iterative algorithm based on decision tree, which has been widely used in life science, computer science and other fields. Given a training dataset $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where x represents samples data, y represents class labels, and $F(x)$ represents the estimation function. The loss function is defined as follows:

$$F' = \underset{F}{\operatorname{argmin}} E_{x,y}[L(y, F(x))] \tag{13}$$

After each gradient enhancement iteration, a negative gradient $\{g_1, g_2, \ldots, g_n\}$ of the loss function on the model output is obtained. Then the decision tree model uses the feature with the largest information gain to partition each node. The information gain in GBDT is defined as follows:

$$V_{j|O}\left(d\right) = \frac{1}{n_o}\left(\frac{\left(\sum_{\{x_i \in O : x_{ij} \le d\}} g_i\right)^2}{n^j_{l|O}\left(d\right)} + \frac{\left(\sum_{\{x_i \in O : x_{ij} > d\}} g_i\right)^2}{n^j_{r|O}\left(d\right)}\right) \tag{14}$$

where, $n_o = \sum I[x_i \in O]$, $n^j_{l|O}(d) = \sum I[x_i \in O : x_{ij} \le d]$, $n^j_{r|O}(d) = \sum I[x_i \in O : x_{ij} > d]$.

Although GBDT has high accuracy and efficiency in low dimensional space. With the increase of dimensions, its advantages no longer exist. To solve this problem, a new method called LightGBM [42] was proposed. By adding two new technologies based on gradient-based unilateral sampling (GOSS) and exclusive function bundling (EFB), LightGBM solves the shortcomings of traditional GBDT, which are time-consuming and inefficient, and has good classification performance [43].

The absolute gradient value of each training instance is calculated and sorted in descending order in the method of GOSS. The training instance with a larger gradient has a greater impact on the information gain, so the first $a \times 100\%$ instances with the larger gradient are selected to form the instance subset A. The remaining instances with smaller gradients are randomly sampled to obtain a subset B of instances of size $b \times |A^C|$. Finally, the examples are separated.

$$\overline{V_j}\left(d\right) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{nl(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i\right)^2}{nr(d)}\right) \tag{15}$$

where, $A_1 = \{x_i \in A : x_{ij} \le d\}$, $A_r = \{x_i \in A : x_{ij} > d\}$, $B_1 = \{x_i \in B : x_{ij} \le d\}$, $B_r = \{x_i \in B : x_{ij} > d\}$.

In the EFB method, one of the features is selected as a vertex, and other features are selected in turn. If other features are not mutually exclusive with the selected feature, edges are added for each feature. And the greedy algorithm is used to produce good bundling

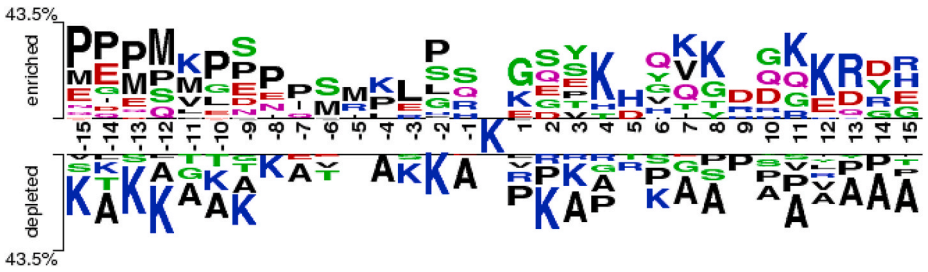

Fig. 2. Comparison of sequence identification of crotonylation and non-crotonylation.

characteristics. Then an offset is added to the original value of the feature, so that the proprietary feature resides can be retained in different bins to construct the feature bundle, and the value of the original function can be identified from the function package. Finally, the features in the same bundle are merged to divide the features into a minimum number of bundles.

2.6. Performance evaluation and model building

The methods of model evaluation mainly include independent dataset test, k-fold cross-validation, and jackknife test. The jackknife test [44] is used to evaluate the performance of the crotonylation sites prediction model. Each sample is taken as the test set, and the remaining samples are taken as the training set. The process is repeated continuously to ensure that each sample can be selected as the test set, and the results of multiple models is used as the final evaluation result. In addition to effective and feasible experimental evaluation method, evaluation indicators are also needed to measure the predictive ability of the model. In this paper, Sn, Sp, ACC and MCC [45] are selected to comprehensively measure the performance of the learner, which can be defined as follows:

$$Sn = \frac{TP}{TP + FN} \tag{16}$$

$$Sp = \frac{TN}{TN + FP} \tag{17}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{18}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{19}$$

where TP, FP, TN, FN represent true positives, false positives, true negatives, and false negatives, respectively. In addition, ROC curve [46] is used to reveal the relationship between sensitivity and specificity, and PR curve [47] is used to reveal the relationship between accuracy and recall. They show the prediction performance of the learner more intuitively.

