An accurate feature-based method for identifying DNA-binding residues on protein surfaces

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · FEBRUARY 2011

Impact Factor: 2.63 · DOI: 10.1002/prot.22898 · Source: PubMed

3 AUTHORS:

Yi Xiong
Shanghai Jiao Tong University
17 PUBLICATIONS **123** CITATIONS
SEE PROFILE

Juan Liu
Wuhan University
130 PUBLICATIONS **707** CITATIONS
SEE PROFILE

Dong-Qing Wei
Shanghai Jiao Tong University
224 PUBLICATIONS **4,646** CITATIONS
SEE PROFILE

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

An accurate feature-based method for identifying DNA-binding residues on protein surfaces

Yi Xiong,[1,2] Juan Liu,[1]* and Dong-Qing Wei[1,2]*

[1] School of Computer, Wuhan University, Wuhan 430072, People's Republic of China

[2] Department of Bioinformatics and Biostatistics, College of Life Science and Biotechnology, Shanghai Jiaotong University, Shanghai 200240, People's Republic of China

ABSTRACT

Proteins that interact with DNA play vital roles in all mechanisms of gene expression and regulation. In order to understand these activities, it is crucial to analyze and identify DNA-binding residues on DNA-binding protein surfaces. Here, we proposed two novel features B-factor and packing density in combination with several conventional features to characterize the DNA-binding residues in a well-constructed representative dataset of 119 protein-DNA complexes from the Protein Data Bank (PDB). Based on the selected features, a prediction model for DNA-binding residues was constructed using support vector machine (SVM). The predictor was evaluated using a 5-fold cross validation on above dataset of 123 DNA-binding proteins. Moreover, two independent datasets of 83 DNA-bound protein structures and their corresponding DNA-free forms were compiled. The B-factor and packing density features were statistically analyzed on these 83 pairs of holo-apo proteins structures. Finally, we developed the SVM model to accurately predict DNA-binding residues on protein surface, given the DNA-free structure of a protein. Results showed here indicate that our method represents a significant improvement of previously existing approaches such as DISPLAR. The observation suggests that our method will be useful in studying protein-DNA interactions to guide consequent works such as site-directed mutagenesis and protein-DNA docking.

Proteins 2011; 79:509–517.

Key words: protein-DNA interaction; structural feature; B-factor; packing density; support vector machine.

INTRODUCTION

Proteins that interact with DNA are involved in a wide range of biological processes such as gene regulation, DNA replication, repair, and rearrangement.[1] A reliable identification of the interaction sites on DNA-binding proteins is the first step to help us understand these activities at the residue level. Experimentally, the binding sites can be identified by analyzing structures of protein-DNA complexes or by site-directed mutagenesis studies. However, acquiring structures of protein-DNA complexes or studying all possible mutations of residues on proteins is a complicated and time-consuming process. Thus, computational methods are needed to assist in finding the distinguished features of DNA-binding residues and identifying the binding sites on protein surfaces.

By statistically analyzing protein-DNA complexes deposited in the PDB,[2] some common features were able to characterize the DNA-binding residues. These characteristics include enrichment of positively charged Arg and Lys residues and polar residues in protein-DNA interfaces.[3] The evolutionary conservation of residues is another important property for the identification of interaction sites on protein surfaces.[4,5] Other features were also utilized such as solvent accessibility, second structure, pK_a, hydrophobicity index, electrostatic potential and dipoles and volumes of the side chains.[5–11] More recently, dynamic-derived features such as residue fluctuations in low frequency[12] and high frequency modes[13] were shown to be correlated with the regions involved in DNA binding.

In recent years, it has becoming apparent that the simple scoring function or linear combination method is not competent to correlate the intricacy of protein-DNA interfaces with a range of features. Instead, the sophisticated machine learning methods, including artificial neural network (ANN) and SVM, have been popularly and successfully applied to the identification of DNA-binding residues using a combination of

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Natural Science Foundation of China; Grant numbers: 60773010, 60970063, 20773085 and 30870476; Grant sponsor: Ph.D. Programs Foundation of Ministry of Education of China; Grant number: 20090141110026; Grant sponsor: National 863 Bioinformatics Projects; Grant number: 2007AA02Z333

*Correspondence to: Juan Liu, School of Computer, Wuhan University, Wuhan 430072, People's Republic of China. E-mail: liujuan@whu.edu.cn and Dong-Qing Wei, College of Life Science and Biotechnology, Shanghai Jiaotong University, Shanghai 200240, Republic of China. E-mail: dqwei@sjtu.edu.cn.

Received 8 April 2010; Revised 15 August 2010; Accepted 15 September 2010
Published online 7 October 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.22898

features as input. Most of these machine learning approaches exploited sequence-based features to construct their classifiers.[6,9,10,14–17] In general, these methods utilized the features of a sequence window rather than a single residue as input. These sequence-based predictors have at least two major limitations. One problem with sequence windows is that amino acids that are in sequence neighbors are not necessarily close in space to confer the DNA-binding function. The other problem is that sequence information provides few clues to the interaction sites and is not sufficient for the accurate prediction of DNA-binding residues.

An alternative to sequence-based prediction is to recognize DNA-binding residues using structure-derived features. The features extracted from DNA-binding protein structures were demonstrated to effectively identify DNA-binding residues in recent studies.[5,8] However, the performance of these methods was still limited by the employed features in their work. Moreover, the increasing availability of high-resolution structures of protein-DNA complexes provides an opportunity towards a more detailed examination of potential structure-based features for their ability to describe and identify DNA-binding residues.

