IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 July 2023, accepted 15 July 2023, date of publication 20 July 2023, date of current version 3 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3297207

METHODS

A Protein-DNA Binding Site Prediction Method Based on Multi-View Feature Fusion of Adjacent Residue

JI YANG AND SHUNING ZHANG
The First Affiliated Hospital of Anhui University of Chinese Medicine, Hefei 230031, China
Corresponding author: Ji Yang (yangji@azyfy.com)

ABSTRACT The interaction between proteins and DNA occurs widely during the replication and transcription of DNA and other life activities. Therefore, the identification of protein- and DNA-binding sites is important for the study of protein function and drug design. Accurate prediction of binding sites has become a challenging and significant task. Although numerous studies have been conducted, prediction is challenging. In this study, a new protein-DNA binding site prediction method was proposed. This method is based on neighboring residue correlations. It uses an improved feature representation method that weighted combines several protein characteristics after a series of processing of the features and chooses a support vector machine as the prediction engine. Experiments on benchmark datasets and independent test datasets show that the proposed method has better predictability than other protein-DNA binding site predictors. This method is complementary to the existing protein-DNA binding site predictors and will be useful in the field of biotechnology.

INDEX TERMS Protein-DNA binding site prediction, neighboring residue correlations, feature processing, multi-view features combining, support vector machine.

I. INTRODUCTION

In living organisms, many biological activities are related to DNA molecules, including gene transcription and replication, DNA transcription, replication and recombination, and other key activities that occur during cell growth. With the help of some specific proteins, at the same time, these life activities will be regulated by the interaction between proteins and DNA [1], [2]. Correctly identified DNA-binding sites on proteins are not only related to the understanding of the mechanism of life activities but also help annotate the function of the protein and help in the design of drugs that promote or inhibit the expression of target genes [3]. The identification methods for DNA-binding proteins and specific DNA-binding sites are primarily based on traditional experimental methods [4]. Although this method has the advantage of high accuracy, it also has the disadvantages of a long experimental period, cumbersome experimental process, and the large amount of manpower and material resources required to complete the entire process. In recent years, with the development of bioinformatics, machine-learning prediction methods have been used to quickly and accurately predict the positions of potential binding sites in DNA-binding proteins.

In this field, many researchers have attempted to develop efficient and accurate forecasting methods. In these methods, features that have been used can be divided into two types: sequence features and structure features. For example, Hwang et al. [47] proposed a prediction method called DP-Bind based on three machine-learning methods and utilized only sequence feature: the Position Specific Scoring Matrix feature. Li et al. [7] proposed an improved method that integrates a structural alignment algorithm and support vector machine–based methods to predict DNA binding sites. Tsuchiya et al. [49] used structure-based features, such as the shape of the molecular surface, to build a predictor. Wang and Brown [6] built a prediction model called BindN, which

The associate editor coordinating the review of this manuscript and approving it for publication was Larbi Boubchir.

extracts three sequence features as input features, including side-chain pKa value, hydrophobicity index, and molecular mass, and utilizes Support Vector Machines(SVM) as classifiers. Zhou et al. [48] proposed a residue-encoding method that utilizes the evolutionary relationships between residues. Song et al. [9] proposed an adjustment algorithm that uses the binding probability between a target residue and its neighboring residues to predict the binding sites. Much progress has been made in binding-site prediction using machine learning. However, the efficiency and accuracy of the prediction remain unsatisfactory, leaving room for improvement. In their study, we found that the binding probability of the target residue could be affected by its adjacent residues on the left and right sides because certain amino acids are important for the interaction between proteins and DNA molecules. In addition, some structural and sequence features can be used together in predictors. Research on combining structural and sequence features based on neighboring residues is still lacking. Inspired by this, we propose a new sequence- and structure-based protein-DNA binding prediction method. This method uses a slide window to obtain the features of neighboring residues. Then, the weighted combination features after feature normalization and dimensionality reduction were extracted from multi-view protein feature sources. These features include structural features such as Accessible Surface Area(ASA) [7], Relative Solvent Accessibility based on structure (RSA-s) [8], Protrusion Index(PI) [9], Depth Index(DI) [10], sequence features such as Position Specific Scoring Matrix(PSSM) [11], and Relative Solvent Accessibility based on sequence (RSA-q) [12]. The experimental results of our method, whether on benchmark datasets or independent datasets, show that it can efficiently improve the accuracy of protein–DNA binding site prediction. Below. We explain how to approach each step individually.

II. MATERIALS AND METHODS

A. DATASETS

To fairly compare our method with other protein–DNA binding site predictors, two benchmark datasets and one independent test dataset were utilized. The following is a brief introduction to the three databases.

1) PDNA-62

The PDNA-62 dataset was first constructed and used by Ahmad and Sarai [13] to distinguish the binding sites using an ANN classifier. This dataset has been used in many studies, including ANN, SVM, Random Forest et al. [5], [17], [18]. PDNA-62 is a non-redundant dataset extracted from the Protein Data Bank (PDB, http://www.rcsb.org/pdb/) database. PDNA-62 was obtained from the structural data of 62 protein-DNA complexes in the PDB. Sequences with more than 25% homology in the protein-DNA complex sequences were removed using CD-HIT [19] software. Using 3.5 Å as the discrimination interval to distinguish whether the residues in the obtained protein sequence were DNA-binding residues or non-binding residues. Accordingly, get 1215 DNA-binding residues and 6948 non-binding residues to build the PDNA-62 dataset. As mentioned above, since this dataset has been used in many studies and has proven to be effective in distinguishing binding sites, we chose this dataset to compare the effectiveness of our method with other existing methods.

2) PDNA-224

PDNA-224 [20] is another dataset used for protein-DNA binding sites. This dataset was first constructed and used by Li et al. in PreDNA [7]. Compared with PDNA-62, PDNA-224 was obtained from the structural data of 224 protein-DNA complexes in PDB. The protein-DNA complex was filtered using the same method as that for PDNA-62, and sequences that were homologous to the PDNA-62 dataset were removed from the results. As a result, get 3778 DNA-binding residues and 53570 non-binding residues to build the PDNA-224 dataset. In PreDNA, which is based on sequence and structure information and other studies, PDNA-224 has a good effect on distinguishing binding sites. Therefore, this dataset was used to compare the performance of the proposed method with that of other existing methods.

3) APO29

To verify the generalizability of the proposed method, an independent test dataset called APO29 was built. The dataset was first constructed by Zhu et al. [21]. The construction method was similar to that of the two benchmark datasets. After the homology screening process, the sequences homologous to the PDNA-62 and PDNA-224 datasets in the screening results were removed to maximize the test accuracy and independence. There were 798 DNA-binding residues and 5979 non-binding residues in the APO29 dataset. According to Zhu et al., APO29 can be used to evaluate the predictive performance of a predictor. Thus, we used this dataset to compare the generalization ability of the proposed method with those of other methods.

B. FEATURE EXTRACTION METHOD

Amino acids in proteins do not exist in isolation. The state and combination of each amino acid in the sequence were affected by the surrounding amino acids, showing different results. Therefore, when extracting relevant features, it is not only possible to extract the currently selected amino acid itself but also to consider the relevant features of its surrounding adjacent amino acids. To solve this problem, the slide-window method [22] is a better choice. According to the size of the set sliding window, the currently selected amino acid and some other adjacent amino acids were combined into a sample with the currently selected amino acid as the center. The class label of the combined sample was the same as that of the currently selected amino acid class label. If the currently selected amino acid is not a protein-DNA binding site, the combined sample centered on this amino acid is in the non-binding state, that is, it is a negative sample; otherwise,

it is a positive sample. Because the size of the sliding window is different, it has a certain impact on the performance of the built prediction model. Therefore, the selection of its size is more important, and the specific selection process will be described later.

III. PROTEIN FEATURE EXTRACTION

A. PROTEIN STRUCTURE FEATURE EXTRACTION

1) ACCESSIBLE SURFACE AREA FEATURE

The (ASA) was first proposed by Lee and Richards [23]. Further research has shown that it plays an important role in predicting protein-DNA interactions [7]. ASA is the surface area of residues accessible to solvent molecules when atoms of a protein or DNA molecule are present in solution. The operational definition is to use a solvent molecule to probe the ball and roll along the protein surface, and all possible trajectory points at the center of the probe can outline a surface. The surface area was defined as the accessible surface area. The calculation formula is as follows:.

$$D = \Delta Z/2 + \Delta' Z \tag{1}$$

$$A = \sum (R/\sqrt{R^2 - Z_i^2}) \times D \times L_i \tag{2}$$

where R refers to the sum of the van der Waals radii of the atoms in the measured and solvent molecules. Parameter L_i refers to the arc length through which the easy molecule rolls in the specified area i. Parameter Z_i refers to the vertical distance from the center of the ball to the i-th area, and parameter ΔZ indicates the distance between the areas. Parameter $\Delta' Z$ is smaller than the $R - Z_i$ and $\Delta Z/2$ parameters. For a given atom, the sum of all scrollable arcs must be computed. We calculated ASA features using the NACCESS program [24].

