BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 6 2013, pages 678–685
doi:10.1093/bioinformatics/btt029

Sequence analysis

Advance Access publication January 17, 2013

PreDNA: accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information

Tao Li[1], Qian-Zhong Li[1,*], Shuai Liu[2], Guo-Liang Fan[1], Yong-Chun Zuo[1,3] and Yong Peng[1]

[1]Laboratory of Theoretical Biophysics, School of Physical Sciences and Technology, [2]College of Computer Science and [3]The National Research Center for Animal Transgenic Biotechnology, Inner Mongolia University, Hohhot, 010021, China

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Protein–DNA interactions often take part in various crucial processes, which are essential for cellular function. The identification of DNA-binding sites in proteins is important for understanding the molecular mechanisms of protein–DNA interaction. Thus, we have developed an improved method to predict DNA-binding sites by integrating structural alignment algorithm and support vector machine–based methods.

Results: Evaluated on a new non-redundant protein set with 224 chains, the method has 80.7% sensitivity and 82.9% specificity in the 5-fold cross-validation test. In addition, it predicts DNA-binding sites with 85.1% sensitivity and 85.3% specificity when tested on a dataset with 62 protein–DNA complexes. Compared with a recently published method, BindN+, our method predicts DNA-binding sites with a 7% better area under the receiver operating characteristic curve value when tested on the same dataset. Many important problems in cell biology require the dense non-linear interactions between functional modules be considered. Thus, our prediction method will be useful in detecting such complex interactions.

Availability: The PreDNA webserver is freely available at: http://202.207.14.178/predna/index.aspx

Contact: qzli@imu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online

Received on October 22, 2012; revised on January 9, 2013; accepted on January 12, 2013

1 INTRODUCTION

Protein–DNA interactions play essential roles in a variety of vital biological processes, such as DNA replication, transcription, packaging (Luscombe *et al.*, 2000). For example, the interactions between DNA and transcription factors are important for modulating the process of gene replication and transcription (Ptashne *et al.*, 2005). The interactions of DNA and histones are involved in chromosome packaging in the cell nucleus (Kornberg, 1974). Moreover, Bullock and Fersht (2001) have shown that mutations of DNA-binding residues, such as those on the tumor repressor protein P53, may predispose individuals to cancer. Thus, the accurate identification of DNA-binding sites in proteins is not only important for understanding the mechanisms of these biological processes but also helpful for annotating the function of proteins. Protein–DNA interaction sites can be detected by using biochemical techniques, such as DNase I foot-printing (Brenowitz *et al.*, 1986), Gel retardation (Lane *et al.*, 1992) *in vivo* foot-printing (Dumitru and McNeil, 1994) *in vitro* binding assay (Diekmann and Hall, 1995) and methylation (Baldwin *et al.*, 2001). However, traditional biochemical methods are time-consuming and laborious to carry out for the rapidly increasing number of protein–DNA complexes. Therefore, developing an objective and effective computational approach to precisely predict DNA-binding sites would be necessary.

Several computational methods have been developed to identify DNA-binding sites in proteins based mainly on protein sequence or protein structure. The sequence-based method can be further divided into two classes. The first class is based on the following: side chain pKa value, hydrophobicity index and molecular mass of the amino acid residues and DNA-binding residues predicted by support vector machine (SVM) (Wang and Brown, 2006; Wang *et al.*, 2010) and random forests classifier (Wang *et al.*, 2009). The second class uses evolutionary and other sequence information to identify DNA-binding sites in proteins (Ahmad and Sarai, 2005; Huang *et al.*, 2009; Hwang *et al.*, 2007; Kuznetsov *et al.*, 2006; Ofran *et al.*, 2007; Wu *et al.*, 2009). Structure-based methods are further classified into a structural alignment-based method and a structural alignment-free method. For the structural alignment-based method, the DNA-binding sites are recognized from a structural alignment between the query sequence and a template known to bind DNA (Gao and Skolnick, 2008; Holm and Sander, 1993). In the structural alignment-free method, DNA-binding sites are predicted using electrostatic potentials and the shape of the molecular surface (Jones *et al.*, 2003; Tsuchiya *et al.*, 2004, 2005).

With the rapid increase in the number of high-quality protein structures in the Protein Data Bank (PDB), developing new and efficient techniques for predicting DNA-binding sites using geometric structures has become feasible. In this article, we introduce a two-stage machine-learning strategy to predict DNA-binding sites. In the first stage, DNA-binding sites are predicted by a structural alignment algorithm using the geometric structure. In the second stage, DNA-binding sites are predicted by SVM using evolutionary information, torsion angles (φ, ψ) present in the backbone structure and solvent accessibility. The final results are obtained by a consensus of SVM predictions and geometric structure-based predictions.

*To whom correspondence should be addressed.

2 MATERIALS AND METHODS

2.1 Dataset preparation

In this study, two datasets, PDNA-62 and PDNA-224, were used to evaluate the performance of our method. A summary of these datasets are shown as follows:

PDNA-62: This is a non-redundant database of representative protein–DNA complexes from the PDB (http://www.rcsb.org/pdb/), which was constructed by Ahmad and Sarai (2005) and used by several other studies (Kuznetsov *et al.*, 2006; Wang and Brown, 2006; Wang *et al.*, 2009, 2010). The sequence identity in the dataset was ≤25% and the resolution of the structures was 2.5 Å or better. Based on the cutoff distance of 3.5 Å, the PDNA-62 dataset contained 1215 DNA-binding residues and 6948 non-binding residues (Ahmad and Sarai, 2005; Ahmad *et al.*, 2004).