The aim of our study is to present an approach that can accurately predict DNA-binding sites on protein surface, given the DNA-free structure of a protein. We focused on novel structure-based features for characterizing and identifying DNA-binding residues. One structure descriptor is the B-factor, which measures the atomic flexibility in crystalline state.[18] The B-factor has been investigated to characterize enzyme catalytic site residues[19] and protein-protein binding sites.[20,21] Here, we extended the B-factor feature to analyze DNA-binding residues to examine whether the feature can describe the binding sites on DNA-bound or DNA-free protein structures. On the other hand, several studies have focused on analyzing the residue packing density, which was shown to be closely associated with protein structures[22] and protein-protein interaction hot spots.[23,24] In a recent study,[25] the packing density of clustered-conserved residues in protein-DNA interfaces was also well elucidated. Here, we presented an analysis of the packing density distribution of DNA-binding sites on protein structure surface in both holo and apo forms. These two novel descriptors with the conventional features, including the evolutionary profile, solvent accessibility and side chain pK_a, were combined to characterize and predict DNA-binding residues on different datasets using statistical methods and SVM technologies. An accurate SVM model was then developed to predict DNA-binding residues on the surface of a DNA-free protein structure. In addition, we demonstrated that our approach achieved better performance than other state-of-the-art methods such as DISPLAR. The high level of prediction performance suggests that our method will be useful in guiding consequent works such as site-directed mutagenesis and protein-DNA docking.

METHODS

Datasets

A representative set of protein-DNA complex structures was selected by the following procedure: A complete list of 1334 protein-DNA complexes, determined by X-ray crystallography with a resolution better than 3Å, was downloaded from the PDB (September 2009 release, http://www.rcsb.org/). The 1046 complexes containing double-stranded DNA of at least six base pairs were retained. These complex structures were further partitioned into 2197 chains. Protein chains with less than five residues within 4.5 Å of the DNA molecule were removed. The resulting 2094 protein chains were culled using PISCES sequence culling server,[26] which aims to provide the longest list possible of the best resolution structures that fulfill the sequence identity, length and structural quality cutoffs. The resultant dataset consists of 206 DNA-binding protein (DBP) chains with mutual sequence identity no more than 25% and a minimum of 40 amino acids. All the structures have R factor no higher than 0.3.

We further collected a total of 83 holo-apo pairs of DNA-binding protein structures determined both with DNA bound and unbound from the 206 DBP chains. The aligned sequences of bound proteins and their unbound counterparts have identity more than 95%.

Here, the set of 206 DBP chains was divided into two subsets. Set 1 (DBP-123) were composed of 123 DBP chains, which had the holo structures but not the corresponding apo forms. Set 2 (HOLO-83) included the remaining 83 DBP chains, all of which had their unbound counterparts. The corresponding unbound structures of the dataset HOLO-83 were collected as Set 3 (APO-83).

Set 1 was used as the training set, and Set 2 and Set 3 were taken as the independent test sets. The detailed lists of these three data sets were provided in supplementary tables.

Definition of surface residues and DNA-binding residues

In this study, we defined surface residues and DNA-binding residues following the definitions of Tjong and Zhou.[5] A residue is taken as a surface residue if its solvent accessible surface area (SASA) is at least 10% of maximum values in a tri-peptide state.[27] The SASA of residues in each protein multimer in the absence of DNA were calculated using the program NACCESS.[28] A surface residue is labeled as a binding residue if it contains at least one heavy atom that falls within the distance of 4.5Å from any heavy atoms of the DNA molecule. According to the criterion, the training dataset DBP-123 contained 18323 surface residues, about 15.8% of which were DNA-binding residues.

Features of DNA-binding residues

Evolutionary profile

A position specific scoring matrix (PSSM), representing the evolutionary profile of a protein sequence, was generated by using the PSI-BLAST program.[29] The PSSM was constructed by three iterations of PSI-BLAST searches against NCBI nonredundant database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) with the BLOSUM62 substitution matrix.[30] Both parameters of e and h were set to 0.001. Since the matrix elements had a wide range, they were scaled to the range [0, 1] by a standard logistic function.[10]

Relative solvent accessibility

Relative solvent accessibility of a residue was calculated as the ratio of its SASA to the nominal maximum area of its residue type in a tri-peptide state.

B-factor

B-factors are highly related to the flexibility of atoms and residues in a protein, and are determined by X-ray crystallographic experiments.[31] B-factor of C_α atom was used to represent its residue flexibility and obtained from its PDB file and. For each protein chain, the *B*-factor of each C_α atom was normalized as follows:

$$NB = \frac{B - \mu(B)}{\sigma(B)} \quad (1)$$

where B is the *B*-factor value of a given residue, $\mu(B)$ and $\sigma(B)$ are the average value and the standard deviation of the B-factors for the selected chain, respectively.

