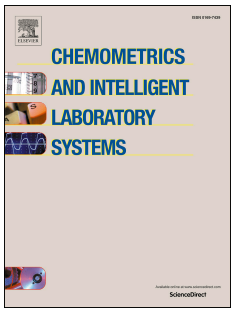
Accepted Manuscript



UbiSitePred: A novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components

Xiaowen Cui, Zhaomin Yu, Bin Yu, Minghui Wang, Baoguang Tian, Qin Ma

PII: S0169-7439(18)30522-7

DOI: https://doi.org/10.1016/j.chemolab.2018.11.012

Reference: CHEMOM 3714

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 8 September 2018

Revised Date: 8 November 2018

Accepted Date: 17 November 2018

Please cite this article as: X. Cui, Z. Yu, B. Yu, M. Wang, B. 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, *Chemometrics and Intelligent Laboratory Systems* (2018), doi: https://doi.org/10.1016/j.chemolab.2018.11.012.

UbiSitePred: a novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components

Xiaowen Cui[a,b,1], Zhaomin Yu[a,b,1], Bin Yu[a,b,c,d,1,*], Minghui Wang[a,b], Baoguang Tian[a,b], Qin Ma[e]

[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

[d] Department of Biochemistry & Molecular Biology, Medical Genetics, and Oncology, University of Calgary, Calgary T2N 4N1, Canada

[e] Bioinformatics and Mathematical Biosciences Lab, Department of Agronomy, Horticulture and Plant Science, South Dakota State University, Brookings, SD 57007, USA

ABSTRACT

Ubiquitination is an essential process in protein post-translational modification, which plays a crucial role in cell life activities, such as proteasomal degradation, transcriptional regulation, and DNA damage repair. Therefore, recognition of ubiquitination sites is a crucial step to understand the molecular mechanisms of ubiquitination. However, the experimental verification of numerous ubiquitination is time-consuming and costly. To alleviate these issues, a computational approach is needed to predict ubiquitination sites. This paper proposes a new method called UbiSitePred for predicting ubiquitination sites combined least absolute shrinkage and selection operator (LASSO) feature selection and support vector machine. First, we use binary encoding (BE), pseudo-amino acid composition (PseAAC), the composition of k-spaced amino acid pairs (CKSAAP), position-specific propensity matrices (PSPM) to extract the sequence feature information; thus, the initial feature space is obtained. Secondly, LASSO is applied to remove the feature redundancy information and selects the optimal feature subset. Finally, the optimal feature subset is input into the support vector machine (SVM) to predict the ubiquitination sites. Five-fold cross-validation shows that UbiSitePred model can achieve a better prediction performance compared with other methods, the AUC values for Set1, Set2, and Set3 are 0.9998, 0.8887, and

[*] Corresponding author.
E-mail address: yubin@qust.edu.cn (B. Yu).
[1] These authors contributed equally to this work.

0.8481, respectively. Notably, the UbiSitePred has overall accuracy rates of 98.33%, 81.12%, and 76.90%, respectively. The results demonstrate that the proposed method is significantly superior to other state-of-the-art prediction methods and provides a new idea for the prediction of other post-translational modification sites of proteins. The source code and all datasets are available at https://github.com/QUST-AIBBDRC/UbiSitePred/.

Keywords: Ubiquitination sites; Binary encoding; Pseudo-amino acid composition; Composition of k-spaced amino acid pairs; Position-specific propensity matrices; Least absolute shrinkage and selection operator.

1. Introduction

Protein post-translational modification (PTM) is the main mode of regulating protein structure and function, which plays a significant role in regulating many cellular processes such as various signaling pathways or networks in cells, gene expression, inactivation and activation of enzymes, and protein-protein interaction [1]. Post-translational modification is also closely related to various pathological states, once a modification abnormality occurs, it is likely to cause disease. As the post-translational modification of proteins is present in dynamically changing living organisms, the type and degree of modification will transform with changes in the internal environment of the organism, and even some of the modifications will be fleeting. Therefore, it is crucial for the further study of protein post-translational modification sites, and also important to help research and design the novel drugs to treat the relevant diseases. At present, the major types of protein post-translational modifications include methylation [2-3], nitrotyrosine [4], phosphorylation [5], SUMOylation [6], prenylation [7], ubiquitination [8], methyladenosine [9,10,11], pseudouridine [12], phosphothreonine [13], crotonylation [14]. Ubiquitination is a process in the most common post-translational modification, which plays a crucial role in the growth and development of organisms, such as protein localization, metabolism, function, regulation, and degradation. At the same time, ubiquitination is also closely related to regulatory function such as cell cycle, apoptosis, transcriptional regulation, signal transduction, and DNA damage repair [15,16]. Besides, ubiquitination imbalance can lead to several human diseases such as cancer, neurodegenerative diseases, muscular dystrophy, immunity diseases and metabolic syndrome [17].

With the life science research entering the post-genome era, the protein sequence data accumulated in the protein database has increased exponentially. Identifying the post-translational modification sites of the protein is of great significance for understanding the post-translational modification process and its functions. Predicting ubiquitination sites provides not only valuable opinion into grasping the ubiquitination molecular mechanisms but also affords useful information for further study of biological sciences and drug development, because of the critical regulatory

role of ubiquitination. Currently, the methods for identifying ubiquitylation sites include site-directed mutagenesis [18] and mass spectrometry [19]. Ubiquitination is a rapid and reversible post-translational modification of proteins; thus, the traditional experimental methods are time-consuming and labor-intensive. Bioinformatics methods combined with machine learning algorithms can efficiently, and large-scale identify the ubiquitination sites [20-21].

Protein sequence feature extraction is an important part of post-translational modification sites prediction, and effective feature extraction methods have a positive effect on the recognition of modification sites. The current feature extraction methods of protein sequences are mainly based on sequence features, physicochemical and biochemical properties features, predicted structural features and evolutionary information features. Qiu et al. [22] used the position weight amino acid composition (PWAA) to extract the sequence position information of amino acid residues to reveal the sequence information around the crotonylation sites. The PWAA feature coding method is also used to identify phosphorylation sites [23] and methylation sites [24]. The composition of k-spaced amino acid pairs (CKSAAP) was widely used to predict post-translational modification sites, such as O-glycosylation sites [25], palmitoylation sites [26], phosphorylation sites [27]. Tung and Ho [21] used 31 physicochemical features to identify ubiquitination sites in protein sequences. Wuyun et al. [28] used the prediction tool PSIPRED [29] to extract secondary structure information of protein sequences for predicting lysine acetylation sites. The position-specific scoring matrix (PSSM) was employed to calculate the evolutionary information of protein sequences through multiple sequence alignments. Abdollah Dehzangi et al. [30] predicted succinylation lysine residues based on PSSM. Jia et al. [31] integrated the sequence coupling information into a pseudo-amino acid composition (PseAAC) to predict the succinylation sites. They also designed the predictor iSuc-PseOpt [32] to process the training dataset using K-nearest neighbor cleansing (KNNC) and insert hypothesis training samples (IHTS) to predict lysine succinylation sites. Ju et al. [33] incorporated the CKSAAP coding into Chou's PseAAC to predict crotonylation sites. Liu et al. [34] identified lysine phosphorylation sites in proteins by incorporating four different levels of amino acid pair coupling information into PseAAC. Qiu et al. [35] proposed a protein phosphorylation site predictor iPhos-PseEn, by fusing different pseudo-components into a set classifier. Xu et al. [36] designed the cysteine S-nitrosylation sites prediction tool ISNO-PseAAC by incorporating position-specific amino acid propensity into PseAAC. They also coupled the amino acid pairing into the PseAAC designed cysteine S-nitrosylation site predictor iSNO-AAPair [37]. Qiu et al. [38] proposed that the protein methylation site predictor iMethyl-PseAAC extracts protein sequence features by the PseAAC algorithm. Protein hydroxylation is closely related to lung cancer and gastric cancer. To understand the mechanism of hydroxylation and help drug development, Xu et al. [39] developed

the predictor iHyd-PseAAC based on the positional specificity of dipeptides into the general form of PseAAC. Hydroxylation of proline and lysine is predicted. Jia et al. [40] integrated the sequence coupling effect into the general PseAAC and identified the carbonylation sites in the protein by Monto Carlo sampling. Based on the covariance discriminant algorithm, they developed the protein SUMO site predictor pSumo-CD [6], which combined the sequence coupling effect into the general PseAAC. Huang et al. [41] developed a method called PredSulSite that incorporated three types of encoding algorithms-secondary structure, grouped weight and autocorrelation function-digging features from tyrosine sulfation proteins, for the identification of tyrosine sulfation sites. Wang et al. [42] constructed a novel malonylation sites online prediction tool, called MaloPred, which can predict malonylation sites by combining sequence-based features, evolutionary-derived information, and physicochemical properties. Liu et al. [43] predicted N-methyladenosine sites by extracting the physicochemical properties of RNA sequences. Qiu et al. [5] incorporated evolutionary information into the general form of PseAAC and applied grey system theory to predict human protein phosphorylation sites.

Feature fusion will bring about redundant information and produce dimension disaster, which also causes troubles for calculation and even affect the forecasting results. Therefore, it is necessary to select the optimal feature subset of the fusion information, reduce noise and eliminate redundant information. At the same time, it can maximumly retain valuable features, improve the efficiency, performance, and robustness of the prediction model [44]. In 2007, Liu et al. [45] introduced the concept of feature extraction and selection, using properties sequential forward selection (PSFS) to extract effective properties of amino acids and a novel computational method was developed for SUMO modification sites prediction based on support vector machine (SVM) algorithm. The research team also used maximum relevance minimum redundancy (mRMR) [46], incremental feature selection (IFS) [46] and feature forward selection (FFS) [47] to select features. Cai et al. [48] created a method to predict N-formylation sites based on the maximum relevance minimum redundancy (mRMR) and incremental feature selection method (IFS) to screen the optimal feature subset. Ju et al. [49] constructed a novel bioinformatics tool named PropPred for predicting lysine propionylation sites combined with the F-score feature method and the incremental feature selection algorithm to remove the redundant features, using support vector machine as a classifier and the prediction accuracy rate reached 75.02%. Wang et al. [50] proposed PrAS to predict amidation sites, which incorporated position-based features, physicochemical and biochemical properties features, predicted structure-based features and evolutionary information features, then used positive contribution feature selection (PCFS) to form the optimized features, finally based on support vector machine classifier, PrAS achieved AUC of 0.96, accuracy of 92.1%, sensitivity of 81.2%, specificity of 94.9% and MCC of 0.76 on

the independent test set. Tung and ho [21] proposed an informative physicochemical property mining algorithm (IPMA), the 31 features selected by IPMA from 531 physicochemical properties for ubiquitylation sites prediction. Qiu et al. [51] designed the methylation sites prediction tool PSSMe based on the optimization method of information gain (IG). Wuyun et al. [28] established the lysine acetylation sites prediction tool KA-predictor which used the Pearson correlation coefficient (PCC) and the stepwise feature selection (SFS) method to select the optimal feature subset.

In the past decades, with the rapid development of proteomics technology, the modified protein sequences related to sites with constant output, which greatly promoted the study of post-translational modification sites of proteins. The identification of these sites is of great significance for understanding the mechanism of protein function. Different types of machine learning methods are widely used for sites prediction because of their learning model and predictive power. The current mainstream machine learning prediction algorithms are logistic regression (LR) [52], Naïve Bayes (NB) [53-54], neural network (NN) [55-56], K-nearest neighbor (KNN) [57-59], random forest (RF) [31,61,62], support vector machine (SVM) [63,64], etc. Logistic regression is a regression analysis algorithm based on logical functions. In 2014, Hou et al. [52] proposed logistic regression classifier LAceP to predict acetylation sites. Naïve Bayes algorithm [52,53] is a powerful probabilistic network model learning method. Xue et al. [54] developed a novel computational method NBA-Palm based on Naïve Bayes to predict palmitoylation sites. The artificial neural network is a simulation of the biological neural system, whose main characteristics include its parallel information processing capabilities, as well as its self-adaptive, self-organizing and fault-tolerant characteristics in the learning process. In 1999, Nikolaj Blom et al. [55] in Denmark first realized the prediction of non-specific protein phosphorylation sites and the effectiveness of the model was verified by using the neural network algorithm. Tang et al. [56] developed GANNPhos to predict phosphorylation sites based on genetic algorithm integrated neural network (GANN). The KNN [57,58] is a commonly supervised learning algorithm according to the similarity between the test sample and the training samples. In 2005, Li et al. [59] designed kinase-specific phosphorylation sites prediction with KNN algorithm. Hu et al. [60] constructed the S-nitrification sites prediction model based on the nearest neighbor algorithm (NNA). Random forest [61] is a simple and effective ensemble learning classification algorithm, which has an excellent classification effect on data with more features. Hasan et al. [62] designed the predictor SulCysSite using the random forest algorithm, to identify protein S-sulfenylation sites with an AUC value of 0.817. Jia et al. [31] developed the predictor pSuc-Lys based on RF algorithm to recognize lysine succinylation sites in proteins with an accuracy of 90.83%. Support vector machine (SVM) is a supervised learning model that map

input samples to high-dimensional space by kernel functions and searches for optimal hyperplane for classification. Zhao et al. [63] developed a new bioinformatics tool named PGluS based on the SVM algorithm for S-glutathionylation sites. The performance of PGluS was measured with an accuracy of 71.41% and an MCC of 0.431. Chen et al. [64] created the prediction tool GSHSite with support vector machine classifier to identify S-glutathionylation sites.

Given the critical regulatory role of ubiquitination, more and more researchers have invested in the prediction of ubiquitination sites and have made significant progress. Tung and Ho [21] developed the UbiPred ubiquitination sites prediction tool, by using SVM with the feature set of 31 informative physicochemical properties selected by IPMA, which can improve the accuracy from 72.19% to 84.44%. Radivojac et al. [65] used amino acid components and physicochemical properties to extract 586 protein sequence features and designed the ubiquitination sites prediction tool UbPred, the accuracy of UbPred reached 72% and an AUC value of 0.8. Cai et al. [66] encoded protein sequences based on PSSM conservation scores, amino acid factors and disorder scores of the surrounding sequence. The mRMR was employed to select optimal features, and the nearest neighbor algorithm (NNA) was chosen as a classifier. The experiments indicated that Mathews correlation coefficient (MCC) of their method was higher than the values of the prediction tools UbPred and UbiPred. Chen et al. [67] by using the composition of k-spaced amino acid pairs (CKSAAP) for feature extraction and developed the predictive tool CKSAAP_UbSite in combination with a support vector machine. Accuracy and MCC of CKSAAP_UbSite reached 73.40% and 0.4694, respectively. Because the application of CKSAAP_UbSite is limited to the proteome of yeast, they also developed the human ubiquitination prediction tool hCKSAAP_UbSite [68], with an AUC value of 0.770. In 2013, Chen et al. [69] systematically analyzed the features of pupylation sites sequential, structural and evolutionary in prokaryotic proteins, the ubiquitination sites of prokaryotes and eukaryotes were compared in detail. In 2014, the research team analyzed the algorithm and feature of different predictive tools in detail, and ubiquitination sites in Saccharomyces cerevisiae, Homo sapiens, Mus musculus, and Arabidopsis thaliana were analyzed, discussing the necessity of species-specific ubiquitination sites prediction [70]. Nguye et al. [71] used amino acid composition (AAC), amino acid pair composition (AAPC) and evolutionary information to extract the features from the protein sequences. The support vector machine (SVM) was applied to generate the prediction model for ubiquitination sites identification, and five-fold cross-validation showed that the SVM model has better generalization ability. Wang et al. [72] proposed an evolutionary screening algorithm (ESA) to extract the physicochemical properties of protein sequences. The SVM was used to establish a prediction model ESA-UbiSite, prediction accuracy reached 92%. Lee et al. [73] established the UbSite of ubiquitination sites prediction using SVM, improved the prediction accuracy of ubiquitination

sites.

