Sequence-Based Prediction of DNA-Binding Residues in Proteins with Conservation and Correlation Information

Xin Ma, Jing Guo, Hong-De Liu, Jian-Ming Xie, and Xiao Sun

Abstract—The recognition of DNA-binding residues in proteins is critical to our understanding of the mechanisms of DNA-protein interactions, gene expression, and for guiding drug design. Therefore, a prediction method DNABR (DNA Binding Residues) is proposed for predicting DNA-binding residues in protein sequences using the random forest (RF) classifier with sequence-based features. Two types of novel sequence features are proposed in this study, which reflect the information about the conservation of physicochemical properties of the amino acids, and the correlation of amino acids between different sequence positions in terms of physicochemical properties. The first type of feature uses the evolutionary information combined with the conservation of physicochemical properties of the amino acids while the second reflects the dependency effect of amino acids with regards to polarity-charge and hydrophobic properties in the protein sequences. Those two features and an orthogonal binary vector which reflect the characteristics of 20 types of amino acids are used to build the DNABR, a model to predict DNA-binding residues in proteins. The DNABR model achieves a value of 0.6586 for Matthew's correlation coefficient (MCC) and 93.04 percent overall accuracy (ACC) with a 68.47 percent sensitivity (SE) and 98.16 percent specificity (SP), respectively. The comparisons with each feature demonstrate that these two novel features contribute most to the improvement in predictive ability. Furthermore, performance comparisons with other approaches clearly show that DNABR has an excellent prediction performance for detecting binding residues in putative DNA-binding protein. The DNABR web-server system is freely available at http://www.cbi.seu.edu.cn/DNABR/.

Index Terms—DNA-binding residues, random forest, physicochemical property, evolutionary information

1 Background

Protein-DNA interactions play an essential role in biological processes and have attracted extensive attention and investigation in recent years [1], [2], [3]. For instance, a great deal of research has focused on the interaction between transcription factors (TF) and DNA [4], [5], [6], while other studies have concentrated on whether a novel protein could bind to DNA [7], [8], [9], [10]. More and more studies have focused on the identification of DNA-binding residues in proteins because it would yield great insight into the mechanisms that underlie DNA-protein interactions [11]. Therefore, the recognition of DNA-binding residues has increasingly attracted more attention and has become a central theme in research related to the function of proteins.

There are several methods through which DNA-binding residues can be identified and these can be categorized in two main types of methodology: experimental methods and computational methods. Experimental methods such as X-ray crystallography and nuclear magnetic resonance (NMR) are useful for obtaining the 3D structures of the DNA-protein complexes and for calculating the DNA-binding residues through this structural information. However, recognizing binding residues through these experimental techniques is expensive and time consuming. Thus, reliable computational methods have been developed that allow an automated prediction of DNA-binding residues based on the information derived from the sequence alone and also from the sequence and structure collectively [12], [13], [14], [15], [16], [17]. When the structural information for the DNA-binding protein is used to predict DNA-binding residues, the problems limiting the application of these experimental techniques to capture 3D structure still exist. Therefore, in this study computational methods for predicting the DNA-binding sites directly from amino acid sequence have been developed.

Several computational-based methods have been used for to construct a predictive model for DNA-binding residues in proteins [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]. These neural network models [30] were developed to predict DNA-binding residues using sequence information. A Naïve Bayes classifier [21] was developed to predict DNA-binding residue based on their identity and on the identities of the neighboring sequences. A support vector machine (SVM) classifier is also an effective tool with the ability to distinguish DNA-binding residues from nonbinding ones and it is has been widely used in recent research [19], [20], [22], [24], [26], [29]. Ofran et al. developed a model named DISIS (see http://cubic.bioc.columbia.edu/services/disis) for predicting DNA-binding residues by using the SVM classifier technology with physicochemical features combining local structure and evolutionary conservation data [24].

- *X. Ma is with the State Key Laboratory of Bioelectronics, School of Biological Science & Medical Engineering, Southeast University and Nanjing Audit University, Nanjing, P.R. China. E-mail: maxin@seu.edu.cn.*
- *J. Guo, H.-D. Liu, J.-M. Xie, and X. Sun are with the State Key Laboratory of Bioelectronics, School of Biological Science & Medical Engineering, Southeast University, Nanjing, P.R. China. E-mail: guojing_srtp@126.com, {liuhongde, xiejm, xsun}@seu.edu.cn.*

Manuscript received 22 Oct. 2011; revised 27 May 2012; accepted 16 July 2012; published online 1 Aug. 2012.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2011-10-0266.
Digital Object Identifier no. 10.1109/TCBB.2012.106.

1545-5963/12/$31.00 © 2012 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

Huang et al. have also predicted sequence-specific and nonspecific binding residues by SVM coupled with evolutionary information [26]. The BindN model [20] proposed by Wang et al. used sequence feature including the biochemical property of amino acids, including the side chain pKa value, hydrophobicity index, and molecular mass with SVM. BindN+ method [29] is also based on SVM algorithm and is encoded with evolutionary information in terms of position specific scoring matrix and same three physical-chemical properties which are used in the BindN model. Random forest (RF) classifiers also offer a strong alternative and provide high accuracy for the prediction of DNA-binding residues. The BindN-rf system was proposed by Wang et al. as a means of predicting DNA-binding residues using a RF classifier, combining biochemical features with several descriptors of evolutionary data [27]. Recently, a meta web server named MetaDBsite [31] was developed to predict DNA-binding residues used sequence information of proteins. MetaDBSite integrates the prediction results from six available online web servers: DISIS, DNABindR, BindN, BindN-rf, DP-Bind, and DBS-PRED for protein DNA-binding residues prediction.

In this study, we propose a novel method for predicting DNA-binding residues using the RF algorithm combining sequence-based features. Feature selection is the most important factor when aiming to improving the computational capacity of classifiers. In previous work, the conservation information about the amino acids at the level of physicochemical properties was not considered. In this study, to improve the classifier for DNA-binding residues, a novel matrix, position specific scoring matrices combining physicochemical properties (PSSM-PP) was put forth, which not only contains the evolutionary information captured by PSSM, but also contains information about the physicochemical properties of the amino acids. Therefore, it captures more information than PSSM and has a better predictive performance. With in models developed by previous work, each amino acid is considered to act independently, which may be incongruous with the idea that each amino acid within a protein sequences closely relies upon its neighboring amino acids. Therefore, in the new system we have developed a novel feature that collects information about one residue and its sequential neighbors. Compared with the initial system, which had an independent amino acid input feature, the predictive ability of the new system is significantly improved when the relationship between adjoining amino acids within protein sequences is considered as an input features. Moreover, the problem of imbalance between data sets brought about by the fact that the number of DNA-binding residues in proteins is less than that of nonbinding residues has also been considered and resolved in this study. The results show that our DNABR model scores very highly (0.6586) when tested using Matthew's correlation coefficient (MCC), with a 93.04 percent overall accuracy (ACC). The DNABR model was also found to have a sensitivity of 68.47 percent (SE) and specificity of 98.16 percent (SP).