Packing density

The definition of packing density is as described earlier[24,25]: packing density is designed as the number of noncovalently bonded residues whose C_α position falls within a sphere of 6 Å radius from the C_α position of a target residue. Two types of packing density were used in our work. The first type of packing density is calculated as the number of residues from the corresponding protein chain, whereas the second type includes the residues from both the protein chain and the DNA molecule if the protein is in holo form. In the packing density definition for DNA nucleotides, any backbone phosphate atom is taken as the C_α position. In following sections, we usually referred to the first type of packing density (i.e., as the input feature for training) if it was not specially designed.

pK_a

The side chain pK_a values we used here were taken from the reference book,[32] and the pK_a value was set to 7.0 for the amino acids with no side chain pK_a values.[9]

Classifiers construction

SVM classifiers were applied to the identification of DNA-binding residues in our experiments. In SVM classifiers, input data are nonlinearly mapped into a high dimensional feature space and optimally separated by a hyper-plane into two classes.[33] In the classification tasks, SVM classifiers were constructed using aforementioned features of structural windows as input. Each window was centered on a target residue and its ten spatially nearest surface residues.

In our study, SVM classifiers were implemented using Weka LibSVM package[34,35] with the radial basis function as a kernel. The optimal values of γ and regularization parameter C were set to be 0.1 and 10.0, respectively.

Evaluation methods

A five-fold cross validation was used to evaluate the classification performance. In this procedure, the dataset DBP-123 was randomly divided into five subsets with an approximately equal number of chains. In each of the five validations, four of the five parts were used for training and the remaining one for testing, which was repeated for five times. In DBP-123 dataset, the ratio of the number of DNA-binding residues to that of the nonbinding sites approximates to 1:5. To overcome the unbalanced problem in the training set, we randomly sampled an equal number of nonbinding residues to that of the DNA-binding ones for training the model. This sampling and training was performed five times in each run, and the final decision for the testing result was made by majority vote.

In the independent test, the SVM model was trained on DBP-123 and then tested on every DBP chain in HOLO-83 and APO-83, respectively. The strategy of sampling and training was the same as that in five-fold cross validation.

Performance measures

The sensitivity (SN, also called recall), specificity (SP), precision (PR), and F-measure (F_1) were used for assessment of the performance of SVM classifiers. These indicators can be measured by the numbers of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each classifier. The performance measures are defined as the following:

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (2)$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \quad (3)$$

$$\mathrm{P}R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (4)$$

$$F_1 = \frac{2 \times \mathrm{SN} \times \mathrm{P}R}{\mathrm{SN} + \mathrm{P}R} \quad (5)$$

The receiver operating characteristic (ROC) curve is a plot of the sensitivity versus (1-specificity) for a binary classifier at varying threshold from 0 to 1 (the threshold is assigned as the probability of the target residue to be a binding site in our work). The area under the curve (AUC) was used as a measure of classifiers performance throughout our work. In subsequent sections, the measures of SN, SP, PR and F_1 were reported at the threshold of 0.5 if it was not specially designed.

Statistical analysis

Statistical analysis was used to examine the roles of individual features. The Kolmogorov-Smirnov test is a nonparametric test for determining whether two samples of observations come from the same distribution. The result of this test was further confirmed by Mann-Whitney *U* test and Student's *t*-test. In following sections, only the results of Kolmogorov-Smirnov test were reported, and the other two tests gave the similar conclusions and hence their data was not shown.

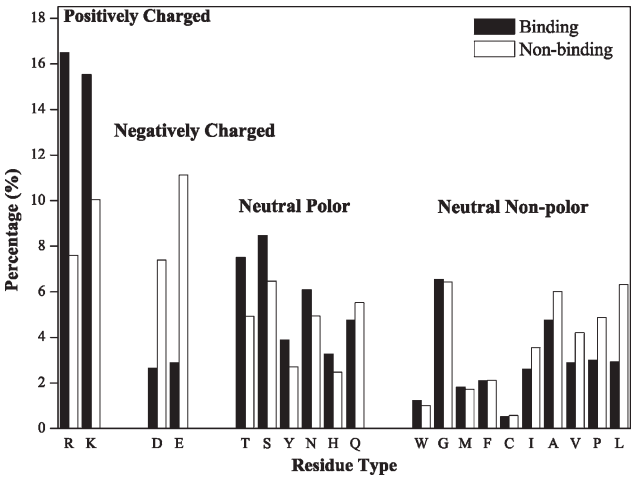

Figure 1

Compositions of the 20 types of residues between DNA-binding and nonbinding groups on DBP-123.

RESULTS AND DISCUSSION

Analysis of DNA-binding residues characteristics

The goal of this section is to analyze statistically how various features are distributed in DNA-binding and nonbinding groups on the dataset DBP-123. Figure 1 shows the compositions of the 20 types of residues between DNA-binding and nonbinding groups. It is observed that DNA-binding interfaces are highly enriched in positively charged residues (Arg and Lys) and devoid of negatively charged ones (Asp and Glu). The observations are in agreement with previous findings.[3,5] It is also clear that residues (Thr, Ser, Tyr, Asn and His) with polar side chains are relatively enriched in the DNA-binding group, whereas residues (Phe, Ile, Ala, Val, Pro, and Leu) with nonpolar side chains are depleted. This reflects the fact that the polar nature of these residues can result in hydrogen bonds between protein and DNA. The charge and polar preference of side chain in DNA-binding residues is partly encoded in their side chain pK_a values.