Although a series of research achievements have been obtained in the prediction of protein ubiquitination sites by statistical and machine learning methods, there is still much room for improvement. First of all, the influence of the feature information of protein sequences on the recognition of ubiquitination sites has not been expounded. The prediction methods based on amino acid sequence feature information still has excellent development potential. Secondly, the fusion of multiple features will generate redundancy and noise information. How to choose the appropriate dimension-reduction method to retain different features information effectively is also one of the challenges we face. Finally, the data of the experimental identification ubiquitination sites has been significantly increased, and there is no effective prediction method and tool.

Inspired by this, we propose a novel method for protein ubiquitination sites prediction, called UbiSitePred. First, binary encoding (BE), pseudo-amino acid composition (PseAAC), the composition of k-spaced amino acid pairs (CKSAAP) and position-specific propensity matrices (PSPM) are used to extract protein sequence features. The best model parameters λ , k and m values are determined by five-fold cross-validation. Thus, we can obtain the initial sequence information to distinguish ubiquitination sites from non-ubiquitination sites. Secondly, compared with Mutual information, Elastic net, Extra-trees, MRMD and LASSO feature selection methods, LASSO is used to determine the optimal feature subset, which could remove redundant and uncorrelated features to provide important feature information for the input classifier. Finally, a prediction model for ubiquitination sites based on support vector machine is constructed and compared with five classifiers: Naïve Bayes, K-nearest neighbor, LibD3C, AdaBoost and random forest. The experimental results show that the UbiSitePred method proposed in this paper can significantly improve the predictive power of ubiquitination sites.

According to a recent series of publications [11,74-86], a truly useful sequence-based statistical predictor has been developed for biological systems and should follow the Chou's 5-step rule [87]: (i) construct a baseline dataset to train and test the predictor; (ii) formulating a biological sequence sample with valid mathematical expressions that can truly and adequately reflect the intrinsic correlation with the target to be predicted; (iii) introducing or developing a robust algorithm to calculate predictions; (iv) Perform cross-validation tests correctly and objectively assess the expected accuracy; (v) Establish a user-friendly web server that the public can access. Below, we will explain how to implement these steps one by one.

2. Materials and Methods

2.1. Datasets

To fairly evaluate the prediction model performance of the lysine ubiquitination sites and compare with other literatures, it is necessary to select an objective and representative dataset.

Protein sequences are usually composed of 20 common amino acids, lysine (K) is an essential amino acid that binds to ubiquitin and affects protein function through ubiquitination [88]. To identify whether the lysine (K) is a ubiquitination site, we need to get information about the amino acids around the lysine (K) residue. In this paper, three different datasets of protein ubiquitination sites were selected. Data Set 1, Data Set 2 and Data Set 3 were established by Cai et al. [20]. Data Set 1 was collected from the UniProt database [21], consisting of 157 lysine ubiquitination sites from 105 protein sequences, and the protein sequence containing the ubiquitination site was used as a positive sample. At the same time, for the protein sequence of 3676 lysines without annotated ubiquitination sites, they were used as negative samples. Regardless of whether it is a positive sample protein sequence or a negative sample protein sequence, the sample window size is 13. Finally, 300 protein sequences with central lysine K sites were obtained, and the number of positive and negative samples each accounted for half. Data Set 1 can download via http://iclab.life.nctu.edu.tw/ubipred/. Data Set 2 and Data Set 3 were from the independent testing dataset and training dataset [68], respectively. The redundant sequences were removed using the Blastclust program [89] (ftp://ftp.ncbi.nih.gov/blast/documentsblastclust.html) with a 30% identity cutoff. Data Set 2 and Data Set 3 were composed of 9537 ubiquitination sites from 3852 proteins, then randomly chose the equal number of non-ubiquitination sites as negative samples, and the distance between lysine in the negative sample and ubiquitination sites in the same protein should not be less than 50 amino acids. Data Set 2 consisted of 6838 sequence fragments, 12236 sequence fragments in Data Set 3, and this sample window size is 27. Data Set 2 and Data Set 3 can be downloaded from http://protein.cau.edu.cn/cksaap_ubsite/download/DataSetForhCKSAAP_UbSite.rar. To ensure the unified length of each peptide, a virtual residue ‘X' was used to fill the corresponding positions where there were no sufficient residues. To facilitate the follow-up work, Data Set 1, Data Set 2 and Data Set 3 are represented by Set1, Set2, and Set3, respectively.

2.2. Binary encoding

Binary encoding (BE) mainly reflects the types and relative positions of amino acids around ubiquitination sites and non-ubiquitination sites in protein sequences. Binary encoding is a simple encoding scheme that transforms the substrate sequences character signals into numerical signals by using an orthonormal encoding scheme, which extracts the features information of 20 common amino acid residues and residue X, according to the order of ACDEFGHIKLMNPQRSTVWYX [49]. Every amino acid residue in the sample sequence fragment is transformed into a 21-dimensional binary feature vector. For example, alanine A is represented by (100000000000000000000), tyrosine Y is represented by (000000000000000000010), the virtual residue X is encoded as the vector (000000000000000000001). Therefore, for each sequence

fragment with the window size of n, resulting in a $21\times n$-dimensional feature vector.

2.3 Pseudo-amino acid composition

Using the high-throughput tools that have been developed, it is possible to extract information on newly discovered protein sequences in time for basic research and drug development. Based on the amino acid composition method, Chou et al. [90] fused the sequence information of amino acids with the physicochemical information of amino acids to propose a method of pseudo-amino acid composition. This method maps protein sequences to the following feature vectors:

$$P=[p_1,p_2,\cdots,p_{20},p_{20+1},\cdots,p_{20+\lambda}]^T \tag{1}$$

Each of these components is given as follows:

$$P_u=\begin{cases}\dfrac{f_u}{\sum_{u=1}^{20}f_u+\omega\sum_{k=1}^{\lambda}\tau_k}, & 1\le u\le 20\\ \dfrac{\omega_{\tau_{u-20}}}{\sum_{u=1}^{20}f_u+\omega\sum_{k=1}^{\lambda}\tau_k}, & 20+1\le u\le 20+\lambda\end{cases} \tag{2}$$

where ω is the weight factor, which was set at 0.05 in [90], f_u is expressed as the frequency of occurrence of the amino acid in the u sequence of the protein sequence in the sequence. It can be seen from the above formula that the first 20 dimensions of the feature vector are the amino acid composition, and the latter λ dimension is the sequence correlation factor reflecting the different levels of the amino acid sequence information. Sequence related factors are obtained by the physicochemical properties of amino acids. At present, researchers have applied the pseudo-amino acid composition method vary widely, and this method is widely used in proteomics [91-97]. In particular, Chou et al. built a very powerful web server called "Pse-in-One" [98] and its updated version "Pse-in-One 2.0" [99], which converts protein, peptide, DNA and RNA sequences into the required numerical vector. In this paper, the feature extraction of protein sequences was performed using the PseAAC online server developed by Chou et al. [100]. On this server, the optimal λ can be determined from the accuracy of the prediction result by selecting a different parameter λ.

2.4. Composition of k-spaced amino acid pairs

The CKSAAP encoding strategy means that it calculates the frequencies of the k-spaced amino acid pairs for each given peptide fragment, and amino acid pairs distance information and amino acid pairs composition information was taken into account, which reflects the biological characteristics near the ubiquitination modification sites of the protein. At present, CKSAAP encoding is not only used for the prediction of phosphorylation sites [27] but also applied to the study of pupylation sites [101] and N-formylation sites [102]. For example, AxxxG represents an

amino acid pair composed of alanine and glycine, separated by three amino acids of any type, and $k=3$ indicates the space between residue pairs. For different k, there are 441 k-spaced amino acid residue pairs for AA, AC, ..., AX, ..., XA, XC, ..., XX. For any given protein sequence with the window size of w, the feature vector can be defined as:

$$(\frac{N_{AA}}{N_{total}},\frac{N_{AC}}{N_{total}},\cdots\cdots,\frac{N_{XX}}{N_{total}}) \tag{3}$$

where $N_{i,j}$ represents the number of amino acid pairs at space of k,and N_{total} represents the number of amino acid pairs with distance k in the protein sequence with the window size of w, so we can know $N_{total}=w-k-1$. In this paper, k is 0, 1, 2, …, 11, and the optimal parameter k is 6. For Set1, Set2, and Set3, the total dimension of the CKSAAP-based feature vectors is 3,087.

2.5. Position-specific propensity matrices

Position-specific propensity matrices (PSPM), proposed by Xu et al. [103] in 2013, uses the position-specific propensity of amino acid pairs to construct vectors. The dataset is divided into the positive dataset and negative dataset according to whether it contains ubiquitination sites. When the sample fragment length of the positive dataset is n, and the space between the amino acid pairs is 0, we will get a position-specific dipeptide composition matrix of $441\times(n-1)$ for A^+. The $j-th$ column of A^+ is $A_j^+=(a_{1,j}^+,a_{2,j}^+,a_{3,j}^+,\cdots\cdots,a_{441,j}^+)'$, $a_{i,j}^+$ denotes the frequency of the $i-th$ dipeptide in $j-th$ column of the positive dataset. Similarly, we can obtain the frequency matrix A^- corresponding to the negative dataset, the position-specific propensity matrix with the size of $441\times(n-1)$ is given by the formula:

$$Z_{i,j}^{m=0}=a_{i,j}^+-a_{i,j}^- \tag{4}$$

Where m represents the space among residue pairs. Repeat the above steps to calculate the position-specific propensity matrix between two amino acids pairs separated by m, correspondingly we obtain a matrix $Z_{m=1},Z_{m=2},\cdots,Z_{m=n-2}$ of size $441\times(n-2),441\times(n-3),\cdots,441\times1$. The matrix $Z_{m=0},Z_{m=1},Z_{m=2},\cdots,Z_{m=n-2}$ obtained above is calculated as follows to obtain a position-specific propensity matrix with size $441\times\frac{n\times(n-1)}{2}$.

$$Z=Z_{m=0}\oplus Z_{m=1}\oplus Z_{m=2}\oplus\cdots\oplus Z_{m=n-2}=\begin{pmatrix} z_{1,1}^{m=0} & z_{1,2}^{m=0} & \cdots & z_{1,n-1}^{m=0} & z_{1,1}^{m=1} & \cdots & z_{1,n-2}^{m=1} & \cdots & \cdots z_{1,1}^{m=n-2} \\ z_{2,1}^{m=0} & z_{2,2}^{m=0} & \cdots & z_{2,n-1}^{m=0} & z_{2,1}^{m=1} & \cdots & z_{2,n-2}^{m=1} & \cdots & \cdots z_{2,1}^{m=n-2} \\ \vdots & \vdots & & & \vdots & & \vdots & & \vdots \\ z_{441,1}^{m=0} & z_{441,2}^{m=0} & \cdots & z_{441,n-1}^{m=0} & z_{441,1}^{m=1} & \cdots & z_{441,n-2}^{m=1} & \cdots & \cdots z_{441,1}^{m=n-2} \end{pmatrix} \tag{5}$$

According to the above steps, the position-specific propensity matrix Z is obtained, for any given protein sequence, the corresponding feature vector can be obtained after being compared with Z.

2.6. LASSO

Given the dataset $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$, where $x\in R^d, y\in R$, subject to the square error as the loss function, the optimization objective is

$$\min_w \sum_{i=1}^{m}(y_i - w^T x_i)^2 \tag{6}$$

Eq. (6) is a general linear regression. To reduce the over-fitting risk, Tibshirani [104] proposed the least absolute shrinkage and selection operator (LASSO) in 1996. The basic idea is to introduce ℓ_1 norm regularization from minimizing residual sum of squares. The LASSO sparse representation coefficient w can be described as follows:

$$J(w)=\min_w \sum_{i=1}^{m}(y_i - w^T x_i)^2 + \gamma\|w\|_1 \tag{7}$$

where the regularization parameter γ controls the penalty of sparse coefficient estimation. $\|w\|_1$ is ℓ_1 norm, and the sparse solution of w means that only the non-zero component w corresponding to the initial feature will appear in the final model. $\gamma\geq 0$ is an adjustable parameter, when the γ value is large enough, there will be more "sparse" solutions, some low-correlation coefficients will be compressed to 0 to remove these variables and achieve the purpose of feature selection; when the value of γ is small, the impact of the regularization constraint is relatively small, in which case all attributes will be selected. In this paper, we set $\gamma=0.005$ and use the coordinate gradient descent method for optimization.

2.7. Support vector machine

Support vector machine (SVM) is an effective machine learning algorithm based on statistical learning theory, which was first proposed by Vapnik [105]. It is widely used in various fields of bioinformatics research, including prediction subcellular localization [106-109], prediction of protein submitochondrial locations [110], prediction of protein structural [111], protein-protein interactions prediction [112], protein fold recognition [113], prediction of protein post-translational modification sites [70,73], prediction of membrane protein types [114] and other protein function research [115]. The basic idea is to find the hyperplane in the samples space and make the samples of different classes linearly separable. At the same time, we can find the optimal classifying hyperplane that can correctly divide the samples into maximal margin and minimal error. If the original sample space is nonlinearly separable, the SVM maps the input space to the high-dimensional feature space through kernel function, so that the sample data becomes linearly separable. At this time, the optimal classification hyperplane needs to satisfy:

$$\begin{aligned} &\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ &s.t.\ \ y_i(w^T\phi(x_i)+b)\geq 1-\xi_i, i=1,2,\cdots,m \end{aligned} \tag{8}$$

where C is the penalty factor, the above problem is solved by the Lagrange method to obtain the

final classification discriminant function:

$$f(x) = \mathrm{sgn}\{\sum_{i=1}^{m} a_i y_i K(x_i, x) + b\} \tag{9}$$

where a_i is the Lagrange multiplier, b is the classification threshold, and $K(x_i, x) = <\varphi(x_i) \cdot \varphi(x)>$ is the kernel function. The commonly used kernel functions include linear kernel function, polynomial kernel function, radial basis kernel function, and the sigmoid kernel function. In particular, the radial basis kernel function can better solve the problem of nonlinear, whether small sample or large sample dataset, high dimensional or low dimensional, etc., which shows excellent prediction performance. The SVM with radial basis kernel functions is also widely used in sites prediction, such as the S-palmitoylation sites [116], the cysteine prenylation sites [117] and lysine phosphoglycerylation sites [118]. This paper uses the support vector machine algorithm in Scikit-learn [119].