2 Methods

2.1 Data Collection

The data set, DBP-337 was used in this work and contained 337 DNA-interacting proteins extracted from all the protein-DNA complexes (released by November 17, 2010) in the Protein Data Bank [32]. (The data set DBP-337 can be found in the Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2012.106). Those protein-DNA complexes were determined by X-ray crystallography with a resolution better than 3.5 Å. Redundant proteins with $>25\%$ sequence identity were removed using the blastclust program within the BLAST package [33] available from NCBI and only the longest amino acid sequences which created nonredundant DBP-337 data set were selected in each cluster.

Herein, an amino acid in the protein chain is defined as having DNA-binding residues if one or more atoms on its side chain or backbone are within the cutoff distance of 3.5 Å from any atoms of the DNA molecule in the complex. This is the same as previous research [18], [19], [20], [27], [30]. Among the 72,196 residues in the 337 DNA-binding protein sequences, 5,084 residues were defines as DNA-binding residues and the remaining 67,112 residues were defined as nonbinding residues.

2.2 Feature Descriptors

Each data instance is obtained by sliding a window along the protein, when predicting whether or not the central residue binds to DNA. Thus, a data instance is defined as positive if the central residue is a DNA-binding residue or negative if the central residue is nonbinding. Then, the data instances, each of which contain the information of L consecutive residues of the proteins being trained, were used to construct the RF classifier. Taking into consideration the previous works [18], [19], [20], [21], [24], [26], [28], [29] in prediction of DNA-binding residues in proteins, we chose different lengths of a data instance L from five to 19 residues and compared prediction performances based on those lengths using DNABR method. We calculated the three performance evaluations, i.e., ACC, BM, and MCC, to determine optimal window size. Compared with other window sizes (See Appendix B, available in the online supplemental material), the RF classifiers constructed with $L = 9$ presented the best performance.

In this work, we considered two novel descriptors, including position specific scoring matrices combining physicochemical properties and amino acid correlation (AAC) with regards to polarity-charge and hydrophobicity propriety.

2.2.1 Position Specific Scoring Matrices Combining with Physicochemical Properties

A novel matrix called PSSM-PP is proposed based on the information of PSSM with six physicochemical properties in this research. The PSSM scores are generated by PSI-BLAST [34] to search against the nonredundant (nr) data set of amino acid sequences at NCBI, and 20 values are obtained for each sequence position. The standard logistic function [35] was used to rescale the value of PSSM within 0 and 1:

$$f(x) = \frac{1}{1 + \exp(-x)}. \quad (1)$$

Six physicochemical properties are considered for each amino acid: the pKa values of amino group, the pKa values of carboxyl group [36], the electron-ion interaction potential (EIIP) [37], the number of lone electron pairs(LEPs), Wiener index [38] and the molecular mass [39]. A new parameter $d_a(i)$ is defined by normalizing six quantitative properties with the following formula:

$$\begin{aligned} d_a(i) = \{P_a(i) - \min\{P_a(1), P_a(2), \ldots, P_a(20)\}\}/ \\ \{\max\{P_a(1), P_a(2), \ldots, P_a(20)\} \\ - \min\{P_a(1), P_a(2), \ldots, P_a(20)\}\}, \end{aligned} \tag{2}$$

where $d_a(i)$ represents the normalized property values that range from 0 to 1, a is the index of the property, and i indicates the ith amino acid. $P_a(i)$ is the value of property a of the ith amino acid. Thus, PSSM-PP is generated by merging 20 amino acid columns of the PSSM into a single column containing the information of a certain physicochemical property. In a PSSM-PP, the entry m_{ak} of position k for a certain physicochemical property a is calculated with:

$$m_{ak} = \sum_{i=1}^{20} \sqrt{d_a(i) f_k(i)}, \tag{3}$$

where a is the index of a certain physicochemical property, k is the index of position, i is the index of the type of amino acids, $f_k(i)$ is the normalized value of the ith type of amino acid in the position k of the PSSM calculated by (1) and $d_a(i)$ is normalized physicochemical property values of a for the ith type amino acids calculated by (2). A process diagram for PSSM-PP can be found in the Appendix C, available in the online supplemental material. According the definition, six values are captured for each sequence position. Therefore, the vector size of PSSM-PP feature is 6×9.

2.2.2 Amino Acid Correlation

Protein-DNA interactions are strongly influenced by electrostatic and hydrophobic interactions. Therefore, two types of AAC were considered based on the polarity-charge and hydrophobic properties. They reflect the information about dependency of amino acids with regards to polarity-charge and hydrophobic properties.

Let $(C_1, C_2, C_3, C_4) \equiv$ (polar amino acid with positive charge, polar amino acid with negative charge, noncharged polar amino acid, nonpolar amino acid). For a detailed definition, see [36]. The AAC of polarity-charge is defined as follows:

$$AAC\_PC(i) = \sum_{k=1}^{L-1} \frac{C_i(k)}{L-k} \log_2\left(\frac{C_i(k)/(L-k)}{N_i^2/L^2}\right), \quad i = 1, 2, 3, 4, \tag{4}$$

where N_i is the number of certain C_i in the instances, $C_i(k)$ means the number of two C_is at a distance of k, L is the sliding window size, and $AAC\_PC(i)$ represents the relevance of the two C_is with different gaps from 1 to L-1 for the charge and polarity property.

The definition of AAC of hydrophobicity property is similar with the definition of ACC of charge and polarity properties. Let $(H_1, H_2, H_3, H_4) \equiv$ (strong hydrophobic residue, weak hydrophobic residue, strong hydrophilic residue, weak hydrophilic residue). For a detailed definition, see [40]. Hence, the AAC of hydrophobic property is defined by

$$AAC\_H(i) = \sum_{k=1}^{L-1} \frac{H_i(k)}{L-k} \log_2\left(\frac{H_i(k)/(L-k)}{M_i^2/L^2}\right), \quad i = 1, 2, 3, 4, \tag{5}$$

where M_i is the number of certain H_i in the sample, $H_i(k)$ means the number of two H_is at a distance of k, and $HC(i)$ denotes the correlation of the two H_is with different gaps from 1 to L-1 for the hydrophobic property. When $C_i(k)$ equals to 0, the problem $0 \log_2 0$ appeared in the (4). To solve the problem, the (4) is transformed to (6) by using a Taylor series

$$\begin{aligned} AAC\_PC(i) &= \sum_{k=1}^{L-1} \frac{C_i(k)}{L-k} \log_2\left(\frac{C_i(k)/(L-k)}{N_i^2/L^2}\right) \\ &= \sum_{k=1}^{L-1} \frac{1}{\mathrm{In}\ 2} \left[\left(\frac{C_i(k)}{(L-k)} - \frac{N_i^2}{L^2}\right) + \frac{(C_i(k)/(L-k) - N_i^2/L^2)^2}{2N_i^2/L^2} \right. \\ &\quad \left. + \mathrm{O}\left(\frac{\mathrm{C_i(k)}}{\mathrm{(L-k)}} - \frac{\mathrm{N_i^2}}{\mathrm{L^2}}\right)^3 \right]. \end{aligned} \tag{6}$$