It is generally established that functionally important residues are usually more evolutionarily conserved than the rest of the protein surface. Here, the residue conservation scores were calculated using the method of Amhad *et al.*[25] As shown in Figure 2(A), the DNA-binding residues were distinguishable from the nonbinding ones on protein surface by their higher degree of conservation. However, conservation score is not sufficient for the description of DNA-binding sites. Some of the binding sites were highly conserved, and some were not conserved (shown in Supporting Information Fig. 1). In addition, a part of non-DNA-binding sites were also highly conserved. They may be other functional residues, such as protein-protein and small ligand interaction sites.

Figure 2(B) shows the differences of solvent accessibility between DNA-binding and nonbinding groups. Similar to the previous study,[5] the positively charged residues Arg and Lys were more exposed in the binding group than in the nonbinding group, giving resultant more binding propensity, whereas for negatively charged residues Asp and Glu it was opposite.

Previous studies have demonstrated that the B-factors of the interfaces were lower than the noninterfaces in protein-protein complexes.[20,36] We calculated the B-factor feature to analyze its distributions in DNA-binding and nonbinding groups. From Figure 2(C), the average values of the B-factors of 20 types of residues in DNA-binding group were also significantly lower than the nonbinding group. This result can be explained by the fact that the atoms of DNA-binding residues are less exposed to solvent and experience less fluctuation, resulting in relatively lower B-factors. Generally, the B-factor is a distinguished feature to characterize the binding residues of biological macromolecules in their bound states.

As shown in Figure 2(D), except for Asn, Asp and Cys, the packing density of DNA-binding residues were slightly lower than the rest residues of the protein surface for the 20 types of residues. Statistical analysis indicates that the packing density average for DNA-binding residues is 2.77 compared to 3.04 of the rest surface. Conducting a Kolmogorov-Smirnov test on packing density data for all surface residues, the P-value was 4.0×10^{-11}, confirming the statistical significance of the difference. In addition, we provided the P-values of the comparison of

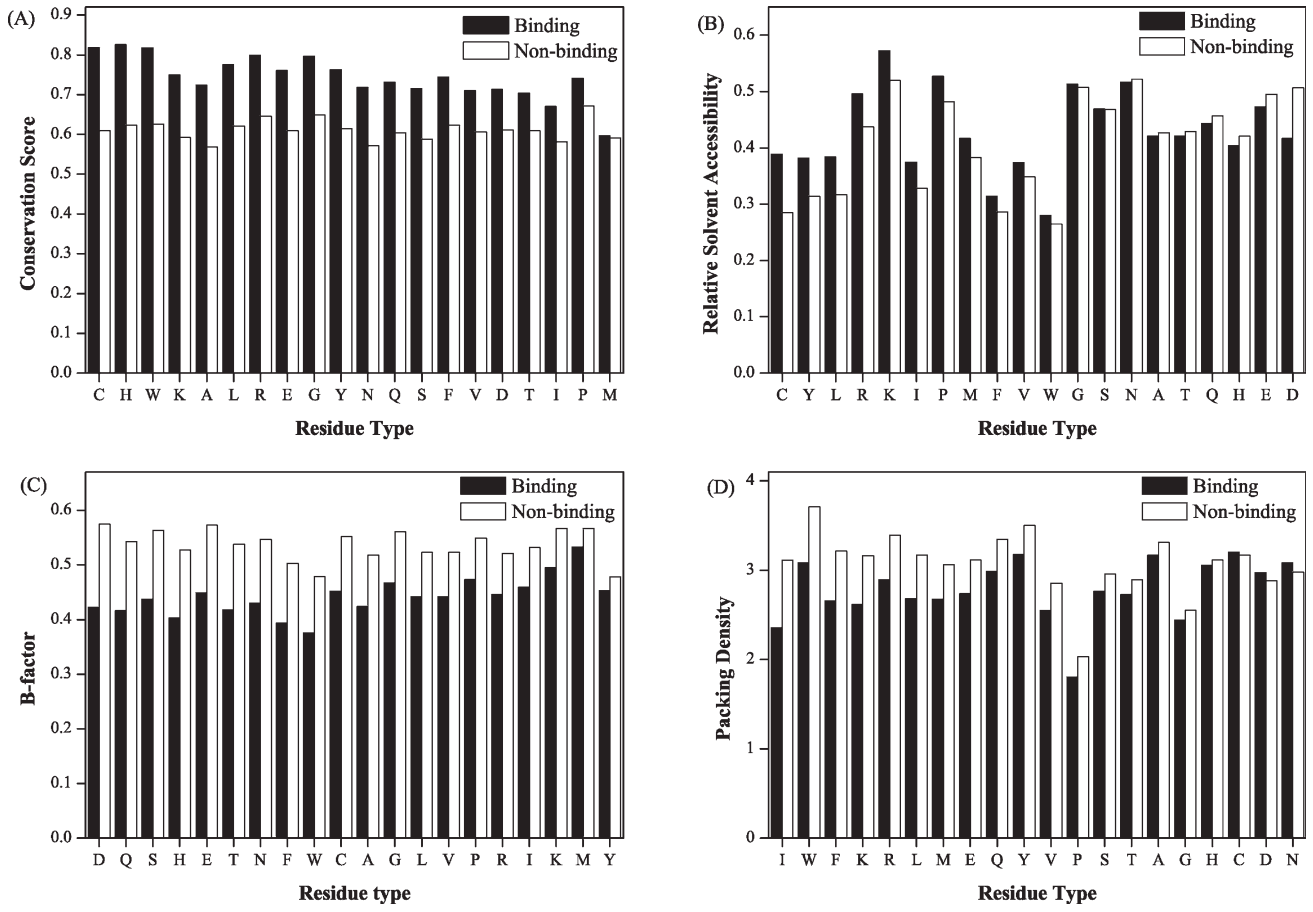

Figure 2

Feature distribution of the 20 types of residues between DNA-binding and nonbinding groups on DBP-123. **A:** Conservation score. **B:** Relative solvent accessibility. **C:** B-factor. **D:** Packing density.

packing density at different cutoff distance of definition of DNA-binding residues and packing density in Supporting Information Table IV.

