Predicting Golgi-Resident Protein Types Using Conditional Covariance Minimization With XGBoost Based on Multiple Features Fusion

Article *in* IEEE Access · August 2019

DOI: 10.1109/ACCESS.2019.2938081

CITATIONS
22

READS
304

5 authors, including:

Minghui Wang
Qingdao University of Science and Technology
166 PUBLICATIONS **4,240** CITATIONS

SEE PROFILE

Qin Ma
The Ohio State University
296 PUBLICATIONS **4,878** CITATIONS

SEE PROFILE

Bin Yu
Qingdao University of Science and Technology
83 PUBLICATIONS **2,269** CITATIONS

SEE PROFILE

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 8, 2019, accepted August 21, 2019, date of publication August 28, 2019, date of current version October 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2938081

Predicting Golgi-Resident Protein Types Using Conditional Covariance Minimization With XGBoost Based on Multiple Features Fusion

HONGYAN ZHOU[1,2], CHENG CHEN[1,2], MINGHUI WANG[1,2], QIN MA[3], AND BIN YU[1,2,4,5]

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

Corresponding author: Bin Yu (yubin@qust.edu.cn)

This work was supported in part by the National Nature Science Foundation of China under Grant 61863010 and Grant 11771188, in part by the Natural Science Foundation of Shandong Province of China under Grant ZR2018MC007, in part by the Project of Shandong Province Higher Educational Science and Technology Program under Grant J17KA159 and Grant J16LI51, in part by the Scientific Research Fund of Hunan Provincial Key Laboratory of Mathematical Modeling and Analysis in Engineering under Grant 2018MMAEZD10, in part by the Key Research and Development Program of Shandong Province of China under Grant 2019GGX101001, and in part by the Extreme Science and Engineering Discovery Environment, which is supported by the National Science Foundation under Grant ACI-1548562.

ABSTRACT The Golgi apparatus is a key organelle for protein synthesis in eukaryotic cell. Any dysfunction of Golgi-resident proteins can lead to different diseases, especially neurodegenerative and inherited diseases, such as diabetes, cancer, and cystic fibrosis, and so on. Therefore, the accurate classification of Golgi-resident proteins may contribute to drug development and further to drug therapy. This paper presents a novel Golgi-resident protein types prediction method called Golgi-XGBoost. First, the feature vectors of protein sequence are extracted by fusing pseudo-amino acid composition (PseAAC), dipeptide composition (DC), pseudo-position specific scoring matrix (PsePSSM) and encoding based on grouped weight (EBGW). Secondly, the conditional covariance minimization (CCM) is used to reduce the dimension of the feature vectors. Then, we adopt the synthetic minority over sampling technique (SMOTE) to balance the samples. Finally, the optimal feature vectors are input into the extreme gradient boosting (XGBoost) classifier to predict the type of Golgi-resident protein. The overall prediction accuracy is 92.1% on training set via jackknife test, which achieves better performance than other state-of-the-art methods. The accuracy of independent testing dataset is 86.5%. And the results show that this paper provides a new method for predicting the type of Golgi-resident protein. The source code and all datasets are available at https://github.com/QUST-AIBBDRC/Golgi-XGBoost/.

INDEX TERMS Golgi-resident protein, multi-information fusion, conditional covariance minimization, synthetic minority over sampling technique, extreme gradient boosting.

I. INTRODUCTION

The Golgi apparatus is an important subcellular organelle for protein synthesis within eukaryotic cell, which is composed of a pile of membrane-bounded cisternae located between the endoplasmic reticulum and the cell surface. The main function of the Golgi apparatus is to process, compare, classify and package proteins synthesized by the endoplasmic reticulum, and then send them to specific parts of the cell or secrete them out of the cell [1]. The Golgi apparatus has three elements, namely, cis-Golgi, medial-Golgi, and trans-Golgi [2]. The cis-Golgi completes the receiving jobs, which is closer to the endoplasmic reticulum and receives the vesicles for sorting and further processing before they are transferred to the trans-Golgi. The medial-Golgi accomplishes the embellishment of glycosylation and the synthesis of polysaccharides and lipids. The trans-Golgi is responsible for the release of tagged and processed proteins into plasma membranes or lysosomes by secretory vesicles.

The associate editor coordinating the review of this article and approving it for publication was Nuno Garcia.

Existing studies have shown that any functional deviation of the Golgi apparatus in cells may trigger inheritable and neurodegenerative diseases, such as diabetes [3], cancer [4], Alzheimer's disease [5] and Parkinson's disease [6]. At present, these diseases are mainly treated with different chemical drugs, such as anti-inflammatory and neuroprotective therapies, but in most cases, these treatments cannot provide a permanent cure [7]. In order to better understand Golgi dysfunction, it is very important to detect abnormalities and injuries in time. Therefore, correctly identifying the type of Golgi-resident proteins can help researchers better understand the role of Golgi proteins for the above problems.

Many machine learning methods have been presented to build prediction tools for protein sub-cellular localization [8]–[19]. However, only few have been specifically designed to study the Golgi-resident proteins. Van Dijk *et al.* [20] proposed a method to predict Golgi-resident protein types for type II membrane portent proteins, in which they adopted support vector machine (SVM) with linear kernel as the classifier. Ding *et al.* [21] used PseAAC with customized Mahalanobis discriminant to identify Golgi protein types. The overall accuracy was 74.7%. And then they improved their methods using g-gapped dipeptide compositions [22], the maximum overall prediction accuracy was improved to 85.4%. Jiao and Du [23] carried out another relevant study, in which they used position specific physicochemical properties (PSPCP) of the amino acid residues to establish a computational model. The prediction accuracy was 86.9% after reducing the feature vector through analysis of variance. Jiao and Du [24] combined PSPCP with Chou's pseudo amino acid composition, and used the minimal redundancy maximum relevance (mRMR) to reduce the feature vector. The average prediction accuracy was 87.1% by using jackknife test.

The discussed methods developed for discriminating Golgi-resident protein types have employed discrete and evolutionary methods for the extraction of relevant information. However, the results reveal that more rigorous technique may be developed which can distinguish the Golgi resident proteins with better accuracy.

This paper proposes a new prediction method called Golgi-XGBoost for Golgi-resident protein types prediction. Firstly, considering the comprehensiveness of feature information extraction, we integrate four feature extraction methods (PseAAC, DC, PsePSSM and EBGW) to extract the feature vectors. Then, comparing different dimension reduction methods for prediction results, the optimal CCM is used to select the feature vectors. Thirdly, since the dataset is imbalanced, SMOTE is used to balance the dataset. Finally, after comparing multiple classifiers, we select the optimal XGBoost as the classifier for prediction. In order to evaluate the performance of the prediction model, jackknife test is carried out on the training dataset. Different parameters are selected and the results are analyzed to determine the best parameters of the prediction model. In order to further test the generalization ability of the prediction model, the prediction model is analyzed in the independent test set and compared with the existing methods. The experimental results show that the proposed method can significantly improve the accuracy of the Golgi resident protein types.