PDNA-224: In addition to the latest PDB (released by January 25, 2011), a new DNA-binding protein dataset was constructed from 978 protein–DNA complexes in this work. The 978 protein–DNA complexes were determined by radiographic crystallography with a resolution better than 3.0 Å. Redundancy among the amino acid sequences was removed using the PISCES software (Wang and Dunbrack, 2003) with a threshold of 25% set for sequence identity. Any proteins homologous to those in the PDNA-62 were also removed by the PISCES software. Finally, 224 non-redundant protein chains containing 57 348 amino acids were obtained. According to an identical criterion suggested in previous studies, a residue is regarded as interacting with DNA if the distance between an atom of the residue and an atom of base is <3.5 Å (Ahmad and Sarai, 2005; Ahmad *et al.*, 2004). Using this criterion, 3778 interacting residues and 53 570 non-interacting residues were projected to be present in the PDNA-224 dataset. The PDNA-224 dataset and the binding sites information are listed in the Supplementary Data S1.

2.2 Methods for prediction

Our method consisted of a SVM predictor and a geometric structure-based predictor. For the SVM predictor, a SVM decision value (*sdv*) for each site in a protein is obtained by SVM using the evolutionary information, solvent accessible surface area and the protein backbone structure (PBS). For the geometric structure-based predictor, another geometric decision value (*gdv*) for each protein site is obtained by the structural alignment method using the geometric structure information. The final result for each site is obtained by combining the two decision values of SVM-based predictor and geometric structure-based predictor. The whole prediction procedure is illustrated in Figure 1.

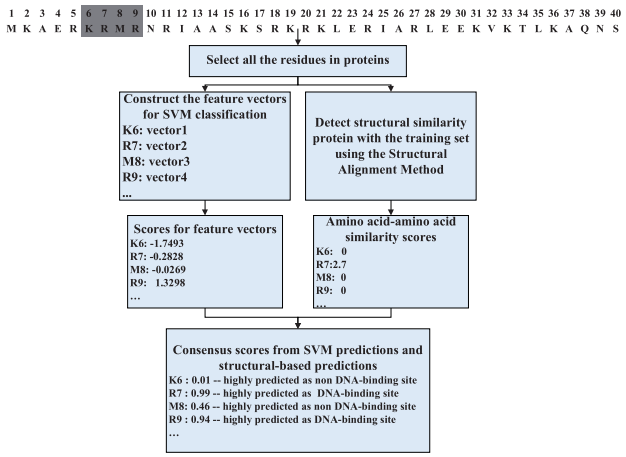

Fig. 1. Overall framework of DNA-binding sites prediction

2.3 SVM-based predictor

Owing to the fact that the DNA-binding site is determined not only by the central residue, but also by other residues present in its neighboring sites, a sliding window with K residues is selected to represent the information of the central site. The information of each residue in the sliding window is constructed using evolutionary information, the torsion angles (φ, ψ) in the PBS and the solvent accessible surface (Li and Li, 2012). These features and the encoding scheme are described in Part 1 of the Supplementary Data S2. Then, the encoded features are selected as the input parameters of the SVM, which is a popular machine-learning approach mainly used in pattern recognition and orientation (Cai *et al.*, 2002, 2004; Kuznetsov *et al.*, 2006; Shu *et al.*, 2008). In this work, the publicly available LibSVM software (version 3.0) was used (http://www.csie.nut.edu.tw/~cjlin/libsvm) (Chang and Lin, 2011). The SVM outputs (SVM decision value) have been converted into conditional probabilities using a sigmoid function. The probability interpretation of $sdv(x)$ is used for evaluating the performance of the SVM predictor. The sigmoid function is defined as follows:

$$P(Y = 1|x) = 1/(1 + \exp(A \times sdv(x) + B)) \quad (1)$$

where x is the SVM input of each feature for position x, $sdv(x)$ denotes the decision value of the test feature for position x, $P(Y=1|x)$ is the probability of DNA-binding prediction, and A and B are the slope and offset, respectively, to be learned by a 3-fold cross-validation method suggested in Platt (2000) (see Part 2 in Supplementary Data S2 for more details) from the training set for the sigmoid function. Empirically, one could use $A=-2.0$, $B=-0.5$ for DNA-binding prediction.

To combine linearly the scores of SVM predictions with the scores of geometric structure-based predictions (see section 2.5 for more detail), we converted conditional probabilities back to SVM decision value by,

$$sdv(x) = \{\ln[(1 - P)/P] - B\}/A \quad (2)$$

where $sdv(x)$ is the decision value of the test feature for position x, P is the conditional probability for DNA-binding and A and B are the same as used in the sigmoid function. Equation (2) is an inverse of Equation (1).

2.4 Geometric structure-based predictors

Generally speaking, if a protein has a geometric structure similar to a DNA-binding protein, the probability that this protein is also DNA-binding protein is much higher. In our study, we use a structural alignment method (see Section 2.5 for more details) to detect structural similarities between a query protein and a template known to bind DNA. The quality of the structural alignment method is determined by an interface alignment score (IAS), which provides a quantitative measure of the similarity between two protein–DNA complexes in the geometry. Therefore, for a complex in the test set, a protein–DNA complex in the training set with the highest IAS is kept, including the PDB code and the DNA-binding information. This information is used to calculate the decision values for all matched residues of the target chain according to the following strategies: If the matched residue in the detected chain is a DNA-binding site, a positive amino acid–amino acid similarity score is assigned to the site in the target chain. On the other hand, if the matched residue in the detected chain is a non-DNA–binding site, a zero score is assigned to the site in the target chain.

2.5 Structural alignment method for detecting structural similarity

Structural similarities between two protein–DNA complexes were calculated using the structural alignment method described in a previous study (Siggers *et al.*, 2005).

Step 1: Construct coordinate system

Based on this method, local coordinate systems are constructed on both the amino acid residue and the nucleic acid (see Fig. 2A).