2.8. Performance evaluation and model construction

The methods for evaluating the effectiveness of the model include self-consistency, independent test, and k-fold cross-validation. Five-fold cross-validation was carried out in order to evaluate the performance of the model, the datasets were randomly divided into five mutually exclusive subsets of similar size, each time one of them is used as a test set, and the other four are used as training sets for the training classifier, and the cross-validation process was repeated five times. The average value of five-fold cross-validation tests was used as the verification result of the performance of the classifier.

To assess the performance of the predictive model more intuitively, sensitivity (Sn), specificity (Sp), overall accuracy (ACC) and Matthews correlation coefficient (MCC) were used to evaluate the prediction results. Sensitivity and specificity represent the ability of the model to predict positive and negative samples correctly, and overall accuracy is the ratio of the number of samples correctly classified to the total number of samples, the Matthews correlation coefficient balances the predictive performance of the metric model even for different amounts of datasets. The four evaluation metrics which are formulated as [36,40,43,74-76,86]:

$$Sn = 1 - \frac{N_-^+}{N^+} \tag{10}$$

$$Sp = 1 - \frac{N_+^-}{N^-} \tag{11}$$

$$ACC = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \tag{12}$$

$$MCC = \frac{1 - \frac{N_-^+ + N_+^-}{N^+ + N^-}}{\sqrt{(1 + \frac{N_+^- - N_-^+}{N^+})(1 + \frac{N_-^+ - N_+^-}{N^-})}} \tag{13}$$

where N^+ denotes the number of ubiquitination sites to be investigated, N^- represents the number of non-ubiquitination sites to be investigated, N_-^+ indicates the number of true ubiquitination sites which are incorrectly predicted as non-ubiquitination sites, N_+^- represents the number of non-ubiquitination sites which are incorrectly predicted as ubiquitination sites.

According to (10), (11), (12), and (13), we can see: when $N_-^+ = 0$, it means that none of the true ubiquitination sites was mispredicted to be of non-ubiquitination sites, we have the sensitivity $Sn = 1$. When $N_-^+ = N^+$, it means that all the true ubiquitination sites were incorrectly predicted to be non-ubiquitination sites, we have the sensitivity $Sn = 0$. Likewise, when $N_+^- = 0$, it means that none of the non-ubiquitination sites was mispredicted to be of ubiquitination sites, we have the specificity $Sp = 1$; Whereas $N_+^- = N^-$, it means that all the non-ubiquitination sites were incorrectly predicted to be of true ubiquitination sites, we have the specificity $Sp = 0$. When $N_-^+ = N_+^- = 0$, it means that none of the true ubiquitination sites in the positive dataset and none of the non-ubiquitination sites in the negative dataset was incorrectly predicted, we have the overall accuracy $ACC = 1$ and $MCC = 1$; whereas $N_-^+ = N^+$ and $N_+^- = N^-$ means that all the true ubiquitination sites in the positive dataset and none of the non-ubiquitination sites in the negative dataset were incorrectly predicted, we have the overall accuracy $ACC = 0$ and $MCC = -1$; Whereas $N_-^+ = N^+/2$ and $N_+^- = N^-/2$, we have the overall accuracy $ACC = 0.5$ and $MCC = 0$ means no better than random prediction. As we can see from the discussion above, it is much more intuitive and easier to understand when using formula (10), (11), (12), and (13) to examine a predictor for its sensitivity, specificity, overall accuracy, and Mathews correlation coefficient.

Either the set of traditional metrics copied from math books or the intuitive metrics derived from the Chou's symbols [120] is valid only for the single-label systems (where each sample solely belongs to one class). For the multi-label systems (where a sample may simultaneously belong to several classes), whose existence has become more frequent in system biology [80,121-126], system medicine [77,78] and biomedicine [127], an entirely different set of metrics as defined in [128] is needed.

Also, the receiver operating characteristic (ROC) curve based on Sn and 1-Sp is commonly used to assess the discrimination ability of a classifier. The area under the ROC curve (AUC) is an indicator to measure the robustness of the prediction model, the closer the AUC value is to 1, the better the model performs.

For convenience, the ubiquitination sites prediction method we propose in this paper is called UbiSitePred, and the calculation flow is shown in Fig. 1. The experimental environment is Windows Server 2012R2 Intel (R) Xeon (TM) CPU E5-2650 @ 2.30GHz 2.30GHz with 32.0GB of RAM, MATLAB2014a and Python 3.6 programming implementation.

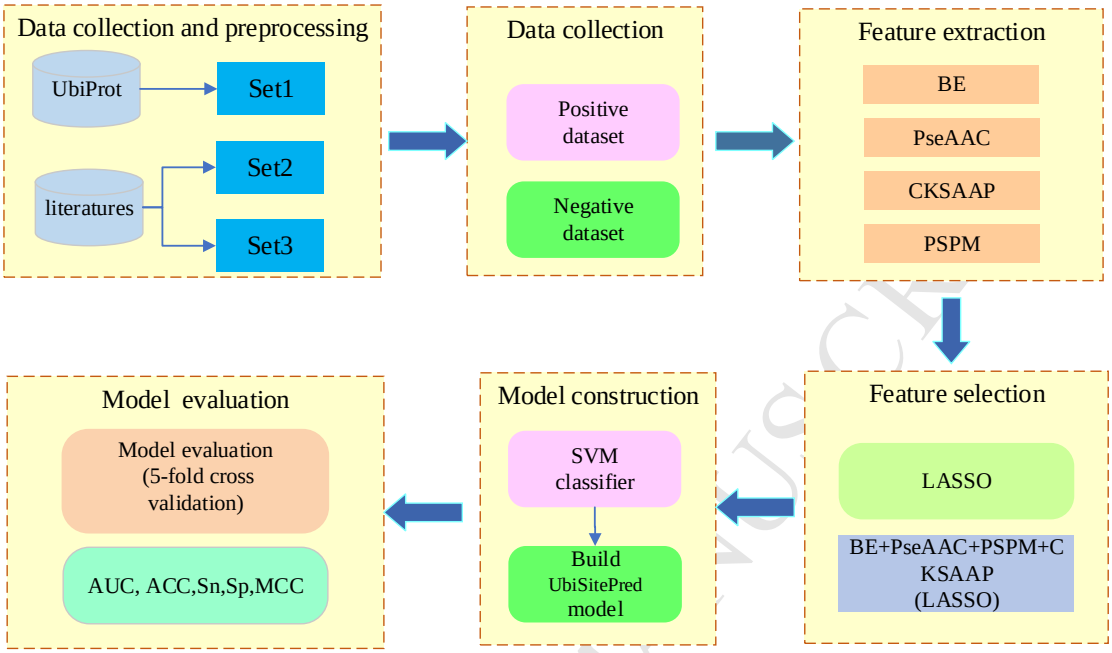


Fig. 1. Flowchart of the UbiSitePred prediction method.

The prediction steps of the UbiSitePred method can be described as:

1) Get the Set1, Set2, and Set3, enter the positive and negative samples of the protein sequences and the corresponding class labels in the model.
2) Feature extraction. The protein sequence is treated as a special string, and the character signal is converted into a numerical signal by coding. (a) Use the BE to extract protein sequences features. b) Generating $20+\lambda$ dimensional feature vectors using the PseAAC algorithm. (c) Feature extraction of protein sequence information using the CKSAAP. (d) Encoding protein sequences using PSPM. Then the four extracted features are combined, each protein sequence in the Set1 with the 3462-dimensional vector; each of the sequence in Set2 and Set3 constructs a 3910-dimensional vector space.
3) Feature selection. For the extracted protein feature vector, LASSO is used to remove redundancy and noise information, and the optimal feature subset is filtered through five-fold cross-validation to provide good feature information for the SVM classifier.
4) According to steps 2) and 3), the selected optimal feature subset and the corresponding class labels are input into the SVM classifier to predict the post-translational the ubiquitination sites of the protein sequences.

5) Model performance evaluation. Five-fold cross-validation was used to evaluate and calculate AUC, ACC, Sn, Sp, MCC, draw ROC curve, and evaluate the model's prediction performance.

3. Results and discussion

3.1. Analysis of sequence characteristics

In this paper, we use the Two-Sample Logos [129] (http://www.twosamplelogo.org/cgi-bin/tsl/tsl.cgi) to obtain a comparison of the double sequence identifiers of the datasets Set1, Set2, and Set3. The frequency-based method revealed the amino acid patterns around the ubiquitination site, more clearly elucidating the residues near the ubiquitination site, and clarifying the statistical significance and significant differences in the residues surrounding the ubiquitination site. Two Sample Logos analysis showed significant differences between the protein sequence of the ubiquitination site and the protein sequence of the non-ubiquitination site for the datasets Set1, Set2, and Set3, as shown in Fig. 2, Fig. 3, and Fig. 4, respectively.

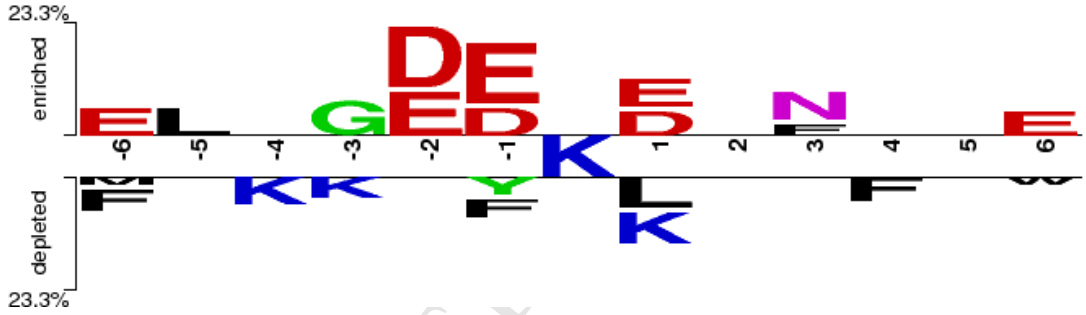


Fig. 2. Comparison of two sequences of amino acid sequences near the ubiquitination and non-ubiquitination sites of Set1.

As presented in Fig. 2, there is a significant difference between the ubiquitination and non-ubiquitination sites near the lysine residue in the Set1. Near the ubiquitinated lysine, the negatively charged glutamic (E) residue tends to enrich at positions -6, -2, -1, 1, 6, and the negatively charged aspartic (D) residue tends to enrich at position -2, -1, 1. There are no significant amino acids in the enrichment positions -4, 2, 4 and 5, and there are also no significant amino acids in the exhaustion positions -5, -2, 2, 3 and 5. Glutamic acid (E) and aspartic acid (D) are more important in the upstream of the positive dataset, while glutamic acid (E) is also more important in the downstream of the positive dataset. According to these characteristics, it can be inferred that the frequency difference in the appearance of various amino acids in Set1 at different positions near lysine significantly affects the ubiquitination process of the lysine site.

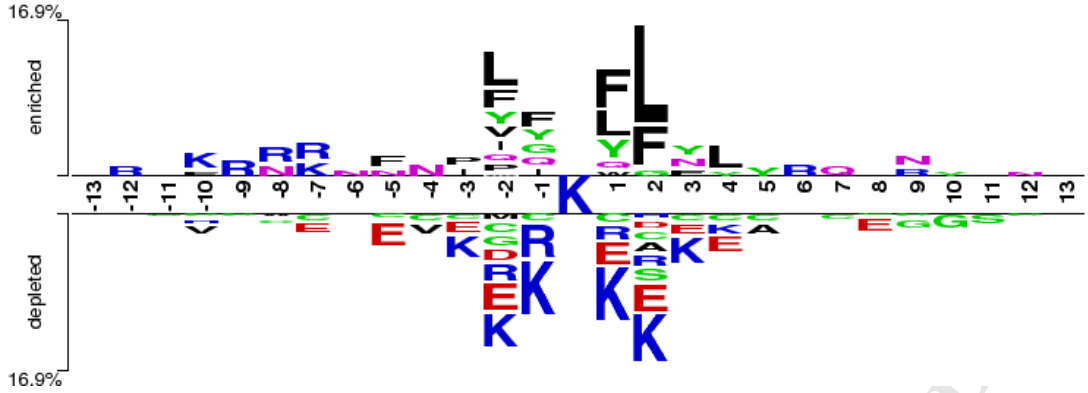


Fig. 3. Comparison of two sequences of amino acid sequences near the ubiquitination and non-ubiquitination sites of Set2.

As can be seen from Fig. 3, there is a significant difference between the ubiquitination sites and non-ubiquitination sites near the lysine residue in the Set2. Compared to other amino acids, the non-polar amino acid leucine (L) is enriched at positions -2, 1, 2, 4, and non-polar phenylalanine (F) frequency of occurrence is higher at positions -5, -2, -1, 1, 2, and the positively charged arginine (R) is significantly higher at positions -12, -9, -8, -7, 6 and 9. There are no significant amino acids at the enrichment positions -13, -11, 8, 11, 13, and there are no considerable amino acids at the depletion positions -13, -12, -11, -9, -6, 6, 12 , 13 as well . Besides, arginine (R) and phenylalanine (F) are of more importance in the upstream of the positive dataset, while leucine (L) and phenylalanine (F) are equally important in the downstream of the positive dataset. Based on these features, it can be inferred that there is a significant difference between the amino acids near the ubiquitination and non-ubiquitination sites in Set2.