II. MATERIALS AND METHODS

A. DATASETS

In order to make a reasonable comparison with the predicted results of Golgi-resident protein types, Ding's dataset [22] was applied in this paper. Using the following methods to extract the training database: 1) Only those proteins, which have been clearly annotated with cis-Golgi or trans-Golgi, have been selected. 2) Only those proteins with experimentally verified annotations were retained, while proteins with ambiguous annotations, such as ''uncertain'', ''predicted'', ''inferred from homology'' and so on, were excluded. 3) The sequences which are fragments of other proteins and the sequences containing nonstandard letters, such as ''B'', ''X'' or ''Z'', were excluded. 4) The sequence similarity level was controlled at less than 25%.

After all the above filtering procedures, a benchmarking dataset was obtained. It contains 42 cis-Golgi proteins and 95 trans-Golgi proteins. And then we collected 13 cis-Golgi and 51 trans-Golgi proteins from Uniprot, which constituted the testing set and were independent of the training set. All of the training data and independent testing data can be found at http://lin.uestc.edu.cn/server/SubGolgi/data.

B. PSEUDO-AMINO ACID COMPOSITION

In order to avoid loss of the order information of proteins, Chou [25] proposed the PseAAC method to extract the feature vectors. And it has been widely used in protein–protein interaction prediction and subcellular locations [26]–[32]. The method maps protein sequences to the following vectors:

$$X = [x_1, x_2, \ldots, x_{20}, x_{20+1}, \ldots, x_{20+\lambda}]^T \tag{1}$$

Each of these components is given as follows:

$$x_u = \begin{cases} \dfrac{f_\mu}{\sum_{\mu=1}^{20} f_\mu + \omega \sum_{k=1}^{\lambda} \tau_k}, & 1 \le \mu \le 20 \\ \dfrac{\omega \tau_{\mu-20}}{\sum_{\mu=1}^{20} f_\mu + \omega \sum_{k=1}^{\lambda} \tau_k}, & 21 \le \mu \le 20+\lambda \end{cases} \tag{2}$$

where X is the feature vector, ω is the weight factor [33], the value is 0.05. And f_μ is the frequency at which the μ-th amino acid in the protein appears. It can be seen from the above formulas, the first 20 dimensions are the frequency of occurrence in the sequence of the conventional 20 kinds of amino acids, and the latter λ dimensions are sequence related factors that reflect different levels of amino acid sequence information, which are obtained through the physicochemical properties of amino acids.

For the selection of parameter λ of PseAAC, $\lambda_{\max}$ is shorter than the length of the shortest sequence in the protein datasets.

In this paper, λ ranges from 0 to 50. By selecting different parameters, the optimal λ can be determined by the accuracy of prediction results.

C. DIPEPTIDE COMPOSITION

The dipeptide composition [34] model calculates the frequency of amino acid pairs, that is, the frequency of dipeptides. There are $20 \times 20 = 400$ combinations of amino acid pairs composed of 20 amino acids. The feature vector elements extracted by dipeptide composition model are shown in equation (3):

$$f_i = \sum_{j=1}^{n} K_j/(L-1), \quad i = 1, 2, \cdots, 400 \tag{3}$$

where K_j is the frequency of amino acid pairs and L is the length of the protein sequence.

D. PSEUDO-POSITION SPECIFIC SCORING MATRIX

In order to express the characteristic information in the amino acid sequence as reasonably as possible, the pseudo-PsePSSM are used to encode the evolution and sequence information of the protein sequence.

For a target protein sequence with L amino acid residues, we use the position specific scoring matrix (PSSM) as its descriptor introduced by Jones etc [35]. The PSSM can be expressed for each protein sequence as the following:

$$X_{PSSM} = \begin{bmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,j} & \cdots & M_{1,20} \\ M_{2,1} & M_{2,2} & \cdots & M_{2,j} & \cdots & M_{2,20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ M_{i,1} & M_{i,2} & \cdots & M_{i,j} & \cdots & M_{i,20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ M_{L,1} & M_{L,2} & \cdots & M_{L,j} & \cdots & M_{L,20} \end{bmatrix} \tag{4}$$

where $M_{i,j}$ is the score of the residue of the i-th position in the amino acid sequence being mutated to the j-th amino acid residue, which can be searched using the PSI-BLAST [36]. In this paper, the parameters of PSI-BLAST are set as: the maximum number of iterations for multiple searches is 3, the threshold of e-value is 0.001, and the rest of the parameters are set by default.

The indicators in the matrix are normalized to the interval (0, 1) using the following equation:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

Since each protein sequence generates a matrix $L \times 20$ and the length L of the protein sequence is different, the PSSM matrices of different protein sequences need to be transformed into vectors with uniform dimensions by the following formula:

$$\bar{X}_{PSSM} = \left[\bar{M}_1, \quad \bar{M}_2, \quad \cdots, \quad \bar{M}_{20} \right] \tag{6}$$

$$\bar{M}_j = \frac{1}{L} \sum_{i=1}^{L} M_{i,j} (j = 1, 2, \cdots, 20) \tag{7}$$

where $\bar{M}_j$ represents the average score of a protein sequence. However, this method only considers the average score of the residue of the i-th position in the amino acid sequence being mutated to the j-th amino acid residue, without considering the sequence information of amino acid residues in the protein sequence. In order to overcome this shortcoming, PsePSSM [37] was proposed. The feature extraction process is:

$$X^{\xi}_{PsePSSM} = \left[\theta^{\xi}_1, \quad \theta^{\xi}_2, \quad \cdots, \quad \theta^{\xi}_j, \quad \cdots, \quad \theta^{\xi}_{20} \right] \tag{8}$$

$$\theta^{\xi}_j = \frac{1}{L - \xi} \sum_{i=1}^{L-\xi} \left[M_{i,j} - M_{(i+\xi),j} \right]^2 \quad (j = 1, 2, \cdots, 20, \xi < L, \xi \neq 0) \tag{9}$$

where θ^{ξ}_j is the correlation factor.

In summary, a protein sequence can be expressed as equation (10) using PsePSSM and generate a $20 + 20 \times \xi$ dimensional vector.

$$X_{PsePSSM} = \left[\bar{M}_1, \bar{M}_2, \cdots, \bar{M}_{20}, \theta^{\xi}_1, \theta^{\xi}_2, \cdots, \theta^{\xi}_{20} \right] \tag{10}$$

The PsePSSM algorithm converts the protein sequences with different lengths in the dataset into vectors with the same dimension after feature extraction. In this paper, ξ is set to 10 after executing the optimization program for the training sample by jackknife (see in section III). So, each protein sequence generates $20 + 20 \times 10 = 220$ dimensional feature vector.

E. ENCODING BASED ON GROUPED WEIGHT

The basic unit of protein is amino acid, which forms various protein sequences through a series of biochemical reactions. Studies have shown that different amino acids have different physical and chemical properties. Zhang *et al.* [38] proposed the EBGW method. Based on the physicochemical properties of amino acids, they reduced the protein sequences to binary feature sequences using the idea of "coarsening" in physics and introduced normal weight function to realize the grouping weight coding of protein sequences. These amino acids are classified into four categories:
neutral and hydrophobic amino acids

$$C1 = \{G, A, V, L, I, M, P, F, W\}$$

neutral and polarity amino acids $C2 = \{Q, N, S, T, Y, C\}$
acidic amino acids $C3 = \{D, E\}$
alkaline amino acids $C4 = \{H, K, R\}$.