2) RELATIVE SOLVENT ACCESSIBILITY FEATURE (BASED ON STRUCTURE)

Relative Solvent Accessibility(RSA) has been widely used in many fields, including three-dimensional protein structures, protein-DNA interactions, and protein-related ligand interactions [25]. The calculation formula is as follows:

$$RSA = \frac{ASA}{MaxASA} \tag{3}$$

The parameter *ASA* is solvent accessible. The parameter *MaxASA* is the maximum possible amount of solvent accessible to the residue.

Five pairs of features related to solvent accessibility and relative solvent accessibility were constructed using an algorithm [26]. These features include all atoms on an amino acid, main-chain or backbone atoms, side-chain atoms, polar side-chain atoms, and nonpolar side-chain atoms.

3) PROTRUSION INDEX AND DEPTH INDEX FEATURE

The protrusion index (PI) [27] and Depth Index(DI) [28] are typically used to describe and distinguish internal spatial structures of proteins. PI was first used in the study of protein-protein interactions, antigenic determinants, and proteolytic cleavage. With the deepening of research in recent years, its important role in the interactions between DNA and proteins has become apparent. The calculation formula is as follows:.

$$V_{ext} = V_{sphere} - V_{int} \tag{4}$$

$$V_{int} = N_{atom} \times V_{atom} \tag{5}$$

$$\mathrm{PI} = V_{ext}/V_{int} \tag{6}$$

A probe sphere was formed with a protein nonhydrogen atom as the center and a fixed distance R as the radius. N_{atom} Here, refers to the number of non-hydrogen atoms in this sphere, and the default radius of the probe sphere is 10 Å. V_{atom} is the average volume of a heavy atom in a protein; here, the value was 20.1Å^3. PI is a six-dimensional vector consisting of the mean, standard deviation, and maximum and minimum protrusion values of all atoms in the residue. Mean and standard deviation of the protrusion values of the side-chain atoms. Each element of this vector is normalized, and its range is–0-1.

DI has a wide range of applications such as rate analysis of amino hydrogen/deuterium exchange in nuclear magnetic resonance (NMR), protein nuclear assembly, and alignment analysis. In addition, this feature is helpful in predicting the interaction sites between proteins and DNA. DI refers to the distance between atom i and the adjacent solvent-accessible atom j (i.e., an atom with ASA value > 0). The calculation formulae are as follows:

$$DI = min(d_1, d_2, d_3, \ldots d_n,) \tag{7}$$

$d_1, d_2, d_3, \ldots d_n$ is the distance between atom i and all solvent-accessible atoms. Therefore, for solvent-accessible atoms, the depth index was 0. For internal atoms, the depth index is proportional to distance.

PSAIA [29] is a cross-platform program that encapsulates computational DI, PI, and other tools. Obtaining PI and DI feature values using this program is required.

B. PROTEIN SEQUENCE FEATURE EXTRACTION

1) POSITION SPECIFIC SCORING MATRIX FEATURE

The position-specific scoring matrix (PSSM) [30] can reflect the evolutionary information of protein sequences through multiple sequence alignments. Many bioinformatics studies have shown that PSSM plays an important role in the prediction of protein structure and function [31]. Consider a protein sequence, P, containing L amino acids as an example. PSSM was obtained using the PSI-BLAST [32] program with an E value of 0.001. PSI-BLAST searched the Swiss-Prot database to perform multiple protein sequence alignments and three iterations. The PSSM has L rows and 20 columns as follows:

$$P_{pssm}^{original} = \begin{bmatrix} O_{1,1} & O_{1,2} & \cdots & O_{1,20} \\ O_{2,1} & O_{2,2} & \cdots & O_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ O_{k,1} & O_{k,2} & \cdots & O_{k,20} \\ O_{L,1} & O_{L,2} & \cdots & O_{L,20} \end{bmatrix}_{L \times 20} \tag{8}$$

where $O_{k,j}$ represents the score of amino acid K in protein sequence P mutated to amino acid J during the evolutionary process. A positive score indicated that the mutation was more likely to occur than expected and a negative score indicated that the mutation was less likely to occur than expected. After obtaining the PSSM of the original protein, normalization was performed. The 20 natural amino acids are represented by 1-20 and sorted according to the alphabetical order of their first letters. First, the mean and standard deviation of each row of the original PSSM were calculated. Consider the kth row as an example, where μ_k and σ_k represent the mean and standard deviation, respectively, of the kth row. The calculation formulae are as follows:

$$\mu_k = \frac{1}{20}\sum_{t=1}^{20} O_{k,t} \tag{9}$$

$$\sigma_k = \sqrt{\frac{1}{20}\sum_{t=1}^{20}(O_{k,t} - \mu_k)^2} \tag{10}$$

The PSSM matrix obtained after performing the normalization operation is denoted as P_{pssm}, and the calculation method for the element values of rows K and J is as follows:

$$P_{k,j} = \frac{O_{k,j} - \mu_k}{\sigma_k} \tag{11}$$

After the final normalization, the PSSM matrix representation of a protein sequence containing L-amino acids was as follows:

$$P_{pssm} = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,20} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ P_{k,1} & P_{k,2} & \cdots & P_{k,20} \\ P_{L,1} & P_{L,2} & \cdots & P_{L,20} \end{bmatrix}_{L\times 20} \tag{12}$$

The PSSM feature vector corresponding to each amino acid was extracted using the PSSM matrix.

2) RELATIVE SOLVENT ACCESSIBILITY FEATURE (BASED ON SEQUENCE)

In addition to the structure-based Relative Solvent Accessibility(RSA) mentioned above, the algorithm [33] was used to extract RSA features based on protein sequence information. The steps of the algorithm are as follows.

Step one: Screen the protein data contained in the (Protein Data Bank, PDB) by setting the X-ray crystal diffraction index to 3.0 Å and the R-factor value of the crystal structure correction to 0.3. Among the screened results, proteins with fewer than 50 amino acids, as determined by multidimensional nuclear magnetic resonance, were removed, and the homology between proteins was guaranteed to be less than 25%. A dataset containing 5717 protein chains and 1242356 amino acids was obtained. Using the aforementioned structural method, the RSA was calculated to construct an RSA database.

Step two: Set the sliding window size to 15, and calculate the distance D between the amino acids. The formula for calculating the distance D^{AB} between amino acids A and B is as follows:

$$D^{AB} = \sum_{i,j} w_i |P_{ij}^A - P_{ij}^B| \tag{13}$$

P_{ij}^A is an amino acid in a sliding window. w_i represents the weight value obtained based on its position in the sliding window (position i at the center of the window was set to 8). The weight value w_i is calculated as follows:.

$$w_i = (8 - |8 - i|)^2 \tag{14}$$

Step three: According to the obtained distance value, select the K-nearest neighbor amino acids from all amino acids in the database, and calculate the value of z.

$$z = (D_{ave} - D)/\sigma \tag{15}$$

where D_{ave} represents the average distance between the target amino acid and all amino acids in the database. where, σ is the standard deviation. Based on the above, the value of sequence-based RSA was calculated as follows:

$$\mathrm{RSA} = \frac{\sum_{i=1}^{k} RSA_i z_i^{\alpha}}{\sum_{i=1}^{k} z_i^{\alpha}} \tag{16}$$

where RSA_i represents the ith nearest-neighbor amino acid in the database. The parameter α adjusts the relative importance of each adjacent amino acid. The optimal values of parameters α and K were determined using the grid search method. The values of K and α used in this study are 64 and 6.31, respectively.

IV. FEATURE PROCESSING METHOD

A. FEATURE SELECTION

Many features are related to proteins and DNA and different features have different effects. Therefore, it is very important to select appropriate features for the prediction of binding sites between proteins and DNA. Currently, there are many commonly used feature selection methods, including breadth-first [34], beam-first [35], and best-first [36]. In this study, the Best First feature selection method was used to select features. The search steps for this method were as follows:

Step one: Obtain the individual prediction performance of each feature in the dataset using the classification algorithm.

Step two: Arrange the prediction performance of all obtained features or feature combinations from low to high.

Step three: Select the feature or feature combination with the best prediction effect from all the prediction results. Then, it is combined with one of the remaining unselected features, and its prediction performance on the dataset is obtained through the classification algorithm.

Step four: Compare The prediction performance of the combination features in the third step on the dataset is compared with the prediction performance obtained in the second step. If it is not as good as the prediction performance in the second step, the experiment ends, and the feature or feature combination obtained in the second step is the final selected feature. Otherwise, it jumps to the second step to continue execution.

B. WEIGHTED FEATURE FUSION

Different protein features describe their relevant properties from different perspectives, and the fusion of these features can compensate for the lack of protein information caused by single-view features. At present, there are two main methods of multifeature fusion: serial feature fusion and parallel feature fusion [37].