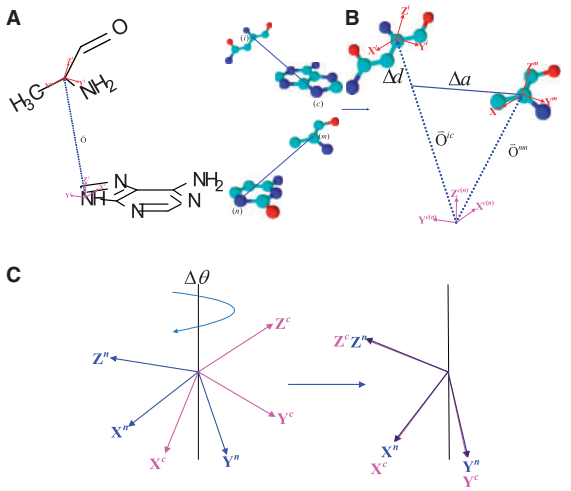

Fig. 2. Description of the geometric parameters used to compare two amino acid–nucleotide pairs. (**A**) Orthonormal coordinate systems constructed on the amino acid backbone (*i*) and DNA base (*m*). The vector $\vec{O}$ connects the origin of the base system to the origin of the amino acid system. The three coordinate vectors of the amino acid frame and the $\vec{O}$ vector are transformed into the base reference frame. These four vectors describe the relative spatial orientation of the amino acid backbone with respect to the base. (**B**) Geometric parameters Δd and $\Delta\alpha$ for (*ic*) and (*mn*) amino acid–nucleotide pairs. Amino acid–nucleotide-pair geometric relationships are described in the base reference frame. In the base reference frame, the radial displacement Δd is the difference between the magnitude of the $\vec{O}^{ic}$ and $\vec{O}^{mn}$ vectors: $\Delta d\ \{(i,c)(m,n)\} = \text{abs}(||\vec{O}^{ic}|| - ||\vec{O}^{mn}||)$. The angle displacement $\Delta\alpha$, which is analogous to an arc length, is represented by the base of an isosceles triangle where the shorter of the two $\vec{O}$ vectors determines the length of the two equivalent sides. (**C**) Geometric parameter $\Delta\theta$ for (*ic*) and (*mn*) amino acid–nucleotide pairs. $\Delta\theta$ (rotation angle) is the acute angle required to superimpose the two transformed amino acid coordinate frame when their origins are coincident

According to the local coordinate systems, the vector ($\vec{O}$) connecting the nucleotide–amino acid pair (base to amino acid) is determined as set out in Figure 2A. The magnitude of the $\vec{O}$ vector for all nucleotide–amino acid pairs is calculated in our datasets. As in the previous study (Siggers *et al.*, 2005), a pair is regarded as 'contacting' if it had a $\vec{O}$ vector whose magnitude was <16 Å. Therefore, this value is chosen as the distance cutoff used to define an 'interfacial' amino acid–nucleotide pair when determining amino acid–amino acid similarity scores.

Step 2: Calculating amino acid–nucleotide geometric similarity score

Based on the coordinate system, the indices *i* and *c* are introduced to represent an amino acid and nucleotide subunit, and *m* and *n* for the other, respectively. By comparing the geometric similarity between two amino acid–nucleotide subunits *ic* and *mn*, the geometric parameters such as the rotation angle and the distance displacement are obtained. The rotation angle is in the form of $\Delta\theta\ \{(i,c)(m,n)\}$. The distance displacement contains a radial displacement, $\Delta d\ \{(i,c)(m,n)\}$, and an angular displacement, $\Delta\alpha\ \{(i,c)(m,n)\}$. The geometric parameters are described in Figure 2. Because the geometric parameters $\Delta d\ \{(i,c)(m,n)\}$, $\Delta\alpha\ \{(i,c)(m,n)\}$ and $\Delta\theta\ \{(i,c)(m,n)\}$ are discrete distributions, these parameters are transformed into corresponding geometric parameters $s^d\{(i,c)(m,n)\}$, $s^\alpha\{(i,c)(m,n)\}$ and $s^\theta\{(i,c)(m,n)\}$ so that the transformed geometric parameters are continuous distributions in the range of ~0–1. In this article, $s^d\{(i,c)(m,n)\}$, $s^\alpha\{(i,c)(m,n)\}$ and $s^\theta\{(i,c)(m,n)\}$ are calculated by using the following logistic function:

$$s^d\{(i,c)(m,n)\} = \exp(A^d(B^d - \Delta d))/(1+\exp(A^d(B^d - \Delta d))) \quad (3)$$

$$s^a\{(i,c)(m,n)\} = \exp(A^a(B^a - \Delta a))/(1+\exp(A^a(B^a - \Delta a))) \quad (4)$$

$$s^\theta\{(i,c)(m,n)\} = \exp(A^\theta(B^\theta - \Delta\theta))/(1+\exp(A^\theta(B^\theta - \Delta\theta))) \quad (5)$$

where A^d, B^d, A^α, B^α, A^θ and B^θ are the constant coefficients, respectively, to be obtained by the method suggested in Siggers *et al.* (2005) (see Part 3 in Supplementary Data S2 for more details). Therefore, the constant coefficients used here are 25.0 (A^d) and 0.5 (B^d) for $s^d\{(i,c)(m,n)\}$, 5.0 (A^α) and 2.6 (B^α) for $s^\alpha\{(i,c)(m,n)\}$ and 0.45 (A^θ) and 25.0 (B^θ) for $s^\theta\{(i,c)(m,n)\}$. The final stage in this step is to calculate the amino acid–nucleotide geometric similarity score; the amino acid–nucleotide geometric similarity score can be written in the form,

$$s\{(i,c)(m,n)\} = \sqrt[3]{s^d\{(i,c)(m,n)\}\, s^a\{(i,c)(m,n)\}\, s^\theta\{(i,c)(m,n)\}} \quad (6)$$

where $s^d\{(i,c)(m,n)\}$, $s^\alpha\{(i,c)(m,n)\}$ and $s^\theta\{(i,c)(m,n)\}$ are the transformed geometric parameters.