Thus, we can get three combinations, each of which can divide the 20 amino acid residues into two disjoint group: $\{C1, C2\}$ vs $\{C3, C4\}$, or $\{C1, C3\}$ vs $\{C2, C4\}$, and $\{C1, C4\}$ vs $\{C2, C3\}$. Let $P : R_1R_2 \cdots R_L$ be a protein sequence, we can transform it into three binary sequences by three homomorphic maps $\Phi_i(P) = \Phi_i(R_1), \Phi_i(R_2), \ldots, \Phi_i(R_L)(i = 1, 2, 3)$ which are defined as follows:

$$\Phi_1(R_j) = \begin{cases} 1 & if\ R_j \in \{C1, C2\} \\ 0 & if\ R_j \in \{C3, C4\} \end{cases} \quad (j = 1, 2, \ldots, L) \tag{11}$$

$$\Phi_2(R_j) = \begin{cases} 1 & if\ R_j \in \{C1, C3\} \\ 0 & if\ R_j \in \{C2, C4\} \end{cases} \quad (j = 1, 2, \ldots, L) \quad (12)$$

$$\Phi_3(R_j) = \begin{cases} 1 & if\ R_j \in \{C1, C4\} \\ 0 & if\ R_j \in \{C2, C3\} \end{cases} \quad (j = 1, 2, \ldots, L) \quad (13)$$

Then $\Phi_1(P)$, $\Phi_2(P)$, $\Phi_3(P)$ are three binary sequences of length L. These sequences are divided into a number of sub-sequences of increasing length successively. A fixed parameter N is set and the length of sub-sequence can be expressed as $\left\lfloor \overline{kL/N} \right\rfloor (k = 1, 2, .., N)$, where $\lfloor . \rfloor$ represents the integer operator. Calculate the frequency of 1 in each sub-sequence, each $\Phi_i(P)$ can be converted into an N-dimensional feature vector. To sum up, for a protein sequence P with length L, a $3N$-dimension vector can be obtained.

In this paper, N is set to 30 after executing the optimization program for the training sample using jackknife test (see in section III). So, each protein sequence generates 90 dimensional feature vectors.

F. CONDITIONAL COVARIANCE MINIMIZATION

Feature selection is an important issue in machine learning, and can lead to both computational benefits and statistical benefits. After the conditional covariance operator was proposed by Baker Charles [39], Fukumizu *et al.* [40], [41] have further studied it and applied it into dimensionality reduction, which provides a conditional dependence measure for random variables. Chen *et al.* [42] recently proposed a new feature selection method called CCM, which is based on minimizing the trace of the conditional covariance operator for feature selection.

For a binary vector $\omega \in \{0, 1\}^d$, the optimization model is

$$\begin{aligned} &\min_{\omega} y^T \left(G_{\omega \odot X} + n\varepsilon_n I_n\right)^{-1} y \\ &\text{subject to } \omega_i \in \{0, 1\}, \quad i = 1, \cdots, d \\ &\qquad 1^T \omega = m \end{aligned} \quad (14)$$

where $y = (y_1, y_2, \cdots, y_n)^T$ is an n-dimensional vector, $x \in R^d$, $\odot$ denotes the Hadamard product between two vectors and $G_{\omega \odot X}$ is the centralized version of the kernel matrix $K_{\omega \odot X}$ with $(K_{\omega \odot X}) = k_1(\omega \odot x_i, \omega \odot x_j)$.

And then approximate (14) by relaxing the domain of ω to the unit hypercube $[0, 1]^d$ and replacing the equality constraint with an inequality constraint:

$$\begin{aligned} &\min_{\omega} y^T \left(G_{\omega \odot X} + n\varepsilon_n I_n\right)^{-1} y \\ &\text{subject to } 0 \le \omega_i \le 1, \quad i = 1, \cdots, d \\ &\qquad 1^T \omega \le m \end{aligned} \quad (15)$$

This objective can be optimized using projected gradient descent, and represents the approximation to (14). By setting the m largest values of ω to 1 and remaining values to 0, a solution to (15) is converted back into a solution for the original problem.

In this paper, several dimension reduction methods are compared, and finally the CCM method is chosen to select the feature vectors according to the prediction accuracy.

G. SMOTE METHOD

Because there is a high degree of imbalance in the samples used in this study and the number of trans-Golgi proteins is significantly lower than that of cis-Golgi proteins. To handle the problems brought by data resampling, this paper uses the SMOTE method proposed by Chawla *et al.* [43]. The SMOTE method is a method of randomly undersampling a large scale of samples while randomly oversampling a small scale of samples. This algorithm is a common method to deal with imbalanced data [44]–[50]. Its detailed operation is as follows:

(1) Set the multiplier of upward sampling to N.
(2) Find the K-nearest neighbor of the sample x_i from the samples of the interface residue, denoted by $x_{i(near)}$, $near \in \{1, \ldots, k\}$, and choose N samples at random, denoted by $y_1, \ldots, y_N$.
(3) Synthesize a new sample x_{i1},

$$x_{i1} = x_i + \xi_1(y_1 - x_i) \quad (16)$$

where ξ_1 is a random number in (0, 1). Repeat the above procedure N times until we got N new samples: $x_{i(new)}$, $new \in \{1, \ldots, N\}$.

Add these newly synthesized samples to the original samples and form a new, more balanced dataset. The visual display of SMOTE sample synthesis method is shown in Table 6.

H. EXTREME GRADIENT BOOSTING METHOD

In the case of consistent feature vector dimensions, this paper adopts multiple classifiers. Based on the prediction accuracy, XGBoost method is chosen. The method is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples [51].

XGBoost is a monitor model, and its corresponding model is a bunch of CART trees. For a given data set with n examples and m features $D = \{(x_i, y_i), x_i \in R^m, y_i \in R\}$ ($|D| = n$) a tree ensemble model uses K additive functions to predict the output.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \quad (17)$$

where K is the number of trees, $F = \{f(x) = w_{q(x)}, q : R^m \to T, w \in R^T\}$ is the set of all possible CARTs, q represents the structure of each tree that maps an example to the corresponding leaf index, and T is the number of leaves in the tree. The objective function to be optimized is given by

$$obj^{(t)} = \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (18)$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$, $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$.

XGBoost's sharding operation is different from general decision tree. The general decision tree does not consider the complexity of the tree, but relies on the subsequent pruning operation to control. XGBoost already takes into account the complexity of the tree, and it doesn't require a separate pruning operation.

I. JACKKNIFE TEST

In this paper we use Jackknife method to test the prediction results. Jackknife test is a re-sampling method proposed by Maurice Quenouille(1949)[52], which is used for hypothesis testing and interval estimation. The method is based on that one sample is randomly selected as testing set in turn, and the remaining samples are employed as training set.