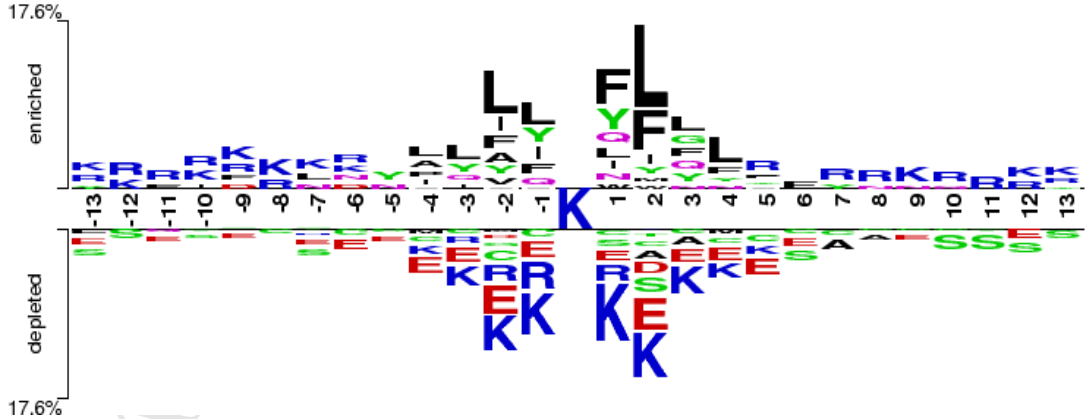


Fig. 4. Comparison of two sequences of amino acid sequences near the ubiquitination and non-ubiquitination sites of Set 3.

As can be seen from Fig. 4, the ubiquitination site of Set3 is near the lysine. Compared with other amino acids, the non-polar amino acid leucine (L) is obviously at the position -2, 1, 2, 4 and the non-polar amino acid phenylalanine (F) occurs more frequently at positions -2, -1, 1, 2, positively charged arginine (R) tends to enrich at position -12, -9, -8, -6, 5, 7, 9, 11, and positively charged lysine (K) is significantly higher at positions -13, -9, -8, -7, 9, 12, 13. Furthermore, lysine

(K) is more important in the upstream of the positive dataset, while leucine (L) and lysine (K) also have a very important role in the downstream of the positive dataset. Based on these characteristics, it can be inferred that there are obvious differences between the amino acids near the ubiquitination and non-ubiquitination sites in Set3.

3.2. Selection of optimal parameters λ value, k value and m value

Extracting effective feature information from protein sequence is a vital step in the prediction model of the protein post-translational modification sites. To better obtain the important feature information, the parameters of the model should be adjusted. We use five-fold cross-validation to determine optimal parameters λ , k and m value of PseAAC, CKSAAP, and PSPM on Set1, Set2, and Set3. The λ value of PseAAC algorithm indicates the proximity of the sequence, i.e., the sequence information of the protein sequence, the CKSAAP, and PSPM algorithm parameter k and m values represent the interval between any two amino acid residues of the protein sequence, which play a crucial role in the construction of the model. If the λ value, the k value, and the m value are set too large, the dimension of the feature vector of the protein sequence will be too high, which will bring more redundant information and affect the predictive performances. If the λ value, the k value, and the m value are set too small, the sequence information contained in the feature vector will be reduced, and the features in the protein sequence cannot be extracted effectively. For the datasets Set1, Set2, and Set3, to find the optimal λ value, k value and m value in the model, the k values and m values are set to 0, 1, 2, 3,4,5, 6, 7, 8, 9, 10, and 11, respectively. Due to the presence of virtual amino acids in the protein sequence, the λ values of 1, 2, 3, 4, 5, 6, and 7 are sequentially set. SVM is employed as a classifier to select optimal λ value, k value, and the m value, using RBF kernel via five-fold cross-validation. We use ACC and AUC to evaluate the prediction performance. The ACC values corresponding to the different parameters λ value in the PseAAC encoding are shown in Table 1. The changes in the ACC values and AUC values corresponding to different λ values in the PseAAC encoding are shown in Fig. 5. The ACC values corresponding to different intervals of k the amino acid residues in the CKSAAP are shown in Table 2. The changes in ACC values and AUC values corresponding to different k values in the CKSAAP are shown in Fig. 6. The ACC values corresponding to different intervals of m the amino acid residues in the PSPM are shown in Table 3, and the changes of the ACC value and the AUC value corresponding to different m values in the PSPM are shown in Fig. 7.

Table 1

ACC values corresponding to different λ values in PseAAC.

ACC(%)	$\lambda=1$	$\lambda=2$	$\lambda=3$	$\lambda=4$	$\lambda=5$	$\lambda=6$	$\lambda=7$
Set1	66.00	65.33	65.67	**68.00**	**68.00**	66.33	67.00
Set2	63.10	63.28	63.50	63.04	**63.60**	63.12	63.16

Set3	65.12	65.64	65.85	65.90	**67.97**	66.14	65.89

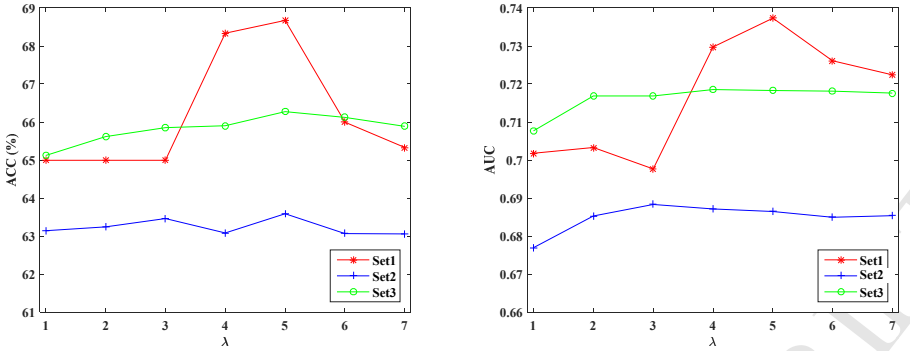


Fig. 5. ACC values and AUC values corresponding to different λ values in PseAAC.

As can be seen from Table 1, by changing the parameter λ values, different prediction effects are obtained. For the dataset Set1, the highest prediction accuracy is 68% at the parameter λ=4 and λ=5, which is 2.67% higher than the parameter λ=2. For the dataset Set2, the highest prediction accuracy is 63.60% at the parameter λ=5, and different parameters have little effect on the prediction accuracy, and both fluctuate around 63%. For the dataset Set3, when the parameter value λ is set to 5, the prediction accuracy is 67.97%. Fig. 5 shows the changes in ACC values and AUC values corresponding to different parameter values in PseAAC encoding. As can be seen from Fig. 5, as the parameter values change, the ACC and AUC values of the Set1, Set2, and Set3 also change. To obtain the optimal parameter λ of the PseAAC algorithm and the parameter λ of the unified model in the post-translational modified ubiquitination site prediction model, the optimal λ value is selected 5 in the model. Therefore, the PseAAC algorithm is used to extract the features of the protein sequence, and each protein sequence generates a 25-dimensional feature vector.

Table 2

ACC values obtained of different k values in the CKSAAP.

ACC (%)	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Set1	65.33	66.00	67.67	66.33	66.33	67.33
Set2	60.34	61.63	62.21	61.89	61.70	62.72
Set3	63.03	64.56	64.34	64.74	64.79	65.21
ACC (%)	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$	$k=11$
Set1	**68.33**	67.00	65.33	66.00	67.67	61.00
Set2	62.81	**63.16**	62.47	62.65	62.33	62.36
Set3	65.25	65.27	**65.45**	65.18	64.91	65.12

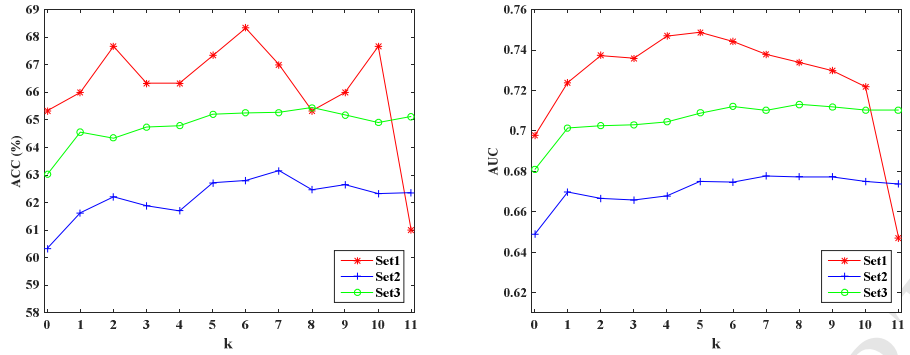


Fig. 6. ACC and AUC values for different k values in the CKSAAP.

Since the corresponding dimension of the CKSAAP encoding of the protein sequence fragment is relatively large, and the window size of a sample is 13. Therefore, the interval k value of the amino acid pair in the CKSAAP code is set to 0 to 11 in order. From Fig 6, it can be seen intuitively that with the increasing interval of k amino acid residues in Set1, the values of ACC and AUC constantly change. The fluctuation of ACC value is more obvious, and the highest prediction accuracy is achieved when the interval is $k = 6$. The change of the AUC value is in a state of rising first and then decreasing, which reaches the maximum value when the interval is $k = 5$. Set2 and Set3 increase the value of ACC and AUC in a relatively stable growth state with the increase of amino acid residue interval k. Set2 achieves the highest prediction accuracy when the interval is $k = 7$, and Set3 achieves the highest prediction accuracy when the interval is $k = 8$. For the AUC value, Set2 and Set3 reach the maximum at intervals of 7 and 8, respectively. Although the three datasets have different values for achieving the highest prediction accuracy and maximum AUC value, to unify the parameter of the model, we choose the optimal parameter $k = 6$. From Table 2, the value of ACC in Set1 reaches a maximum of 68.33% at an interval value of $k = 6$. In the dataset Set2, the accuracy ACC value increases from 60.34% corresponding to $k = 0$ at the beginning to 62.81% when $k = 6$. For the dataset Set3, the ACC value was 65.25% at an interval value of $k = 6$, which was an increase of 2.22% compared to the initial amino acid interval of $k = 0$. Considering the effect of the CKSAAP coded median on the datasets Set1, Set2, and Set3, we select the amino acid pair interval of 6. The CKSAAP encode considers the additive effect, for example, $k = 3$ considering the cumulative effect of the amino acid pair interval 0, 1, 2, 3 on the predictive power of the support vector machine. When the amino acid in the CKSAAP code is $k = 6$ for the optimal interval, the CKSAAP coded dimensions corresponding to the datasets Set1, Set2, and Set3 are all 3,087-dimensions.

Table 3

ACC values corresponding to different m values in PSPM.

ACC (%)	$m = 0$	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$
Set1	88.00	91.67	95.00	97.33	96.33	95.33
Set2	57.96	62.06	64.17	65.66	66.88	67.52

Set3	58.39	60.53	62.59	63.48	63.64	64.44
ACC (%)	$m=6$	$m=7$	$m=8$	$m=9$	$m=10$	$m=11$
Set1	97.00	97.67	97.67	98.00	98.33	**99.00**
Set2	68.29	69.67	70.21	71.18	**73.03**	72.61
Set3	64.29	64.75	65.67	66.25	**67.66**	66.17

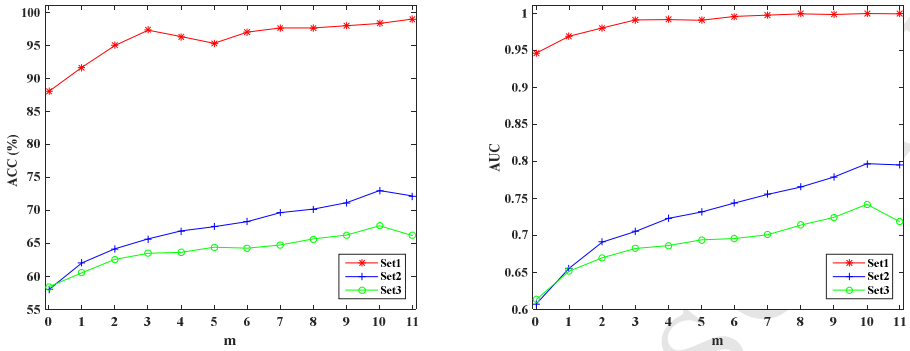


Fig. 7. ACC values and AUC values corresponding to different m values in PSPM.

Different interval values of the amino acid pairs in the PSPM encoding have different effects on the prediction performance of the support vector machine classifier. Considering that the length of the Set1 sample sequence is 13, therefore, the values of amino acid residue intervals in the PSPM coding are set to 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11, respectively. It can be seen from Fig. 7 that the ACC and AUC values of the datasets Set1, Set2, and Set3 change as the amino acid interval m changes, and the prediction accuracy of Set1 slightly falls when the amino acid residue interval is $m=4$ and $m=5$, when the interval is $m=11$ the highest prediction accuracy is achieved. For Set2 and Set3, when the interval is 0 to 10, the corresponding ACC values are growing as the interval increases. When the interval is $m=10$, the highest prediction accuracy is achieved. However, when the interval is $m=11$, the prediction accuracy rate begins to decrease. The AUC value of Set1 increases with the amino acid residue spacing of the protein sequence increases, reaching a maximum at intervals of $m=11$. For Set2 and Set3, when the interval is 0 to 10, the corresponding AUC value keeps increasing with the increase. When the value is $m=10$, the AUC value reaches the maximum, but when it is $m=11$, the AUC value starts to decrease. Considering comprehensively, choose the value corresponding to the maximum value of ACC and AUC as the best parameter in PSPM coding, so choose $m=10$ as the parameter. From Table 3, it can be seen that when the interval value in the dataset Set1 is $m=10$, the value of ACC is 98.33%, which is 10.33% higher than that when the interval of the amino acid is $m=0$. For Set2, the ACC value increases from 57.96% for the interval $m=0$ to 73.03% for the interval $m=10$. For the Set3, the value of ACC increases with the interval m, reaching a maximum of 67.66% at an interval of $m=10$. Similar to CKSAAP, the spacing in PSPM coding also refers to the cumulative effect of amino acid pair spacing on the prediction performance of

the support vector machine. When the amino acid pair interval value is $m = 10$, the datasets Set1, Set2, and Set3 correspond to the dimensions 77, 231, and 231, respectively.

3.3. Feature extraction

Using feature extraction algorithms to extract valuable information from protein sequences is a vital link in the prediction of post-translational modification sites. BE reflects the position-specific information of amino acid residues around the ubiquitination site, PseAAC can avoid completely losing the sequence-pattern information for proteins, CKSAAP mirrors the interaction of residues inside the sequence surrounding the ubiquitination site, PSPM uses position-specific amino acid pairs to extract feature information. In this paper, six feature extraction algorithms have been selected, including four separate feature coding methods (BE, PseAAC, CKSAAP, PSPM) and one hybrid feature coding method called All (BE+PseAAC+CKSAAP+PSPM) and feature encoding based on LASSO dimension reduction Optimal (BE+PseAAC+CKSAAP+PSPM (LASSO)). To verify the effectiveness of hybrid feature, support vector machine was selected as a classifier to predict ubiquitination sites, and AUC, ACC, Sn, Sp, and MCC were used as metrics to evaluate the power of the prediction model. Because the dataset after fusion has a large number of dimensions and redundancy, the Optimal feature extraction method is used as a comparison.