Step 3: Calculating amino acid–amino acid similarity score

The amino acid–amino acid similarity score $S\ \{i,m\}$ is the average of the geometric similarity scores for all successful matches,

$$S\{i,m\} = \sum\nolimits_{c,n} s_{successful}\{(i,c)(m,n)\}/N \quad (7)$$

where N is the number of all successful matches.

Step 4: Calculating the Protein–DNA interface alignment score

Based on an amino acid–amino acid similarity matrix with elements $S\ \{i,m\}$, the protein–DNA IAS is determined by summing the $S\ \{i,m\}$ scores over all aligned residues,

$$\text{IAS} = \sum S\{i,m\} \quad (8)$$

where $S\ \{i,m\}$ is the amino acid–amino acid similarity score, and IAS is the Protein–DNA IAS.

Therefore, for a test instance, a complex in the training set with the highest IAS is kept, including the PDB code and the DNA-binding information. The amino acid–amino acid similarity score is used to calculate the *gdv* for all matched residues of the test instance. However, there are two important differences in the usefulness of the amino acid–amino acid similarity score: (i) In our study, the amino acid–amino acid similarity scores are converted into the average amino acid–amino acid similarity score using the normalized function suggested in a previous study (Chang and Lin, 2011) in order that the average amino acid–amino acid similarity score remains the same as the decision value derived from SVM predictions. (ii) Not all the amino acid–amino acid similarity scores are used in our study. The amino acid–amino acid similarity scores in which the matched residues in the detected chain are non-DNA–binding sites are not included in this study. The remaining amino acid–amino acid similarity scores are regarded as the *gdv*. The scheme is defined as follows:

$$gdv(x,m) = \begin{cases} 0 & \text{if the matched residue in the alignment chain is a nonDNA} - \text{binding} \\ S\{x,m\} & \text{if the matched residue in the alignment chain is a DNA} - \text{binding} \end{cases} \quad (9)$$

where $S\ \{x,m\}$ is the average amino acid–amino acid similarity score for position x and $gdv(x,m)$ is the gdv for position x.

2.6 The combination of the decision values from the SVM and geometry-based predictor

To predict the DNA-binding sites accurately, we used a novel scheme by combining the SVM decision value with the gdv. The scheme is based on the expected value in probability theory; suppose a random variable X can take value x_1 with probability p_1, value x_2 with probability p_2, and so on, up to value x_k with probability p_k. Then the expectation of this random variable X is defined as follows,

$$E(X) = x_1p_1 + x_2p_2 + ... + x_kp_k \quad (10)$$

where $p_1 + p_2 + \ldots + p_k = 1$.

Therefore, in this section, the decision values from the SVM and the geometry-based predictor are combined by a combination scheme as follows:

$$com(x) = (1.0 - p) \times sdv(x) + p \times gdv(x) \quad (x = 1, 2...L) \quad (11)$$

where $sdv(x)$ is the SVM decision value for the position x, $gdv(x)$ is the gdv for the position x, $com(x)$ is the combined decision value for the position x, L is the length of the protein sequence and p is the proportion factor, $p = 0.45$ is used by learning the method from the training set (See Part 4 in Supplementary Data S2 for more details).

Then, the $com(x)$ elements are normalized to fall with the range of O ~1 by the sigmoid function, which is analogous to equation (1) as shown below,

$$pcom(Y = 1|x) = 1/(1 + \exp(A \times com(x) + B)) \quad (12)$$

where $com(x)$ denotes the combined decision value for the position x, $pcom(Y = 1|x)$ is the probability interpretation of the $com(x)$, and A and B are the slope and offset, respectively. In this article, we use $A = -6.0$, $B = -0.2$ for the final prediction.

The final prediction result is obtained using the probability interpretation of the $com(x)$. Given a protein sequence X with l residues, we obtain an l-dimensional feature vector [$pcom(1)$, $pcom(2)$,. . ., $pcom(l)$]. Then, we use a threshold ξ_0 to integrate the vector into a non-linear discriminant function. The threshold ξ_0 gives the decision for sample X. As $pcom(x) > \xi_0$, the sample X is classified into the positive group, or into the negative group as $pcom(x) \leq \xi_0$. Empirically, one could use $\xi_0 = 0.52$ for DNA-binding prediction (See Part 5 in Supplementary Data S2 for more details).

3 RESULTS AND DISCUSSION

In predicting DNA-binding sites, the 5-fold cross-validation test is often used to examine the effectiveness of a predictor (Wang and Brown, 2006; Wang *et al.*, 2009, 2010; Wu *et al.*, 2009). The performance of our predictor was also assessed by the 5-fold cross-validation test. During this test, a dataset is randomly divided into five non-overlapping sets, four of which are used for training the predictor and the accuracy of the predictor is assessed on the remaining sets. This process is repeated five times. The predictive capability of our method was evaluated by the sensitivity (Sn), specificity (Sp), Matthew's correlation coefficient (MCC), overall prediction accuracy (Acc), strength (Str) and false-positive rate (FPR):

$$Sn = TP/(TP + FN) \quad (13)$$

$$Sp = TN/(TN + FP) \quad (14)$$

$$Strength = (Sn + Sp)/2 \quad (15)$$

$$FPR = FP/(TN + FP) = 1 - Sp \quad (16)$$

$$Acc = (TP + TN)/(TP + FP + TN + FN) \quad (17)$$

$$MCC = (TP \times TN - FP \times FN) \Big/ \sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)} \quad (18)$$

where TP represents the number of correctly identified positives, FN represents the number of the positives identified as negatives, TN represents the number of the correctly identified negatives and FP represents the number of negatives identified as positives.