The motivation behind the jackknife is as follows. If $\bar{x}$ denotes the mean for a sample of size n, we can also compute the sample mean when the j-th data point is removed (or jackknifed), and

$$\bar{x}_{-j} = \frac{1}{n-1}\sum_{i\neq j}^{n} x_i.$$

Observe that if we know both $\bar{x}$ x and $\bar{x}_{-j}$ we can compute the value of the j-th data point as

$$x_j = n\bar{x} - (n-1)\bar{x}_{-j}.$$

J. MODEL BUILDING AND PERFORMANCE EVALUATION

The flowchart of the Golgi resident protein types prediction model based on the multi-information fusion proposed in this paper is shown in Figure 1. Our experimental environment is Intel (R) Core (TM) i5-4258U CPU @2.40 GHz 4.00GB of RAM and uses MATLAB2014a programming. The steps of the Golgi resident protein types prediction method based on the Golgi- XGBoost method are described as:

(1) Input protein sequences and the corresponding class labels.
(2) Feature extraction: protein sequences are converted into numerical signals by PseAAC, DC, PsePSSM and EBGW, and their sequence information, physical and chemical information and evolutionary information are obtained and used as initial feature vectors.
(3) Feature selection: the CCM method is performed on the feature vectors obtained by fusion of PseAAC, DC, PsePSSM and EBGW, and the redundant information in the sequence is removed to obtain the optimal feature vectors.
(4) Imbalanced data procession: apply the Smote algorithm to balance the dataset.
(5) Prediction algorithm: using XGBoost classifier to construct a Golgi type prediction model.
(6) Using the model constructed in (1)-(5), the independent test set is used to test the prediction model.

In this paper, sensitivity (SE), specificity (SP), matthew's correlation coefficient (MCC) and overall accuracy (ACC), which are well-established performance metrics [53], [54], are used to evaluate the results of the prediction system. They are defined as follows:

$$SE = \frac{TP}{TP+FN} \tag{19}$$

$$SP = \frac{TN}{TN+FP} \tag{20}$$

$$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \tag{21}$$

$$ACC = \frac{TP+TN}{TP+FN+TN+FP} \tag{22}$$

where the symbols TP, TN, FP, FN respectively denote the number of true positives, true negatives, false positives and false negatives.

Additionally, we use ROC (receiver operating characteristic curve) and PRC (precision recall curve) curves to evaluate the generalization performance of the model. The area under the ROC curve (AUC) and the PRC curve (AUPR) are used as a quantitative indicator of the robustness of the model. In general, the larger AUC and AUPR values are, the better model prediction performance is [55].

III. RESULTS AND DISCUSSION

A. THE CHOICE OF MODEL PARAMETERS λ, ξ AND N

The selection of parameters is very important for a prediction system. How to extract effective feature information from protein sequences is the key to the success of a prediction model for Golgi-resident protein types. In this paper, PseAAC algorithm, PsePSSM algorithm and EBGW algorithm are used to extract feature vectors from protein sequences. During the feature extraction process of these three algorithms, the selection of λ, ξ and N values is crucial to the construction of the model.

Parameter λ represents the sequence proximity, i.e. the sequence information of the protein sequence. Parameter ξ not only considers the sequence information, but also considers the evolution information of protein sequence. If the values of λ and ξ are set too big, the feature vector dimension of the protein sequence will be too high, which will bring more redundant information and affect the prediction effect. If the values of λ and ξ are set too small, we may get too little information about the protein sequence. In order to find the optimal value of λ in the model, since the shortest protein sequence length for the two protein datasets is 50, λ is set from 0 to 50. The XGBoost classifier is used to classify the data set. The results are tested by the method of jackknife, so as to obtain the prediction overall accuracy of the protein data set, as shown in Figure 2. Similarly, in order to find the best parameters of the model, we set the value of ξ to be 2 to 14. For different ξ values, XGBoost is used to classify the data set, and the results are tested by the method of jackknife, so as to obtain the prediction overall accuracy of the protein interaction data set, as shown in Figure 2.

As can be seen from Figure 2, when $\lambda = 30$ prediction overall accuracy reaches the highest, which is 65.69%.

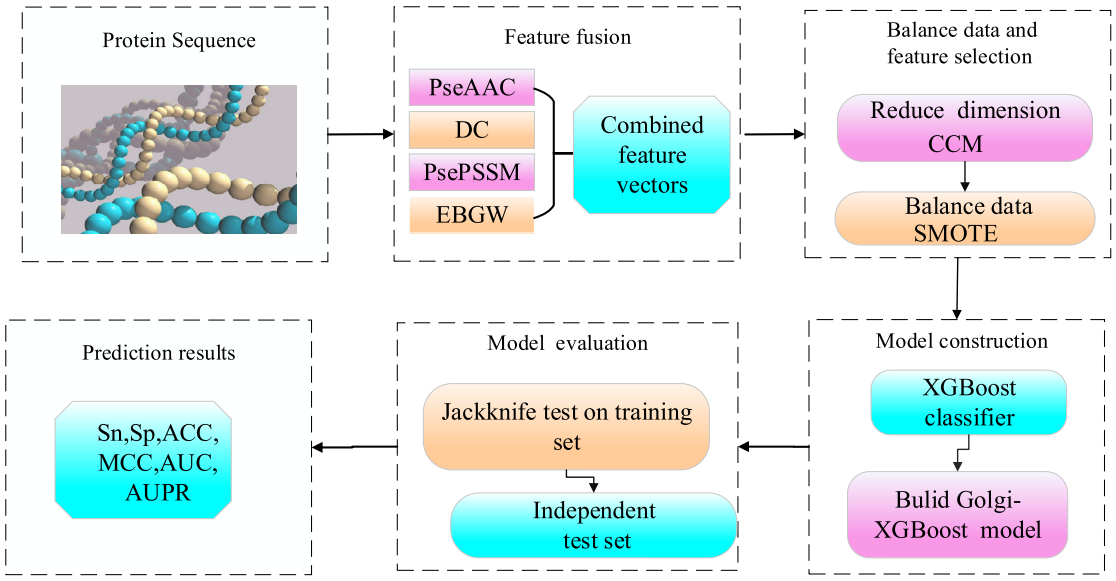

FIGURE 1. Flowchart of Golgi resident types prediction method.