Owing to the large difference in dimensionality between the protein sequence and structural features extracted in this study, the dimensionality of the PSSM feature vector was 220, whereas the dimensionality of the DI feature vector was 66; therefore, we chose the weighted serial feature fusion method. The specific steps are as follows:

Assuming that the two feature vectors to be fused are $\mathbf{A}_1$ and $\mathbf{A}_2$, these two features are different, the dimensions are m_1 and m_2 respectively, their weight indices are set to P_1 and P_2 respectively, and the calculation formulas of P_1 and P_2 are as follows:

$$P_1 = \frac{P_{\mathbf{A}_1}}{P_{\mathbf{A}_1} + P_{\mathbf{A}_2}} \tag{17}$$

$$P_2 = \frac{P_{\mathbf{A}_2}}{P_{\mathbf{A}_1} + P_{\mathbf{A}_2}} \tag{18}$$

$P_{\mathbf{A}_1}$ represents the prediction accuracy obtained by the single-view feature vector $\mathbf{A}_1$ under the classification algorithm. $P_{\mathbf{A}_2}$ represents the prediction accuracy obtained by the single-view feature vector $\mathbf{A}_2$ under the classification algorithm. These two index values represent the importance of the two features to the research problem. The new eigenvector $\mathbf{A}$ obtained after serial fusion is as follows:

$$\mathbf{A} = [P_1\mathbf{A_1}, P_2\mathbf{A_2}] \tag{19}$$

The new feature vector $\mathbf{A}$ obtained after weighted fusion contains all the information in the feature vector to be fused. Because the importance of different features in the research problem is different, the weights obtained are also different.

However, if the protein features from different perspectives are simply combined without other processing, it may lead to the inability to fully utilize the multi-view feature information of the protein. To avoid this problem, we set the weights for the different perspective features to ensure that the features of each perspective can fully play their role in the prediction model. The step of setting weights is further discussed in the ''Results and Discussion'' section.

C. FEATURE NORMALIZATION

Because the value ranges of the eigenvalues obtained by different feature extraction algorithms may be different, this may lead to imbalance problems. For example, the role of a feature vector that is much larger than those of other feature vectors in classification prediction may be exaggerated, whereas the role of a feature vector with a small value may be ignored. Thus, we used the linear function transformation method to normalize the data as follows:

$$y = (x - Min)/(Max - Min) \tag{20}$$

D. FEATURE DIMENSIONALITY REDUCTION

Higher feature dimensions complicate the prediction of protein-DNA binding sites. Therefore, it is necessary to use a data dimensionality reduction method to process the obtained features. Current traditional data dimensionality reduction methods include Linear Discriminant Analysis(LDA) [39] and Principal Component Analysis(PCA) [40]. Although PCA has many advantages, it has a major disadvantage: it may not perform well for high-dimensional data. Therefore, in this study, we chose a method of generalization based on PCA: Generalized Principal Component Analysis(GPCA) [37]. Steps are as follows.

If it is considered that eigenvector $\mathbf{A}$ is in a single space, the total number of pattern classes is denoted by M, and the prior probability corresponding to pattern class i is expressed as $\mathrm{P}(\omega_i)$. The corresponding average eigenvector calculation formula is as follows:

$$\overline{\mathbf{A}_i} = E\left\{(\mathbf{A}|\omega_i)\right\} \tag{21}$$

The formula for calculating the mean vector of all eigenvectors is as follows.

$$\overline{\mathbf{A}_i} = E\left\{\mathbf{A}\right\} = \sum\nolimits_{i=1}^{M} \mathrm{P}(\omega_i) \cdot \overline{\mathbf{A}_i} \tag{22}$$

According to the data obtained from the above calculation, the total scattering matrix S_t can be expressed as follows:

$$S_t = E\left\{\left(\mathbf{A} - \overline{\mathbf{A}}\right)\left(\mathbf{A} - \overline{\mathbf{A}}\right)^H\right\} \tag{23}$$

S_t is Hermite matrix. According to Ding and Cai [50] and Liu and Wechsler [37], the GPCA can be explained as follows:

$\mathbf{C}_1, \mathbf{C}_2, \ldots \mathbf{C}_m$ is the orthogonal eigenvectors corresponding to S_t. $\lambda_1, \lambda_2, \ldots \lambda_m$ are the corresponding eigenvalues, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$. The first feature with a maximum value of n is filtered out and used as the projection axis. Then, for each eigenvector $\mathbf{A}$, an n-dimensional projection vector can be obtained; let it be $\mathbf{B}$. The calculation formula is as follows:

$$\Phi = (\mathbf{C}_1, \mathbf{C}_2, \ldots \mathbf{C}_n) \tag{24}$$

$$\mathbf{B} = \Phi^H \mathbf{A} \tag{25}$$

The $\mathbf{B}$ vector is called a dimensionality reduction vector, through which the original feature vector $\mathbf{A}$ is replaced.

The size of the dimension m after dimension reduction is different, which also affects the prediction performance. The procedure for selecting the optimal value of m is described in the ''Results and Discussion'' section.

V. CLASSIFIER SELECTION

The classification and prediction performance of the prediction model are not only related to the classification ability of the feature but also to the classifier selected by the model. In this study, three classifiers are selected and used, which are Support Vector Machine(SVM) [41], Radial Basis Function(RBF) [42], and Random Forest(RF) [43].

SVM: SVM is a classification method based on statistical theory in data mining, and is established based on the two principles of VC dimension and structural risk minimization of statistical learning theory. The SVM classifier can obtain better generalization ability under limited sample conditions and is suitable for many linear and nonlinear problems, such as pattern recognition and regression. This solves the problem that classifiers require large-capacity samples and can prevent problems such as the curse of dimensionality. It has been widely used in the identification and prediction of protein spots and protein-DNA binding sites [44], [45], etc.

SVMs have two forms: support vector classification (SVC) and support vector regression (SVR). In this study, because the research problem was to identify the binding site, SVC was selected as the classifier of the model.

Gaussian kernel function is one of the most commonly used SVM kernel functions. The calculation formula is as follows:

$$K(\mathbf{x_i}, \mathbf{x_j}) = e^{-\|x_i - x_j\|^2/2\sigma^2} \tag{26}$$

The construction of the predictive model needs to determine the optimal values of the regulation parameter γ and the kernel parameter σ. In; to obtain the best value of these two parameters, we used the grid search strategy provided in LIBSVM software through ten-fold cross-validation.

RBF: Lowe [46] and others were the first to combine radial basis functions with neural network applications. The RBF network has a wide range of applications and can be used not only for classification recognition and prediction but also for function approximation. RBF has the characteristics of a simple structure and excellent classification prediction performance.

RBF is a three-layer network that is currently used frequently. The first layer is the input layer, which is mainly composed of signal-source nodes, and the input is the feature value of the predicted sample. The second layer was the hidden layer. This layer transforms the input samples and uses the kernel function to convert the input data that are inseparable in the low-dimensional space into a high-dimensional space, and converts it into separable data for classification and recognition. The third layer is the output layer, which outputs the category of the predicted samples often using a linear activation function. The Gaussian function is defined as follows:

$$\mathrm{g}\,(\mathrm{x}) = \exp\left(\frac{\|\mathbf{X} - \mathbf{C}\|^2}{2\sigma^2}\right) \tag{27}$$

where x denotes the input training sample point. c is the center of the radial basis function. σ is the variance, which represents the width of the kernel function and controls the radial range of the basis function. The RBF network has three adjustable parameters: the center of the radial basis function, variance of the radial basis function, and weight from the hidden layer to the output layer of the radial basis function network. The optimal values of these three parameters were obtained as follows.

Supervised learning with error-correction algorithms from training samples. First, the three parameters of the basis function center, variance, and weight are randomly initialized. Then, they were gradually adjusted and corrected using the gradient descent method until the optimal solution was obtained. Based on these parameters, we built an RBF network to identify and predict protein-DNA binding sites.

RF: Breiman [43] and others proposed a new machine learning method in 2001, which has a better generalization effect and classification ability than traditional decision trees, and named it Random Forest. RF integrates multiple trees through ensemble learning, and its basic unit is a decision tree. RF can construct different numbers of decision subtrees according to different problems. It uses multiple decision trees to train, classify, and predict the samples. The random forest method has been widely used in various bioinformatic fields. Because of its good performance in the prediction of many protein-DNA-related interactions, we chose it as a classifier for the prediction of protein-DNA-binding sites. For the sample set S, the specific training steps of random forest are as follows:

Step one: Random forest uses a self-service sampling method to obtain the training sample set of each tree. Randomly generate K subsample sets from the original sample set S. Each self-service sample set has N samples, and then, the K sample sets are trained as a single classification tree.

Step two: During the training process, if the internal nodes of the tree must be split, m (mM) is randomly selected as candidate attributes according to the M attributes of each sample. Then, the principle of minimum node impurity is adopted to select an attribute from m candidate attributes for splitting.

Step three: During the growth of the classification tree, each node must be split according to step two until it is completely split. The entire process does not require pruning.

Step four: Each trained tree classifier is formed into a random forest. Each classification tree in the random forest predicts new data, and voting is then used to obtain the final classification prediction result.

There are two main types of random forests: classification and regression. This study used a random forest classification algorithm. We need to predetermine the number of samples (*mTry*) randomly selected at each split before using the RF. According to Breiman's suggestion, if the total number of features is M, the number of samples selected can be $sqrt\,(M)$, $sqrt\,(M)/2$, $2 * sqrt\,(M)$. In this study, according to the experimental results, the number of samples (*mTry*) that select $sqrt\,(M)$ can obtain the best prediction result. Another important parameter was the number of trees (*nTree*). According to the experimental results, when *nTree* is set to M, the prediction effect is the best.