Using the same method, (7) was obtained from (5) as follows:

$$\begin{aligned} AAC\_H(i) = \sum_{k=1}^{L-1} \frac{1}{\ln 2} \left[\left(\frac{H_i(k)}{L-k} - \frac{M_i^2}{L^2}\right) \right. \\ + \frac{(H_i(k)/(L-k) - M_i^2/L^2)^2}{2N_i^2/L^2} \\ \left. + \mathrm{O}\left(\frac{H_i(k)}{L-k} - \frac{M_i^2}{L^2}\right)^3 \right] \quad i = 1, 2, 3, 4. \end{aligned} \tag{7}$$

Then, we concatenate the vector spaces of AAC_PC and AAC_H to represent the AAC feature vector ((6), (7)), and the size of this feature is 8D

$$\{AAC\} = \{AAC\_PC\} \oplus \{AAC\_H\}. \tag{8}$$

2.2.3 Orthogonal Binary Vector (OBV)

Work produced by Shen et al. [41] indicates the importance that can be reflected by the dipoles and volumes of the side chains of amino acids, respectively. Therefore a 6D OBV was used to code the amino acids in each class. The result from our previous study indicates that an OBV is an important feature for distinguishing DNA-binding residues from nonbinding ones. The OBV reflect the information about single amino acid and it could be used to complement the AAC feature. The size of the OBV feature is 54D.

In this study, each data instance is coded with sequence-based features by combing position specific scoring matrices incorporating physicochemical properties, AAC, and OBV. The vector in each instance has a dimension 116 and is less than that of previous studies.

2.3 Classification with Random Forests

The RF classifier builds an ensemble of decision trees for classification [42]. The RF R package [43] is used to implement the RF algorithm.

To evaluate the performance of the classifier, a fivefold cross-validation procedure was used in this research. During the procedure, the data instances obtained from the data set were randomly divided into five parts. In each of the five round steps, four of these parts were used as a training set to construct a classifier, while the remaining one was used as test set to evaluate the performance.

2.4 The Algorithm to Balance the Data Set

The data instances obtained from DBP-337 contained 4,754 DNA-binding residues as positive instances and 61,451 nonbinding residues as negative ones. Thus, the imbalanced data set problem should be considered when improving the performance of classifier. The method to solve the problem of an imbalanced data set was exactly the same as previous study [44], which presented an algorithm to downsize the majority class by selecting the safe instances. After processing the algorithm to balance the data set, the processed data set (PDBP-337) containing all positive instances (4,754) and safe negative instances (28,374) can be obtained as the new data set for constructing the RF classifier.

2.5 Performance Assessment and Reliability Index (RI)

The following performance assessments are used in this study: the overall prediction accuracy, sensitivity, specificity and MCC, and a balanced measure (BM) defined as harmonic mean of sensitivity and specificity

$$BM = \frac{2SE \times SP}{SE + SP}. \tag{9}$$

The receiver operating characteristic curve (ROC) [45] is a robust approach for measure the performance, which is drawn by plotting the true positive rate (SE) against false positive rate (1-SP). The area under the ROC curve (AUC) [46] which is a reliable measure of classifier performance, was also used in our research. The prediction reliability index plays an important role in evaluating the quality of prediction and the RI was adopted to show the level of reliability for the prediction results of each amino acid. RI is defined as:

$$RI = \begin{cases} \mathrm{int}\left(\frac{500}{9} D\right) & if\ D < 0.18 \\ 10 & if\ D \geq 0.18 \end{cases} \tag{10}$$

$$D = |F_+ - \Phi|, \tag{11}$$

where $_{F_+}$ is the percentage of the tree votes for the positive class in each instance, and Φ is the threshold by which to classify instances according to tree votes for the positive instances, and in this study was set to 0.2 because the ratio of positive to the whole instances for training is nearly 1: 6.

3 Results and Discussion

3.1 Performance Comparison to Other Methods

BindN (http://bioinformatics.ksu.edu/bindn/), BindN+ (http://bioinfo.ggc.org/bindn+/), and BindN-rf (http://bioinfo.ggc.org/bindn-rf/) were all proposed by Wang et al. to prediction DNA-binding residues in proteins and performed on the same PDNA-62 data set. These three models all use the 3.5 Å cut off distance defined for DNA-binding residues, which is the same as with DNABR.

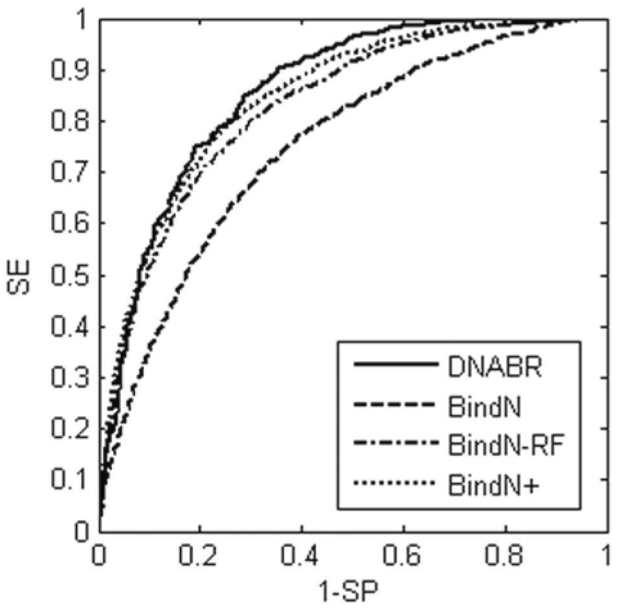

Fig. 1. Performance comparisons of DNABR, BindN, BindN-rf, and BindN+ using ROC curves. The AUC value for each predictor is 0.8669, 0.7488, 0.8257, and 0.8445, respectively.

To compare with those three methods, an independent test data set TS-72 with 72 protein chains was used to evaluate the performance of each model. The 72 protein chains were randomly selected from the DBP-337 data set without the protein chains of the PDNA-62 data set [20] used to construct the BindN, BindN+, and BindN-rf. We then trained the DNABR model on the remaining 265 proteins in the TR265 data set using the same strategy as the original DNABR model, and applied the new model to predict DNA-binding residues in the TS-72 test data set. We also applied the BindN model, the BindN+, and BindN-rf model to predict the putative DNA-binding residues in the same TS-72. These three methods could predict potential DNA-binding residues on the web servers, therefore the prediction results of the TS-72 data set were obtained by submitting the sequences to the web server. Fig. 1 shows the comparison of the ROC curves for the four methods on the independent test TS-72 data set. The results reveal that the AUC values are 0.8669, 0.7488, 0.8257, and 0.8445 for DNABR, BindN, BindN-rf, and BindN+ method, respectively. Therefore, DNABR significantly outperforms BindN, BindN-rf, and BindN+. Compared with the previous works, it is clearly shown that our DNABR model achieves the best performance. The number of instances obtained from DBP-337 to construct the DNABR model is far more than that obtained from PDNA-62 to construct BindN+, BindN, and BindN-rf, which proves that DNABR is more reliable than BindN+, BindN or BindN-rf.