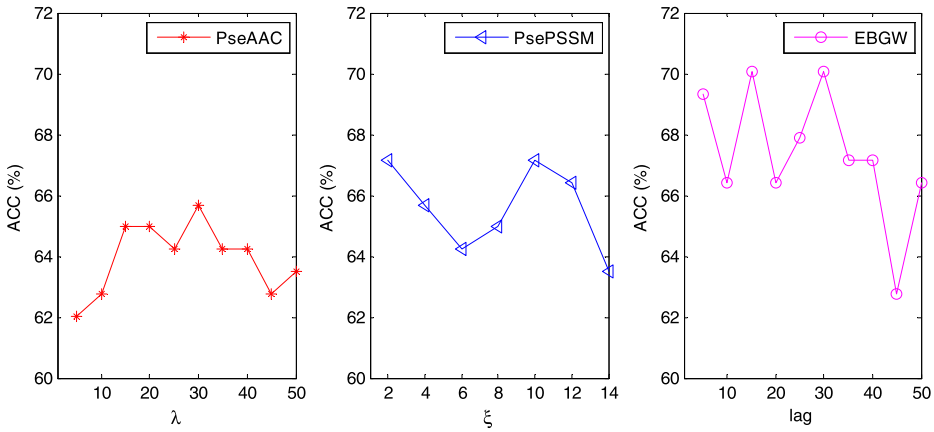

FIGURE 2. The prediction overall accuracy of Golgi resident protein types by selecting different parameters values.

PseAAC is used to extract features of protein sequences, and each protein sequence generates $20 + 30 = 50$ dimensional feature vector. When $\xi = 10$, prediction overall accuracy reaches the highest, which is 67.15%. PsePSSM is used to extract features of protein sequences, and each protein sequence generates $20 + 20 \times 10 = 220$ dimensional feature vector.

Features of protein sequences are extracted by EBGW, and each protein sequence in the data set is transformed into a $3N$ dimensional numerical sequence. With the EBGW algorithm, the impact of parameter N on model construction needs to be taken into account. For different N values, XGBoost is used to classify the data set, and the results are tested by the method of jackknife, so as to obtain the prediction overall accuracy of the protein interaction data set, as shown in Figure 2.

It can be seen from Figure 2 that with different values of N, the results of the prediction results using the EBGW method are also different. When $N = 30$, prediction overall accuracy reaches the highest, which is 70.07%. The optimal value of N in the EBGW feature extraction method is 30. It means that the EBGW is used to extract features of protein sequences, and each protein sequence generates $3 \times 30 = 90$ dimensional feature vector.

B. FUSION OF MULTIPLE FEATURE EXTRACTION METHODS

Since only one feature extraction method is used for feature extraction of protein sequences, some information will be lost. Currently, multiple feature extraction methods are often used to characterize protein sequences [49]–[60]. In this paper, four feature extraction methods are adopted, namely PseAAC, DC, PsePSSM and EBGW. The values of the parameters are $\lambda = 30, \xi = 10$, and $N = 30$ respectively. Table 1 is the feature extraction scheme of the dataset.

TABLE 1. Feature extraction schemes on Golgi resideng protein dabates.

Feature extraction method	PseAAC	DC	PsePSSM	EBGW
Feature vector dimensions	50	400	220	90

TABLE 2. Evaluation results of different dimension reduction methods.

Methods	ACC (%)	SE (%)	SP (%)	MCC
EN	88.49	88.10	88.89	0.7699
ET	87.70	92.06	83.33	0.7569
LASSO	90.87	88.89	92.86	0.8181
MI	90.48	88.89	92.06	0.8099
CCM	92.06	89.68	94.44	0.8422

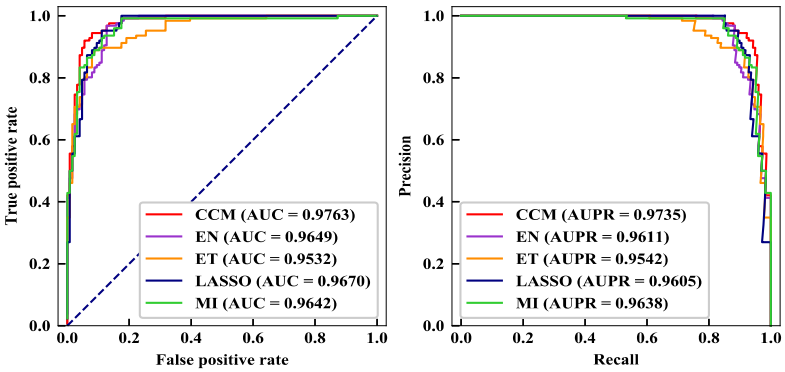

FIGURE 3. ROC and PRC curves of training dataset with different dimension reduction methods.

Considering the comprehensiveness of feature information extraction, this paper fuses four feature extraction methods, and the fused feature vector dimension is 760.

C. THE CHOICE OF DIMENSION REDUCTION METHODS

In the case of consistent feature vector dimensions, this paper adopts the methods of LASSO [61], elastic net (EN) [62], mutual information (MI) [63], extremely randomized trees (ET) [64] and CCM to select feature, and the prediction effects of these methods are compared. XGBoost is used to classify the data set, and the results are tested by the method of jackknife, so as to obtain the prediction accuracy, sensitivity, specificity, and Matthews correlation coefficient of the protein interaction data set, as shown in Table 2.

Table 2 shows that the prediction results of the Golgi resident protein types under five dimension reduction methods. It can be seen from Table 2 that the highest values of ACC, SE, SP and MCC obtained by the CCM method are 92.60%, 89.68%, 94.44% and 0.8422, respectively. Its ACC is 3.57%, 4.36%, 1.19%, and 1.58% higher than that obtained by EN, ET, LASSO and MI methods, respectively. The MCC of CCM method is 7.23%, 8.53%, 2.41, and 3.23% higher than the EN, ET, LASSO and MI methods, respectively.

Figure 3 shows the ROC and PRC curves of the training dataset under different dimension reduction methods. It can be seen from Figure 3 that the area under the ROC curve obtained by the CCM method has the largest AUC value of 0.9763, which is 1.14%, 2.31%, 0.93% and 1.21% higher than that of the EN, ET, LASSO and MI methods, respectively. The area under the PRC curve obtained by the CCM method has a maximum AUPR value of 0.9735, which is 1.24%, 1.93%, 1.30%, and 0.97% higher than that of the EN, ET, LASSO and MI methods, respectively.

In summary, we use the CCM method to obtain the highest ACC, which indicates that it can fully consider the protein sequence information and improve the performance of the Golgi resident protein types prediction model. The dimension of feature vector is reduced from 760 to 200 by using CCM.

D. THE CHOICE OF METHOD FOR UNBALANCED PROBLEM

Since the dataset is imbalanced, the number of trans-Golgi proteins is significantly lower than that of cis-Golgi proteins. This paper adopts synthetic minority over-sampling technique (SMOTE) to balance the dataset. In the training data set, the comparison of the predicted results before and after using SMOTE method is shown in Table 3.

It can be found from Table 3, the prediction results of ACC, SE, SP and MCC by using the SMOTE method are 90.48%, 92.86%, 88.10% and 0.8104, which are 24.30%, 7.75%, 64.29% and 70.23% higher than without the SMOTE

TABLE 3. Evaluation results of prediction with and without SMOTE method on training dataset.

Evaluation	ACC (%)	SE (%)	SP (%)	MCC
Without SMOTE	66.18	85.11	23.81	0.1081
With SMOTE	90.48	92.86	88.10	0.8104

TABLE 4. Evaluation results of prediction under different classification algorithms.