For convenience, the crotonylation sites prediction method proposed in this paper is called LightGBM-CroSite, and the pipeline is shown in Fig. 1. The experimental environment: Windows Server 2012R2 Intel(R) Xeon(TM) CPU E5-2650 @ 2.30 GHz 2.30 GHz with 32.0 GB of RAM, MATLAB2014a and Python3.7.

The steps of LightGBM-CroSite can be described as follows:

(1) The training dataset is obtained, and the crotonylation sites sequences, non-crotonylation sites sequences, and corresponding labels are input.
(2) Feature extraction. BE, PWAA, EBGW, KNN, and PsePSSM are used to extract sequence feature, physicochemical feature and evolutionary information of protein sequences.
(3) Feature selection. The elastic net is used to remove redundant and irrelevant information in the original feature space.
(4) The SMOTE algorithm is used to balance the 159 positive samples and 847 negative samples.
(5) The balanced feature vectors are input into the LightGBM classifier to predict the crotonylation sites.
(6) Model performance evaluation. Jackknife test is used to calculate the Sn, Sp, ACC and MCC, and PR curves and ROC curves are drawn to evaluate the performance of the model.

3. Results and discussion

3.1. Sequence analysis of protein lysine crotonylation sites

In order to better analyze the difference between the crotonylation sites and non-crotonylation sites of samples, this paper uses Two Sample Logos [48] (http://www.twosamplelogo.org/cgi-bin/tsl/tsl.cgi) to analyze the protein sequences, and studies the frequency and location differences of 20 common amino acids near crotonylation sites and non-crotonylation sites.

In this paper, the dataset contains 31 samples, among which are 1 central lysine, 15 upstream amino acids and 15 downstream amino acids. It can be seen from Fig. 2 that the positively charged residual lysine (K) is mainly concentrated in the upstream segments at positions + 1, +4, +7, +8, +9, +11, +12 for the positive sample. Proline (P) is mainly concentrated in the downstream segment at - 15, - 14, - 13, - 12, - 10, - 9, - 8, - 7, - 4, - 2. But the result of the negative sample is opposite to the positive sample. The positively charged residual lysine (K) is mainly concentrated in the downstream segment at - 15, - 14, - 13, - 12, - 10, - 9, - 8, - 3, - 2. Proline (P) is mainly concentrated in the upstream segments at + 1, +2, +4, +6, +8, +9, +10, +11, +13, +14, +15. In addition, the percentages of residues in the positive and negative samples are also different. For the positive samples, the proportion of residues such as methionine (M), glutamic acid (E), serine (S) and glycine (G) are rich. For the negative samples, alanine (A) appears more frequently. Therefore, we conclude that the frequency and position of amino acid residues near crotonylation sites and non-crotonylation sites are significantly different.

3.2. Selection of model parameter

3.2.1. Selection of parameter L in EBGW

EBGW is an important feature extraction method, which is often used to extract sequence features of proteins. The selection of parameter L in the EBGW affects the model performance. In order to choose the value with the best prediction result, we set the parameter L to 1, 6, 11, 16, 21, 26, 31. For each L value, the obtained features are input into LightGBM for classification and the ACC, Sn, Sp, MCC and AUC are used to evaluate the prediction performance. The prediction results of different L values are shown in Table 1.

As can be seen from Table 1, different L values lead to different prediction results. As the increase of L, ACC value increases first and then decreases. When $L = 11$, the ACC obtains the maximum value of 95.52%, MCC obtains the maximum value of 0.8284, and the AUC reaches the value of 0.9555, which is 0.66% lower than that at $L = 31$.

Table 1
Prediction results obtained with different L values.

L	ACC (%)	Sn (%)	Sp (%)	MCC	AUC
$L = 1$	90.00	59.74	95.74	0.6015	0.8578
$L = 6$	95.12	83.64	97.28	0.8156	0.9554
$L = 11$	95.52	83.01	97.87	0.8284	0.9555
$L = 16$	95.32	80.50	98.11	0.8188	0.9515
$L = 21$	95.12	81.76	97.63	0.8131	0.9582
$L = 26$	95.22	81.76	97.75	0.8166	0.9589
$L = 31$	95.42	82.38	97.87	0.8242	0.9621

Table 2
Prediction results obtained with different ξ values.

ξ	ACC (%)	Sn (%)	Sp (%)	MCC	AUC
$\xi = 1$	95.72	84.27	97.87	0.8367	0.9739
$\xi = 3$	96.02	83.64	98.34	0.8467	0.9784
$\xi = 5$	96.12	84.90	98.22	0.8514	0.9802
$\xi = 7$	95.92	84.27	98.11	0.8437	0.9823
$\xi = 9$	95.82	83.64	98.11	0.8396	0.9818

After comprehensive analysis, the prediction performance is the best at $L = 11$. Therefore, the optimal parameter L of EBGW is selected to 11 to extract the sequences feature, and each protein sequence finally obtains the 33 dimensions feature vector.