To proved that DNABR is a useful tool for predicting DNA-binding residues using amino acid sequences information, PDB ID: 1C9B, which was not used for training the RF classifier, was selected to examine the prediction performance using Pymol software [47]. 1C9B is a human

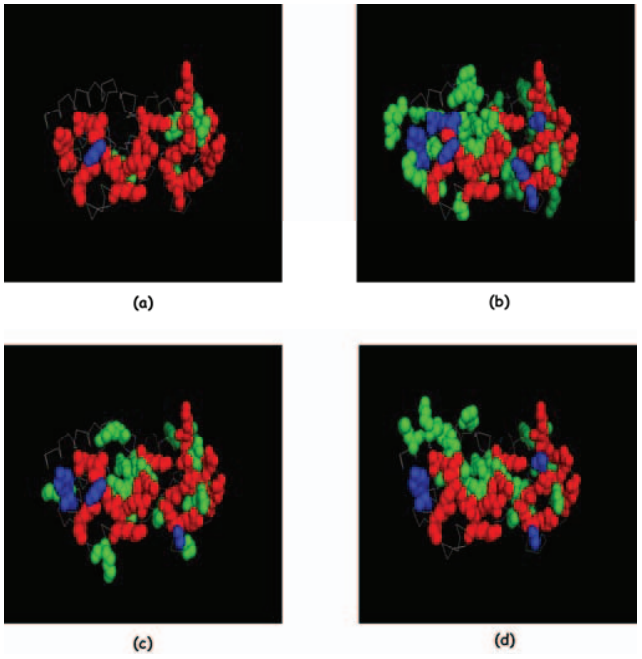

Fig. 2. Predicted results for the A chain of 1C9B using (a) DNABR, (b) BindN, (c) BindN-rf, and (d) BindN+.

TATA-box binding protein (TBP) core domain-human transcription factor IIB (TFIIB) core domain complex bound to extended, modified adenoviral major late promoter (ADMLP). Figs. 2a, 2b, 2c, and 2d show the results of predicted DNA-binding sites for DNABR to comparing with BindN, BindN-rf, and BindN+ with 3.5 Å distance cutoff, and those four methods were all tested on the A chains in 1C9B. Each sphere denotes an atom in Figs. 2a, 2b, 2c, and 2d, Red sphere represents true DNA-binding residue that is correctly predicted (TP). Blue sphere indicates DNA-binding residue that is predicted as non-binding one (FN) and green sphere represents nonbinding residue that is predicted as DNA-binding one (FP). As shown from Figs. 2a, 2b, 2c, and 2d, it's obviously that DNABR outperforms other three models from the results of comparison. Table 1 shows the detail predicted results for the A chain of 1C9B. DNA-binding residues were predicted by the our DNABR method at 95.94 percent overall accuracy with MCC of 0.8124, and with a sensitivity of 95 percent and a specificity of 96.06 percent, while the low overall accuracy was 77.66, 89.34, and 87.81 percent with the

TABLE 1
The Prediction Performance for the A Chain of 1C9B Using DNABR and Other Three Methods

Method	ACC(%)	SE(%)	SP(%)	BM(%)	MCC
DNABR	95.54	95	96.06	95.52	0.8124
BindN	77.66	65	79.09	71.35	0.3060
BindN-rf	89.34	85	89.83	87.34	0.5912
BindN+	87.81	85	88.12	86.53	0.5598

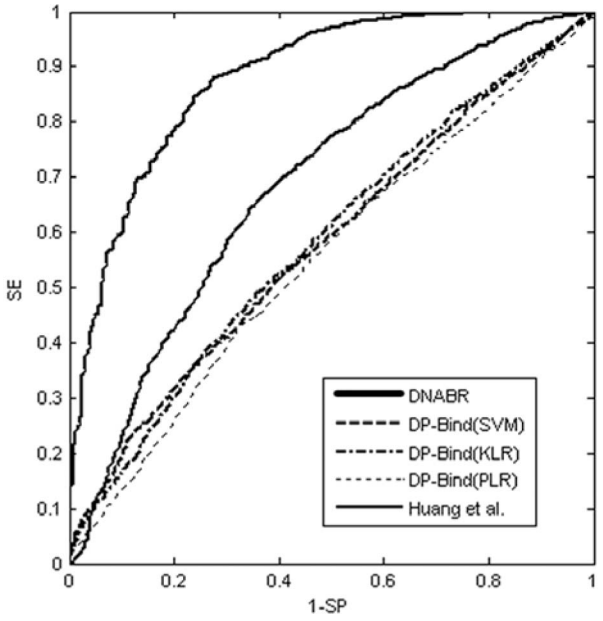

Fig. 3. A comparison of the prediction performance with DNABR, Huang et al. and three methods in DP-Bind using a cutoff distance of 4.5 Å for the binding residue. The AUC value AUC are 0.8808, 0.6876, 0.5778, 0.5832, and 0.5526 for DNABR, Huang et al., DP-Bind(SVM), DP-Bind(KLR), and DP-Bind(PLR), respectively.

Matthew's correlation coefficient 0.3060, 0.5912, and 0.5598 for the BindN, BindN-rf, and BindN+.

The DP-Bind (http://lcg.rit.albany.edu/dp-bind/) [23] predicts DNA-binding residues using feature of evolutionary conservation in the form of PSI-BLAST position-specific scoring matrix with the models constructed onto the PDNA-62 data set. DP-Bind uses a web server to predict the DNA-binding residues and three machine-learning methods can be used in DP-Bind for prediction. After submitting the query protein to the servers, the putative DNA-binding residues of TS-72 can be received by E-mail. The Huang et al. model predicts DNA-binding residues using a SVM classifier with evolutionary information about amino acid sequences in terms of their PSSM. To use the Huang et al. method for predicting the putative DNA-binding residues in data set TS-72, we repeated the process of constructing a prediction model based on the training data set TR265. As DP-Bind and Huang et al. uses a 4.5 Å cutoff distance definition for DNA-binding residues, we also reconstructed a model on the training TR265 data set with the 4.5 Å cutoff distance and predicted DNA-binding residues for TS-72. As shown in Fig. 3, DNABR achieved an AUC of 0.8808, which is much better than the predictive ability of Huang et al. (AUC 0.6876) and DP-Bind using three machine-learning models (AUC are 0.5778, 0.5832, and 0.5526 for DP-Bind(SVM), DP-Bind(KLR), and DP-Bind(PLR), respectively). These results clearly demonstrate that DNABR performs significantly better than those two methods based on the same criterion of a cutoff distance of 4.5 Å. Therefore, the results demonstrate that whatever cutoff distance for the binding residues we choose, our DNABR method still has an excellent prediction performance.