Classifier	ACC (%)	SE (%)	SP (%)	MCC
RF	86.51	84.13	88.89	0.7310
KNN	87.30	80.95	93.65	0.7521
SVM	78.57	76.98	80.16	0.5717
NB	78.57	76.19	80.95	0.5721
XGBoost	92.06	89.68	94.44	0.8422

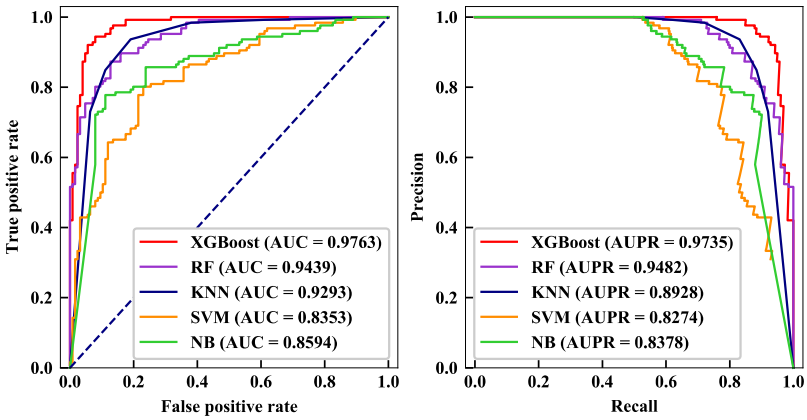

FIGURE 4. ROC and PRC curves of training dataset under different classification algorithms.

method, respectively. It shows that the SMOTE method can achieve a relatively good prediction effect.

E. SELECTION OF CLASSIFICATION ALGORITHMS

This paper focuses on five classification algorithms: random forest (RF), K-nearest neighbor (KNN), support vector machine (SVM), naïve Bayes (NB), and extreme gradient boosting (XGBoost). Following the principle of selecting the best through repeated tests, K is 5 in KNN, SVM selects the linear kernel function, and RF uses 200 trees. The specific results of the prediction under the five classification algorithms are tested with the method of jackknife, as shown in Table 4. In order to more intuitively see the comparison of prediction effects, Figure 4 shows the ROC and PRC curves of the training dataset.

Table 4 shows that the prediction results of the Golgi resident protein types under five classification algorithms. It can be seen from Table 4 that the highest values of ACC, SE, SP and MCC obtained by the XGBoost algorithm are92.60%, 89.68%, 94.44% and 0.8422, respectively. Its ACC is 5.55%, 3.57%, 13.49% and 13.49% higher than that obtained by RF, KNN, SVM and NB algorithms, respectively. The MCC of XGBoost algorithm is 11.12%, 9.01%, 27.05% and 27.01% higher than the RF, KNN, SVM and NB algorithms, respectively.

Figure 4 shows the ROC and PRC curves of the training dataset under different classification algorithms. It can be seen from Figure 4 that the area under the ROC curve obtained by the XGBoost algorithm has the largest AUC value of 0.9763, which is 3.24%, 4.70%, 14.10% and 11.69% higher than that of the RF, KNN, SVM and NB algorithms, respectively. The area under the PRC curve obtained by the CCM method has a maximum AUPR value of 0.9735, which is 2.53%, 4.29%, 14.61% and 13.57% higher than that of the RF, KNN, SVM and NB algorithms, respectively.

As the prediction algorithm of the model, the results of the XGBoost algorithm are significantly better than other classification algorithms. Therefore, we choose XGBoost classification algorithm as the classification algorithm of the model.

F. COMPARISON WITH OTHER METHODS

Several prediction methods have been proposed for the prediction of Golgi resident protein types. In order to prove the effectiveness of the model presented in this paper, based on jackknife test, we compare its prediction accuracy with other models on the same dataset. The comparison results of the four methods of the training dataset are shown in Table 5.

It can be seen from Table 5 that the ACC of the Golgi-XGBoost prediction model proposed on the training dataset

TABLE 5. Comparison of prediction methods for Golgi resident protein types on training dataset.

	ACC (%)	SE (%)	SP (%)	MCC
Ding et al.[22]	85.4	90.5	73.8	0.652
Jiao and Du[23]	86.9	92.6	73.8	0.684
Jiao and Du[26]	86.9	93.7	71.4	0.682
Golgi-XGBoost	92.1	89.7	94.4	0.842

TABLE 6. Comparison of prediction methods for Golgi resident protein types on independent testing dataset.

	ACC (%)	SE (%)	SP (%)	MCC
Ding et al.[22]	85.9	69.2	90.2	0.658
Jiao and Du[23]	85.9	69.2	90.2	0.820
Golgi-XGBoost	86.5	98.1	75.0	0.751

is 92.1%, which is 2.5%–6.7% higher than other methods. The results show that our method significantly improves the prediction accuracy of Golgi resident protein types. Similarly, our proposed prediction model has the highest SP and MCC, which are 94.4% and 0.842, respectively. Among them, MCC is 19% higher than Ding's model [22] and 15.8%, 16% higher than the Jiao's model [23], [26]. Overall, the method in this paper has obvious advantages in the training dataset.

In order to further verify the actual prediction ability of our model, we collected 13 cis-Golgi and 51 trans-Golgi proteins from Uniprot as the independent testing dataset. The results of comparison with others are shown in Table 6. We can see that the ACC of the Golgi-XGBoost prediction model proposed is 86.5%, which is 0.6% higher than other methods.

In a word, the method proposed in this paper significantly improves the prediction accuracy of Golgi- resident protein types, and the prediction results are satisfactory.

IV. CONCLUSION

With the advent of the era of big data, protein sequence data in biological databases are increasing rapidly. It is far from enough to obtain sequence information by experimental methods. Therefore, it is more and more important to study the prediction of protein structure and function based on machine learning. In this paper a new method to predict Golgi resident protein types, called Golgi-XGBoost, is proposed. At first, we fuse PseAAC, DC, PsePSSM and EBGW to extract the feature vectors. Then CCM is used to reduce the dimension of the feature vectors. Thirdly, SMOTE is used to deal with the imbalance of dataset. Finally, select XBoost as the classifier. Among them, PseAAC feature extraction contains the frequency and order information for amino acids. PsePSSM feature extraction can get the evolutionary information and sequence information. The physicochemical information of protein sequences can be extracted by EBGW feature extraction. The information of amino acid pairs in a protein sequence can be extracted by using DC feature extraction. Using CCM dimensionality reduction method can effectively eliminate redundant information in protein sequences. XGBoost classification algorithm can handle high dimensional data, avoid over-fitting and it is efficient, flexible, and convenient. The experimental results show that the proposed method can effectively improve the prediction accuracy of Golgi-resident protein types. We expect this method to be useful not only for the prediction of Gogli-resident protein types but also as a powerful tool in related fields such as bioinformatics, proteomics and molecular biology. In future work, we will strive to provide a web server for the prediction method proposed in this paper.