3.2.2. Selection of parameterξ in PsePSSM

PsePSSM is an effective feature extraction method, which can extract the evolutionary information of protein sequences. The selection of parameter ξ in the PsePSSM is very important. In order to find the optimal value, we set the parameter ξ to 1, 3, 5, 7, 9. For each ξ value, the obtained features are input into LightGBM for classification and the ACC, Sn, Sp, MCC and AUC are used to evaluate the prediction performance. The prediction results corresponding to different ξ values are shown in Table 2.

As can be seen from Table 2, different ξ values will have different effects on the prediction results. With the increase of ξ, ACC value increases first and then decreases, and the maximum ACC value is 96.12% at $\xi = 5$. MCC obtains the maximum value of 0.8514 at $\xi = 5$ and AUC reaches the highest value of 0.9823 at $\xi = 7$. After comprehensive analysis, the prediction result of $\xi = 5$ is the best. Therefore, the optimal parameter ξ of PsePSSM is determined to be 5, and each protein sequence finally obtains the 120-dimensions feature vector $P_3 = (x_1, x_2, \ldots, x_{20}, \theta_1^1, \theta_2^1, \ldots, \theta_{20}^1, \ldots, \theta_1^5, \theta_2^5, \ldots, \theta_{20}^5)$.

3.3. The effect of feature extraction algorithm

The information obtained by the single feature extraction method is often not comprehensive enough, and the prediction results are not ideal. Multi-feature fusion can use different type features to improve the prediction performance of the model. In this paper, we use BE, PWAA, EBGW, KNN and PsePSSM to encode protein sequences, and obtain 651, 21, 33, 8 and 120 dimensions feature vectors, respectively. We fuse the five types of feature vectors and obtain the fused feature set ALL. The prediction results of different feature extraction methods on ACC, Sn, SP, MCC and AUC are shown in Table 3 through Jackknife test.

As can be seen from Table 3, the prediction results have some differences corresponding different feature extraction methods. Among the five single feature extraction methods, ACC, MCC, and AUC of PsePSSM reach the highest values, which are 96.12%, 0.8514, and 0.9802, respectively. The ACC, MCC and AUC of BE is the lowest, which are 95.32%, 0.8258 and 0.9679 respectively. After fusing the five features, the ACC, MCC and AUC of ALL are 96.62%, 0.8712, and 0.9814, which

Table 3
The prediction results of different feature extraction methods.

Feature extraction methods	ACC (%)	Sn (%)	Sp (%)	MCC	AUC
BE	95.32	86.16	97.04	0.8258	0.9679
PWAA	95.72	84.27	97.87	0.8367	0.9652
EBGW	95.52	83.01	97.87	0.8284	0.9555
KNN	95.82	84.27	97.99	0.8402	0.9697
PsePSSM	96.12	84.90	98.22	0.8514	0.9802
ALL	96.62	87.42	98.34	0.8712	0.9814

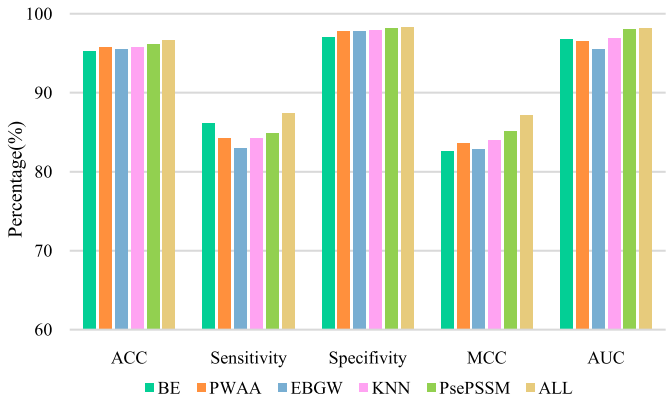

Fig. 3. Comparison of prediction results on different feature extraction methods.

are 0.5%, 1.98%, and 0.12% higher than PsePSSM respectively. The results indicate multi-information fusion can improve the prediction accuracy of various metrics. In order to better analyze the effect of different feature extraction methods on the crotonylation sites prediction, the column chart of six feature extraction methods on ACC, Sn, SP, MCC and AUC are shown in Fig. 3.

As can be seen from Fig. 3, the six feature extraction methods have different effects for different evaluation indexes, where the six feature extraction methods have excellent performance in ACC, Sp and AUC, but have poor performance in Sn and MCC. Compared the six feature encode methods, the ACC, Sn, Sp, AUC and MCC of the fusion feature ALL are all higher than the five feature extraction methods of BE, PWAA, EBGW, KNN and PsePSSM. Multi-feature fusion can make the information more comprehensive and improve the predictive power. Therefore, we apply multi-feature fusion to extract protein sequences features for crotonylation sites prediction.