VI. EVALUATION INDICES AND VALIDATION STRATEGIES

A. EVALUATION INDICES

In this study, we discuss a binary classification problem. The classification results were Positive or Negative. In a

binary classification problem, the prediction results can be represented by a 2 × 2 confusion matrix. The confusion matrix was divided into the following four cases: (1) correctly predicted protein-DNA binding sites (True Positive, TP); (2) correctly predicted protein-DNA non-binding sites (True Negative, TN); (3) mispredicted protein-DNA binding sites (False Negative, FN); and mispredicted protein-DNA non-binding sites (False Positive, FP). The four situations are listed in Table 1.

TABLE 1. Confusion matrix contingency table.

Predict \ Actual	Positive	Negative
Negative	FN	TN
Positive	TP	FP

Using the four parameters TP, TN, FP, and FN introduced in Table 1, the values of the four evaluation indicators can be calculated: sensitivity (Sen), specificity (Spe), accuracy (ACC), Matthew's Correlation Coefficient (MCC) [51], [52], [53], and Youden's index. The specific calculation method is as follows:

$$Sen = \frac{TP}{TP + FN} \tag{28}$$

$$Spe = \frac{TN}{TN + FP} \tag{29}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{30}$$

$$MCC = \frac{TP \times TN + FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \tag{31}$$

$$YoudensIndex = Sen + Spe - 1 \tag{32}$$

The two indices of ACC and MCC take into account the comprehensive predictive ability of positive and negative samples, and both can evaluate the predictive ability of the prediction model well. However, for the sample imbalance problem, the MCC evaluation index is more representative than ACC. For example, when 95% of the sample set was negative, if the ACC was 95%, it seemed very high. In fact, the model only defines all samples as negative samples, and is not predictive of positive samples. However, MCC is an index that can normally reflect the prediction performance of the prediction model when the number of samples of the two types is unbalanced. Therefore, the prediction ability of the model in this study mainly refers to the MCC value of the prediction result followed by the ACC value.

B. CROSS-VALIDATION

A ten-fold cross-validations method was used to test the method proposed in this study and compare it with the current mainstream prediction methods. In this way, both over- and under-learning can be avoided, making the final result more convincing.

VII. RESULTS AND DISCUSSION

A. SLIDING WINDOW SIZE SELECTION

As mentioned above, the sizes of the sliding windows are different, which has a significant impact on the predictive performance of the built model. To select the optimal sliding window size, we used the SVM classification algorithm on the PDNA-224 dataset based on PSSM features to perform ten-fold cross-validations. The sliding window was set to a range from 3 to 21. Because the sliding window needs to select the adjacent amino acid information on both sides of the amino acid simultaneously, the test was performed with a step size of 2. The test results are presented in Table 2.

TABLE 2. The results of ten-fold cross-validations under different sliding windows using the SVM classifier on the PDNA-224 dataset based on PSSM features.

Sliding window size	Sen(%)	Spe(%)	ACC(%)	MCC	Youden's Index
3	65.7	74.8	74.1	0.231	0.405
5	66.6	76.5	75.8	0.249	0.431
7	65.8	76.7	76.0	0.247	0.425
9	67.6	76.4	75.7	0.254	0.440
11	67.8	76.9	76.3	0.261	0.447
13	67.8	76.9	76.2	0.260	0.447
15	66.7	76.5	75.8	0.250	0.432
17	66.2	77.0	76.2	0.251	0.432
19	65.4	76.9	76.1	0.246	0.423
21	67.2	76.3	75.7	0.252	0.435

As shown in Table 2, when the step size was 2 and the sliding window ranged from 3 to 21, Sen, Spe, ACC, MCC, and Youden's index showed a trend of rising and decreasing fluctuations after reaching the peak. Sen increased from 65.7% to 67.8% and then decreased to 67.2%. After Spe increased from 74.8% to 76.9%, it continued to decrease to 76.3%. The ACC fluctuated from 74.1% to 76.3%, then decreased, and finally reached 75.7%. Youden's index fluctuated from 74.1% to 76.3%, decreased, and finally reached 75.7%. MCC fluctuates from 0.23 to 0.26, and then fluctuated to 0.25, as shown in Figure 1.

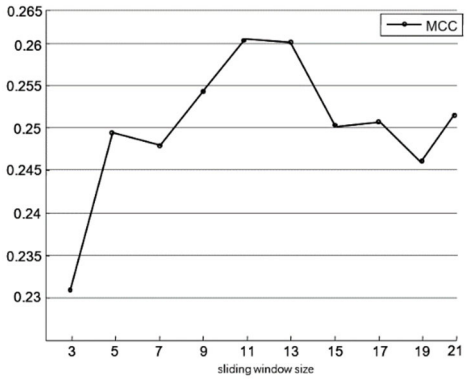

FIGURE 1. Based on PSSM features and SVM classification algorithm, on the PDNA-224 dataset, ten-fold cross-validations are performed, the step size is 2, the sliding window size is from 3 to 21, and the change curve of the evaluation index MCC.

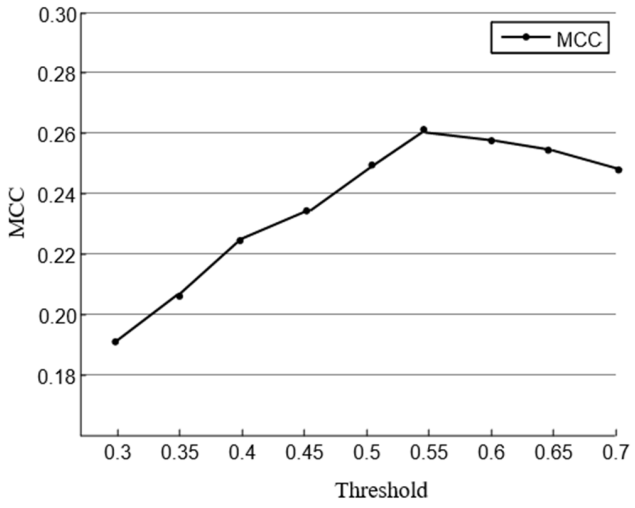

FIGURE 2. Based on PSSM features and SVM classification algorithm, ten-fold cross-validations are performed on the PDNA-224 dataset, the step size is 0.05, the threshold size is from 0.3 to 0.7, and the change curve of the evaluation index MCC.

Figure 1 and Table 2 show that the sizes of the sliding windows were different, which had a greater impact on the classification results. ACC and MCC achieved maximum values of 76.3% and 0.26, respectively, when the sliding window size was 11. From the above analysis, it can be concluded that when extracting features, it is most appropriate to use a sliding window with a size of 11 to extract the features of the target protein residue and its surrounding adjacent residues, and to maximize the performance of the prediction model.

B. THRESHOLD SIZE SELECTION

When the prediction model predicts whether a protein residue is a protein-DNA binding site, it first calculates the probability value of the residue being a binding site and the probability value of not being a binding site. However, it cannot be considered that the residue is the binding site if the probability value is greater than 0.5. Different classification thresholds had a significant impact on the prediction performance of the constructed prediction model. On the PDNA-224 dataset, based on the PSSM feature, the SVM classifier was used to perform ten-fold cross-validation to determine the optimal threshold size and illustrate the impact of different threshold sizes on the performance of the prediction model constructed in this study.

TABLE 3. The results of ten-fold cross-validations at different thresholds using the SVM classifier on the PDNA-224 dataset based on PSSM features.

Threshold size	Sen(%)	Spe(%)	ACC(%)	MCC	Youden's Index
0.30	90.0	46.5	49.6	0.188	0.365
0.35	86.6	53.8	56.1	0.206	0.404
0.40	82.1	61.0	62.4	0.223	0.431
0.45	77.3	67.0	67.7	0.236	0.443
0.50	72.9	71.7	71.9	0.247	0.446
0.55	67.2	77.5	75.7	0.260	0.447
0.60	61.3	82.9	74.3	0.258	0.442
0.65	52.9	86.0	73.7	0.253	0.389
0.70	46.5	89.6	71.4	0.249	0.361

It can be seen from Table 3 that when 0.05 is the step size and the threshold value ranges from 0.3 to 0.7 for the experiments, the results obtained using different thresholds are quite different. Overall, Sen exhibited a downward trend from 90% to 46.5%. Overall, Spe increased from 46.5% to 89.6%. ACC, MCC, and Youden's index showed a trend of rising first and then falling after reaching the peak. The ACC increased from 49.6% to 75.7% and then dropped to 71.4%. MCC increased from 0.188 to 0.260 before falling to 0.249, as shown in Figure 2. Youden's index rose from 0.365 to 0.447 and then dropped to 0.361.

After analyzing Table 3 and Figure 2, it can be seen that the classification was performed according to different threshold values, and the results were significantly affected by the threshold value. At the same time, observing the change in the MCC value, it can be seen that its change is small in the interval 0.55 to 0.6, and the maximum value is obtained in this interval. Based on the optimization principle, we further divided the area and refined the threshold area with a step size of 0.01. Table 4 presents the results. It can be seen that when the threshold is 0.56, the maximum MCC value can be obtained, which is 0.261.

From the above analysis of the results, it can be seen that when using the proposed method to predict protein-DNA binding sites on the PDNA-224 dataset, the optimal threshold used was 0.56, and the best threshold was obtained under this threshold. predictive performance.

TABLE 4. The results of ten-fold cross-validations using the SVM classifier on the PDNA-224 dataset based on PSSM features at different thresholds.