TABLE 2
The Prediction Performance of the RF Model Based on Various Features, Which Was Evaluated by Fivefold Cross Validation on PDBP-337

Features	ACC±SD(%)	SE±SD(%)	SP±SD(%)	BM±SD(%)	MCC
A	91.10±0.40	62.36±1.82	96.27±0.35	75.69±1.29	0.5928
A+B	92.34±0.55	64.44±1.17	97.70±0.39	77.66±0.71	0.6331
A+C	91.55±0.39	60.55±1.34	97.18±0.29	74.61.±0.84	0.6125
A+B+C	93.04±0.47	68.47±2.92	98.16±0.60	80.67±1.79	0.6586

A: Position specific scoring matrices combining with physicochemical properties (PSSM-PP) B: Amino acid correlation (AAC) C: Orthogonal binary vectors (OBV)

3.2 Prediction Performance of Random Forest Model Based on Various Features

All the features described in the material and method section was used to build an RF-based prediction model (DNABR). The performance of DNABR was evaluated by fivefold cross validation on a processed training data set (PDPB-337). The predicted results obtained by using all the features are shown in Table 2. A balanced performance was achieved with an ACC of 93.04 percent and MCC of 0.6586 for the fivefold cross validation. The corresponding SE, SP, and BM were 68.47, 98.16, and 80.67 percent, respectively. As shown in Fig. 4, the ROC curve for the DNABR model has an AUC value of 0.9405, which is significantly higher than random guessing (0.5). Because the SVM has been successfully applied to many fields, this classifier was also applied to compare the performance to RF classifier in this work. Both classifiers were used on the same data set, with the same sequence-based features and were evaluated by fivefold cross validation. The ROC analysis shows that the SVM classifier achieves very similar performance with the RF classifier (Fig. 4), and the AUC values are 0.9405 and 0.9345 for the RF and SVM classifiers, respectively. Thus, we concluded that the sequence-based features and the processed training data set allow one to obtain a perfect prediction performance. The prediction RI is an important measure for evaluating the quality of the prediction and it represents the level of prediction performance of an unknown protein. Fig. 5 shows the fraction of samples with a different RI and the distribution of the accuracy of prediction fraction of samples with a different RI using the RF classifier with fivefold cross validation. For example, about 76.1 percent of all samples have an RI ≥ 5 and about 87.60 percent of these samples were correctly predicted by DNABR.

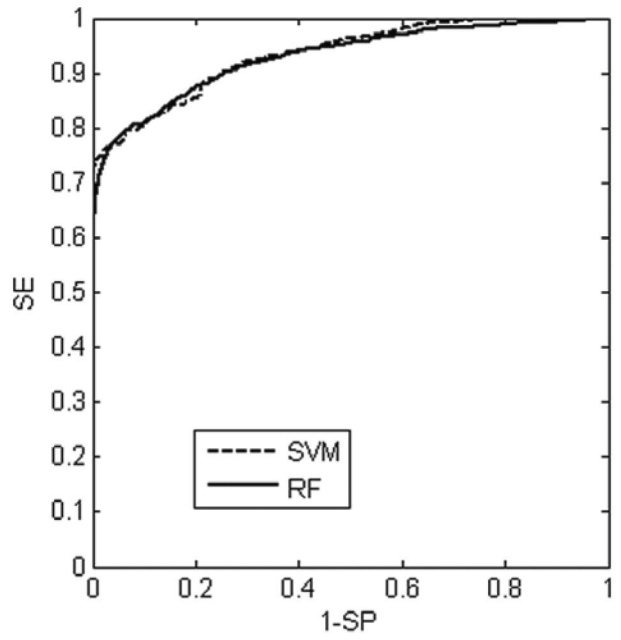

Fig. 4. Performance comparisons of the ROC graph with the RF and SVM classifiers on the same PDBP-337 data set using the same hybrid feature (PSSM-PP+AAC+OBV). The AUC values are 0.9405 and 0.9345 for the RF and SVM classifiers, respectively.

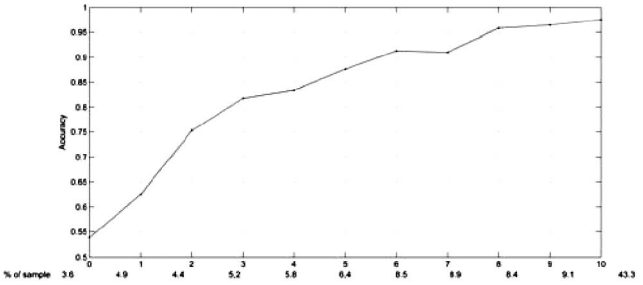

Fig. 5. The fraction of samples and prediction accuracy with each RI using the DNABR method. For example, 43.3 percent of all samples have a $\mathrm{RI} = 10$ and of these samples 97.49 percent were predicted correctly.

3.3 The Importance of the Novel Features

A PSSM is a useful feature and has contributed to improving the prediction performance of DNA-binding residues in proteins in previous studies [18], [23], [26], [28]. Based on the PSSM, we proposed a novel matrix named PSSM-PP, which not only contains the evolutionary information captured by PSSM, but also the conservation information about the amino acids at the level of their physicochemical properties. Many physicochemical properties relevant to DNA-protein interactions were considered in this research. For instance, the side chain pKa value determines the ionization state of a residue. Because the phosphate groups of nucleic acids are negatively charged, the ionization state of amino acid side chains affects the interaction with DNA molecules. The Wiener index is topological index of a molecule, which has the effect on the process by which the protein interacts with the DNA. In this research, we first chose 35 descriptors from AAindex database which are relevant to DNA-protein interactions and other nine physicochemical properties which are also relevant for DNA-protein interactions and mentioned in previous works [20], [27], [28], [29]: isoelectric point, pKa for α-COOH (pKa1), the pKa for α-NH3 (pKa2), the number of lone electron pairs, molecular mass, EIIP, the Wiener index, the Balaban index and the lowest free energy. (Details see the Appendix D, available in the online supplemental material) The key to choose each physicochemical property is based on how well the PSSM-PP constituted by it and PSSM distinguishes the binding residues from nonbinding ones. We used a two-sample t-test to test each PSSM-PP constituted by one physicochemical property and PSSM for

positive samples and negative samples and calculate the p-values, small p-value indicating greater separation and large p-values indicating less separation. We used training data set TR265 to obtained p-value of each physicochemical property. Each p-value of physicochemical property could be calculated by the t-test to test the PSSM-PP feature constituted by it and PSSM for positive samples and negative samples in TR265. The 10 best ranked physicochemical properties can be seen from the Appendix E, available in the online supplemental material. We found that the molecular mass, pKa1, pKa2 and the Wiener index are on the top. The remaining six physicochemical properties indicate the structure information of amino acid. We preferred to select physicochemical properties based on sequence information, therefore we first selected these top four physicochemical properties, i.e., pKa1, pKa2, the molecular mass, and the Wiener index.