3.4. Comparison of different dimension reduction methods

Multi-information fusion comprehensively extracts various information of protein sequences, but it also generates redundant and noise information which will affect the prediction effect of the model. Dimension reduction method can retain the important features for classification, and improve the computational efficiency of model. In this paper, elastic net, Lasso [38], extra-trees (ET) [49], singular value decomposition (SVD) [50], local linear embedding (LLE) [51] and multiple dimensional scaling (MDS) [52] are used to reduce the dimensions of fusion feature dataset, and six feature subsets with dimensions of 217, 200, 218, 217, 217 and 217 are obtained respectively. The six feature subsets are input into LightGBM for classification. Based on jackknife test, the prediction results of different dimension reduction methods on ACC, Sn, Sp, MCC and AUC are shown in Table 4.

It can be seen from Table 4 that different dimension reduction methods have different effects on the prediction accuracy. Compared with Lasso, ET, SVD, LLE, and MDS methods, the elastic net has a greater prediction performance. Its ACC, Sn, SP, MCC and AUC are 97.02%, 88.68%, 98.58%, 0.8864 and 0.9815, respectively. The ACC of elastic net is 0.3%, 0.7%, 0.8%, 1.49%, and 1.4% higher than the ACC of Lasso,

Table 4
Prediction results of different dimension reduction methods.

Dimension reduction methods	ACC (%)	Sn (%)	Sp (%)	MCC	AUC
Elastic net	97.02	88.68	98.58	0.8864	0.9815
Lasso	96.72	87.42	98.47	0.8748	0.9809
ET	96.32	86.79	98.11	0.8602	0.9785
SVD	96.22	83.02	98.70	0.8537	0.9708
LLE	95.53	86.16	97.28	0.8324	0.9692
MDS	95.62	82.39	98.11	0.8314	0.9690

ET, SVD, LLE, and MDS, respectively. The MCC value of elastic net is 1.16%, 2.62%, 3.27%, 5.4%, and 5.5% higher than the corresponding value of Lasso, ET, SVD, LLE, and MDS, respectively. In order to compare the six dimension reduction methods more comprehensively, the ROC and PR curves of different feature selection methods are compared as shown in Fig. 4.

Compared the ROC and PR curves of different dimension reduction methods, the curve of one method is completely covered by the curve of another method, and the performance of the later is better than the former. As can be seen from Fig. 4, the ROC and PR curves of the six dimension reduction methods intersect with each other, and we cannot obtained the accurate conclusions through curves. So we compare the area under the ROC curves and PR curves called AUC and AUPR. The higher the value, the better the prediction performance. The AUC value of elastic net is 0.9815, which is 1.25%, 1.23%, 1.07%, 0.3%, 0.06% higher than the AUC values of MDS, LLE, SVD, ET, and Lasso, respectively. The AUPR value of elastic net is 0.9535, which is 3.23%, 6.1%, 2.52%, 0.61% and 0.31% higher than MDS, LLE, SVD, ET and Lasso, respectively. We conclude that the prediction performance of elastic net is better than the other five feature selection methods, so we choose elastic net as the optimal dimension reduction method to predict crotonylation sites.

3.5. The effectiveness of the SMOTE algorithm

The dataset selected in this paper consists of 159 positive samples and 847 negative samples, which exists a serious imbalance problem that will lead to the deviation of model prediction. In this paper, the SMOTE is used to deal with the imbalance dataset. The oversampling times and undersampling times in SMOTE are set to 403.14 and 124.8 respectively and the balance dataset of 795 positive samples and 793 negative samples is obtained. Then, the balanced dataset and imbalanced dataset are input into LightGBM for classification, respectively. Through Jackknife test, the prediction results of ACC, MCC, AUC, Sn and Sp on imbalanced dataset and balanced dataset are obtained and shown in Table 5.

As can be seen from Table 5, the SMOTE can affect the final prediction results on imbalanced dataset. ACC reflects the overall proportion of correct prediction and the ACC values before and after SMOTE are 97.01% and 98.99%, which increased by 1.98%. Sn and Sp represent the percentage of correct predictions for positive and negative samples, and the imbalance of samples may increases the difference between Sn and Sp. After using SMOTE to synthesize new samples, the disparity between Sn and Sp is shrunk. Sn increases from 88.67% to 98.86%, and Sp increases from 98.58% to 99.11%. In addition, MCC and AUC also increased to some extent. MCC increases from 0.8864 to 0.9798, and AUC increases from 0.9815 to 0.9996. These results show that SMOTE can improve the prediction performance of the minority samples, and improve the overall prediction accuracy of crotonylation sites. Therefore, we select SMOTE to deal with the imbalance between positive dataset and negative dataset.