Threshold size	Sen(%)	Spe(%)	ACC(%)	MCC	Youden's Index
0.55	67.2	77.5	75.7	0.260	0.447
0.56	66.7	78.0	75.7	0.261	0.447
0.57	64.9	79.6	75.2	0.260	0.445
0.58	63.4	81.0	75.0	0.259	0.444
0.59	61.9	81.8	74.6	0.258	0.437
0.60	61.3	82.9	74.3	0.258	0.442

TABLE 5. The results of ten-fold cross-validations using three classifiers on the PDNA-224 dataset based on each single-view feature. RSA-s[a] means RSA features based on structure, RSA-q[b] means RSA features based on sequence, D_P[c] means DI and PI features.

Feature	Classification algorithm	Sen (%)	Spe (%)	ACC(%)	MCC	Youden's Index
PSSM	SVM	66.7	78.0	75.7	0.26	0.447
	RBF	63.5	75.3	74.2	0.24	0.388
	RF	62.9	74.8	73.4	0.23	0.377
ASA	SVM	67.8	63.1	63.5	0.16	0.309
	RBF	66.7	60.3	61.2	0.15	0.270
	RF	63.6	58.6	60.1	0.13	0.222
RSA-s[a]	SVM	67.8	64.3	65.1	0.18	0.321
	RBF	66.9	63.6	64.6	0.17	0.305
	RF	63.8	63.3	63.9	0.17	0.271
RSA-q[b]	SVM	63.0	58.7	60.0	0.11	0.217
	RBF	61.8	57.3	59.7	0.11	0.191
	RF	60.2	57.0	58.9	0.10	0.172
D_P[c]	SVM	65.6	60.1	60.5	0.13	0.257
	RBF	63.4	59.8	60.1	0.13	0.232
	RF	61.3	58.1	58.9	0.11	0.194

C. FEATURE SELECTION

1) COMPARISON OF SINGLE-VIEW FEATURE RESULTS

First, we extracted the single-view features introduced above and used the three classification algorithms described above as classifiers on the PDNA-224 standard dataset to perform ten-fold cross-validations. The experimental results are presented in Table 5.

Analyzing Table 5, it can be concluded:

1. All of these sequence and structural features have a certain effect on the prediction of protein and DNA binding sites, and the overall prediction accuracy exceeds 60%. Overall, the predictive effects of the structural features were relatively close, and the best predictive effect was the RSA feature.
2. Combining Table 5 and Figure 3, it can be seen that the prediction results of the three classifiers introduced above (SVM, RBF, and RF) show little difference. However, it can be observed that in terms of single-view feature prediction, the prediction performance of the SVM classifier is better.
3. Overall, the PSSM feature had the best effect; its prediction accuracy exceeded 75%, and the MCC value of its prediction result was also the highest. It can be seen that the relative evolution information of sequences has a great influence on the prediction of protein-DNA binding sites.

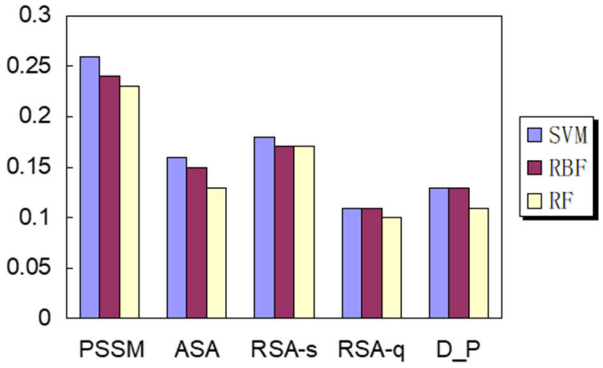

FIGURE 3. The value of MCC in the results of ten-fold cross-validations using three classifiers on the PDNA-224 dataset based on each single-view feature.

2) COMPARISON OF MULTI-VIEW FEATURE COMBINATION RESULTS

After obtaining the prediction results for the single-view features introduced above, we combined the obtained single-view features according to the Best First method. For the PDNA-224 standard dataset, the three classification algorithms introduced above were used as classifiers to perform ten-fold cross-validations to obtain the best combination of features. The experimental results are presented in Table 6.

From Table 6, we can see:

1. According to the above experimental results, the features extracted in this study have a certain effect on the prediction of protein-DNA binding sites. Thus, we attempted to combine the features obtained from a single perspective to observe the prediction effect using the first-best algorithm. After combining PSSM features with D_P features, it was found that the predictive performance of this combination was better than that of a single-view feature. Compared to using PSSM features alone, Sen, Spe, ACC, MCC, and Youden's index increased by 3.6%, 1.1%, 1.4%, 3%, and 4.7%, respectively. Based on this, we attempted to combine different single-view features to observe this effect. After combining the ASA features with the previous PSSM and D_P features, it was found

TABLE 6. The results of ten-fold cross-validations using three classifiers on the PDNA-224 dataset based on multi-view feature. D_P[a] means DI and PI features, RSA-s[b] means RSA features based on structure, RSA-q[c] means RSA features based on sequence.

Feature	classification algorithm	Sen (%)	Spe (%)	ACC (%)	MCC	Youden's Index
PSSM+D_P[a]	SVM	70.3	79.1	77.1	0.29	0.494
	RBF	69.4	77.9	73.6	0.28	0.473
	RF	68.0	76.5	72.9	0.28	0.445
PSSM+D_P[a]+ASA	SVM	76.9	81.3	81.0	0.35	0.582
	RBF	75.3	80.6	80.1	0.33	0.559
	RF	73.0	79.8	79.3	0.32	0.528
PSSM+D_P[a]+ASA+RSA-s[b]	SVM	78.1	83.8	83.3	0.39	0.619
	RBF	77.6	82.1	81.5	0.37	0.597
	RF	76.3	81.3	80.1	0.36	0.576
PSSM+D_P[a]+ASA+RSA-s[b]+RSA-q[c]	SVM	77.6	86.3	85.8	0.41	0.639
	RBF	74.1	85.6	84.1	0.39	0.597
	RF	73.2	84.1	81.9	0.38	0.573

that the feature combination performed better than the previous PSSM and D_P feature combinations, and the prediction performance was further improved. Sen, Spe, ACC, MCC, and Youden's index increased by 6.6%, 2.2%, 3.9%,6%, and 8.8%, respectively. Continue to combine the features. After combining RSA features based on structure with PSSM, D_P, and ASA features, the prediction performance of this feature combination is improved. Sen, Spe, ACC, and MCC increased by 1.2%, 2.5%, 2.3%, 4%, and 3.7%, respectively, compared to the previous feature combination. After combining the last feature RSA features based on the sequence with the PSSM, D_P, ASA, and RSA features, the MCC and ACC of the feature combination were 2% and 2.5% higher, respectively, than before. Through these, it can be seen that the prediction effect of the multi-view feature combination is higher than that of the single-view feature. The main reason for this is that there is a certain degree of complementary and interrelated relationship between different protein features.

2. The feature combination of RSA based on structure, PSSM, D_P, ASA, and RSA based on sequence features had the best predictive performance, and its predictive ability was greatly improved compared to the features extracted from a single perspective.
3. From Table 6 and Figure 4, it can be seen intuitively that the SVM classifier performed well in single-view feature prediction, and its prediction performance was still ahead of the other two predictors in multi-view feature combinations.

According to the single-view feature prediction results and related multi-view feature combination prediction results, the best feature combination can be selected using the Best First algorithm, and the results are shown in Table 7.

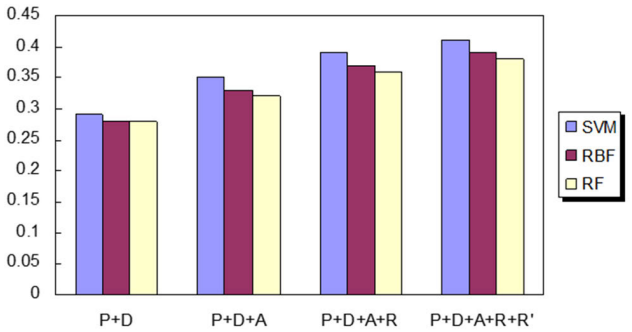

FIGURE 4. The value of MCC in the results of ten-fold cross-validations using three classifiers on the PDNA-224 dataset based on each multi-view feature P stands for PSSM, D stands for D_P, A stands for ASA, R stands for RSA based on structure, R′ stands for RSA based on sequence.

TABLE 7. The best feature combination after the Best First algorithm selection. D_P[a] means DI and PI features, RSA-s[b] means RSA features based on structure, RSA-q[c] means RSA features based on sequence.

Features	Acc (%)	MCC
PSSM	75.7	0.26
ASA	63.5	0.16
RSA-s	65.1	0.18
D_P	60.5	0.13
RSA-q	60.0	0.11
PSSM+D_P[a]	77.1	0.29
PSSM+D_P[a]+ASA	81.0	0.35
PSSM+D_P[a]+ASA+RSA-s[b]	83.3	0.39
PSSM+D_P[a]+ASA+RSA-s[b]+RSA-q[c]	85.8	0.41

D. FEATURE FUSION

There are differences between the features obtained from different perspectives and the functions they perform are also different. By simply combining them, we cannot show the different roles played by the features from different perspectives. After the best feature combination is selected by the Best First algorithm, it must be processed by the weighted feature fusion method to improve the prediction accuracy of the constructed model.