For the datasets Set1, Set2, and Set3, six feature-encoded feature sets are respectively input into the support vector machine classifier for ubiquitination sites prediction, and five-fold cross-validation method is used for evaluation. The prediction results of the different feature extraction algorithms for the datasets Set1, Set2, and Set3 are shown in Table 4, Table 5, and Table 6, respectively.

Table 4

Prediction results of different features of the dataset Set1.

Set1	AUC	ACC (%)	Sn (%)	Sp (%)	MCC
BE	0.6446	61.67	59.33	64.00	0.2343
PseAAC	0.7373	67.67	65.33	70.00	0.3584
CKSAAP	0.7442	68.33	60.00	76.67	0.3753
PSPM	**0.9998**	**98.33**	**99.33**	**97.33**	**0.9674**
All	0.9080	83.00	78.00	88.00	0.6643
Optimal	**0.9998**	**98.33**	**98.67**	**98.00**	**0.9672**

It can be seen from Table 4 that there are four separate feature encoding modes in the Set1. Compared with the BE, PseAAC and CKSAAP encoding methods, the PSPM encoding method has a greater influence on the ubiquitination site prediction. The values of AUC, ACC, Sn, Sp, and MCC corresponding to the PSPM are 0.9998, 98.33%, 99.33%, 97.33%, and 0.9674, respectively. To extract more protein sequence information, four feature coding methods were combined to obtain the feature All, and the corresponding AUC and ACC values were 0.9080 and 83.00%,

respectively. However, the values of AUC, ACC, Sn, Sp, and MCC corresponding to the optimal feature set Optimal are 0.9998, 98.33%, 98.67%, 98.00%, and 0.9672, respectively. The AUC and prediction accuracies are 9.18% and 10.33 % higher than the corresponding values of All, respectively.

Table 5

Prediction results of different features of the dataset Set2.

Set2	AUC	ACC (%)	Sn (%)	Sp (%)	MCC
BE	0.7124	65.12	64.67	65.57	0.3030
PseAAC	0.6865	63.50	64.58	62.42	0.2701
CKSAAP	0.6748	62.81	60.4	65.22	0.2567
PSPM	**0.7967**	**73.03**	**72.19**	**73.88**	**0.4608**
All	0.7819	71.03	69.44	72.62	0.4213
Optimal	**0.8887**	**81.12**	**79.56**	**82.68**	**0.6232**

As can be seen from Table 5 that there are four separate feature encoding methods in the Set2. The PSPM coding method extracts protein sequences as feature vectors and achieves better performance compared with the BE, PseAAC and CKSAAP encoding methods. The values of AUC, ACC, Sn, Sp, and MCC are 0.7967, 73.03%, 72.19%, 73.88%, and 0.4608, respectively. The value of ACC in the PSPM is 7.91%, 9.53%, 10.22% higher than the value of ACC in the BE, PseAAC, CKSAAP, respectively. The feature All corresponding the AUC value is 0.7819, the accuracy is 71.03%, the sensitivity value is 69.44%, and the specificity value is 72.62%, and the MCC is 0.4213. However, the values of the evaluation indexes corresponding to the optimal feature set Optimal to exceed the index values of the four individual codes and the All features, and the values of AUC, ACC, Sn, Sp, and MCC are 0.8887, 81.12%, 79.56%, 82.68% and 0.6232 respectively. The AUC value is 10.68% higher than the index corresponding to the All feature, and the MCC value is 20.19% higher than the index corresponding to the All feature.

Table 6

Prediction results of different features of the dataset Set3.

Set3	AUC	ACC (%)	Sn (%)	Sp (%)	MCC
BE	0.7323	67.42	67.65	67.18	0.3484
PseAAC	0.7183	66.27	68.39	64.15	0.3258
CKSAAP	0.7120	65.25	64.24	66.26	0.3051
PSPM	**0.7421**	**67.66**	**66.74**	**68.59**	**0.3534**
All	**0.7732**	**70.33**	**69.78**	**70.89**	**0.4068**
Optimal	**0.8481**	**76.90**	**76.46**	**77.35**	**0.5382**

From Table 6, we can see that for the dataset Set3, the values of AUC, ACC, Sn, Sp, and MCC corresponding to PSPM encoding are 0.7421, 67.66%, 66.74%, 68.59%, and 0.3534, respectively. PSPM coding method extracts protein sequences as feature vectors and achieves better performance, compared with BE, PseAAC and CKSAAP coding methods. The value of AUC in PSPM coding is 0.98% higher than BE, 2.38% higher than the value of AUC in PseAAC,

and 3.01% higher than the value of AUC in CKSAAP. However, the feature All with an AUC value of 0.7732, an accuracy rate of 70.33%, a sensitivity value of 69.78%, and a specificity value of 70.89%, and the MCC is 0.4068. The values of the evaluation indexes corresponding to the feature set All exceed the index values of the four encode modes, and it shows that the feature fusion can improve the predictive performance of the model to a certain extent. The AUC and ACC of the optimal feature set Optimal are 0.8481, 76.90% respectively, the AUC value is 7.49% higher than the feature set All, and the ACC value is 6.57% higher than the index value corresponding to the All feature.

In summary, for the datasets Set1, Set2, and Set3, the influence of the six feature encoding methods on the robustness of the prediction model is compared. In the single feature encoding method, the PSPM encoding can effectively extract the feature information of the amino acid pair. For the fusion feature encoding method, the All feature extraction method in Set3 not only reflects the positional information of amino acid residues around the ubiquitination sites but also effectively utilizes the amino acid pair information and the physicochemical information of amino acids, the mode information improves the prediction performance to a certain extent. But for Set1 and Set2, the prediction result is not as good as the ideal result, because hybrid feature brings uncorrelated feature information to reduce the prediction of the model performance, so choose LASSO to remove the fusion of the redundancy and noise information and achieve a good ubiquitination sites prediction result.

3.4. The effect of feature selection algorithm

We obtain the initial sequence feature information by fusing BE, PseAAC, CKSAAP and PSPM four feature encoding methods on Set1, Set2, and Set3. But more unrelated features are generated at the same time. In this paper, we choose the optimal feature subset to improve the prediction accuracy of the model. The feature selection methods include Mutual information (MI) [130,131], Elastic net [132], Extra-trees (ET) [133], MRMD [134,135] and LASSO. Selecting different parameter values in LASSO has a different influence on the dimensionality reduction effect. Therefore, to select the optimal feature subset from the fusion feature set All, the parameter γ of LASSO are set to 0.001, 0.002, 0.005, 0.01, 0.015 and 0.02 respectively through five-fold cross-validation. The SVM is a classifier, and the prediction accuracy is the evaluation criterion. The results show that LASSO shows excellent dimensionality reduction effect when the parameter γ value is 0.005. When using the Mutual information, MRMD to reduce the dimension of the features dataset All, the dimension corresponding to the feature subset is consistent with LASSO. The dimensional comparison of the fusion features datasets with Set1, Set2, and Set3 for the five feature selection methods is shown in Fig. 8.

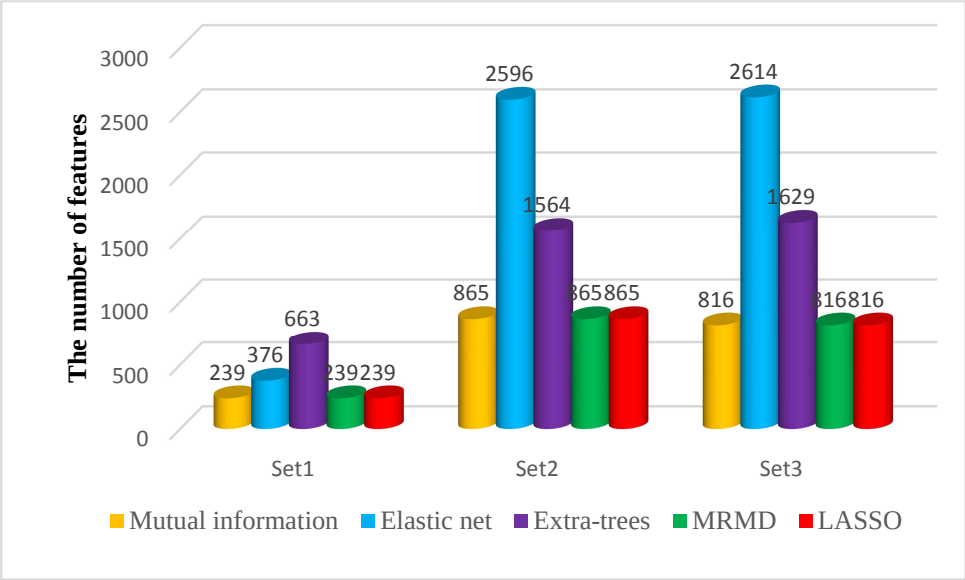


Fig. 8. Comparison of dimensions with five feature selection methods.

We can intuitively compare the five feature selection methods from Fig. 8, and it is found that LASSO can remove more redundant information and select feature subsets more effectively than MI, Elastic net, ET and MRMD. The Set1 initial feature space contains 3462 feature vectors, after MI, Elastic net, ET, MRMD and LASSO dimensionality reduction, the dimensions are 239, 376, 663, 239 and 239, respectively. MI, MRMD, and LASSO significantly remove redundant features information. The Set2 fusion feature datasets contain 3910-dimensional feature vectors, after MI, Elastic net, ET, MRMD and LASSO dimensionality reduction, the dimensions are 865, 2596, 1564, 865 and 865, respectively, and LASSO significantly removes redundant features information. The Set3 contains 3910-dimensional feature vectors, after MI, Elastic net, ET, MRMD and LASSO dimensionality reduction, the dimensions are 816, 2614, 1629, 816 and 816, respectively, and LASSO can effectively retain features.

Although dimension reduction can effectively remove the redundant information in the protein sequence, at the same time, it is hoped that the optimal feature subsets can improve the prediction performance of the model, and then select the optimal feature selection algorithm by comparing the prediction results of different feature selection methods. The support vector machine is selected as a classifier to predict the reduced dimension of MI, Elastic net, ET, MRMD and LASSO, and different dimension reduction methods were measured using AUC, ACC, Sn, Sp, and MCC, as shown in Table 7.

As shown in Table 7, we use five methods to select the optimal feature subset for Set1, the corresponding overall accuracy for MI, Elastic net, ET, MRMD, and LASSO are 97.00%, 98.67%, 95.33%, 64.67%, and 98.33%, respectively. In addition to the MRMD method, the other four feature selection methods AUC value, sensitivity, specificity, and MCC all achieved good prediction results. According to AUC value, sensitivity, specificity, and MCC, LASSO in the Set2

dataset has a good dimensionality reduction effect. The AUC value, ACC, sensitivity, specificity and the MCC is 0.8887, 81.12%, 79.56%, 82.68%, and 0.6232, respectively. The value of AUC is 8.07%, 6.35%, 6.9%, and 16.4% higher than the corresponding values of MI, Elastic net, ET and MRMD. After the Set3 dataset is selected by the LASSO, the AUC value is 0.8481, the overall accuracy rate ACC value is 76.90%, the sensitivity value is 76.46%, the specificity with 77.35% and the MCC is 0.5382. The value of ACC is 6.3%, 3.71%, 4.2%, and 7.81% higher than the value corresponding to dimension reduction through MI, Elastic net, ET and MRMD, respectively.

Table 7

Comparison of prediction results of five feature selection methods.

Datasets	Feature selection methods	Performance metrics				
		AUC	ACC (%)	Sn (%)	Sp (%)	MCC
Set1	Mutual information	0.9980	97.00	96.67	97.33	0.9413
	Elastic net	**1.0000**	**98.67**	**99.33**	**98.00**	**0.9743**
	Extra-trees	0.9971	95.33	95.33	95.33	0.9086
	MRMD	0.6934	64.67	68.00	61.33	0.2946
	LASSO	0.9998	98.33	98.67	**98.00**	0.9672
Set2	Mutual information	0.8080	72.68	70.66	74.70	0.4542
	Elastic net	0.8252	74.58	73.21	75.96	0.4922
	Extra-trees	0.8197	74.22	73.79	74.64	0.4847
	MRMD	0.7247	67.04	67.45	66.63	0.3411
	LASSO	**0.8887**	**81.12**	**79.56**	**82.68**	**0.6232**
Set3	Mutual information	0.7778	70.60	70.12	71.09	0.4122
	Elastic net	0.8062	73.19	72.70	73.68	0.4639
	Extra-trees	0.7993	72.70	72.59	72.82	0.4541
	MRMD	0.7581	69.09	69.42	68.76	0.3819
	LASSO	**0.8481**	**76.90**	**76.46**	**77.35**	**0.5382**

For the datasets Set1, Set2, and Set3, we compare the five dimensionality reduction methods of MI, Elastic net, ET, MRMD, and LASSO. Using Elastic net methods significantly reduces the dimensionality reduction dimension of the dataset Set1, which brings convenience to the calculation, but the prediction performance is not improved, and the datasets Set2 and Set3 have no apparent dimensionality reduction effect. ET method is not the effective removal of irrelevant variables, to provide the best feature subset for classification. To facilitate the comparison of feature selection effects, the MI and MRMD remain the same as LASSO, but the predicted performance is significantly lower than that of LASSO. The LASSO dimension reduction method integrates the feature process with the learner training process by introducing ℓ_1 regularization parameters, which can reduce redundancy and noise information. Five-fold cross-validation shows that LASSO is superior to the predictive performance of MI, Elastic net, ET and MRMD, indicating that the optimal feature subsets determined by LASSO reject not only irrelevant features and redundant features, but also preserve essential features of protein sequences,

effectively reduce the complexity of the model, accelerate the calculation speed, improve the model's prediction accuracy rate, and provide good feature information for the support vector machine classifier. Therefore, we use LASSO to select the optimal feature subsets.

3.5. Selection of classification algorithms

To construct an efficient ubiquitination sites prediction model, the selection of the classification algorithm is of great importance. In this paper, for the feature subset called Optimal, Naïve Bayes (NB), K-nearest neighbor (KNN), LibD3C [136], AdaBoost [137], random forest (RF), and support vector machines (SVM) six classification algorithms are employed to predict the ubiquitination sites of the post-translational modification on the Set1, Set2, and Set3 datasets. The SVM algorithm applies radial basis kernel function. Both the NB algorithm and the AdaBoost algorithm use the default parameters. The Euclidean distance is used in the KNN algorithm, and the number of neighbors is 10. The number of decision trees selected in the RF is 500. The feature subsets of Optimal are input into the classifiers NB, KNN, LibD3C, AdaBoost, RF, and SVM, respectively, and the predictable outcome of the datasets Set1, Set2, and Set3 under different classifiers are shown in Table 8. To analyze the prediction performance of datasets Set1, Set2, and Set3 for different classifiers more intuitively, five-fold cross-validation is employed in the experiment to draw the line for the ACC and AUC values of the ubiquitination sites for the six classifiers, the figures are shown in Fig. 9. In addition, receiver operating characteristic (ROC) curves are used to compare the robustness of different prediction models. Fig. 10, Fig. 11 and Fig. 12 are the ROC curves of the datasets Set1, Set2, and Set3 obtained by the six classifier methods, respectively.