3.6. Comparison of different classification methods

Classifier is the most critical part of the prediction model, which plays the crucial role in the final prediction results. In order to construct classification model with strong prediction performance, NB [53], AdaBoost [54], KNN [55], XGBoost [56], SVM [20], RF [18] and LightGBM [42] are used to identify crotonylation sites in this paper. AdaBoost and NB adopt default parameters. The number of neighbors of KNN is 10. The learning rate of XGBoost is set to 0.01 and the number of iterations is 500. The SVM chooses the radial basis kernel function. RF selects the gini coefficient to divide the nodes, and the number of decision trees is set to 500. The maximum depth of the LightGBM tree is 15, and the learning rate is set to 0.2. Through jackknife test, the prediction results of seven classifiers on ACC, Sn, Sp, MCC and AUC are shown in Table 6.

It can be obtained from Table 6 that different classifiers have different prediction results. NB has the worst prediction effect, and AdaBoost, KNN, XGBoost, SVM, RF and LightGBM have better prediction results. The ACC of LightGBM is 98.99%, which is 29.91%, 4.22%, 1.26%, 0.69%, 0.19% and 0.06% higher than NB, AdaBoost, KNN, XGBoost, SVM and RF, respectively. The MCC of LightGBM is 0.9798, which is the highest value of seven classifiers. In order to evaluate the prediction performance of classifiers more comprehensively, the ROC and PR curves comparison of different classifiers are shown in Fig. 5.

The area under the ROC curves and PR curves can evaluate the model

Table 5
Comparison of predict results before and after SMOTE method.

Datasets	ACC (%)	Sn (%)	Sp (%)	MCC	AUC
Imbalance dataset	97.01	88.67	98.58	0.8864	0.9815
Balance dataset	98.99	98.86	99.11	0.9798	0.9996

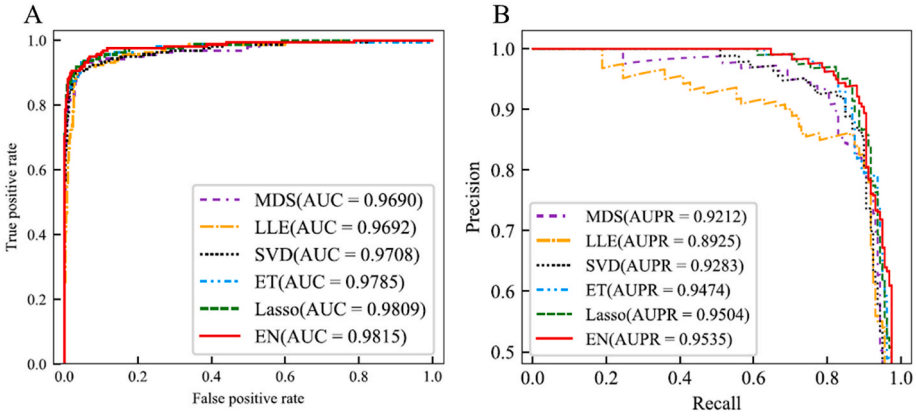

Fig. 4. The ROC and PR curves of different dimension reduction methods. The (A) graph is the ROC curve, and the (B) graph is the PR curve.

Table 6
Prediction results for different classifiers.

Classifiers	ACC (%)	Sn (%)	Sp (%)	MCC	AUC
NB	69.08	99.37	38.71	0.4793	0.9600
AdaBoost	94.77	95.35	94.20	0.8955	0.9907
KNN	97.73	99.62	95.84	0.9553	0.9920
XGBoost	98.30	98.49	98.11	0.9660	0.9979
SVM	98.80	99.25	98.36	0.9761	0.9968
RF	98.93	98.99	98.87	0.9786	0.9989
LightGBM	98.99	98.86	99.11	0.9798	0.9996

prediction performance of different classifiers on the dataset. It can be seen from Fig. 5 that the ROC and PR curves of NB, AdaBoost, KNN, SVM, XGBoost, and RF are completely covered by the ROC and PR curves of LightGBM. The AUC of the seven classifiers are 0.9600, 0.9907, 0.9920, 0.9972, 0.9979, 0.9990 and 0.9996 respectively, and the AUPR of the seven classifiers are 0.9551, 0.9916, 0.9926, 0.9979, 0.9980, 0.9991 and 0.9996 respectively. Compared with the AUC and AUPR of the seven classifiers, the AUC and AUPR of LightGBM reach the maximum values. The AUC of LightGBM method is 3.96%, 0.89%, 0.76%, 0.24%, 0.17% and 0.06% higher than NB, AdaBoost, KNN, SVM, XGBoost and RF, respectively. The AUPR is 4.45%, 0.8%, 0.7%, 0.17%, 0.16%, and 0.05% higher than NB, AdaBoost, KNN, SVM, XGBoost and RF, respectively. Therefore, LightGBM outperforms the other six classifiers, and it is selected as the final classifier to predict crotonylation sites.

3.7. Comparison with other state-of-the-art methods

At present, CrotPred [28], Qiu's method [29]. and CKSAAP_CrotSite [32] have been proposed to predict crotonylation sites. In order to further evaluate the performance of LightGBM-CroSite, we compare it with CrotPred, Qiu's method and CKSAAP_CrotSite. The prediction results of four different methods on the same dataset are shown in Table 7.