As mentioned above, owing to the large dimensional difference between the extracted protein sequence and the structural features, the obtained feature vectors were processed using the weighted serial feature fusion method. As the value ranges of the feature values obtained from different perspectives may differ, it is necessary to normalize them

before fusing the features from different perspectives. The normalization method chosen in this study is the linear function conversion method, and its calculation formula has been previously introduced.

After several experiments, it was found that when the feature weight was set to 0.8, the effect was optimal. On the PDNA-224 dataset, weighted serial fusion was performed on the feature combinations obtained through feature selection, and the fused results were used to perform ten-fold cross-validations using the SVM, RBF, and RF classification algorithms.

The result is shown in Table 8.

TABLE 8. Prediction results after ten-fold cross-validations using three classifiers on the PDNA-224 dataset before and after feature fusion. D_Pa means DI and PI features, RSA-sb means RSA features based on structure, RSA-qc means RSA features based on sequence.

Feature	Classification algorithm	Sen (%)	Spe (%)	ACC (%)	MCC	Youden's Index
Before	SVM	77.68	86.30	85.81	0.410	0. 639
	RBF	76.25	85.79	85.27	0.403	0. 620
	RF	74.91	83.81	82.36	0.391	0. 587
PSSM+ $w \cdot$ D_P^{a}+ASA+ $w \cdot$ RSA-s^{b}+RSA-q^{c}	SVM	78.89	86.33	85.86	0.417	0. 652
	RBF	77.10	85.09	85.32	0.405	0. 622
	RF	75.65	84.96	84.61	0.394	0. 606

Table 8 shows that after using the weighted serial fusion algorithm to perform feature fusion processing on the best feature combination, the fused feature set performed better in protein-DNA binding site prediction than the simple linear combination feature set. After normalization and fusion, the feature groups improved by approximately 1.2%, 0.03%, 0.05%, 0.7%, and 1.3% for Sen, Spe, ACC, MCC, and Youden index, respectively. Therefore, it can be said that the weighted serial fusion process for the best feature combination is effective.

From Table 8 and Figure 5, it can be seen intuitively that, when predicting the best feature combination before and after fusion, the classifier with the best effect was still the SVM classifier.

E. FEATURE DIMENSIONALITY REDUCTION

1) M VALUE SIZE

After serial fusion, the total dimension of the features is the sum of all feature dimensions to be fused, which may lead to excessively large feature group dimensions. This not only greatly increases the training time of the model, but may also cause redundancy, leading to a decrease in prediction accuracy. Therefore, as previously mentioned, we used the GPCA method to reduce the dimensionality of the fused features.

In the GPCA algorithm, the choice of parameter m, that is, the total dimension after dimension reduction, is critical. Thus, we experimented with the PDNA-224 dataset

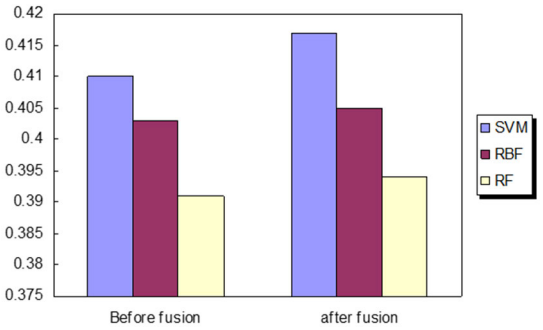

FIGURE 5. On the PDNA-224 dataset before and after feature fusion, the value of MCC in the prediction results after ten-fold cross-validations using three classifiers.

to reduce the dimensionality of fused features. According to the previous experiments, it can be seen that the SVM classifier performs better on the research questions in this study; therefore, the SVM classifier was used to perform ten-fold cross-validation to determine the size of the optimal m value. Based on previous experimental experience and the total dimensions of the features after fusion in this study, the experiment was carried out in the range of 360-400 with a step size of 5 to determine the optimal value of m. The experimental results are presented in Table 9.

TABLE 9. The results obtained under different m values after performing ten-fold cross-validations on the PDNA-224 dataset based on the fused feature set and the SVM classifier.

m value	Sen(%)	Spe(%)	ACC(%)	MCC	Youden's Index
360	77.28	85.44	84.92	0.3959	0.6272
365	78.09	85.66	85.18	0.4027	0.6375
370	77.68	86.08	85.54	0.4063	0.6376
375	77.28	86.69	86.09	0.4132	0.6397
380	77.13	86.82	86.26	0.4167	0.6395
385	78.42	87.71	87.62	0.4261	0.6613
390	77.18	86.71	86.12	0.4153	0.6389
395	77.68	86.20	85.65	0.4081	0.6388
400	75.53	86.14	85.16	0.3968	0.6167

From Table 9, in the interval–360-400, as the value of m continues to increase, the accuracy of the prediction and the value of MCC both show a trend of increasing first and then decreasing after reaching the peak. ACC increased from 84.92% to 87.62%, dropped to 85.16%, MCC increased from 0.395 to 0.426, and then dropped to 0.396, as shown in Figure 6.

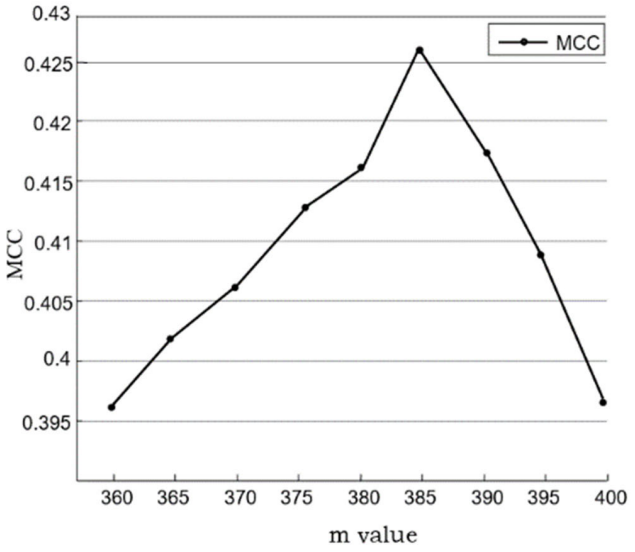

FIGURE 6. Based on the feature group after feature fusion and the SVM classification algorithm, on the PDNA-224 data set, the step size is 5, and the m value ranges from 360 to 400. The change curve of MCC in the prediction results.

It can be observed that the size of the feature group dimension m after dimensionality reduction has a significant impact on the prediction performance of the model, and an inappropriate value of m may lead to loss or redundancy of feature group information after dimensionality reduction. By analyzing Table 8 and Figure 6, it can be seen that when the value of m is less than 385, as the value of m increases, the prediction effect of the model increases with the increase of the value of m. When the value of m was 385, the dimensionality was reduced, and the processed features were predicted the best. When m was greater than 385, the prediction effect of the model decreased as m increased. Therefore, this study selected 385 as the m value and performed dimensionality reduction processing on the fused feature group to obtain the best prediction effect.

2) DIMENSIONALITY REDUCTION EFFECT

After obtaining the optimal m-value size for the PDNA-224 dataset, the obtained weighted and fused feature groups were subjected to a dimensionality reduction. The SVM, RBF, and RF classifiers were used to perform ten-fold cross-validations to select the classifier with the best effect for the construction of the prediction system in this study.

After reducing the feature dimension and performing classification prediction, the Spe, ACC, MCC, and Youden's index increased by 1.4%, 1.8%,0.9%, and 1%, respectively. Although Sen is slightly lower than before, MCC has improved because Spe is higher than before. Although the dimension reduction process does not significantly improve the prediction accuracy and MCC, the total dimensions of the features are significantly reduced after the dimension reduction. On the one hand, this reduces the redundancy between features; on the other hand, it also greatly reduces the training

TABLE 10. Prediction results after ten-fold cross-validations using three classifiers on the PDNA-224 dataset before and after feature dimensionality reduction.

Feature	Classification algorithm	Sen (%)	Spe (%)	ACC (%)	MCC	Youden's Index
Before dimensionality reduction	SVM	78.8	86.3	85.8	0.417	0.651
	RBF	77.1	86.0	85.3	0.405	0.631
	RF	75.6	84.9	84.6	0.394	0.605
After dimensionality reduction	SVM	78.4	87.7	87.6	0.426	0.661
	RBF	76.2	86.0	85.6	0.406	0.622
	RF	76.1	85.6	85.5	0.398	0.617

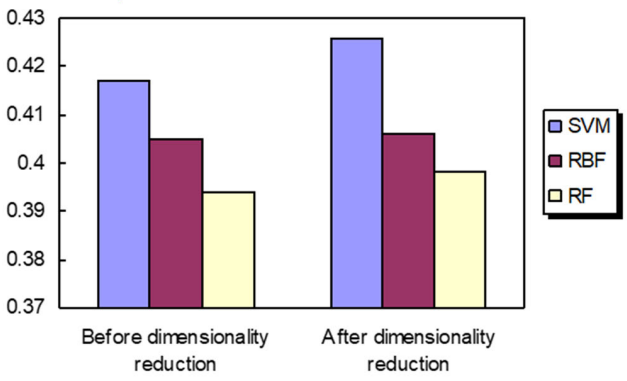

FIGURE 7. On the PDNA-224 dataset before and after dimensionality reduction, the value of MCC in the prediction results after ten-fold cross-validations using three classifiers.