Table 8

Comparison of the prediction results of the six classifiers.

Datasets	Classifiers	AUC	ACC (%)	Sn (%)	Sp (%)	MCC
Set1	NB	0.7161	60.67	74.67	46.67	0.2218
	KNN	0.9924	95.33	97.33	93.33	0.9088
	LibD3C	**1.0000**	**99.67**	**99.33**	**100.00**	**0.9934**
	AdaBoost	0.9993	99.00	98.67	99.33	0.9803
	RF	**1.0000**	**99.67**	**99.33**	**100.00**	**0.9934**
	SVM	0.9998	98.33	98.67	98.00	0.9672
Set2	NB	0.7113	61.33	83.30	39.37	0.2538
	KNN	0.6379	59.37	61.27	57.48	0.1894
	LibD3C	0.7945	72.64	72.71	72.57	0.4528
	AdaBoost	0.8425	75.75	76.08	75.43	0.5152
	RF	0.8421	76.02	77.30	74.73	0.5208
	SVM	**0.8887**	**81.12**	**79.56**	**82.68**	**0.6232**
Set3	NB	0.7099	66.20	54.68	77.72	0.3353
	KNN	0.6633	60.62	50.26	70.99	0.2175
	LibD3C	0.7765	70.63	71.67	69.58	0.4126

AdaBoost	0.7952	72.14	72.93	71.35	0.4429
RF	0.7977	72.37	74.50	70.24	0.4479
SVM	**0.8481**	**76.90**	**76.46**	**77.35**	**0.5382**

As can be seen from Table 8, the LibD3C and RF have the best prediction performance for Set1, the AUC value is 1, the prediction accuracy is 99.67%, the AUC value of the support vector machine is 0.9998, and the prediction accuracy ACC is 98.33%. The AUC value of LibD3C and RF are 0.02% higher than SVM, and the ACC value of LibD3C and RF are 1.34% higher than SVM. Although the prediction performance of SVM does not reach the expected effect, the difference with the random forest classifier is tiny. Naïve Bayes has the lowest forecast accuracy, which is 39% and 38.33% lower than those of RF and SVM, respectively. For the Set2, the best predictive performance is the model established using the support vector machine classifier. The AUC value is 0.8887, which is higher than the AUC values of the other five classifiers,and the accuracy rate is 81.12%. The poor prediction performance is based on the K-nearest neighbor classifier model. The AUC value is 0.6379, and the accuracy rate is 59.37%, 25.08%, and 21.75% lower than support vector machines respectively. The MCC of the SVM model was 0.6232, which was 36.94%, 43.38%, 17.04%,10.8%, and 10.24% higher than those of NB, KNN, LibD3C, AdaBoost, and RF, respectively. For the Set3, the best performance is support vector machine classifier, the AUC value is 0.8481, and the accuracy is 76.90%, which is higher than the AUC and ACC values of the other five classifiers. The poor predictive performance is based on the K-nearest neighbor classifier, and the AUC value is 0.6633, the accuracy rate is 60.62%. The MCC of the SVM model was 0.5382, which was 20.29%, 32.07%, 12.56%, 9.53%, and 9.03% higher than those of NB, KNN, LibD3C, AdaBoost, and RF, respectively. By comparing the AUC value, ACC value and MCC value of the datasets Set1, Set2 and Set3 with six classifiers, the support vector machine has excellent prediction performance and can effectively predict ubiquitination sites.

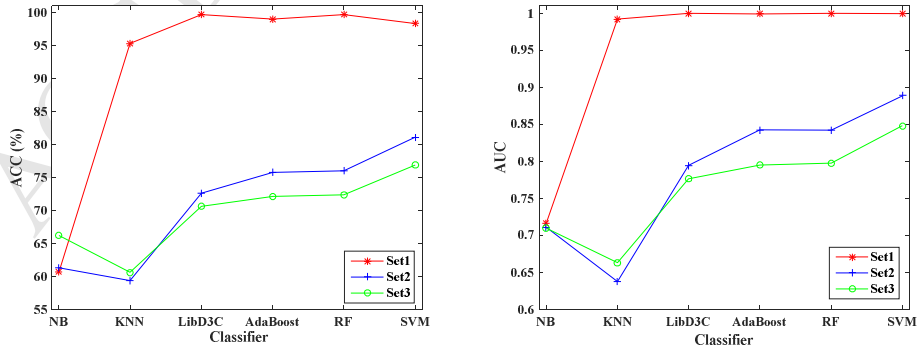


Fig. 9. ACC and AUC values obtained from different classifiers for the Set1, Set2, and Set3.

It can be seen intuitively from Fig. 9 that the datasets Set1, Set2, and Set3 change the ACC values and AUC values of six classifiers, such as Naïve Bayes, K-nearest neighbor, LibD3C, AdaBoost, random forest, and support vector machine. For the accuracy rate ACC varies from 59%

to 99%, the AUC value varies from 0.63 to 1. For the AUC and ACC values, the prediction performance of the classifier SVM is superior to the other five classifiers, and the values of the classifier AdaBoost and the random forest are not much different. The classifier KNN has poor performance in predicting protein ubiquitination sites, and its correlation value changes are relatively significant. The prediction performance of the Naïve Bayes and LibD3C is relatively stable, but the performance is weaker than the support vector machine. For the dataset Set1, the accuracy of the LibD3C and random forest reaches 100%, the Naïve Bayes prediction accuracy rate is the lowest. For Set2 and Set3, the prediction performance of the classifier SVM is better than the other five classifiers, and K-nearest neighbor has the lowest prediction accuracy. Overall, the classifiers AdaBoost, random forest and support vector machine have better prediction performance, while LibD3C is an ensemble classifier with clustering and dynamic selection strategy, and the prediction performance is also good. The classifiers Naïve Bayes and K-nearest neighbor have poor prediction performance for protein ubiquitination sites.

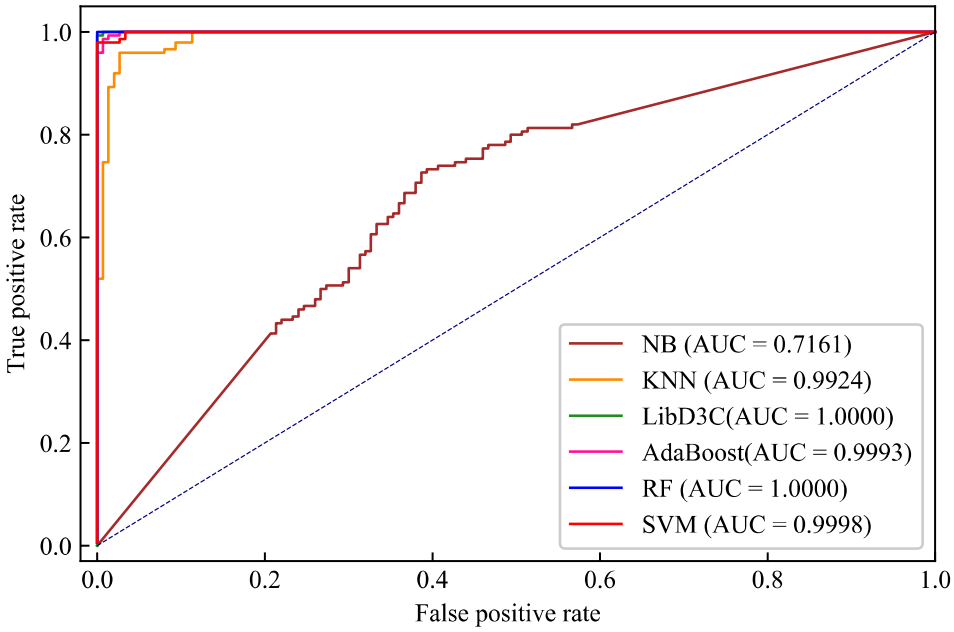


Fig. 10. ROC curve of Set1 about six classifiers.

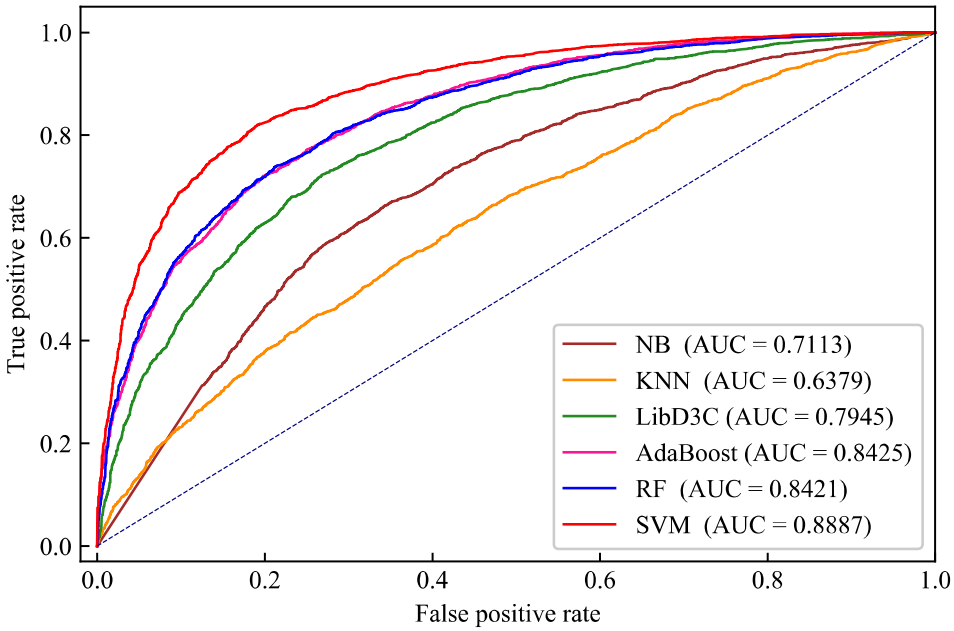


Fig. 11. ROC curve of Set2 about six classifiers.

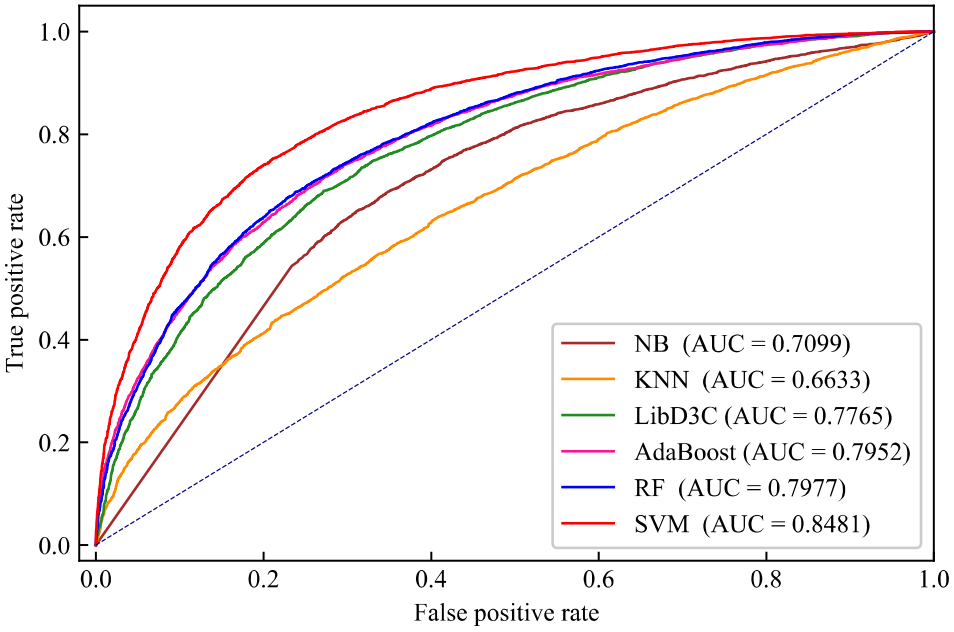


Fig. 12. ROC curve of Set3 about six classifiers.

To evaluate the performance of different methods for ubiquitination sites prediction, the ROC is used as the performance criterion. If the ROC curve of one classifier is entirely enveloped by the curve of another classifier, the latter has better prediction performance than the former. As can be seen from Fig. 10, for the dataset Set1, the ROC curve of LibD3C and RF includes ROC curves corresponding to the classifiers NB, KNN, AdaBoost, and SVM. The AUC values of LibD3C and RF are 28.39% higher than NB. As can be seen from Fig. 11, the Set2 which uses the SVM classifier model has better robustness, and its ROC completely encapsulates the ROC curve corresponding to the classifiers NB, KNN, LibD3C, AdaBoost, and RF. The ROC curves of

AdaBoost and RF coincide, indicates that the predictive performance of the two classifiers is similar. The AUC value of the SVM is 4.62% and 4.66% higher than AdaBoost and RF, respectively. As can be seen from Fig. 12, we select the SVM classifier build model in Set3 which has the best robustness. The AUC value of the ROC curve reaches 0.8481 which is significantly higher than the other five methods, whose AUC value is 18.48% higher than KNN.

We select the best classifier according to five-fold cross-validation on Set1, Set2, and Set3 dataset. For the three datasets, the K-nearest neighbor classifier is simple and effective, but the amount of calculation is relatively large, and the prediction performance is not stable. There are few parameters needed to estimate the Naïve Bayes algorithm, but the assumption of independence between attributes is often not established, and there is still room for improvement. As an ensemble classifier, LibD3C improves prediction performance to some extent. Both AdaBoost and random forest are tree-based ensemble classifiers, so the prediction performance of the two classifiers is not much different. There is still a gap compared to the model of the support vector machine classifier. The computational complexity of the support vector machine depends on the number of support vectors, rather than the dimensions of the sample space. In a sense, it avoids "dimensional disasters," seizes essential samples, eliminates a large number of redundant samples, and is superior to the other five classifiers. Considering comprehensively, the radial basis support vector machine is chosen as the best classifier to predict the protein ubiquitination sites in this paper.

3.6. Comparison with other methods

To more objectively evaluate the predictive performance of the learning model established in this paper, five-fold cross-validation is employed in the same datasets Set1, Set2, and Set3. We compare the proposed method UbiSitePred with Cai et al. [20] which contains the Efficient Bayesian Multivariate Classifier (EBMC), Naïve Bayes (NB), Feature Selection NB (FSNB), Model Averaged NB (MANB), Support Vector Machine (SVM), Logistic Regression (LR), Least Absolute Shrinkage and Selection Operator (LASSO). The AUC values are shown in Table 9 and Fig. 13.