Classifiers performance on five-fold cross validation

In previous section, it was demonstrated that individual features were possible to differentiate, to some degree, the binding residues from nonbinding ones on protein surfaces. In this section, different combinations of these features were utilized as input for SVM classifiers to evaluate their performance using five-fold cross validation on the dataset DBP-123. From Table I, the AUC value of 0.817 was given by the model created from PSSM alone, which was the major feature and contributed most to the performance of our model. Models generated from the PSSM with combinations of RSA, pK_a, PD or B-factor all gave the AUC higher than 0.817, indicating that the additional features do improve the performance of the model with PSSM alone. The best performance was achieved by integrating PSSM with all other four features. The result demonstrates that these five features are capable of providing complementary information for discriminating DNA-binding residues from nonbinding ones in DBP-123.

B-factor and DNA-binding residues

In this section, we used HOLO-83 and APO-83 datasets to examine the distribution of the residue B-factor values in both bound and unbound states. In HOLO-83, the distribution of the normalized B-factor values between the interfaces and noninterfaces is similar with that in DBP-123 (data not shown). It is concluded that the DNA-binding sites are more rigidly held in place than nonbinding residues in the protein-DNA complexes. In DNA-free proteins of the data set APO-83, the results show that the B-factor values of binding sites varied over the 20 types of residues [See Fig. 3(A)]. As shown in Figure 3(B), the B-factors of binding sites in free proteins were ranging from relatively lower values (rigidity) to higher values

Table I
Performance Measures of the Five-Fold Cross-Validation with a Different Combination of Features Using DBP-123

Feature	SN (%)	SP (%)	PR (%)	F_1	AUC
PSSM+pK_a+RSA+PD+B-factor	76.6 ± 4.0[a]	76.2 ± 3.3	38.7 ± 6.1	0.511 ± 0.061	0.836
PSSM+pK_a+RSA+PD	76.7 ± 4.3	74.9 ± 3.7	37.5 ± 6.7	0.500 ± 0.068	0.822
PSSM+pK_a+PD	76.1 ± 4.1	74.9 ± 3.5	37.3 ± 6.3	0.497 ± 0.064	0.821
PSSM+PD	76.3 ± 4.3	74.4 ± 3.7	36.9 ± 6.5	0.494 ± 0.067	0.820
PSSM+pK_a+RSA	75.8 ± 4.5	74.5 ± 3.5	36.9 ± 6.3	0.493 ± 0.064	0.820
PSSM+pK_a	75.7 ± 4.9	74.4 ± 3.2	36.7 ± 6.3	0.491 ± 0.066	0.820
PSSM+RSA+PD	75.5 ± 4.3	74.9 ± 3.7	37.2 ± 6.8	0.495 ± 0.069	0.819
PSSM+RSA	75.5 ± 4.5	74.5 ± 4.0	36.9 ± 6.5	0.492 ± 0.065	0.818
PSSM	75.8 ± 4.6	74.3 ± 3.9	36.8 ± 6.5	0.492 ± 0.066	0.817

[a]The data is standard deviation for five-fold.

(flexibility), which suggests that the binding sites have dual character about mobility. The result is in agreement with the previous finding of Luque and Freire.[37]

Packing density and DNA-binding residues

The aim of this section is to investigate the packing density distribution in HOLO-83 and APO-83. From Figure 4(A), the average packing density of DNA-binding residues was moderately lower than nonbinding in APO-83 when the packing density measure was only derived from the protein chain. A similar curve is also obtained in DBP-123 and HOLO-83 (data not shown). The lesser packed regions may be critical for protein binding site flexibility.[24] However, as shown in Figure 4(B), the average packing density of DNA-binding residues was