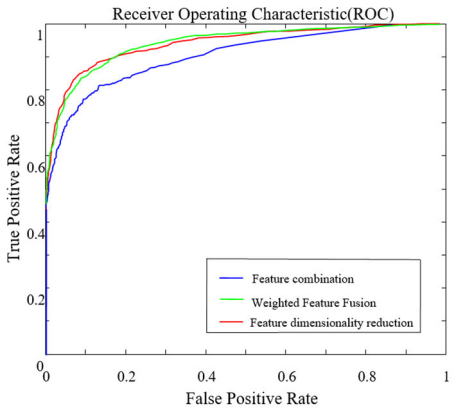

FIGURE 8. ROC curve graph after Feature combination, Weighted Feature Fusion, and Feature dimensionality reduction.

speed of the model and the time required for prediction, and improves the execution efficiency of the model.

It can be observed from Table 10 and Figure 7 that this sequence of operations is valid. The area under each curve comprehensively reflected the effectiveness of the proposed method. Therefore, it can be said that the feature dimensionality reduction method used in this study is effective. A protein-DNA binding site prediction model based on multi-

view feature fusion and a support vector machine designed in this study was constructed. Next, we compared the prediction performance of existing popular prediction systems on the PDNA-224 dataset to illustrate the effectiveness of the prediction model constructed in this study.

F. PERFORMANCE COMPARISON WITH OTHER RELATED PREDICTION SYSTEMS

As mentioned above, many achievements have been made in the field of protein-DNA binding site prediction. To test the effectiveness of the proposed method, we compared the performance of our method with four other state-of-the-art methods using a benchmark dataset. These methods include PreDNA [7], DP-Bind [47], EL_PSSM-RT [48], and Novel_Predict [9]. The comparative results for PDNA-224 and PDNA-62 are presented in Tables 11 and 12, respectively.

TABLE 11. On the PDNA-224 dataset, the proposed method compares with other predictors.

Method	Sen(%)	Spe(%)	ACC(%)	MCC	Youden's Index
PreDNA	76.1	82.2	81.8	0.35	0.583
DP-Bind	69.4	77.6	76.2	N/A	0.470
EL_PSSM-RT	79.6	78.0	78.1	0.34	0.576
Novel_Predict	54.1	91.6	88.9	0.37	0.457
Proposed Method	78.4	87.7	87.6	0.43	0.661

1) ON PDNA-224 DATASET

It can be seen from the table that the proposed method has better prediction performance than other popular protein-DNA binding site predictors. Compared with the second-best method, Novel_Predict, in the prediction results, the proposed method was 24.3%, 6%, 6.72%, and 20.4% higher in Sen, MCC, and Youden's index, respectively. Although the Spe and ACC of Novel_Predict are 3.9% and 1.3% higher than those of the proposed method, respectively, the value of Sen is 24.3% lower than that of the proposed method. Considering the PDNA-224 database (containing 3778 DNA-binding residues and 53570 non-DNA-binding residues), the number of DNA-binding residues was much smaller than that of the non-binding residues. It shows that this method has the problem of wrongly predicting the binding site as a non-binding site when predicting the binding site, and this result does not reflect the effectiveness of the prediction method very well. Therefore, the method proposed in this study is superior to the Novel_Predict method for the overall prediction.

TABLE 12. On the PDNA-62 dataset, the proposed method compares with other predictors.

Method	Sen(%)	Spe(%)	ACC(%)	MCC	Youden's Index
PreDNA	76.8	79.7	79.4	0.42	0.565
DP-Bind	76.4	76.6	77.2	N/A	0.530
EL_PSSM-RT	85.0	80.1	80.8	0.51	0.651
Novel_Predict	73.6	89.5	87.5	0.59	0.631
Proposed Method	84.6	86.7	88.1	0.63	0.713

TABLE 13. On the APO29 dataset, the proposed method compares with other predictors.

Method	Sen(%)	Spe(%)	ACC(%)	MCC	Youden's Index
PreDNA	58.6	89.3	86.1	0.41	0.479
DP-Bind	60.3	79.1	76.7	N/A	0.394
Novel_Predict	52.6	86.5	83.9	0.39	0.391
Proposed Method	62.9	87.1	86.9	0.43	0.502

2) ON PDNA-62 DATASET

Table 12 lists the experimental results on PDNA-62 dataset. The results clearly show that the method in this study is also effective on the PNDA-62 dataset, and its prediction accuracy is higher than that of other existing protein-DNA binding site prediction algorithms. Compared with the novel prediction method, the proposed prediction method improves Sen, ACC, MCC, and Youden's index by 11%, 0.6%, 4%, and 8.2%, respectively. Although the value of Spe was slightly lower than that of Novel_Predict, Sen was higher than the Novel_Predict method to a greater extent. Therefore, overall, the prediction accuracy of the proposed method was better than that of the Novel_Predict method.

3) ON APO29 DATASET

To further compare the predictive performance of the proposed method with that of other existing methods, we evaluated the proposed method on the independent test dataset APO29. In addition, to further verify the generalization ability of the prediction system and to avoid the model overfitting phenomenon that may occur when only standard datasets are used for training. Table 13 shows the difference in prediction performance between the proposed method and the other

three state-of-the-art methods: PreDNA [7], DP-Bind [47], and Novel_Predict [9].

The proposed method achieved the highest ACC, MCC, and Youden index values of 86.9%, 0.43, and 0.502, respectively. Prediction results on the APO29 independent test dataset demonstrated that our method generalizes well to protein-DNA binding residues.

VIII. CONCLUSION

The identification and analysis of binding sites between proteins and DNA are of great significance for studying the mechanism of protein function. With the continuous development of bioinformatics, the use of bioinformatics methods to predict protein-DNA binding sites will be one of the work centers in related fields in the future. We have completed some related research work on the prediction of protein-DNA binding sites and proposed a new prediction method, and through a series of works, the prediction accuracy of protein-DNA binding sites has been improved to a certain extent. The prediction results on the two benchmark datasets over ten-fold cross-validations demonstrated that the proposed method was effective in achieving better performance than other prediction methods. At the same time, the results on the independent test dataset also show that the method proposed in this paper has the best prediction performance. All experimental results demonstrate that our proposed method is highly competitive for predicting protein-DNA binding sites. Our study is complementary to existing protein-DNA binding site predictors.

Although the proposed method achieved certain results, there are still some deficiencies that need to be improved and will be further studied in the future.

We used a single classification algorithm to build a predictor, which may have overfitting in some cases and has certain limitations. Therefore, we consider different classification algorithms for fusion to further improve the prediction performance and avoid overfitting.

We selected some features of proteins and achieved certain results, but the selection of features was still weak. However, many factors affect the binding process of proteins and DNA, such as binding free energy and entropy. These are related to the biological mechanisms of protein-DNA interactions. Therefore, in future research, this special information should be explored further to improve the effect of prediction.

Although our proposed method achieved promising results on two benchmark datasets and an independent test set, real-world applications are still lacking. Next, we will further explore the effect of the method proposed in this study when facing real-world cases.