Table 9

Comparison of UbiSitePred with other methods based on AUC.

Datasets	UbiSitePred	EBMC	NB	FSNB	MANB	SVM	LR	LASSO
Set1	0.9998	0.6714	0.5289	0.5613	0.5545	0.6597	0.7244	0.6933
Set2	0.8887	0.6467	0.5330	0.5582	0.5502	0.6039	0.6140	0.6041
Set3	0.8481	0.6667	0.5141	0.5633	0.5192	0.6102	0.6476	0.6129

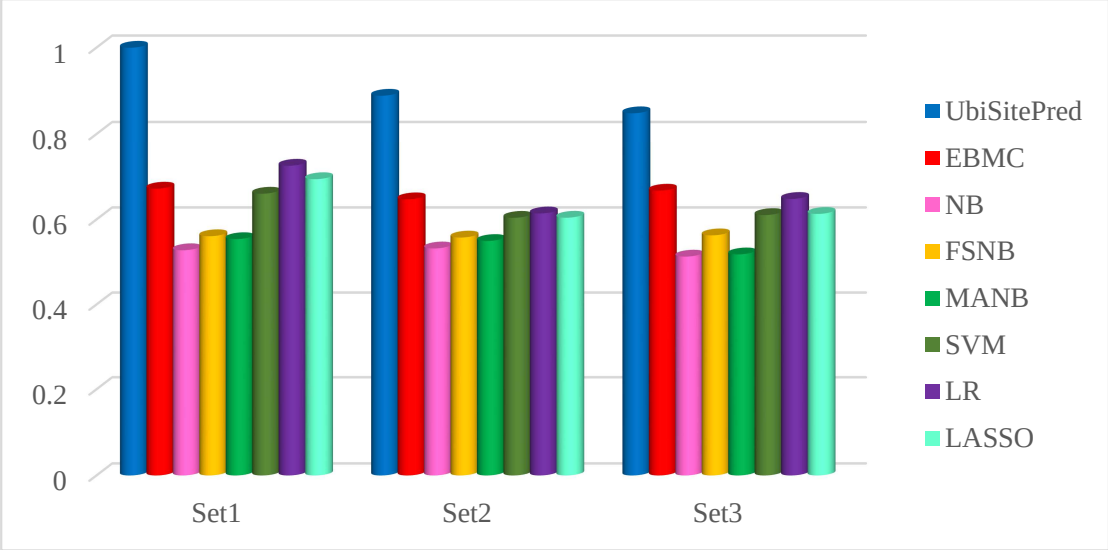


Fig. 13. Comparison of UbiSitePred with other methods AUC.

From Table 9, we can see that for the dataset Set1, the AUC value of the prediction model UbiSitePred proposed by us is 0.9998, which is higher than the AUC value of other methods. The AUC value of LR method is 0.7244, and the AUC value of the NB method is 0.5289. It can be seen that for the dataset Set1, the model UbiSitePred achieves better prediction performance, which is 27.54%-47.09% higher than the AUC of other methods. For the dataset Set2, the AUC value of UbiSitePred, EBMC, and NB is 0.8887, 0.6467 and 0.5330, respectively. It can be seen that for the dataset Set2, the model UbiSitePred achieves a better prediction performance, which is 24.20%-35.57% higher than the AUC of other methods. For the dataset Set3, the AUC value is 0.8481 of the prediction model UbiSitePred proposed by us, which is significantly higher than the AUC value of other methods in Set3. The AUC value is 0.6667 in the EBMC method, and the AUC value corresponding to the NB method is 0.5141. It can be seen that for the dataset Set3, the model UbiSitePred achieves better prediction performance, and the AUC is 18.14%-33.4% higher than Cai et al.[20].

From Fig. 13 for the datasets Set1, Set2, and Set3, it can be seen intuitively that Cai proposed seven methods for predicting ubiquitination sites of proteins, the AUC values of these methods ranged from 0.51 to 0.73. However, the UbiSitePred and AUC values of the prediction model established in this paper are all higher than 0.8, which is better than the other 7 methods. The AUC value is significantly improved, and satisfactory prediction results are obtained.

4. Conclusion

Protein is an important supporter of human physiological activity and physiological function, and protein post-translational modification plays a significant role in the cell's life activities. The study of PTM can help reveal the function oristel and the laws of cell activity such as growth and development, metabolism, signal transduction, differentiation, and apoptosis. Ubiquitination is

very important in a variety of cellular life activities. Using machine learning methods to identify ubiquitination sites is of great significance for the further understanding of the life activities of cells. In this paper, UbiSitePred for protein ubiquitination site prediction is proposed. For the Set1, Set2, and Set3, BE can intuitively acquire the type and positional features of amino acid residues around the ubiquitination and non-ubiquitination sites of lysine, PseAAC fuses the sequence information of amino acids with the physicochemical information of amino acids, CKSAAP takes into account the compositional information of the amino acid pairs, and PSPM extracts the positional features from protein sequence. To improve the performance of ubiquitination sites prediction model, we select the optimal parameters of PseAAC, CKSAAP, and PSPM by five-fold cross-validation. Then, by fusing the four coding feature methods to obtain the feature set All. The LASSO significantly removes redundancy and noise information and effectively retains important feature information for identifying ubiquitination sites. Support vector machine shows good performance in solving non-linear problems and effectively avoid overfitting. Five-fold cross-validation shows the AUC values of Set1, Set2 and Set3 reached 0.9998, 0.8887 and 0.8481, respectively. Compared with other methods, the prediction results of UbiSitePred proposed in this paper are higher than other methods, which indicate that UbiSitePred not only can effectively predict ubiquitination sites, but also it can provide new ideas for the construction of other post-translational modification sites prediction tools for proteins.

As pointed out in [138], user-friendly and publicly accessible web-servers represent the future direction for reporting various important computational analyses and findings [10,12,74,139-142]. They have significantly enhanced the impacts of computational biology on medical science [143], driving medical science into an unprecedented revolution [97]. In our future work, we shall strive to establish a web-server for the new method presented in this paper.

Acknowledgments

The authors sincerely thank the anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China (Nos. 61863010 and 11771188), the Natural Science Foundation of Shandong Province of China (No. ZR2018MC007), and the Project of Shandong Province Higher Educational Science and Technology Program (No. J17KA159). This work was supported by National Science Foundation/EPSCoR Award No. IIA-1355423, the State of South Dakota Research Innovation Center and the Agriculture Experiment Station of South Dakota State University (SDSU). This work was also supported by Hatch Project: SD00H558-15/project accession No. 1008151 from the USDA National Institute of Food and Agriculture. This work used the Extreme Science and Engineering Discovery

Environment (XSEDE), which is supported by the National Science Foundation (grant number ACI-1548562).

References

[1] M. Mann, O.N. Jensen, Proteomic analysis of post-translational modifications, Nat. Biotechnol. 21 (2003) 255-261.

[2] W.R. Qiu, S.Y. Jiang, B.Q. Sun, X. Xiao, X. Cheng, iRNA-2methyl: identify RNA 2′-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier, Med. Chem.13 (2017) 734-743.

[3] W.R. Qiu, S.Y. Jiang, Z.C. Xu, X. Xiao, iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition, Oncotarget 8 (2017) 41178-41188.

[4] Y. Xu, X. Wen, L.S. Wen, L.Y. Wu, N.Y. Deng, iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition, PLoS One 9 (2014) e105018.

[5] W.R. Qiu, B.Q. Sun, X. Xiao, D. Xu, iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory, Mol. Inform. 36 (2017) 1600010.

[6] J. Jia, L. Zhang, Z. Liu, X. Xiao, pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC, Bioinformatics 32 (2016) 3133-3141.

[7] Y. Xu, C. Li, iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC, Med. Chem. 13 (2017) 544-551.

[8] W.R. Qiu, X. Xiao, W.Z. Lin, iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model, J. Biomol. Struct. Dyn. 33 (2015) 1731-1742.

[9] W. Chen, P. Feng, H. Ding, H. Lin, iRNA-Methyl: identifying N6-methyladenosine sites using pseudo nucleotide composition, Anal. Biochem. 490 (2015) 26-33.

[10] W. Chen, H. Ding, X. Zhou, H. Lin, iRNA(m6A)-PseDNC: identifying N6-methyladenosine sites using pseudo dinucleotide composition, Anal. Biochem. (2018), https://doi.org/10.1016/j.ab.

[11] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, iDNA6mA-PseKNC: identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC, Genomics (2018), https://doi.org/10.1016/j.ygeno.

[12] W. Chen, H. Tang, J. Ye, H. Lin, iRNA-PseU: identifying RNA pseudouridine sites, Mol. Ther.-Nucl. Acids 5 (2016) e332.

[13] Y.D. Khan, N. Rasool, W. Hussain, S.A. Khan, iPhosT-PseAAC: identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC, Anal. Biochem.550 (2018) 109-116.

[14] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, J.H. Jia, iKcr-PseEns: identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier, Genomics 110 (2018) 239-246.

[15] T.S. Gao, Z.X. Liu, Y.B. Wang, H. Cheng, Q. Yang, A.Y. Guo, J. Ren, Y. Xue, UUCD: a family-based database of ubiquitin and ubiquitin-like conjugation, Nucleic. Acids Res. 41 (2013) D445-D451.

[16] K. Haglund, I. Dikic, Ubiquitylation and cell signaling, Embo. J. 24 (2014) 3353-3359.

[17] D. Hoeller, C.M. Hecker, I. Dikic, Ubiquitin and ubiquitin-like proteins in cancer pathogenesis, Nat. Rev. Cancer 6 (2006) 776-788.

[18] M.S. Gentry, C.A. Worby, J.E. Dixon, Insights into lafora disease: malin is an e3 ubiquitin ligase that ubiquitinates and promotes the degradation of laforin, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 8501-8506.

[19] J. Peng, D. Schwartz, J.E. Elias , C.C. Thoreen, D. Cheng, G. Marsischky, J. Roelofs, D. Finley, S.P. Gygi, A proteomics approach to understanding protein ubiquitination, Nat. Biotechnol. 21 (2003) 921-926.

[20] B. Cai, X. Jiang, Computational methods for ubiquitination site prediction using physicochemical properties of protein sequences, BMC Bioinform. 17 (2016) 116.

[21] C.W. Tung, S.Y. Ho, Computational identification of ubiquitylation sites from protein sequences, BMC Bioinform. 9 (2008) 310.

[22] 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. 83 (2017) 75-81.

[23] S.Y. Huang, S.P. Shi, J.D. Qiu, M.C. Liu, Using support vector machines to identify protein phosphorylation sites in viruses, J. Mol. Graph. Model. 56 (2015) 84-90.

[24] S.P. Shi, J.D. Qiu, X.Y. Sun, S.B. Suo, S.Y. Huang, R.P. Liang, PMeS: prediction of methylation sites based on enhanced feature encoding scheme, PLoS One 7 (2012) e38772.

[25] Y.Z. Chen, Y.R. Tang, Z.Y. Sheng, Z. Zhang, Prediction of mucin-type o-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs, BMC Bioinform. 9 (2008) 101.

[26] X.B. Wang, L.Y. Wu, Y.C. Wang, N.Y. Deng, Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs, Protein Eng. Des. Sel. 22 (2009) 707-712.

[27] X. Zhao, W. Zhang, X. Xin, Z. Ma, M. Yin, Prediction of protein phosphorylation sites by using the composition of k-spaced amino acid pairs, PLoS One 7 (2012) e46302.

[28] Q. Wuyun, Z. Wei, Y. Zhang, J. Ruan, H. Gang, Improved species-specific lysine acetylation site prediction based on a large variety of features set, PLoS One 11 (2016) e0155370.

[29] L.J. Mcguffin, K. Bryson, D.T. Jones, The PSIPRED protein structure prediction server, Bioinformatics 16 (2000) 404-405.

[30] A. Dehzangi, Y. López, S.P. Lal, G. Taherzadeh, J. Michaelson, A. Sattar, T. Tsunoda, A. Sharma, PSSM-Suc: accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction, J. Theor. Biol. 425 (2017) 97-102.

[31] J. Jia, Z. Liu, X. Xiao, B. Liu, K.C. Chou, pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach, J. Theor. Biol. 394 (2016) 223-230.

[32] J. Jia, Z. Liu, X. Xiao, B. Liu, 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.

[33] 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.

[34] 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.

[35] W.R. Qiu, X. Xiao, Z.C. Xu, iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier, Oncotarget 7 (2016) 51270-51283.

[36] 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.

[37] Y. Xu, X.J. Shao, L.Y. Wu, N.Y. Deng, iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins, PeerJ 1 (2013) e171.

[38] W.R. Qiu, X. Xiao, W.Z. Lin, iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach, Biomed Res. Int. 2014 (2014) 947416.

[39] Y. Xu, X. Wen, X.J. Shao, N.Y. Deng, iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition, Int. J. Mol. Sci. 15 (2014) 7594-7610.

[40] J. Jia, Z. Liu, X. Xiao, B. Liu, iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC, Oncotarget 7 (2016) 34558-34570.

[41] S.Y. Huang, S.P. Shi, J.D. Qiu, X.Y. Sun, S.B. Suo, R.P. Liang, Predsulsite: prediction of protein tyrosine sulfation sites with multiple features and analysis, Anal. Biochem. 428 (2012) 16-23.

[42] 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 (2017) 1457-1463.

[43] Z. Liu, X. Xiao, D.J. Yu, J. Jia, W.R. Qiu, pRNAm-PC: predicting N-methyladenosine sites in RNA sequences via physical-chemical properties, Anal. Biochem. 497 (2016) 60-67.

[44] Y. Saeys, I. Inza, P. Larranaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507-2517.

[45] B. Liu, S. Li, Y. Wang, L. Lu, Y. Li, Y. Cai, Predicting the protein sumo modification sites based on properties sequential forward selection (PSFS), Biochem. Bioph. Res. Co. 358 (2007) 136-139.

[46] S. Niu, T. Huang, K. Feng, Y. Cai, Y. Li, Prediction of tyrosine sulfation with mRMR feature selection and analysis, J. Proteome Res. 9 (2010) 6490-6497.

[47] Y.D. Cai, L. Lu, Predicting N-terminal acetylation based on feature selection method, Biochem. Bioph. Res. Co. 372 (2008) 862-865.

[48] Y. Zhou, T. Huang, G. Huang, N. Zhang, X.Y. Kong, Y.D. Cai, Prediction of protein N-formylation and comparison with N-acetylation based on a feature selection method, Neurocomputing 217 (2016) 53-62.