To further evaluate the capability of using our method on an unbalanced dataset, the receiver operating characteristic (ROC) curve (Swets, 1988), one of the most robust approaches for classifier evaluation, was also used. It is obtained by plotting the true-positive rate (sensitivity) on the y-axis against the FPR (1-specificity) on the x-axis. The area under the ROC curve (AUROC) (Bradley, 1997) can be used as a reliable measure for the prediction performance. The maximum value of AUROC (1) denotes a perfect prediction. A random guess receives an AUROC value close to 0.5.

3.1 Effect of the sliding window size in the construction of the SVM input vectors

To obtain the best performance for predicting DNA-binding sites, the sliding window size is optimized with respect to the overall accuracy, MCC and Strength values. The optimized sliding window size is obtained by testing the performance of different sliding window sizes from 3 to 21 with the default parameters c, g and weight in SVM. Table 1 shows the results of different sliding window sizes on the dataset PDNA-62. It was found that a sliding window of 11 amino acids achieves the best predicting performance. Therefore, the optimized sliding window size was set to 11 in this study.

3.2 The predicted results on the PDNA-62 dataset

Table 2 shows the predicted results for the PDNA-62 dataset with various combinations of features. We show that the lowest overall prediction accuracy is 62.6% when only using the *PBS* as the input parameter. By using the *PSSM* profiles as the input vector, the accuracy of the SVM classifier is significantly improved up to 79.4%. However, the prediction performance slightly increases when the *PSSM* profiles were combined with *PBS* or normalized solvent accessible surface area (*NSASA*) are used as the input vectors. The best results (Acc of 83.6%, MCC of 0.50 and Strength of 82.2%) were obtained when all the

Table 1. The test results for the PDNA-62 dataset with respect to different window sizes based on the 5-fold cross-validation test

K	c	g	Weight	Sn (%)	Sp (%)	Acc (%)	MCC	Str (%)
3	4.0	0.0078125	7.04	76.4	76.6	76.6	0.38	76.5
5	1.0	0.0078125	7.04	77.2	77.9	77.8	0.40	77.5
7	0.5	0.0078125	7.04	76.8	78.6	78.4	0.40	77.7
9	0.25	0.0078125	7.04	76.8	78.3	78.2	0.40	77.6
11	**0.25**	**0.0078125**	**7.04**	**76.8**	**79.7**	**79.4**	**0.42**	**78.3**
13	0.125	0.0078125	7.04	76.3	78.3	78.1	0.40	77.3
15	0.25	0.003906	7.04	76.9	78.2	78.1	0.40	77.6
17	0.25	0.003906	7.04	76.3	79.4	79.1	0.41	77.9
19	0.25	0.003906	7.04	75.4	80.3	79.7	0.41	77.9
21	0.25	0.003906	7.04	74.2	80.8	80.0	0.41	77.5

The values in bold indicate that they are the best values.

Table 2. The prediction performances for the PDNA-62 dataset based on various features in the 5-fold cross-validation test

Feature vector	c	g	Weight	Sn (%)	Sp (%)	Acc (%)	MCC	Str (%)
NPSSM	0.25	0.0078125	7.04	76.8	79.7	79.4	0.42	78.3
NSASA	0.25	0.5	7.04	68.3	67.5	67.6	0.24	67.9
PBS	0.25	0.015625	7.04	57.4	63.7	62.6	0.14	60.4
NSASA + PBS	0.25	0.015625	7.04	76.4	63.3	64.9	0.26	69.8
NPSSM + NSASA	0.25	0.0078125	7.04	79.7	81.8	81.6	0.46	80.7
NPSSM + PBS	0.25	0.0078125	7.04	77.3	82.6	82.0	0.46	80.0
NPSSM + NSASA + PBS	**0.25**	**0.0078125**	**7.04**	**80.2**	**84.1**	**83.6**	**0.50**	**82.2**

NPSSM, normalized PSSM score; *NSASA*, normalized solvent accessible surface area; *PBS*, protein backbone structure. The values in bold indicate that they are the best values.

features, including *PSSM* profiles, *PBS* and *NSASA* were combined. Therefore, in our work, the SVM-based prediction model was constructed by combining all the features. By integrating SVM with geometric structure-based predictor, we obtained a final ROC curve for the final prediction shown in Figure 3. It is clear that the AUROC values in our predictions are ~5–10% higher than those obtained in other studies. The best values for Acc, MCC, strength and AUROC were 85.5%, 0.55, 85.3% and 0.926, respectively, in our study.

3.3 The predicted results on the new PDNA-224 dataset

To further evaluate the performance of our method in predicting DNA-binding sites, we applied it to a new dataset PDNA-224 generated in this work. The test results of the SVM models using various features are listed in Table 3. Our results show that the prediction performance increases significantly while using the *PSSM* profiles as the parameter of the SVM model. When the *PSSM* profile was combined with one or more additional features as input vectors, the performances were slightly improved. The best performance was achieved with an MCC of 0.35, overall accuracy of 81.8% and strength of 79.2% (with a sensitivity of 76.1% and specificity of 82.2%) by combining the *PSSM* profiles, *PBS* and *NSASA*. In addition, the ROC curves of predicting DNA-binding sites for the PDNA-224 dataset (Fig. 4) was obtained by combining the SVM predictor using different parameters with the geometric structure-based predictor. The AUROC values show that the combined results are better than the results of the SVM models using various features. The best AUROC value was 0.898 in the 5-fold cross-validation test.