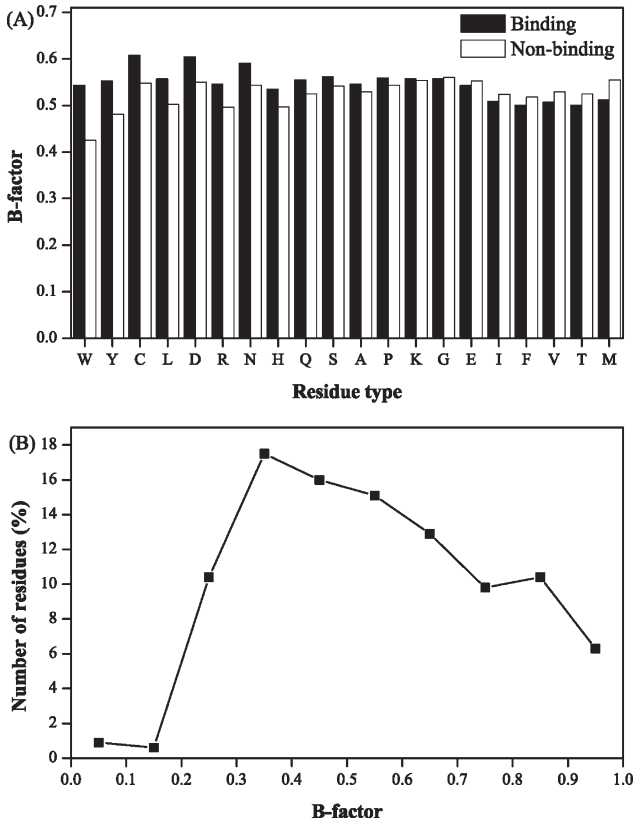

Figure 3
A: B-factor distribution of the 20 types of residues between DNA-binding and nonbinding groups on APO-83. **B:** The relative frequency of DNA-binding residues in different range of B-factor values on APO-83.

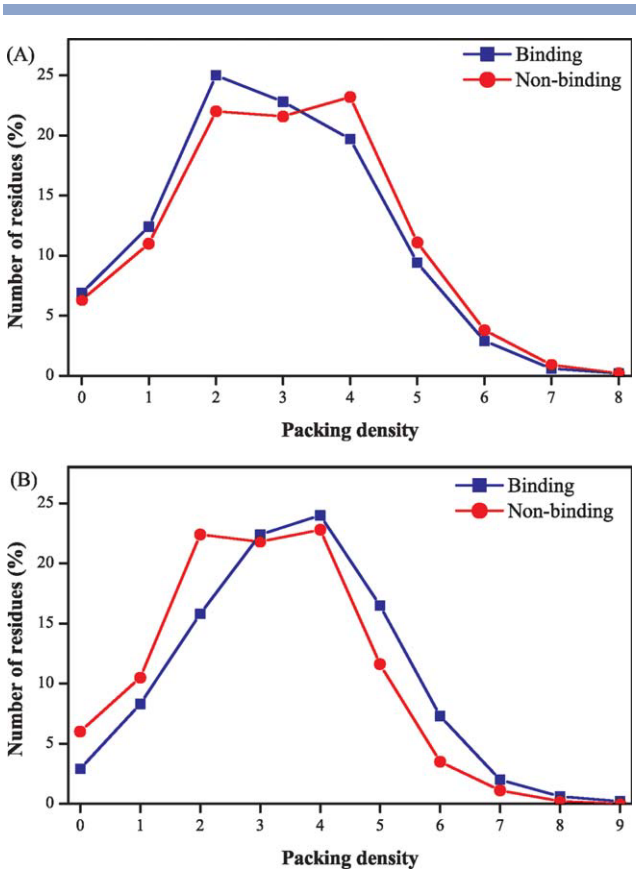

Figure 4
A: The relative frequency of DNA-binding and nonbinding residues in different range of packing density values on APO-83. The packing density was computed from only the DNA-free proteins. **B:** The relative frequency of DNA-binding and nonbinding residues in different range of packing density values on HOLO-83. The packing density was derived from both the protein and DNA sides.

Table II
Performance Measures of the Independent Tests with a Different Combination of Features on HOLO-83 and APO-83

Feature	Dataset	SN (%)	SP (%)	PR (%)	F_1	AUC
PSSM+pK_a+RSA+PD+B-factor	HOLO-83	69.9 ± 18.0[a]	80.1 ± 11.6	43.1 ± 20.0	0.507 ± 0.182	0.820
	APO-83	61.6 ± 21.1	79.5 ± 11.9	38.1 ± 18.3	0.440 ± 0.172	0.782
PSSM+pK_a+RSA+PD	HOLO-83	68.9 ± 18.5	77.9 ± 13.8	41.1 ± 20.0	0.483 ± 0.177	0.807
	APO-83	70.1 ± 18.0	77.8 ± 15.1	41.1 ± 19.8	0.482 ± 0.174	0.810
PSSM+pK_a+RSA	HOLO-83	69.1 ± 18.5	77.8 ± 13.3	41.1 ± 20.7	0.485 ± 0.184	0.807
	APO-83	69.1 ± 18.4	78.0 ± 14.2	40.8 ± 19.8	0.476 ± 0.173	0.807
PSSM+pK_a+PD	HOLO-83	69.5 ± 17.4	77.1 ± 13.5	40.4 ± 20.1	0.479 ± 0.176	0.805
	APO-83	68.7 ± 17.9	77.7 ± 13.6	39.8 ± 19.4	0.471 ± 0.174	0.808
PSSM+RSA+PD	HOLO-83	69.2 ± 19.4	77.3 ± 14.1	40.6 ± 20.0	0.481 ± 0.179	0.802
	APO-83	68.7 ± 19.2	77.1 ± 14.8	39.4 ± 19.1	0.468 ± 0.174	0.804

[a]The data is standard deviation for 83 protein chains.

significantly higher in HOLO-83 than nonbinding sites when it was derived from both the protein and DNA sides. The result was similar to the previous study,[25] in which it was demonstrated that the packing density of clustered-conserved residues in protein-DNA interfaces was significantly higher than the rest of interfaces.