REFERENCES

[1] N. M. Luscombe, S. E. Austin, H. M. Berman, and J. M. Thornton, "An overview of the structures of protein-DNA complexes," *Genome Biol.*, vol. 1, no. 1, pp. 1–37, 2000.
[2] M. Ptashne, "Regulation of transcription: From lambda to eukaryotes," *Trends Biochem. Sci.*, vol. 30, no. 6, pp. 275–279, Jun. 2005.
[3] H. Kono and A. Sarai, "Structure-based prediction of DNA target sites by regulatory proteins," *Proteins, Struct., Function, Genet.*, vol. 35, no. 1, pp. 114–131, Apr. 1999.
[4] G. B. Ruvkun and F. M. Ausubel, "A general method for site-directed mutagenesis in prokaryotes," *Nature*, vol. 289, no. 5793, pp. 85–88, Jan. 1981.
[5] L. Wang, C. Huang, M. Q. Yang, and J. Y. Yang, "BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features," *BMC Syst. Biol.*, vol. 4, no. S1, pp. 1–20, May 2010.
[6] L. Wang and S. J. Brown, "BindN: A web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences," *Nucleic Acids Res.*, vol. 34, no. Web Server, pp. W243–W248, Jul. 2006.
[7] M. P. Asghari and P. Abdolmaleki, "Prediction of RNA- and DNA-binding proteins using various machine learning classifiers," *Avicenna J. Med. Biotechnol.*, vol. 11, no. 1, pp. 104–111, Mar. 2019.
[8] Y. Liu, W. Gong, Z. Yang, and C. Li, "SNB-PSSM: A spatial neighbor-based PSSM used for protein-RNA binding site prediction," *J. Mol. Recognit.*, vol. 34, p. e2887, Jun. 2021.
[9] J. Song, G. Liu, and J. Jiang, "A novel prediction method for protein DNA-binding residues based on neighboring residue correlations," *Biotechnol. Biotechnolog. Equip.*, vol. 36, no. 1, pp. 865–877, Dec. 2022.
[10] T. Li, Q.-Z. Li, S. Liu, G.-L. Fan, Y.-C. Zuo, and Y. Peng, "PreDNA: Accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information," *Bioinformatics*, vol. 29, no. 6, pp. 678–685, Mar. 2013.
[11] M. S. Klausen, M. C. Jespersen, H. Nielsen, K. K. Jensen, V. I. Jurtz, C. K. Sønderby, M. O. A. Sommer, O. Winther, M. Nielsen, B. Petersen, and P. Marcatili, "NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning," *Proteins, Struct., Function, Bioinf.*, vol. 87, no. 6, pp. 520–527, Jun. 2019.
[12] S. Zhang, L. Zhao, C.-H. Zheng, and J. Xia, "A feature-based approach to predict hot spots in protein–DNA binding interfaces," *Briefings Bioinf.*, vol. 21, no. 3, pp. 1038–1046, May 2020.
[13] X. Zhu, L. Liu, J. He, T. Fang, Y. Xiong, and J. C. Mitchell, "IPNHOT: A knowledge-based approach for identifying protein-nucleic acid interaction hot spots," *BMC Bioinf.*, vol. 21, no. 1, pp. 1–24, Dec. 2020.
[14] N. Wang, J. Zhang, and B. Liu, "IDRBP-PPCT: Identifying nucleic acid-binding proteins based on position-specific score matrix and position-specific frequency matrix cross transformation," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 4, pp. 2284–2293, Jul. 2022.
[15] M. Torrisi, G. Pollastri, and Q. Le, "Deep learning methods in protein structure prediction," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1301–1310, Jan. 2020.
[16] S. Ahmad and A. Sarai, "PSSM-based prediction of DNA binding sites in proteins," *BMC Bioinf.*, vol. 6, no. 1, p. 33, 2005.
[17] S. Patiyal, A. Dhall, and G. P. S. Raghava, "DBpred: A deep learning method for the prediction of DNA interacting residues in protein sequences," *BioRxiv*, vol. 2021, pp. 1–18, Aug. 2021.
[18] A. Emamjomeh, D. Choobineh, B. Hajieghrari, N. MahdiNezhad, and A. Khodavirdipour, "DNA–protein interaction: Identification, prediction and data analysis," *Mol. Biol. Rep.*, vol. 46, no. 3, pp. 3571–3596, Jun. 2019.
[19] W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006.
[20] Y. Zhang, S. Qiao, S. Ji, N. Han, D. Liu, and J. Zhou, "Identification of DNA–protein binding sites by bootstrap multiple convolutional neural networks on sequence information," *Eng. Appl. Artif. Intell.*, vol. 79, pp. 58–66, Mar. 2019.
[21] X. Zhu, S. S. Ericksen, and J. C. Mitchell, "DBSI: DNA-binding site identifier," *Nucleic Acids Res.*, vol. 41, no. 16, p. e160, Sep. 2013.
[22] X. Du, J. Hu, and S. Li, "Using Chou's 5-step rule to predict dna-protein binding with multi-scale complementary feature," *J. Proteome Res.*, vol. 20, no. 3, pp. 1639–1656, 2021.
[23] B. Lee and F. M. Richards, "The interpretation of protein structures: Estimation of static accessibility," *J. Mol. Biol.*, vol. 55, no. 3, p. 379, Feb. 1971.
[24] J. Ding and E. Arnold, "NACCESS," Tech. Rep., 2006.
[25] Y. Wang, Z. Xue, G. Shen, and J. Xu, "PRINTR: Prediction of RNA binding sites in proteins using SVM and profiles," *Amino Acids*, vol. 35, no. 2, pp. 295–302, Aug. 2008.
[26] S. Ahmad, "ASAView: Database and tool for solvent accessibility representation in proteins," *BMC Bioinf.*, vol. 5, no. 1, p. 51, 2004.

[27] A. Pintar, O. Carugo, and S. Pongor, "CX, an algorithm that identifies protruding atoms in proteins," *Bioinformatics*, vol. 18, no. 7, pp. 980–984, Jul. 2002.
[28] A. Pintar, O. Carugo, and S. Pongor, "DPX: For the analysis of the protein core," *Bioinformatics*, vol. 19, no. 2, pp. 313–314, Jan. 2003.
[29] J. Mihel, M. Šikić, S. Tomić, B. Jeren, and K. Vlahoviček, "PSAIA—Protein structure and interaction analyzer," *BMC Struct. Biol.*, vol. 8, no. 1, pp. 1–11, Dec. 2008.
[30] F. Ali, S. Ahmed, Z. N. K. Swati, and S. Akbar, "DP-BINDER: Machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information," *J. Comput.-Aided Mol. Des.*, vol. 33, no. 7, pp. 645–658, Jul. 2019.
[31] A. A. Schaffer, "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements," *Nucleic Acids Res.*, vol. 29, no. 14, pp. 2994–3005, Jul. 2001.
[32] N. Q. K. Le and B. P. Nguyen, "Prediction of FMN binding sites in electron transport chains based on 2-D CNN and PSSM profiles," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 6, pp. 2189–2197, Nov. 2021.
[33] P. D. Bank, "Protein data bank," *Nature New Biol.*, vol. 233, p. 223, Jan. 1971.
[34] S. Beamer, K. Asanović, and D. Patterson, "Direction-optimizing breadth-first search," *Sci. Program.*, vol. 21, nos. 3–4, pp. 137–148, 2013.
[35] A. M. Rush, Y.-W. Chang, and M. Collins, "Optimal beam search for machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 210–221.
[36] N. B. Karahanoglu and H. Erdogan, "A orthogonal matching pursuit: Best-first search for compressed sensing signal recovery," *Digit. Signal Process.*, vol. 22, no. 4, pp. 555–568, Jul. 2012.
[37] C. Liu and H. Wechsler, "A shape- and texture-based enhanced Fisher classifier for face recognition," *IEEE Trans. Image Process.*, vol. 10, no. 4, pp. 598–608, Apr. 2001.
[38] J. Yang, J.-Y. Yang, D. Zhang, and J.-F. Lu, "Feature fusion: Parallel strategy vs. serial strategy," *Pattern Recognit.*, vol. 36, no. 6, pp. 1369–1381, Jun. 2003.
[39] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis—A brief tutorial," *Inst. Signal Inf. Process.*, vol. 18, pp. 1–8, Jan. 1998.
[40] H. Abdi and L. J. Williams, "Principal component analysis," *WIREs Comput. Statist.*, vol. 2, no. 4, pp. 433–459, Jul./Aug. 2010.
[41] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Apr. 1995.
[42] Y. Ge, L. Jianhong, and L. Zhiyuan, "A study on RBF neural network based online algorithm models adaptive to thermal processes," *Proc.-Chin. Soc. Elect. Eng.*, vol. 24, no. 1, pp. 191–195, 2004.
[43] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[44] L.-C. Shen, Y. Liu, J. Song, and D.-J. Yu, "SAResNet: Self-attention residual network for predicting DNA-protein binding," *Briefings Bioinf.*, vol. 22, no. 5, Sep. 2021, Art. no. bbab101.
[45] Y.-H. Zhu, J. Hu, X.-N. Song, and D.-J. Yu, "DNAPred: Accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines," *J. Chem. Inf. Model.*, vol. 59, no. 6, pp. 3057–3071, Jun. 2019.
[46] D. S. Broomhead and L. Owe, "Multi-variable functional interpolation and adaptive networks," *Complex Syst.*, vol. 2, pp. 321–355, Mar. 1988.
[47] S. Hwang, Z. Gou, and I. B. Kuznetsov, "DP-bind: A web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins," *Bioinformatics*, vol. 23, no. 5, pp. 634–636, Mar. 2007.
[48] J. Zhou, Q. Lu, R. Xu, Y. He, and H. Wang, "EL_PSSM-RT: DNA-binding residue prediction by integrating ensemble learning with PSSM relation transformation," *BMC Bioinf.*, vol. 18, no. 1, pp. 1–16, Dec. 2017.
[49] Y. Tsuchiya, K. Kinoshita, and H. Nakamura, "Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces," *Proteins, Struct., Function, Bioinf.*, vol. 55, no. 4, pp. 885–894, Apr. 2004.
[50] X. Ding and M. Cai, "Matrix theory in engineering," Tianjin Univ. Press, Tianjin, China, Tech. Rep., 1995.
[51] N. Q. K. Le, Q.-T. Ho, V.-N. Nguyen, and J.-S. Chang, "BERT-promoter: An improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection," *Comput. Biol. Chem.*, vol. 99, Aug. 2022, Art. no. 107732.
[52] N.-Q.-K. Le, T.-T.-D. Nguyen, and Y.-Y. Ou, "Identifying the molecular functions of electron transport proteins using radial basis function networks and biochemical properties," *J. Mol. Graph. Model.*, vol. 73, pp. 166–178, May 2017.
[53] N.-Q.-K. Le and Y.-Y. Ou, "Incorporating efficient radial basis function networks and significant amino acid pairs for predicting GTP binding sites in transport proteins," *BMC Bioinf.*, vol. 17, no. S19, pp. 183–192, Dec. 2016.

JI YANG received the bachelor's degree in computer application technology from the Nanjing University of Science and Technology, China. His current research interests include bioinformatics and data-mining.

SHUNING ZHANG received the bachelor's degree from the Anhui University of Chinese Medicine, China. Her current research interests include bioinformatics and neurosciences.