REFERENCES

[1] E. D'Angelo, "The critical role of Golgi cells in regulating spatio-temporal integration and plasticity at the cerebellum input stage," *Frontiers Neurosci.*, vol. 2, no. 1, pp. 35–46, Jul. 2008.
[2] M. S. Ladinsky, D. N. Mastronarde, J. R. McIntosh, K. E. Howell, and L. A. Staehelin, "Golgi structure in three dimensions: Functional insights from the normal rat kidney cell," *J. Cell Biol.*, vol. 144, no. 6, pp. 1135–1149, Mar. 1999.
[3] S. Hoyer, "Is sporadic Alzheimer disease the brain type of non-insulin dependent diabetes mellitus? A challenging hypothesis," *J. Neural Transmiss.*, vol. 105, nos. 4–5, pp. 415–422, Jul. 1998.
[4] J. M. H. Van den Elsen, D. A. Kuntz, and D. R. Rose, "Structure of Golgi α-mannosidase II: A target for inhibition of growth and metastasis of cancer cells," *EMBO J.*, vol. 20, no. 12, pp. 3008–3017, Dec. 2001.
[5] L. J. Su, P. K. Auluck, T. F. Outeiro, E. Yeger-Lotem, J. A. Kritzer, D. F. Tardiff, K. E. Strathearn, F. Liu, S. Cao, S. Hamamichi, K. J. Hill, K. A. Caldwell, G. W. Bell, E. Fraenkel, A. A. Cooper, G. A. Caldwell, J. M. McCaffery, J. C. Rochet, and S. Lindquist, "Compounds from an unbiased chemical screen reverse both ER-to-Golgi trafficking defects and mitochondrial dysfunction in Parkinson's disease models," *Disease Models Mech.*, vol. 3, nos. 3–4, pp. 194–208, Mar./Apr. 2010.
[6] T. Arendt, H. G. Zveuintshva, and T. A. Lkontovich, "Dendritic changes in the basal nucleus of meynert and in the diagonal band nucleus in Alzheimer's disease—A quantitative Golgi investigation," *Neuroscience*, vol. 19, no. 4, pp. 1265–1278, Dec. 1986.
[7] D. D. Elsberry and M. T. Rise, "Techniques for treating neurodegenerative disorders by infusion of nerve growth factors into the brain," U.S. Patents US6042579A, Aug. 5, 1998.
[8] W. Qiu, S. Li, X. Cui, Z. Yu, M. Wang, J. Du, Y. Peng, and B. Yu, "Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition," *J. Theor. Biol.*, vol. 450, pp. 86–103, Aug. 2018.
[9] H. Shi, S. Liu, J. Q. Chen, X. Li, Q. Ma, and B. Yu, "Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure," *Genomics*, to be published.
[10] B. G. Tian, X. Wu, C. Chen, W. Y. Qiu, Q. Ma, and B. Yu, "Predicting protein–protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach," *J. Theor. Biol.*, vol. 462, pp. 329–346, Feb. 2019.
[11] B. Yu, S. Li, W. Qiu, M. Wang, J. Du, Y. Zhang, and X. Chen, "Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction," *BMC Genomics*, vol. 19, Jun. 2018, Art. no. 478.

[12] X. Cui, Z. Yu, B. Yu, M. Wang, B. Tian, and 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 Intell. Lab. Syst.*, vol. 184, pp. 28–43, Jan. 2019.

[13] X. Cheng, X. Xiao, and K. C. Chou, "pLoc_bal-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by quasi-balancing training dataset and general PseAAC," *J. Theor. Biol.*, vol. 458, pp. 92–102, Dec. 2018.

[14] K.-C. Chou, X. Cheng, and X. Xiao, "pLoc_bal-mHum: Predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset," *Genomics*, to be published.

[15] X. Xiao, X. Cheng, G. Chen, Q. Mao, and K.-C. Chou, "pLoc_bal-mGpos: Predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC," *Genomics*, to be published.

[16] K.-C. Chou and H.-B. Shen, "Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms," *Nature Protocols*, vol. 3, no. 2, pp. 153–162, Feb. 2008.

[17] K.-C. Chou and H.-B. Shen, "Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms," *Natural Sci.*, vol. 2, pp. 1090–1103, Jan. 2010.

[18] G.-L. Fan and Q.-Z. Li, "Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition," *Amino Acids*, vol. 43, no. 2, pp. 545–555, Aug. 2012.

[19] D. Yu, X. Wu, H. Shen, J. Yang, Z. Tang, Y. Qi, and J. Yang, "Enhancing membrane protein subcellular localization prediction by parallel fusion of multi-view features," *IEEE Trans. Nanobiosci.*, vol. 11, no. 4, pp. 375–385, Dec. 2012.

[20] A. D. van Dijk, D. Bosch, C. J. ter Braak, A. R. van der Krol, and R. C. van Ham, "Predicting sub-Golgi localization of type II membrane proteins," *Bioinformatics*, vol. 24, no. 16, pp. 1779–1786, Aug. 2008.

[21] H. Ding, L. Liu, F.-B. Guo, J. Huang, and H. Lin, "Identify Golgi protein types with modified Mahalanobis discriminant algorithm and pseudo amino acid composition," *Protein Peptide Lett.*, vol. 18, no. 1, pp. 58–63, Jan. 2011.

[22] H. Ding, S.-H. Guo, E.-Z. Deng, L.-F. Yuan, F.-B. Guo, J. Huang, N. Rao, W. Chen, and H. Lin, "Prediction of Golgi-resident protein types by using feature selection technique," *Chemometrics Intell. Lab. Syst.*, vol. 124, pp. 9–13, May 2013.

[23] Y. S. Jiao and P. F. Du, "Predicting Golgi-resident protein types using pseudo amino acid compositions: Approaches with positional specific physicochemical properties," *J. Theor. Biol.*, vol. 391, pp. 35–42, Feb. 2016.

[24] Y. S. Jiao and P. F. Du, "Prediction of Golgi-resident protein types using general form of Chou's pseudo-amino acid compositions: Approaches with minimal redundancy maximal relevance feature selection," *J. Theor. Biol.*, vol. 402, pp. 38–44, Aug. 2016.

[25] K. C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[26] J. Jia, Z. Liu, X. Xiao, B. Liu, and K. C. Chou, "Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition," *J. Biomol. Struct. Dyn.*, vol. 34, no. 9, pp. 1946–1961, Sep. 2016.

[27] J.-X. Zhai, T.-J. Cao, J.-Y. An, and Y.-T. Bian, "Highly accurate prediction of protein self-interactions by incorporating the average block and PSSM information into the general PseAAC," *J. Theor. Biol.*, vol. 432, pp. 80–86, Nov. 2017.

[28] X. Cheng, X. Xiao, and K. C. Chou, "pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC," *Genomics*, vol. 110, no. 1, pp. 50–58, Jan. 2018.

[29] X. Cheng, S. G. Zhao, X. Xiao, and K. C. Chou, "iATC-mISF: A multi-label classifier for predicting the classes of anatomical therapeutic chemicals," *Bioinformatic*, vol. 33, no. 3, pp. 341–346, Feb. 2017.