We also used test data set TS-72 to obtained p-value of each physicochemical property to verify that the p-values of physicochemical properties are independent of the data set. Forty four-values of physicochemical properties could be calculated by the t-test to test the PSSM-PP feature constituted by each physicochemical properties and PSSM for positive samples and negative samples in TS-72. The rank of physicochemical properties from TS-72 is very similar to that from TR265. According to the result from TS-72, pKa1, pKa2, the molecular mass and the Wiener index are also on the top (Appendix F, available in the online supplemental material). We used the same strategy to obtained p-value of each physicochemical property in fivefold data sets which used in fivefold cross validation. The results of best ranked physicochemical properties obtained from five fold data sets are the same as the results obtained from training data set (TR265) and test data set (TS-72).

Moreover, the EIIP and the number of LEPs are closely related to DNA-protein interactions. EIIP representing the main energy term of the valence electrons, are essential physical parameters of biological molecules determining their long-range properties [48], [49]. The long-range biomolecular interactions represent an important factor which influences biological processes. For this reason, long-range interactions representing intrinsic physicochemical properties of proteins and nucleotide sequences should be included in analysis of protein-DNA interactions. As we described above, the EIIP is an important physical parameter of biological molecules that is important for determining their long-range properties, which represent intrinsic physicochemical properties of proteins and nucleotide sequences. For this reason, EIIP can work well in the analysis of protein-DNA interactions. The number of lone electron pairs is much related to the formation of hydrogen bonds which are the main force for DNA-protein interaction. Therefore, in this research, we selected the six physicochemical properties: pKa1, pKa2, the molecular mass, the Wiener index, EIIP, and LEPs.

To further ensure that the four top properties (pKa1, pKa2, the molecular mass, and the Wiener index) combining EIIP and LEPs works best for all examples in training and test data set, we evaluated the performance of the four top properties combining all possible combinations of two

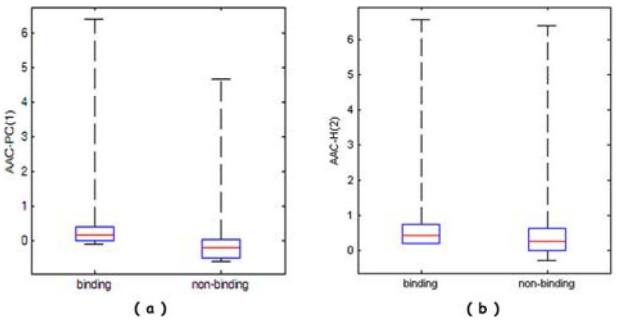

Fig. 6. Box plots of the two components of AAC features for binding and nonbinding residues. (a) AAC_PC (1) feature represent the correlation about polar amino acid with positive charge (b) AAC_H(2) feature represent the correlation for weak hydrophobic residue.

properties each from the remaining six properties in Appendix E, available in the online supplemental material. Compared the results from 15 (C_6^2) combinations of features, we found that four top properties, EIIP and LEPs combining with PSSM achieves the best performance on test data set TS-72 except the performance produced by four top properties, normalized frequency of alpha-helix and unfolding Gibbs energy in water pH9.0 combining with PSSM. (Details see Appendix G, available in the online supplemental material) Although the performance by the latter (AUC 0.8672) is slightly better than the former (AUC 0.8669), we preferred to select physicochemical properties based on sequence information. Therefore, it is reliable and reasonable to select the six physicochemical properties (pKa1, pKa2, the molecular mass, the Wiener index, EIIP, and LEPs) in this research.

The PSSM-PP captures more information and achieves better prediction performance than that of PSSM. Seen from Appendix H, available in the online supplemental material, the use of PSSM-PP for input encoding reaches an MCC of 0.5928 with 91.10 percent ACC, 62.36 percent SE, and 96.27 percent SP, which is better than PSSM for input encoding with the same data set PDPB-337 where the results were a MCC of 0.5753 with 89.71 percent ACC, 56.58 percent SE, and 95.02 percent SP. The value of AUC also denotes the comparison of performance and using PSSM-PP as the feature for classification achieves 0.8991, which improved the value of AUC (0.8764) using the PSSM feature. Therefore, we have already proved that the prediction performance of the classifier based on PSSM-PP is better than that based on PSSM or other two features (AAC and OBV), which was evaluated by fivefold cross validation on PDBP-337.

To ensure that the results are completely independent of the data set and evaluation method of fivefold cross validation, we trained the RF model on the training data set TR265 using PSSM-PP, PSSM, ACC, and OBV, respectively, and applied the new models to predict DNA-binding residues in the TS-72 test data set. Seen from Appendix I, available in the online supplemental material, the value of MCC also denotes the comparison of performance and using PSSM-PP as the feature for RF model achieves 0.4826, which improved the value of MCC (0.4537, 0.4218, 0.3752) using the PSSM, ACC, and OBV, respectively. Therefore, in

our research PSSM-PP is used as a significant feature instead of PSSM.

Noticeable differences in the AAC features were found between the binding and nonbinding residues. Using AAC_PC(1) and AAC_H(2) as examples, Figs. 6a and 6b show that binding and nonbinding residues display contrasting behavior in terms of two components of the AAC feature. The other six components of the AAC features also show significant difference between binding and nonbinding residues (see Appendix J, available in the online supplemental material). We also calculated the p-values of eight ACC components to measure the ability to separate the binding residues from the nonbinding ones. Each of the resulting p-values was less than 0.0001. These results show that the AAC feature carries important information about a residue binding to the DNA and plays a significant role in distinguishing binding and nonbinding residues. The results from Table 2 also highlight this conclusion that when AAC features are combined with PSSM-PP, the value of accuracy significantly increases and achieves 92.34 percent. The propensity of DNA-protein interactions can illuminate the importance of AAC features that represent the dependency of amino acids about polarity-charge and hydrophobicity: 1) from statistical values in the previous research [22], it is obvious that the charge and the polar property of the residues correlates well with its binding probability, and 2) hydrophobic residues tend to cluster on the surface of the protein, which means that hydrophobic residues tend to be binding residues. Moreover, those AAC features contain the information about the correlation between a residue and its neighbor residues having the same type. Therefore, the AAC features capture more information than independent residues.

4 Web Server

The DNABR web server (http://www.cbi.seu.edu.cn/DNABR/) was developed for biological research on the prediction of DNA-binding residues in proteins. Users can submit an amino acid sequence in a FASTA format to the web server for the prediction of binding residues. The DNA-binding residues within the submitted sequence will be predicted by the model constructed by a RF classifier on the processed PDBP-337 training data set using novel sequence-based features. The RF algorithm is computed through the random Forest R package [43]. Because it is time consuming to capture the PSSM-PP feature of the input sequence, an e-mail address is required to send the results. The predicted results will be sent back along with the predicted DNA-binding residues marked with a "+" and the nonbinding residues marked with an "−," along the input sequence (see Appendix K, available in the online supplemental material). The RI value is also calculated to measure the prediction reliability and may range from 0 to 10 for presentation. The higher the value is, the more reliable the prediction is.