3.4 Comparison with other computational methods

DNA-binding sites have been predicted successfully using predictors, such as Dps-pred (Ahmad *et al.*, 2004), Dbs-pssm (Ahmad and Sarai, 2005), BindN (Wang and Brown, 2006), Dp-bind (Kuznetsov *et al.*, 2006), DP-Bind (Hwang *et al.*, 2007), BindN-RF (Wang *et al.*, 2009) and BindN +(Wang *et al.*, 2010). When we compared our method with prior methods using the common dataset, PDNA-62, we obtained the results shown in Table 4. Overall, the best results of our prediction were 85.2% (85.2% sensitivity and 85.3% specificity). The results show that our predictor has a considerably higher sensitivity of 85.2% compared with 40.3% from a previous study (Ahmad *et al.*, 2004). Both the sensitivity and specificity of our predictor were also better than those of other predictors shown in Table 4. Moreover, our method is benchmarked with one of the most successful DNA-binding sites predictors, BindN+. Our method improves the sensitivity, specificity, accuracy, strength, MCC and AUROC values from 77.3%, 79.3%, 79.0%, 78.3%,

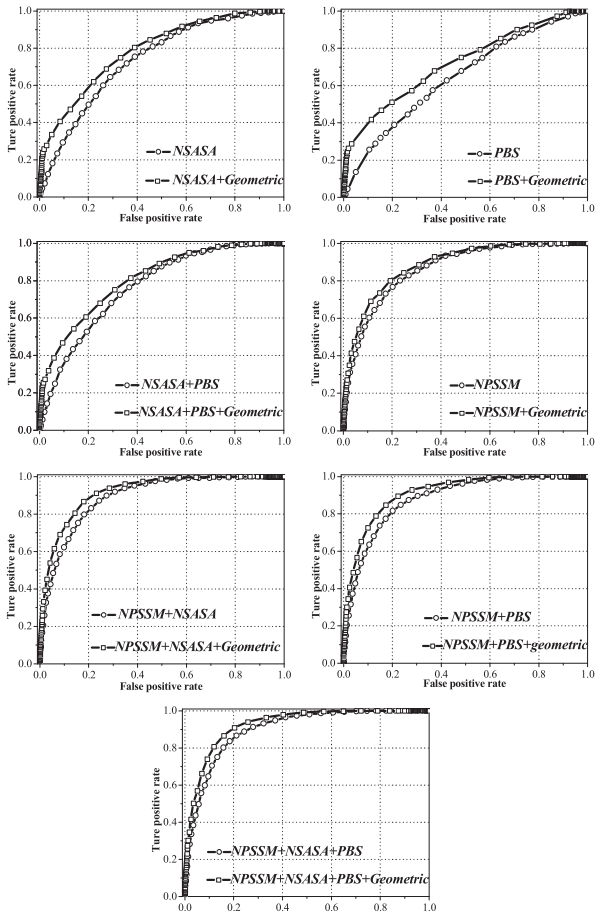

Fig. 3. ROC curves for the DNA-binding sites prediction in PDNA-62 dataset by combining SVM predictor using different parameters (such as *NSASA*, *PBS*, *NPSSM*, *NSASA + PBS*, *NPSSM + NSASA*, *NPSSM + PBS*, *NPSSM + NSASA + PBS*) with geometric structure-based predictor

Table 3. The prediction performances for the PDNA-224 dataset based on various features in the 5-fold cross-validation test

Feature vector	c	g	Weight	Sn (%)	Sp (%)	Acc (%)	MCC	Str (%)
NPSSM	0.25	0.0078125	14.18	69.5	79.8	79.1	0.29	74.6
NSASA	0.125	0.25	14.18	80.6	55.3	56.9	0.18	67.9
PBS	0.125	0.03125	14.18	58.2	62.9	62.6	0.11	60.6
NSASA + PBS	0.125	0.03125	14.18	76.3	62.5	63.4	0.20	69.4
NPSSM + NSASA	0.125	0.0078125	14.18	76.3	80.0	79.8	0.33	78.2
NPSSM + PBS	0.125	0.0078125	14.18	73.5	80.3	79.9	0.32	76.9
NPSSM + NSASA + PBS	**0.125**	**0.0078125**	**14.18**	**76.1**	**82.2**	**81.8**	**0.35**	**79.2**

NPSSM, normalized PSSM score; *NSASA*, normalized solvent accessible surface area, *PBS*: protein backbone structure. The values in bold indicate that they are the best values.

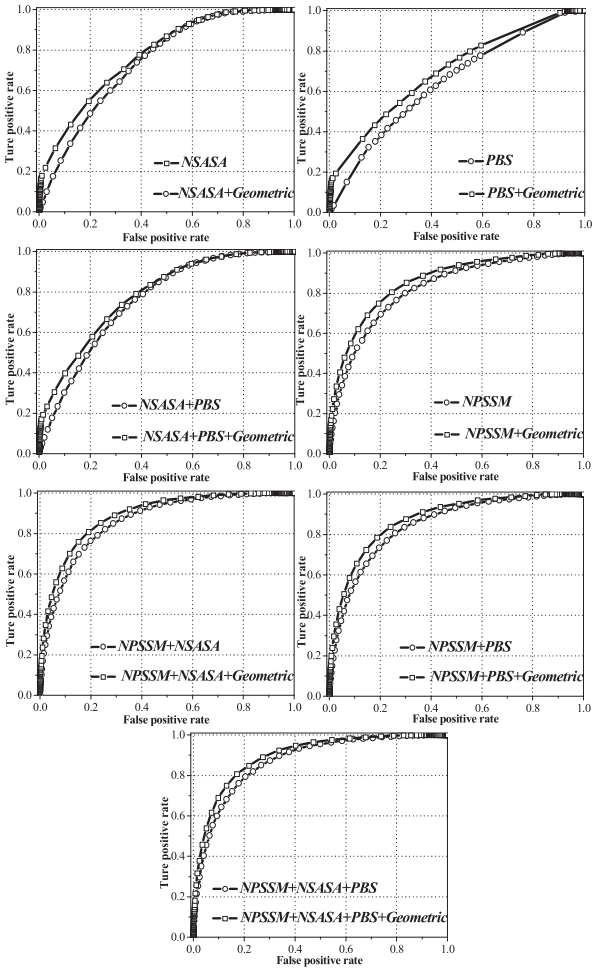

Fig. 4. ROC curves for the DNA-binding sites prediction in PDNA-224 dataset by combining SVM predictor using different parameters (such as *NSASA*, *PBS*, *NPSSM*, *NSASA + PBS*, *NPSSM + NSASA*, *NPSSM + PBS*, *NPSSM + NSASA + PBS*) with geometric structure-based predictor

0.44 and 0.859 to 85.2%, 85.3%, 85.2%, 85.2%, 0.55 and 0.926, respectively. The comparative results indicate that our method has good generalization abilities in recognizing DNA-binding sites.