It can be seen from Table 7 that the ACC, Sn, Sp, and MCC values of the LightGBM-CroSite for crotonylation sites prediction are 98.99%, 98.86%, 99.11%, and 0.9798, respectively. Compared with other methods, the ACC of LightGBM-CroSite is 20.76% higher than that of CrotPred, 4.56% higher than that of Qiu's method and 0.88% higher than that of CKSAAP_CrotSite. The MCC value of LightGBM-CroSite is 45.39%, 20.18% and 5.15% higher than CrotPred, Qiu's method and CKSAAP_CrotSite, respectively. Therefore, the LightGBM-CroSite is significantly better than CrotPred, Qiu's method and CKSAAP_CrotSite, which can significantly improve the prediction performance of crotonylation sites.

4. Conclusion

Lysine crotonylation is an important biological process, which participates in multiple processes such as chromosomal organization, nucleic acid metabolism and gene expression in cells. It is necessary to study the principle and function of lysine crotonylation modification. In order to solve the problem of high cost and low efficiency of traditional experimental methods, this paper proposes a novel crotonylation sites prediction method called LightGBM-CroSite. First, BE, PWAA, EBGW, KNN and PsePSSM are used to extract the features of protein sequences. BE extracts the sequence features of protein sequences, EBGW is used to extract the physicochemical feature of protein sequences, PsePSSM can extract the evolutionary information of protein sequences, KNN extracts the neighbor information of protein sequences, and PWAA extracts the sequence order information of protein sequences. Second, the sequence feature, physicochemical properties feature, and evolutionary information of protein sequences are obtained by fusing five methods. Compared with the five single features extraction methods, the fusion feature achieves the best effect on the dataset, showing the necessity of multi-feature fusion. Third, the elastic net is compared with other dimension reduction methods, and it is used to select the optimal feature subset. It effectively removes redundant information and retains effective features. Then, SMOTE is used to deal with the problem of imbalance, and it reduces the negative impact of imbalance on the model. Finally, the LightGBM is employed to predict the crotonylation sites. By adding GOSS and EFB, the LightGBM method solves the shortcomings of the previous GBDT algorithms. The model constructed by LightGBM has fast training speed and high efficiency. Compared with other state-of-the-art methods, our method can accurately and effectively improve the prediction effect of crotonylation sites. We will build a web server based on the method of this article in our future work and we hope our method can contribute to the research of bioinformatics and proteomics.

Author contributions

Yaning Liu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Validation, Visualization. **Zhaomin Yu:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft,

Table 7
Comparison of the LightGBM-CroSite with other methods.

Methods	ACC (%)	Sn (%)	Sp (%)	MCC
CrotPred [28]	78.23	79.41	77.78	0.5259
Qiu's method [29]	94.43	71.69	98.70	0.7780
CKSAAP_CrotSite [32]	98.11	92.45	99.17	0.9283
LightGBM-CroSite	98.99	98.86	99.11	0.9798

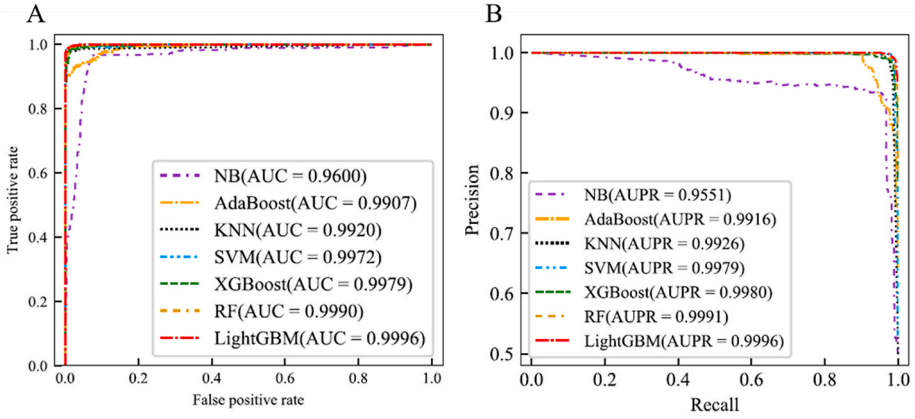

Fig. 5. The ROC and PR curves of different classification methods. The (A) graph is the ROC curve, and the (B) graph is the PR curve.

Validation, Visualization. **Cheng Chen:** Formal analysis, Investigation, Writing - original draft. **Yu Han:** Writing- Reviewing and Editing. **Bin Yu:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Writing- Reviewing and Editing.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

The authors sincerely thank the anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China (No. 61863010), the Key Research and Development Program of Shandong Province of China (No. 2019GGX101001), and the Natural Science Foundation of Shandong Province of China (Nos. ZR2018MC007, ZR2019MEE066).