5 Conclusions

In this report, an approach based on the RF classifier and novel sequence-based features have been described for the prediction of DNA-binding residues in proteins. In addition to the OBV feature, which captures the information about a single amino acid, two novel features, PSSM-PP and ACC, are also proposed in the present study as these reflect the information about conservation and correlation of physicochemical properties. Specifically, PSSM-PP denotes the evolutionary information combining conservation information for the physicochemical properties of the amino acids and AAC reflects the information about the dependency of amino acids within the protein sequences. The PSSM-PP and AAC features are original, and the results of this study indicate that the predictive performance can improve significantly when these two novel features are applied. These new features capture more information about the interaction between amino acids and the DNA and have better capacity to classify binding residues from nonbinding ones than described in previous reports. The best RF classifier achieved a prediction accuracy of 93.04 percent with a MCC of 0.6586 and BM of 80.67 percent. The comparison to other methods indicates that DNABR, using a RF algorithm combined with the novel sequence-based features mentioned above, is an excellent model for predicting DNA-binding residues using only sequence information. The new method has been implemented in a web server named DNABR for biological research on the prediction of DNA-binding residues. DNABR method achieves perfect performance on prediction of DNA-binding residues.

The approach of feature selection and the method of the model establishment could be used in the study on the prediction of RNA-binding residues. For example, PSSMP-PP which we proposed in this research contributed most to improving the prediction performance of DNA-binding residues in proteins. However, considering the difference between mechanism of RNA-protein interaction and that of DNA-protein interaction, the physicochemical properties which constitute PSSM-PP should be selected based on mechanism of RNA-protein interaction. We think that RNA-binding residues in proteins can be also predicted with high accuracy after DNABR method improved.

Acknowledgments

The authors wish to thank Veljko Veljkovic for the suggestion of EIIP. This work is support by National Natural Science Foundation of China (Project No. 61073141 and No. 60971099). Funding to pay the Open Access publication charges for this paper was provided by National Natural Science Foundation of China 61073141.

References

[1] J. Wang and Morigen, "BayesPI - A New Model to Study Protein-DNA Interactions: A Case Study of Condition-Specific Protein Binding Parameters for Yeast Transcription Factors," *BMC Bioinformatics*, vol. 10, article 345, 2009.

[2] L. Zamdborg and P. Ma, "Discovery of Protein-DNA Interactions by Penalized Multivariate Regression," *Nucleic Acids Research*, vol. 37, no. 16, pp. 5246-5254, Sept. 2009.

[3] J.B. Kinney, G. Tkacik, and C.G. Callan Jr., "Precise Physical Models of Protein-DNA Interaction from High-Throughput Data," *Proc Nat'l Academy of Sciences USA*, vol. 104, no. 2, pp. 501-506, Jan. 2007.

[4] U. Singh, E. Bongcam-Rudloff, and B. Westermark, "A DNA Sequence Directed Mutual Transcription Regulation of HSF1 and NFIX Involves Novel Heat Sensitive Protein Interactions," *PLoS One,* vol. 4, no. 4, e5050, pp. 1-12, 2009.

[5] D. Ucar et al., "Predicting Functionality of Protein-DNA Interactions by Integrating Diverse Evidence," *Bioinformatics,* vol. 25, no. 12, pp. i137-i144, June 2009.

[6] J.M. Vaquerizas et al., "A Census of Human Transcription Factors: Function, Expression and Evolution," *Nature Rev. Genetics,* vol. 10, no. 4, pp. 252-263, Apr. 2009.

[7] A. Hoglund and O. Kohlbacher, "From Sequence to Structure and Back Again: Approaches for Predicting Protein-DNA Binding," *Proteome Science,* vol. 2, no. 1, pp. 1-9, June 2004.

[8] Y. Fang et al., "Predicting DNA-Binding Proteins: Approached from Chou's Pseudo Amino Acid Composition and Other Specific Sequence Features," *Amino Acids,* vol. 34, no. 1, pp. 103-109, Jan. 2008.

[9] L. Nanni and A. Lumini, "Combing Ontologies and Dipeptide Composition for Predicting DNA-Binding Proteins," *Amino Acids,* vol. 34, no. 4, pp. 635-641, May 2008.

[10] V.H. Nagaraj, R.A. O'Flanagan, and A.M. Sengupta, "Better Estimation of Protein-DNA Interaction Parameters Improve Prediction of Functional Sites," *BMC Biotechnology,* vol. 8, article 94, 2008.

[11] P. Aloy et al., "Automated Structure-Based Prediction of Functional Sites in Proteins: Applications to Assessing the Validity of Inheriting Protein Function from Homology in Genome Annotation and to Protein Docking," *J. Molecular Biology,* vol. 311, no. 2, pp. 395-408, Aug. 2001.

[12] N. Bhardwaj et al., "Structure Based Prediction of Binding Residues on DNA-Binding Proteins," *Proc. IEEE 27th Int'l Conf. Eng. in Medicine and Biology Soc.,* vol. 3, pp. 2611-4, 2005.

[13] S. Jones et al., "Using Electrostatic Potentials to Predict DNA-Binding Sites on DNA-Binding Proteins," *Nucleic Acids Research,* vol. 31, no. 24, pp. 7189-98, Dec. 2003.

[14] I.B. Kuznetsov et al., "Using Evolutionary and Structural Information to Predict DNA-Binding Sites on DNA-Binding Proteins," *Proteins,* vol. 64, no. 1, pp. 19-27, July 2006.

[15] G. Nimrod et al., "Identification of DNA-Binding Proteins Using Structural, Electrostatic and Evolutionary Features," *J. Molecular Biology,* vol. 387, no. 4, pp. 1040-1053, Apr. 2009.

[16] G. Nimrod et al., "iDBPs: A Web Server for the Identification of DNA Binding Proteins," *Bioinformatics,* vol. 26, no. 5, pp. 692-693, Mar. 2010.

[17] E.W. Stawiski, L.M. Gregoret, and Y. Mandel-Gutfreund, "Annotating Nucleic Acid-Binding Function Based on Protein Structure," *J. Molecular Biology,* vol. 326, no. 4, pp. 1065-1079, Feb. 2003.

[18] S. Ahmad and A. Sarai, "PSSM-Based Prediction of DNA Binding Sites in Proteins," *BMC Bioinformatics,* vol. 6, article 33, 2005.

[19] L. Wang and S.J. Brown, "Prediction of DNA-Binding Residues from Sequence Features," *J. Bioinformatics Computational Biology,* vol. 4, no. 6, pp. 1141-1158, Dec. 2006.

[20] 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 Research,* vol. 34, no. web server issue, pp. W243-W248, July 2006.

[21] C. Yan et al., "Predicting DNA-Binding Sites of Proteins from Amino Acid Sequence," *BMC Bioinformatics,* vol. 7, article 262, 2006.

[22] N. Bhardwaj and H. Lu, "Residue-Level Prediction of DNA-Binding Sites and Its Application on DNA-Binding Protein Predictions," *FEBS Letters,* vol. 581, no. 5, pp. 1058-1066, Mar. 2007.

[23] 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.