3.5 False-positive sites predicted with high confidence

To further illustrate the predictive capability of our predictor, our results for the common PDNA-62 dataset were compared with those of the PDB. The results show that 197 residues that do not bind to DNA in the PDB annotation are predicted as DNA-binding sites with >90% confidence. However, based on deep and comprehensive analyses, the 197 residues can be classified into five groups. We used the pdb_id+chain_letter (e.g. 1B3TA) to name the protein chains.

In the first group, we found 27 false-positive residues with other related functions. Some of the residues are protein–protein interaction sites (Ala46 in 1A02J, Gly82 in 1B3TA, Asp50 in 1CF7A, Gln18 in 1FJLA, Glu128 in 1GDTA, Lys22 in 1HWTD, Gln44 in 1HWTD, His 32 in 1IHFB, Leu37 in 1IHFB, Cys93 in 1MNMA, Val63 in 1PYIA, Asn117 in 1TSRA, Thr118 in 1TSRA, Gln28 in 1YRNA), while others represent water-binding sites (Val99 in 1A02N, Gly158 in 1A74A, Thr9 in 1BHMA, Ala40 in 1CDWA, Trp65 in 1FJLA, Glu30 in 1HCQA, Met 42 in 1HCQA, Leu83 in 1IGNA, Glu101 in 1IGNA, Tyr57 in 1PUEE, Thr30 in 1TSRA). Moreover, Ser49 in 1DP7P is a ligand-binding site and Cys125 in 1A74A is a Zn-binding site. Although these predicted residues are not DNA-binding sites, they do interact functionally with other residues or small molecules. The results seem to indicate that our method is helpful for predicting other functional sites.

The second group contains 21 false-positive residues. Our results show that these residues (Lys42, Phe63, Leu76, Asp77, Lys102 and His113 in 1UBDC) have higher amino acid–amino acid similarity scores with actual DNA-binding residues based on the structural alignment method. Because of structural relatedness to actual DNA binding residues, these false-positives may indeed bind to DNA.

In the third group, seven sites (Gly158 in 1GDTA, Arg10 in 1PERL, Thr27 in 1PERL, Lys23 in 1SRSA, Gly36 in 1SRSA, Arg17 in 1YSAC and Ser19 in 1YSAC) are neither functional nor geometrically similar. But the sequences of 1GDTA, 1PERL, 1SRSA, 1YSAC are exactly the same as 1GDTB, 1PERR, 1SRSB, 1YSAD. And Gly158 in 1GDTB, Arg10 in 1PERR, Thr27 in 1PERR, Lys23 in 1SRSB, Gly36 in 1SRSB, Arg17 in 1YSAD and Ser19 in 1YSAD have been verified in actual DNA-binding experiments (Berman *et al.*, 2000). Therefore, the results show that our method can detect true DNA-binding sites.

Table 4. The predictive results compared with other computational methods on the PDNA-62 dataset

Methods	Sn (%)	Sp (%)	Acc (%)	Str (%)	MCC	Threshold	AUC
Dps-pred[a]	40.3	81.8	79.1	61.1	—	—	—
Dbs-pssm[b]	68.2	66.0	66.4	67.1	—	—	—
BindN[c]	69.4	70.5	70.3	—	—	—	0.752
Dp-bind[d]	79.2	77.2	78.1	—	0.49	—	—
DP-Bind[e]	76.4	76.6	77.2	—	—	—	—
BindN-RF[f]	78.1	78.2	78.2	78.1	—	—	0.861
BindN+[g]	77.3	79.3	79.0	78.3	0.44	—	0.859
Our method	**85.2**	**85.3**	**85.2**	**85.2**	**0.55**	**0.03**	**0.926**

The values in bold indicate that they are the best values.
[a]Dps-pred (Ahmad *et al.*, 2004).
[b]Dbs-pssm (Ahmad and Sarai, 2005).
[c]BindN (Wang and Brown, 2006).
[d]Dp-bind (Kuznetsov *et al.*, 2006).
[e]DP-Bind (Hwang *et al.*, 2007).
[f]BindN-RF (Wang *et al.*, 2009).
[g]BindN+ (Wang *et al.*, 2010).

Fig. 5. The spatial structure of 1B3TA, Lys17 is located in the beginning of α-helices, and Thr130 is located in the end of β-strands

The fourth group has 110 mis-predicted residues that are located in DNA-binding regions (DBRs). A DBR is an independently folded protein region that contains at least one motif that recognizes double- or single-stranded DNA. From the predicted results, it is obvious that some demonstrated binding regions are in reasonable agreement with experimental data (see Supplementary Data S3 for more details). For example, in 1A02F, we observe that Arg7, Asn10, Lys11, Ala13, Lys16, Ser17, Arg18, Arg21 form a DBR. The results indicate that our method predicts DNA-binding with greater accuracy and will be of help to biologists.

Besides the four groups mentioned above, only 32 residues in the fifth group were incorrectly predicted by our method.