References

[1] V.G. Allfrey, B.G.T. Pogo, V.C. Littau, E.L. Gershey, A.E. Mirsky, Histone acetylation in insect chromosomes, Science 159 (1968) 314–316.
[2] Y.D. Khan, N. Rasool, W. Hussain, S.A. Khan, K.C. Chou, iPhosT-PseAAC, Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC, Anal. Biochem. 550 (2018) 109–116.
[3] W. Deng, Y. Wang, L. Ma, Y. Zhang, S. Ullah, Y. Xue, Computational prediction of methylation types of covalently modified lysine and arginine residues in proteins, Brief. Bioinform. 18 (2016) 647–658.
[4] L. Kiemer, J.D. Bendtsen, N.J.B. Blom, NetAcet: prediction of N-terminal acetylation sites, Bioinformatics 21 (2005) 1269–1270.
[5] J. Yu, S. Shi, F. Zhang, G. Chen, M. Cao, PredGly: predicting lysine glycation sites for Homo sapiens based on XGboost feature optimization, Bioinformatics 35 (2019) 2749–2756.
[6] W. Wei, A. Mao, B. Tang, Q. Zeng, S. Gao, X.G. Liu, L. Lu, W. Li, J.X. Du, J. Li, J. Wong, L.J. Liao, Large-scale identification of protein crotonylation reveals its role in multiplecellular functions, J. Proteome Res. 16 (2017) 1743–1752.
[7] L.M. Liu, Y. Xu, K.C. Chou, iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC, Med. Chem. 13 (2017) 552–559.
[8] M.H. Wang, X.W. Cui, B. Yu, C. Chen, Q. Ma, H.Y. Zhou, SulSite-GTB: identification of protein S-sulfenylation sites by fusing multiple feature information and gradient tree boosting, Neural. Comput. Appl. 32 (2020) 13843–13862.
[9] Y. Xu, J. Ding, L.Y. Wu, K.C. Chou, iSNO-PseAAC: predict cysteine S-Nitrosylation Sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition, PloS One 8 (2013), e55844.
[10] L.N. Wang, S.P. Shi, H.D. Xu, P.P. Wen, J.D. Qiu, Computational prediction of species-specific malonylation sites via enhanced characteristic strategy, Bioinformatics 33 (2016) 1457–1463.
[11] Y. Saeys, I. Inza, P. Larranaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507–2517.
[12] X.W. Cui, Z.M. Yu, B. Yu, M.H. Wang, B.G. Tian, Q. Ma, UbiSitePred: a novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components, Chemometr. Intell. Lab. 184 (2019) 28–43.
[13] J.R. Wang, W.L. Huang, M.J. Tsai, K.T. Hsu, H.L. Huang, S.Y. Ho, ESA-UbiSite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives, Bioinformatics 33 (2016) 661–668.
[14] M. Cao, G.D. Chen, J.L. Yu, S.P. Shi, Computational prediction and analysis of species-specific fungi phosphorylation via feature optimization strategy, Brief. Bioinform. 21 (2020) 595–608.
[15] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Curran Associates Inc, 2008, pp. 1322–1328.
[16] X. Hong, S. Chen, C.J. Harris, A kernel-based two-class classifier for imbalanced data sets, IEEE T. Neural Networ. 18 (2007) 28–41.
[17] X.Y. Wang, B. Yu, A.J. Ma, C. Chen, B.Q. Liu, Q. Ma, Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique, Bioinformatics 35 (2019) 2395–2402.
[18] H.D. Ismail, A. Jones, J.H. Kim, R.H. Newman, D.B. Kc1, Rf-Phos, A novel general Phosphorylation site prediction tool based on random forest, BioMed Res. Int. 2016 (2016) 3281590.
[19] J. Jia, Z. Liu, X. Xiao, B.X. Liu, K.C. Chou, iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset, Anal. Biochem. 497 (2016) 48–56.
[20] Z. Ju, H. Gu, Predicting pupylation sites in prokaryotic proteins using semi-supervised self-training support vector machine algorithm, Anal. Biochem. 507 (2016) 1–6.
[21] Z. Ju, J.Z. Cao, Prediction of protein N-formylation using the composition of k-spaced amino acid pairs, Anal. Biochem. 534 (2017) 40–45.
[22] T. Hou, G. Zheng, P. Zhang, J. Jia, J. Li, L. Xie, C.C. Wei, Y.X. Li, LAceP: lysine acetylation site prediction using logistic regression classifiers, PloS One 9 (2014), e89575.
[23] F. Li, C. Li, T.T. Marquez-Lago, A. Leier, T. Akutsu, A.W. Purcell, A.I. Smith, T. Lithgow, R.J. Daly, J. Song, K.C. Chou, Quokka: a comprehensive tool for rapid and accurate prediction of kinase family specific phosphorylation sites in the human proteome, Bioinformatics 34 (2018) 4223–4231.
[24] M.S. Ahmed, M. Shahjaman, E. Kabir, M. Kamruzzaman, Prediction of protein acetylation sites using kernel naive Bayes classifier based on protein sequences profiling, Bioinformation 14 (2018) 213–218.
[25] Y. Xue, H. Chen, C. Jin, Z. Sun, X. Yao, NBA-Palm: prediction of palmitoylation site implemented in Naive Bayes algorithm, BMC Bioinf. 7 (2006), 458-458.