[49] Z. Ju, J.J. He, Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC, J. Mol. Graph. Model. 76 (2017) 356-363.

[50] T. Wang, W. Zheng, Q. Wuyun, Z. Wu, J. Ruan, G. Hu, J. Gao, PrAS: prediction of amidation sites using multiple feature extraction, Comput. Biol. Chem. 66 (2017) 57-62.

[51] P.P. Wen, S.P. Shi , H.D. Xu, L.N. Wang, J.D. Qiu, Accurate in silico prediction of species-specific methylation sites based on information gain feature optimization, Bioinformatics 32 (2016) 3107-3115.

[52] T. Hou, G.Y. Zheng, P.Y. 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.

[53] X. Chen, M. Chen, K. Ning, BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network, Bioinformatics 22 (2006) 2952-2954.

[54] Y. Xue, H. Chen, C. Jin, Z. Sun, X. Yao, NBA-Palm: prediction of palmitoylation site implemented in Naïve Bayes algorithm, BMC Bioinform. 7 (2006) 1-10.

[55] N. Blom, S.S. Gammeltoft, Sequence and structure-based prediction of eukaryotic protein phosphorylation sites, J. Mol. Biol. 294 (1999) 1351-1362.

[56] Y.R. Tang, Y.Z. Chen, C.A. Canchaya, Z. Zhang, GANNPhos: a new phosphorylation site predictor based on a genetic algorithm integrated neural network, Protein. Eng. Des. Sel. 20 (2007) 405-412.

[57] G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer, KNN model-based approach in classification, Lect. Notes Comput. Sc. 2888 (2003) 986-996.

[58] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory. 13 (2002) 21-27.

[59] A. Li, L. Wang, Y. Shi, M. Wang, Phosphorylation site prediction with a modified k-nearest neighbor algorithm and BLOSUM62 matrix, Conf. Proc. IEEE Eng. Med. Biol. Soc. 6 (2005) 6075-6078.

[60] B.Q. Li, L.L. Hu, S. Niu, Y.D. Cai, K.C. Chou, Predict and analyze S-nitrosylation modification sites with the mrmr and IFS approaches, J. Proteomics 75 (2012) 1654-1665.

[61] L. Breiman, Random forest, Mach. Learn. 45 (2001) 5-32.

[62] M.M. Hasan, D. Guo, H. Kurata, Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information, Mol. Biosyst. 13 (2017) 2545-2550.

[63] X. Zhao, Q. Ning, M. Ai, H. Chai, M. Yin, PGluS: prediction of protein S-glutathionylation sites with multiple features and analysis, J. Theor. Biol. 380 (2015) 524-529.

[64] Y.J. Chen, C.T. Lu, K.Y. Huang, H.Y. Wu, Y.J. Chen, T.Y. Lee, GSHSite: exploiting an iteratively statistical method to identify S-glutathionylation sites with substrate specificity, PloS One 10 (2015) e0118752.

[65] P. Radivojac, V. Vhaynes, Identification, analysis, and prediction of protein ubiquitination sites, Proteins 78 (2010) 365-380.

[66] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, Y. Li, Prediction of lysine ubiquitination with mRMR feature selection and analysis, Amino Acids 42 (2012) 1387-1395.

[67] Z. Chen, Y.Z. Chen, X.F. Wang, C. Wang, R.X. Yan, Z. Zhang, Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs, PLoS One 6 (2011) e22930.

[68] Z. Chen, Y. Zhou, J. Song, Z. Zhang, hCKSAAP_Ubsite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties, Biochim. Biophys. Acta 1834 (2013) 1461-1467.

[69] X. Chen, J.D. Qiu, S.P. Shi, S.B. Suo, R.P. Liang, Systematic analysis and prediction of pupylation sites in prokaryotic proteins, PLoS One 8 (2013) e74002.

[70] Z. Chen, Y. Zhou, Z. Zhang, J. Song, Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features, Brief. Bioinform. 16 (2015) 640-657.

[71] V.N. Nguyen, K.Y. Huang, C.H. Huang, K.R. Lai, T.Y. Lee, A new scheme to characterize and identify protein ubiquitination sites, IEEE Acm. T. Comput. Bi. 14 (2017) 393-403.

[72] 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 (2017) 661-668.

[73] T.Y. Lee, S.A. Chen, H.Y. Hung, Y.Y. Ou, Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites, PLoS One 6 (2010) e17331.

[74] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC, Mol. Ther.-Nucl. Acids 7 (2017) 155-163.

[75] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, iRNA-3typeA: identifying 3-types of modification at RNA's adenosine sites, Mol. Ther.-Nucl. Acids. 11 (2018) 468-474.

[76] B. Liu, F. Yang, 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function, Mol. Ther.-Nucl. Acids 7 (2017) 267-277.

[77] X. Cheng, S.G. Zhao, X. Xiao, iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals, Oncotarget 8 (2017) 58494-58503.

[78] X. Cheng, S.G. Zhao, X. Xiao, iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, Bioinformatics 33 (2017) 341-346.

[79] B. Liu, S. Wang, R. Long, iRSpot-EL: identify recombination spots with an ensemble learning approach, Bioinformatics 33 (2017) 35-41.

[80] X. Cheng, X. Xiao, pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC, Gene 628 (2017) 315-321.

[81] B. Liu, K. Li, D.S. Huang, iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach, Bioinformatics (2018), doi:10.1093/bioinformatics/bty458.

[82] Z.D. Su, Y. Huang, Z.Y. Zhang, Y.W. Zhao, D. Wang, W. Chen, H. Lin, iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC, Bioinformatics (2018), doi:10.1093/bioinformatics/bty508.

[83] B. Liu, F. Yang, D.S. Huang, iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC, Bioinformatics 34 (2018) 33-40.

[84] B. Liu, F. Weng, D.S. Huang, iRO-3wPseKNC: identify DNA replication origins by three-window-based PseKNC, Bioinformatics (2018), doi:10.1093/bioinformatics/bty312.

[85] L. Cai, T. Huang, J. Su, X. Zhang, W. Chen, F. Zhang, L. He, Implications of newly identified brain eQTL genes and their interactors in Schizophrenia, Mol. Ther.-Nucl. Acids 12 (2018) 433-442.

[86] H. Yang, W.R. Qiu, G. Liu, F.B. Guo, W. Chen, H. Lin, iRSpot-Pse6NC: identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC, Int. J. Biol. Sci. 14 (2018) 883-891.

[87] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (2011) 236-247.

[88] J. Herrmann, L.O. Lerman, A. Lerman, Ubiquitin and ubiquitin-like proteins in protein regulation, Circ. Res. 100 (2007) 1276-1291.

[89] 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.

[90] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins 43 (2001) 246-255.

[91] M. Arif, M. Hayat, Z. Jan, iMem-2LSAAC: a two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into Chou's pseudo amino acid composition, J. Theor. Biol. 442 (2018) 11-21.

[92] J. Mei, J. Zhao, Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers, Sci. Rep. 8 (2018) 2359.

[93] J. Mei, J. Zhao, Analysis and prediction of presynaptic and postsynaptic neurotoxins by Chou's general pseudo amino acid composition and motif features, J. Theor. Biol. 427 (2018) 147-153.

[94] M.S. Krishnan, Using Chou's general PseAAC to analyze the evolutionary relationship of receptor associated proteins (RAP) with various folding patterns of protein domains, J. Theor. Biol. 445 (2018) 62-74.

[95] L. Zhang, L. Kong, iRSpot-ADPM: identify recombination spots by incorporating the associated dinucleotide product model into Chou's pseudo components, J. Theor. Biol. 441 (2018) 1-8.

[96] S. Zhang, X. Duan, Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC, J. Theor. Biol. 437 (2018) 239-250.

[97] K.C. Chou, An unprecedented revolution in medicinal chemistry driven by the progress of biological science, Curr. Top. Med. Chem. 17 (2017) 2337-2358.

[98] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nucleic Acids Res. 43 (2015) W65-W71.

[99] B. Liu, H. Wu, Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein Sequences, Natural Science 9 (2017) 67-91.

[100] H.B. Shen, K.C. Chou, PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition, Anal. Biochem. 373 (2008) 386-388.

[101] 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.

[102] 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.

[103] Y. Xu, X. Wang, Y. Wang, Y. Tian, X. Shao, L.Y. Wu, N.Y. Deng, Prediction of posttranslational modification sites from amino acid sequences with kernel methods, J. Theor. Biol. 344 (2014) 78-87.

[104] R.J. Tibshirani, Regression shrinkage and selection via the LASSO: a retrospective, J. R. Stat. Soc. B. 58 (1996) 267-288.

[105] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273-297.

[106] B. Yu, S. Li, W.Y. Qiu, C. Chen, R.X. Chen, L. Wang, M.H. Wang, Y. Zhang, Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising, Oncotarget 8 (2017) 107640-107665.

[107] B. Yu, S. Li, C. Chen, J.M. Xu, W.Y. Qiu, X. Wu, R.X. Chen, Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet

denoising and Chou's pseudo amino acid composition, Chemomet. Intell. Lab. 167 (2017) 102-112.

[108] B. Yu, S. Li, W.Y. Qiu, M.H. Wang, J.W. Du, Y.S. Zhang, X. Chen, Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction, BMC. Genomics 19 (2018) 478-494.

[109] S.B. Zhang, Q.R. Tang. Predicting protein subcellular localization based on information content of gene ontology terms. J. Comput. Biol. Chem. 65 (2016) 1-7.

[110] W. Qiu, S. Li, X. Cui, Z. Yu, M. Wang, J. Du, Y. Peng, B. Yu, Predicting proteinsubmitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo amino acid composition, J. Theor. Biol. 45 (2018) 86-103.

[111] B. Yu, L.F. Lou, S. Li, Y.S. Zhang, W.Y. Qiu, X. Wu, M.H. Wang, B.G. Tian, Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising, J. Mol. Graph. Model. 76 (2017) 260-273.

[112] Y. Guo, L. Yu, Z. Wen, M. Li, Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences, Nucleic Acids Res. 36 (2008) 3025-3030.

[113] A. Dehzangi, K. Paliwal, J. Lyons, A. Sharma, A. Sattar, A segmentation-based method to extract structural and evolutionary features for protein fold recognition, IEEE Acm T. Comput. Bi. 11 (2014) 510-519.

[114] C. Ding, L.F. Yuan, S.H. Guo, H. Lin, W. Chen, Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions, J. Proteomics 77 (2012) 321-328.

[115] H. Ding, S.H. Guo, E.Z. Deng, L.F. Yuan, F.B. Guo, Prediction of Golgi-resident protein types by using feature selection technique, Chemom. Intell. Lab. 124 (2013) 9-13.

[116] S.L. Weng, H.J. Kao, C.H. Huang, T.Y. Lee, Mdd-palm: identification of protein S-palmitoylation sites with substrate motifs based on maximal dependence decomposition, PLoS One 12 (2017) e0179529.

[117] Y. Xu, Z. Wang, C. Li, K.C. Chou, iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC, Med. Chem. 13 (2017) 544-551.

[118] Q.Y. Chen, J. Tang, P.F. Du, Predicting protein lysine phosphoglycerylation sites by hybridizing many sequence based features, Mol. Biosyst. 13 (2017) 874-882.

[119] F. Pedregosa, G.Varoquaux, A. Gramfort, V.Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2012) 2825-2830.

[120] K.C. Chou, Prediction of signal peptides using scaled window, Peptides 22 (2001) 1973-1979.

[121] X. Cheng, X. Xiao, pLoc-mPlant: predict subcellular localization of multi-location plant proteins via incorporating the optimal GO information into general PseAAC, Mol. Biosyst. 13 (2017) 1722-1727.

[122] X. Xiao, X. Cheng, G. Chen, Q. Mao, pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC, Genomics (2018), doi:10.1016/j.ygeno.2018.05.017.

[123] X. Cheng, S.G. Zhao, W.Z. Lin, X. Xiao, pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites, Bioinformatics 33 (2017) 3524-3531.

[124] X. Xiao, X. Cheng, S. Su, Q. Nao, pLoc-mGpos: incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins, Natural Science 9 (2017) 331-349.

[125] X. Cheng, X. Xiao, pLoc-mGneg: predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC, Genomics 110 (2018) 231-239.

[126] X. Cheng, X. Xiao, pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC, Genomics 110 (2018) 50-58.

[127] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, iPTM-mLys: identifying multiple lysine PTM sites and their different types, Bioinformatics 32 (2016) 3116-3123.

[128] K.C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. Biosyst. 9 (2013) 1092-1100.

[129] 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.

[130] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E Stat. Nonlin. Soft. Matter. Phys. 69 (2004) 066138.

[131] B.C. Ross, Mutual information between discrete and continuous data sets, PLoS One 9 (2014) e87357.

[132] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Statist. Soc. 67 (2005) 301-320.

[133] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3-42.

[134] Q. Zou, J. Zeng, L. Cao, R. Ji, A novel features ranking metric with application to scalable visual and bioinformatics data classification, Neurocomputing 173 (2016) 346-354.

[135] Q. Zou, S. Wan, Y. Ju, J.J. Tang, X.X. Zeng, Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy, BMC Syst. Biol. 4 (2016) 401-412.

[136] C. Lin, W. Chen, C. Qiu, Y. Wu, S. Krishnan, Q. Zou, LibD3C: ensemble classifiers with a clustering and dynamic selection strategy, Neurocomputing 123 (2014) 424-435.

[137] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to Boosting, J. Comput. Syst. Sci. 55 (1995) 119-139.

[138] K.C. Chou, H.B. Shen, Recent advances in developing web-servers for predicting protein attributes, Natural Science 1 (2009) 63-92.

[139] B. Liu, L. Fang, R. Long, X. Lan, iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, Bioinformatics 32 (2016) 362-369.

[140] X. Cheng, W.Z. Lin, X. Xiao, pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC, Bioinformatics (2018), doi:10.1093/bioinformatics/bty628.

[141] X. Cheng, X. Xiao, pLoc_bal-mGneg: predict subcellular localization of Gram-negative bacterial proteins by quasi-balancing training dataset and general PseAAC, J. Theor. Biol. 458 (2018) 92-102.

[142] K.C. Chou, X. Cheng, X. Xiao, pLoc_bal-mHum: predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset, Genomics (2018), doi:10.1016/j.ygeno.2018.08.007.

[143] K.C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (2015) 218-234.

Highlights

- A new method (Ubi-SVM) to predict the ubiquitination sites.
- Fusing BE, CKSAAP and PSPM methods to extract protein sequence features information.
- LASSO method can effectively remove redundant information in the protein sequences.
- We investigate the effect of the five different classifiers on the results.
- The proposed method increases the prediction performance over several methods.