3.6 Effect of the beginning or the end of α-helices and β-strands

From an analysis on the false-negatives sites with <10% confidence level, we find it difficult to predict DNA-binding sites when these sites are located in the beginning or the end of α-helices and β-strands. For example, Lys17, Thr130 in 1B3TA are DNA-binding residues according to experimental data. But these are mis-predicted in this article. From the spatial structure of 1B3TA (Fig. 5), we find Lys17 is located in the beginning of α-helices and Thr130 is located in the end of β-strands. As alluded to earlier, residues of this nature are difficult to predict, as they occur in the beginning or end of α-helices and β-strands that are usually distorted randomly (Zhou *et al.*, 2010).

4 CONCLUSIONS

The following conclusions can be drawn from this work:

(1) Based on SVM and geometric structure-based algorithms, a novel computational method for predicting DNA-binding sites in proteins is proposed in this work. Overall, method when applied to the common dataset PDNA-62 shows better predictive capability when compared with other methods using the 5-fold cross-validation test.

(2) Our method performs equally well and can detect novel DNA-binding proteins when tested on a new DNA-binding protein dataset generated in this study.

(3) Our method has a better predictive value especially with false-positive sites located in a DBR and can indeed show that these residues have other functions unrelated to DNA binding.

(4) From an analysis on the false-negatives sites with <10% confidence level, it is difficult to predict some binding residues that are located in the beginning or end of α-helices or β-strands.

ACKNOWLEDGEMENTS

We wish to express our gratitude to executive editor and three anonymous reviewers whose constructive comments were very helpful in strengthening the presentation of this article. We thank Dr. Guoli Wang of the Fox Chase Cancer Centre, Philadelphia, PA, for his help in the use of the PISCES software. We would like to thank Zixuan Yuan for her help in language correction of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (grant no: 61063016); doctor subject Foundation of the Ministry of Education of China (grant no: 20101501110004).

Conflict of Interest: none declared

REFERENCES

Ahmad,S. *et al.* (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.

Ahmad,S. and Sarai,A. (2005) PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics*, **6**, 33.

Baldwin,A.S. *et al.* (2001) Methylation and uracil interference assays for analysis of protein-DNA interactions. *Curr. Protoc. Mol. Biol.*, **Chapter 12,** Unit 12.3.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bradley,A.P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.*, **30**, 1145–1159.

Brenowitz,M. *et al.* (1986) Quantitative DNase footprint titration: a method for studying protein-DNA interactions. *Methods Enzymol.*, **130**, 132–181.

Bullock,A.N. and Fersht,A.R. (2001) Rescuing the function of mutant p53. *Nat. Rev. Cancer*, **1**, 68–76.

Cai,Y.D. *et al.* (2002) Support vector machines for predicting HIV protease cleavage sites in protein. *J. Comput. Chem.*, **23**, 267–274.

Cai,Y.D. *et al.* (2004) Application of SVM to predict membrane protein types. *J. Theor. Biol.*, **226**, 373–376.

Chang,C.C. and Lin,C.J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 21–27.

Diekmann,D. and Hall,A. (1995) In vitro binding assay for interactions of Rho and Rac with GTPase-activating proteins and effectors. *Methods Enzymol.*, **256**, 207–215.

Dumitru,I. and McNeil,J.B. (1994) A simple in vivo footprinting method to examine DNA-protein interactions over the yeast PYK UAS element. *Nucleic Acids Res.*, **22**, 1450–1455.

Gao,M. and Skolnick,J. (2008) DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. *Nucleic Acids Res.*, **36**, 3978–3992.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Huang,Y.F. *et al.* (2009) DNA-binding residues and binding mode prediction with binding-mechanism concerned models. *BMC Genomics*, **10** (**Suppl 3**), S23.

Hwang,S. *et al.* (2007) DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics*, **23**, 634–636.

Jones,S. *et al.* (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.*, **31**, 7189–7198.

Kornberg,R.D. (1974) Chromatin structure: a repeating unit of histones and DNA. *Science*, **184**, 868–871.

Kuznetsov,I.B. *et al.* (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins*, **64**, 19–27.

Lane,D. *et al.* (1992) Use of gel retardation to analyze protein-nucleic acid interactions. *Microbiol. Rev.*, **56**, 509–528.

Li,T. and Li,Q.Z. (2012) Annotating the protein-RNA interaction sites in proteins using evolutionary information and protein backbone structure. *J. Theor. Biol.*, **312C**, 55–64.

Luscombe,N.M. *et al.* (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, REVIEWS001.

Ofran,Y. *et al.* (2007) Prediction of DNA-binding residues from sequence. *Bioinformatics*, **23**, i347–i353.

Platt,J.C. (2000) Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola,A. *et al.* (ed.) *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, pp. 61–74.

Ptashne,M. (2005) Regulation of transcription: from lambda to eukaryotes. *Trends Biochem. Sci.*, **30**, 275–279.

Shu,N. *et al.* (2008) Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics*, **24**, 775–782.

Siggers,T.W. *et al.* (2005) Structural alignment of protein-DNA interfaces: insights into the determinants of binding specificity. *J. Mol. Biol.*, **345**, 1027–1045.

Swets,J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Tsuchiya,Y. *et al.* (2004) Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins*, **55**, 885–894.

Tsuchiya,Y. *et al.* (2005) PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces. *Bioinformatics*, **21**, 1721–1723.

Wang,G. and Dunbrack,R.L. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Wang,L. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.*, **34**, W243–W248.

Wang,L. *et al.* (2009) Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genomics*, **10** (**Suppl 1**), S1.

Wang,L. *et al.* (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol.*, **4** (**Suppl 1**), S3.

Wu,J. *et al.* (2009) Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics*, **25**, 30–35.

Zhou,T. *et al.* (2010) A novel mehod for accurate one-dimentional protein structure prediction based on fragment matching. *Bioinformatics*, **26**, 470–477.