Independent test on HOLO-83 and APO-83

In the section, we study our classifier performance on the independent test, which is mimicking a true prediction since the model trained on one dataset is really tested on an unseen dataset. To begin with, the predictions were made for the independent test using the best model trained on DBP-123. As shown in Table II, the model trained on all five features still performed well on the independent test set HOLO-83, although there was a slight decrease of the performance compared to that in five-fold cross validation (i.e., the AUC value decreased from 0.836 to 0.820). However, when the model incorporating the B-factor feature was tested on APO-83, the prediction performance was definitely poorer (the AUC was only 0.782). This result is not strange, since in previous section our analysis showed that the B-factor distribution of binding sites in DNA-free proteins was not similar with that in DNA-bound proteins.

One of the aims of our current work is to present an accurate predictor of DNA-binding residues not only in bound but also in unbound proteins. Therefore, the SVM model integrating all features but B-factor was selected as the final model in our method. As Table II displays, the final model can perform equally well on the unbound structures of APO-83 as did on the bound structures of HOLO-83.

We further examined the model performance change if the packing density was removed. It is clear that the packing density do add the value to the final model, especially tested on the DNA-free proteins in APO-83. The other feature (RSA or pK_a) was also individually removed, resulting in the decrease of performance, which was comparable to that of the packing density feature.

Comparison with DISPLAR

Performance comparisons among different DNA-binding residues prediction approaches are scientifically meaningful only if they use the same definition of the surface and interface residues and test on the same dataset. Accordingly, our approach was compared with the established method DISPLAR web server[5] that predicts DNA-binding residues based on protein structures. In DISPLAR, a different cutoff distance of 5 Å was used to define a binding residue. For a fair comparison, we

Table III
Performance Comparison Between Our Method and DISPALR on Different Test Sets

Dataset	Method	SN (%)	SP (%)	PR (%)	F_1
PDNA-62	DISPLAR[a]	63.1 ± 30.3[b]	78.7 ± 22.2	60.3 ± 25.9	0.575 ± 0.237
	Ours (0.758[c])	63.3 ± 26.7	87.2 ± 12.9	72.3 ± 19.6	0.638 ± 0.201
	Ours (0.574)	82.7 ± 18.4	72.7 ± 19.5	60.3 ± 17.7	0.676 ± 0.144
HOLO-83	DISPLAR	46.2 ± 28.9	90.5 ± 13.1	51.3 ± 29.6	0.451 ± 0.259
	Ours (0.659)	46.2 ± 21.1	89.6 ± 8.8	53.1 ± 25.3	0.454 ± 0.191
	Ours (0.636)	49.6 ± 20.5	87.9 ± 9.8	51.3 ± 24.5	0.467 ± 0.186
APO-83	DISPLAR	40.5 ± 30.0	87.7 ± 19.5	45.2 ± 32.0	0.391 ± 0.267
	Ours (0.700)	40.5 ± 21.8	92.1 ± 7.6	55.8 ± 25.1	0.419 ± 0.187
	Ours (0.532)	65.0 ± 19.7	80.6 ± 13.1	45.3 ± 20.7	0.494 ± 0.172

[a]The DISPALR web server was developed by Tjong and Zhou.[5]
[b]The data is standard deviation for all the protein chains in the testing set.
[c]The threshold of our SVM model for comparison at the equal or comparable sensitivity or precision of DISPLAR.

retrained our model on the DBP-123 dataset using the same cutoff distance of 5 Å in the definition of a binding residue. The comparison was carried out on the same set PDNA-62 (listed in the Supporting Information Table III), composed of DNA-binding protein structures in their holo forms. As shown in Table III, DISPLAR achieved a precision of 60.3% for the level of recall at 63.1%. This result is inferior to ours, i.e., for a recall of 63.2% we obtained precision more than 72%. Since some of the chains in PDNA-62 share more than 25% sequence identity with the training sets of DISPLAR and our methods, the reported performance in PDNA-62 should be checked with caution.

In addition to what these two methods had been done already with the PDNA-62 data set, a comparison of how DISPLAR performs on the data set HOLO-83 and APO-83 was also examined. From Table III, our method achieved a slight better performance than DISPLAR on the dataset HOLO-83, i.e., for the same recall 46.2%, our method had a slight higher precision of 53.1% compared to that of 51.3% in DISPLAR. In APO-83, our method showed a significant improvement of the performance of DISPLAR. On the same recall of 40.5%, our approach achieved about 10% higher than the precision obtained by DISPALR. It is implied that our method is superior in performance to DISPLAR in the prediction of DNA-binding residues on unbound protein structures. The result can be partially attributed to the following factors. Firstly, a well defined benchmark dataset was used to training our SVM model. Secondly, our method utilized the additional features of packing density and pK_a, which were demonstrated to improve the prediction performance in both DNA-bound and DNA-free protein structures. Finally, in our method a structural window based on nearest spatial surface neighbors was used instead of a structural window derived from closest spatial neighbors adopted by DISPALR (The details of comparison between these two types of spatial window were provided in Supporting Information Table V).