[24] Y. Ofran, V. Mysore, and B. Rost, "Prediction of DNA-Binding Residues from Sequence," *Bioinformatics,* vol. 23, no. 13, pp. i347-i353, July 2007.

[25] H. Tjong and H.X. Zhou, "DISPLAR: An Accurate Method for Predicting DNA-Binding Sites on Protein Surfaces," *Nucleic Acids Research,* vol. 35, no. 5, pp. 1465-1477, 2007.

[26] Y.F. Huang et al., "DNA-Binding Residues and Binding Mode Prediction with Binding-Mechanism Concerned Models," *BMC Genomics,* vol. 10, Suppl. 3, article S23, pp. 1-10, 2009.

[27] L. Wang, M.Q. Yang, and J.Y. Yang, "Prediction of DNA-Binding Residues from Protein Sequence Information Using Random Forests," *BMC Genomics,* vol. 10, Suppl. 1, article S1, pp. 1-9, 2009.

[28] J.-S. Wu, X. Ma, H.-D. Liu, X.-N. Yang, J.-M. Xie, and X. Sun, "A SVM-Based Approach for Predicting DNA-Binding Residues in Proteins from Amino Acid Sequences," *Proc. Int'l Joint Conf. Bioinformatics, Systems Biology and Intelligent Computing,* pp. 225-229, 2009.

[29] L. Wang et al., "BindN+ for Accurate Prediction of DNA and RNA-Binding Residues from Protein Sequence Features," *BMC Systems Biology,* vol. 4, Suppl 1, article S3, pp. 1-9, 2010.

[30] S. Ahmad, M.M. Gromiha, and A. Sarai, "Analysis and Prediction of DNA-Binding Proteins and Their Binding Residues Based on Composition, Sequence and Structural Information," *Bioinformatics,* vol. 20, no. 4, pp. 477-486, Mar. 2004.

[31] J. Si et al., "MetaDBSite: A Meta Approach to Improve Protein DNA-Binding Sites Prediction," *BMC Systems Biology,* vol. 5, Suppl 1, article S7, pp. 1-7, 2011.

[32] H.M. Berman et al., "The Protein Data Bank," *Nucleic Acids Research,* vol. 28, no. 1, pp. 235-242, Jan. 2000.

[33] S.F. Altschul et al., "Basic Local Alignment Search Tool," *J. Molecular Biology,* vol. 215, no. 3, pp. 403-410, Oct. 1990.

[34] S.F. Altschul et al., "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs," *Nucleic Acids Research,* vol. 25, no. 17, pp. 3389-3402, Sept. 1997.

[35] Q. Zhang, S. Yoon, and W.J. Welsh, "Improved Method for Predicting Beta-Turn Using Support Vector Machine," *Bioinformatics,* vol. 21, no. 10, pp. 2370-2374, May 2005.

[36] J. Wang, "Biochemistry," *Higher Education (in Chinese),* 2002.

[37] V. Veljkovic et al., "Application of the EIIP/ISM Bioinformatics Concept in Development of New Drugs," *Current Medical Chemistry,* vol. 14, no. 4, pp. 441-453, 2007.

[38] D. Bonchev, "The Overall Wiener Index—A New Tool for Characterization of Molecular Topology," *J. Chemical Information and Computer Sciences,* vol. 41, no. 3, pp. 582-592, May/June 2001.

[39] V.N. Vapnik, *Statisical Learning Theory.* Wiley, 1998.

[40] L.J. Hu Xiu zhen, "Statistical Analysis of Application of Hydrophilicity Hydrophobicity and Molecular Size of Amino Acid (in Chinese)," *J. Inner Mongolia Polytechnic Univ.,* vol. 19, no. 3, pp. 187-191, 2000.

[41] J. Shen et al., "Predicting Protein-Protein Interactions Based Only on Sequences Information," *Proc. Nat'l Academy of Sciences USA,* vol. 104, no. 11, pp. 4337-4341, Mar. 2007.

[42] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[43] A.W. Liaw and M. Weiner, "Classification and Regression by Random Forest," *R. News,* vol. 2, pp. 18-22, 2002.

[44] G. Cohen et al., "Learning from Imbalanced Data in Surveillance of Nosocomial Infection," *Artificial Intelligence Medicine,* vol. 37, no. 1, pp. 7-18, May 2006.

[45] J.P. Egan, *Signal Detection Theory and ROC-Analysis.* Academic Press, 1975.

[46] A.P. Bradley, "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms," *Pattern Recognition,* vol. 30, pp. 1145-1159, 1997.

[47] W.L. DeLano, *The PyMOL Molecular Graphics System.* DeLano Scientific, 2002.

[48] V. Veljkovic, *A Theoretical Approach to Preselection of Carcinogens and Chemical Carcinogenesis.* Gordon & Breach, 1980.

[49] N. Veljkovic et al., "Discovery of New Therapeutic Targets by the Informational Spectrum Method," *Current Protein Peptide Sciences,* vol. 9, no. 5, pp. 493-506, Oct. 2008.

Xin Ma received the BS degree in mathematical science in 2003 from SuZhou University and the PhD degree in biomedical engineering in 2012 from Southeast University. Her research interests mainly focus on Bioinformatics and machine learning. And she published 16 papers in journals and conferences in these areas.

Jing Guo received the BS degree in School of Biological Science and Medical Engineering, Southeast University, Nanjing, China, in 2009. Now, she is working toward the master's degree in the State Key Laboratory of Bioelectronics, Southeast University. And she published four papers in journals and conferences in these areas.

Hong-De Liu received the PhD degree in biomedical engineering in 2006 from northwest normal University. He is currently an associate professor of School of Biological Science & Medical Engineering in Southeast University of China. His research focuses on "nucleosome positioning and its roles in gene regulation." And he published more than 20 paper in journals and conferences in these areas.

Jian-Ming Xie received the B.Eng degree in biomedical engineering from Southeast University, Nanjing, China, in 1993, the BMed degree in clinical medicine from Nanjing Medical University, Nanjing, China, in 1996, and the PhD degree in biomedical engineering from Southeast University, Nanjing, China, in 2009. He is with the School of Biological Science and Medical Engineering at the Southeast University as assistant lecturer (1997), lecturer (1999), and associate professor (2005). His research interests include the systems biology and high-throughput genomic data analysis, especially focus on the integration analysis of gene expression and genome variation data for studying the molecular mechanism of the human complex disease. And he published more than 10 paper in journals and conferences in these areas.

Xiao Sun received the BS degree in computer science in 1984 and the PhD degree in biomedical engineering in 1993 from Southeast University. Since 2001, he has been a professor in the School of Biological Science and Medical Engineering of Southeast University. His major research interests include Bioinformatics and Biomedical Signal Processing. He has been in charge of and completed many projects which were funded by the Natural Science Foundation of China or National High-tech R&D Program (863 Program). He is the author or corresponding author of over 60 papers. He also published two books "*Fundamentals of Bioinformatics*" and "*R & BioConductor and Their Applications in Genome Analysis*."