[30] X. Cheng, S. G. Zhao, X. Xiao, and K. C. Chou, "iATC-mHyb: A hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals," *Oncotarget*, vol. 8, no. 35, pp. 58494–58503, Apr. 2017.

[31] P. P. Zhu, W.-C. Li, Z.-J. Zhong, E.-Z. Deng, H. Ding, W. Chen, and H. Lin, "Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition," *Mol. Biosyst.*, vol. 11, no. 2, pp. 558–563, Feb. 2015.

[32] J. Lin, H. Chen, S. Li, Y. Liu, X. Li, and B. Yu, "Accurate prediction of potential druggable proteins based on genetic algorithm and bagging-SVM ensemble classifier," *Artif. Intell. Med.*, no. 98, pp. 35–47, Jul. 2019.

[33] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins*, vol. 43, no. 3, pp. 246–255, May 2001.

[34] A. Khan, A. Majid, and M. Hayat, "CE-PLoc: An ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition," *Comput. Biol. Chem.*, vol. 35, no. 4, pp. 218–229, Aug. 2011.

[35] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J. Mol. Biol.*, vol. 292, no. 2, pp. 195–202, Sep. 1999.

[36] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids. Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.

[37] H. B. Shen and K. C. Chou, "Nuc-PLoc: A new Web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM," *Protein Eng. Des. Sel.*, vol. 20, no. 11, pp. 561–567, Nov. 2007.

[38] Z. H. Zhang, Z.-H. Wang, Z.-R. Zhang, and Y.-X. Wang, "A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine," *FEBS Lett.*, vol. 580, no. 26, pp. 6169–6174, Oct. 2006.

[39] C. R. Baker, "Joint measures and cross-covariance operators," *Trans. Amer. Math. Soc.*, vol. 186, pp. 273–286, Dec. 1973.

[40] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces," *J. Mach. Learn. Res.*, vol. 5, pp. 73–99, Jan. 2004.

[41] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Kernel dimension reduction in regression," *Ann. Statist.*, vol. 37, no. 4, pp. 1871–1905, Aug. 2009.

[42] J. B. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan, "Kernel feature selection via conditional covariance minimization," in *Proc. NIPS*, Jul. 2017, pp. 1–10.

[43] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[44] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinf.*, vol. 14, no. 1, p. 106, 2013.

[45] J. F. Díez-Pastor, J. J. Rodríguez, C. García-Osorio, and L. I. Kuncheva, "Random balance: Ensembles of variable priors classifiers for imbalanced data," *Knowl.-Based Syst.*, vol. 85, pp. 96–111, Sep. 2015.

[46] L. Ma and S. Fan, "CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests," *Bmc Bioinf.*, vol. 18, no. 1, p. 169, Mar. 2017.

[47] S. Cateni, V. Colla, and M. Vannucci, "A method for resampling imbalanced datasets in binary classification tasks for real-world problems," *Neurocomputing*, vol. 135, pp. 32–41, Jul. 2014.

[48] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, "SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," *Inf. Sci.*, vol. 291, pp. 184–203, Jan. 2015.

[49] A. Nath and K. Subbiah, "Unsupervised learning assisted robust prediction of bioluminescent proteins," *Comput. Biol. Med.*, vol. 68, pp. 27–36, Jan. 2016.

[50] X. Y. Wang, B. Yu, A. J. Ma, C. Chen, B. Q. Liu, and Q. Ma, "Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique," *Bioinformatics*, vol. 35, no. 14, pp. 2395–2402, Jul. 2019.

[51] T. Q. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[52] M. H. Quenouille, "Approximate tests of correlation in time-series," *J. Roy. Stat. Soc.*, vol. 11, no. 1, pp. 68–84, 1949.

[53] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[54] D. G. Altman and J. M. Bland, "Diagnostic tests. 1: Sensitivity and specificity," *BMJ*, vol. 308, no. 6943, p. 1552, Jun. 1994.

[55] C. Chen, Q. M. Zhang, Q. Ma, and B. Yu, "LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion," *Chemometrics Intell. Lab. Syst.*, vol. 191, pp. 54–64, Aug. 2019.

[56] B. Yu, S. Li, C. Chen, J. Xu, W. Qiu, X. Wu, and R. Chen, "Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition," *Chemometrics Intell. Lab. Syst.*, vol. 167, pp. 102–112, Aug. 2017.
[57] B. Yu, S. Li, W.-Y. Qiu, C. Chen, R.-X. Chen, L. Wang, M.-H. Wang, and Y. Zhang, "Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising," *Oncotarget*, vol. 8, no. 64, pp. 107640–107665, Dec. 2017.
[58] B. Yu, L. Lou, S. Li, Y. Zhang, W. Qiu, X. Wu, M. Wang, and B. Tian, "Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising," *J. Mol. Graph. Model.*, vol. 76, pp. 260–273, Sep. 2017.
[59] B. Yu and Y. Zhang, "The analysis of colon cancer gene expression profiles and the extraction of informative genes," *J. Comput. Theor. Nanosci.*, vol. 10, no. 5, pp. 1097–1103, May 2013.
[60] B. Yu and Y. Zhang, "A simple method for predicting transmembrane proteins based on wavelet transform," *Int. J. Biol. Sci.*, vol. 9, no. 1, pp. 22–33, Jan. 2013.
[61] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1418–1429, Dec. 2006.
[62] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statistical Soc., Ser. B, (Stat. Methodol.)*, vol. 2, no. 2, pp. 301–320, Dec. 2005.
[63] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Comput. Linguistics*, vol. 16, no. 1, pp. 22–29, Mar. 2002.
[64] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006.

HONGYAN ZHOU received the Ph.D. degree in physical oceanography from the Ocean University of China, in 2008. She is currently a Lecturer with the College of Mathematics and Physics, Qingdao University of Science and Technology. Her current research interests include bioinformatics, machine learning, and data mining.

CHENG CHEN received the bachelor's degree from the College of Mathematics and Physics, Qingdao University of Science and Technology, in 2017, where he is currently pursuing the master's degree in statistics. His research interests include bioinformatics, machine learning, and data mining.

MINGHUI WANG received the Ph.D. degree in mathematics from East China Normal University, in 2008. He is currently a Professor with the College of Mathematics and Physics, Qingdao University of Science and Technology. His main research interests include computing mathematics and bioinformatics.

QIN MA received the Ph.D. degree in operational research from Shandong University, in 2010. He is currently an Associate Professor with the Department of Biomedical Informatics, The Ohio State University. His main research interests include bioinformatics, computational systems biology, and machine learning.

BIN YU received the M.Sc. degree in computational mathematics from Shanghai University, in 2005. He is currently pursuing the Ph.D. degree in bioinformatics with the University of Science and Technology of China. From 2018 to 2019, he was a Visiting Professor with the Department of Biochemistry and Molecular Biology, Medical Genetics, and Oncology, University of Calgary. He is currently an Associate Professor with the College of Mathematics and Physics, and the Research Director of the Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology. He has published about 70 research articles in various international journals and proceedings of conferences. His current research interests include bioinformatics, systems biology, machine learning, and biomedical big data mining.