[26] J. Jia, Z. Liu, X. Xiao, B.X. Liu, K.C. Chou, pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach, J. Theor. Boil. 394 (2016) 223–230.
[27] Z. Ju, J.J. He, Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection, Anal. Biochem. 550 (2018) 1–7.
[28] G. Huang, W. Zeng, A discrete hidden Markov model for detecting histone crotonyllysine sites, Math. Comput. Chem. 75 (2016) 717–730.
[29] W.R. Qiu, B.Q. Sun, H. Tang, J. Huang, H. Lin, Identify and analysis crotonylation sites in histone by using support vector machines, Artif. Intell. Med. (2017) 75–81.
[30] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, K.C. Chou, iPTM-mLys: identifying multiple lysine PTM sites and their different types, Bioinformatics 32 (2016) 3116–3123.
[31] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, J.H. Jia, K.C. Chou, iKcr-PseEns: identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier, Genomics 110 (2018) 239–246.
[32] Z. Ju, J.J. He, Prediction of lysine crotonylation sites by incorporating the composition of k-spaced amino acid pairs into Chou's general PseAAC, J. Mol. Graph. Model. 77 (2017) 200–204.
[33] S.P. Shi, J.D. Qiu, X.Y. Sun, S.B. Suo, S.Y. Huang, R.P. Liang, A method to distinguish between lysine acetylation and lysine methylation from protein sequences, J. Theor. Biol. 310 (2012) 223–230.
[34] Z.H. Zhang, Z.H. Wang, Z.R. Zhang, Y.X. Wang, A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine, Febs Lett. 580 (2006) 6169–6174.
[35] J. Gao, J.J. Thelen, A.K. Dunker, D. Xu, Musite, a tool for global prediction of general and kinase-specific phosphorylation sites, Mol. Cell. Proteomics 9 (2010) 2586–2600.
[36] D. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292 (1999) 195–202.
[37] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[38] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. Roy. Stat. Soc. B 58 (1996) 267–288.
[39] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Stat. Soc. B. 67 (2005) 301–320.
[40] G.D. Chen, M. Cao, J.L. Yu, X.Y. Guo, S.P. Shi, Prediction and functional analysis of prokaryote lysine acetylation site by incorporating six types of features into Chou's general PseAAC, J. Theor. Biol. 461 (2019) 92–101.
[41] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[42] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.Y. Liu, LightGBM: a highly efficient gradient boosting decision tree, in: 31st Conference Neural Information Processing Systems NIPS, Curran Associates Inc, 57 Morehouse Lane, Red Hook,NY,United States, 2017, pp. 3146–3154.
[43] C. Chen, Q.M. Zhang, Q. Ma, B. Yu, LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion, Chemometr. Intell. Lab. 191 (2019) 54–64.
[44] B. Yu, W.Y. Qiu, C. Chen, A.J. Ma, J. Jiang, H.Y. Zhou, Q. Ma, SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting, Bioinformatics 36 (2020) 1074–1081.
[45] B. Yu, S. Li, W. Qiu, M.H. Wang, J.W. Du, Y. Zhang, X. Chen, Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction, BMC Genom. 19 (2018) 478.
[46] B. Yu, Z.M. Yu, A.J. Ma, C. Chen, B.Q. Liu, B.G. Tian, Q. Ma, DNNAce: prediction of prokaryote lysine acetylation sites through deep neural networks with multi-information fusion, Chemometr. Intell. Lab. 200 (2020) 103999.
[47] H. Shi, S.M. Liu, J.Q. Chen, X. Li, Q. Ma, B. Yu, Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure, Genomics 111 (2019) 1839–1852.
[48] V. Vacic, L.M. Iakoucheva, P. Radivojac, Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments, Bioinformatics 22 (2006) 1536–1537.
[49] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42.
[50] H. Andrews, C. Patterson, Singular value decomposition (SVD) image coding, IEEE Trans. Commun. 24 (1976) 425–432.
[51] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[52] I. Borg, P. Groenen, Modern multidimensional scaling: theory and applications (second edition), J. Educ. Meas. 40 (2003) 277–280.
[53] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (1997) 131–163.

[54] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to Boosting, J. Comput. Syst. Sci. 55 (1997) 119–139.
[55] F. Nigsch, A. Bender, B.V. Buuren, J. Tissen, E. Nigsch, J.B. Mitchell, Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization, J. Chem. Inf. Model. 46 (2006) 2412–2422.
[56] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, United States, 2016, pp. 785–794.