CONCLUSIONS

In this study, we made a systematic attempt to develop an accurate SVM model to identify DNA-binding residues on protein surface, given the unbound structure of a protein and the fact that it does bind to DNA. First, we conducted a statistical analysis of the characteristics of DNA-binding residues on the enlarged datasets of DNA-bound and DNA-free protein structures using the new features of B-factor and packing density and several conventional features including evolutionary profile, solvent accessibility and pK_a. Our analysis implies that the B-factor values vary on DNA-binding residues in DNA-free protein structures, whereas in DNA-bound proteins the average B-factor values of DNA-binding residues are lower than the rest of protein structure surface, reflecting the fact that DNA-binding regions become rigid upon bound to DNA molecules. We also note that the subtle difference of the packing density between DNA-binding and nonbinding residues can help identify DNA-binding sites on unbound structures. Then, we built a final SVM model incorporating PSSM with additional features (RSA, pK_a and RD) trained on DBP-123. The performance of our method is significantly better than existing methods such as DISPLAR on DNA-free protein structures. Therefore, our method will be useful in characterizing and identifying DNA-binding residues on protein structures, which are known to interact with the DNA molecules. In the next step, we make efforts to offer a web-based interface through which our approach will be available to biologists who are devoted to the studies of protein-DNA interactions.

ACKNOWLEDGMENTS

The authors thank Jun-Feng Xia for helpful discussion and encouragement and Tao Zeng for critical reading of the manuscript. They are grateful to the two anonymous reviewers for their insightful suggestions on enhancing the quality and rigor of this article.

REFERENCES

1. Luscombe NM, Austin SE, Berman HM, Thornton JM. An overview of the structures of protein-DNA complexes. Genome Biol 2000;1:REVIEWS001.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
3. Jones S, van Heyningen P, Berman HM, Thornton JM. Protein-DNA interactions: a structural analysis. J Mol Biol 1999;287:877–896.
4. Luscombe NM, Thornton JM. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J Mol Biol 2002;320:991–1009.
5. Tjong H, Zhou HX. DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. Nucleic Acids Res 2007; 35:1465–1477.
6. Wu JS, Liu HD, Duan XY, Ding Y, Wu HT, Bai YF, Sun X. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. Bioinformatics 2009;25:30–35.
7. Wang L, Yang MQ, Yang JY. Prediction of DNA-binding residues from protein sequence information using random forests. BMC Genomics 2009;10 (Suppl 1):S1.
8. Bhardwaj N, Lu H. Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions. FEBS Lett 2007;581:1058–1066.
9. Wang LJ, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 2006;34:W243–W248.
10. Kuznetsov IB, Gou ZK, Li R, Hwang SW. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins 2006;64:19–27.
11. Andrabi M, Mizuguchi K, Sarai A, Ahmad S. Prediction of mono- and di-nucleotide-specific DNA-binding sites in proteins using neural networks. BMC Struct Biol 2009;9:30.
12. Zen A, de Chiara C, Pastore A, Micheletti C. Using dynamics-based comparisons to predict nucleic acid binding sites in proteins:

an application to OB-fold domains. Bioinformatics 2009;25:1876–1883.
13. Ozbek P, Soner S, Erman B, Haliloglu T. DNABINDPROT: fluctuation-based predictor of DNA-binding residues within a network of interacting residues. Nucleic Acids Res 2010;38:W417–W423.
14. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 2004; 20:477–486.
15. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics 2005;6:33.
16. Ofran Y, Mysore V, Rost B. Prediction of DNA-binding residues from sequence. Bioinformatics 2007;23:I347–I353.
17. Hwang S, Gou ZK, Kuznetsov IB. DP-Bind: a Web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics 2007;23:634–636.
18. Cole C, Warwicker J. Side-chain conformational entropy at protein-protein interfaces. Protein Sci 2002;11:2860–2870.
19. Yuan Z, Zhao J, Wang ZX. Flexibility analysis of enzyme active sites by crystallographic temperature factors. Protein Eng 2003;16:109–114.
20. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol 2004;338:181–199.
21. Liu R, Jiang W, Zhou Y. Identifying protein-protein interaction sites in transient complexes with temperature factor, sequence profile and accessible surface area. Amino Acids 2010;38:263–270.
22. Baud F, Karlin S. Measures of residue density in protein structures. Proc Natl Acad Sci USA 1999;96:12494–12499.
23. Cho KI, Kim D, Lee D. A feature-based approach to modeling protein-protein interaction hot spots. Nucleic Acids Res 2009;37: 2672–2687.
24. Keskin O, Ma B, Nussinov R. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. J Mol Biol 2005;345:1281–1294.
25. Ahmad S, Keskin O, Sarai A, Nussinov R. Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. Nucleic Acids Res 2008; 36: 5922–5932.
26. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
27. Miller S, Lesk AM, Janin J, Chothia C. The accessible surface area and stability of oligomeric proteins. Nature 1987;328:834–836.
28. Hubbard SJ, Thornton JM. NACCESS. Department of Biochemistry and Molecular Biology, University College London 1993.
29. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389–3402.
30. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
31. Drenth J. Principles of protein X-ray crystallography. Springer Verlag, New York; 1999.
32. Lehninger A, Nelson D, Cox M. Lehninger principles of biochemistry. WH Freeman, New York; 2004.
33. Cortes C, Vapnik V. Support-vector networks. Machine learning 1995;20:273–297.
34. El-Manzalawy Y, Honavar V. WLSVM: integrating LibSVM into Weka environment. Software available at: http://wwwcsiastateedu/~yasser/wlsvm/ 2005.
35. Chang C, Lin C. LIBSVM: a library for support vector machines. Software available at: http://wwwcsientuedutw/~cjlin/libsvm/ 2001.
36. Jones S, Thornton JM. Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol 1995;63:31–65.
37. Luque I, Freire E. Structural stability of binding sites: consequences for binding affinity and allosteric effects. Proteins 2000;Suppl 4:63–71.