Predicting DNA-binding sites of proteins based on sequential and 3D structural information

ARTICLE *in* MGG MOLECULAR & GENERAL GENETICS · JANUARY 2014

Impact Factor: 2.73 · DOI: 10.1007/s00438-014-0812-x · Source: PubMed

4 AUTHORS, INCLUDING:

Biqing li
Chinese Academy of Sciences
28 PUBLICATIONS **333** CITATIONS
SEE PROFILE

Juan Ding
Harvard Medical School
33 PUBLICATIONS **310** CITATIONS
SEE PROFILE

Yu-Dong Cai
Shanghai University
208 PUBLICATIONS **6,964** CITATIONS
SEE PROFILE

Mol Genet Genomics
DOI 10.1007/s00438-014-0812-x

METHODOLOGY ARTICLE

Predicting DNA-binding sites of proteins based on sequential and 3D structural information

Bi-Qing Li · Kai-Yan Feng · Juan Ding · Yu-Dong Cai

Received: 17 June 2013 / Accepted: 4 January 2014

Abstract Protein–DNA interactions play important roles in many biological processes. To understand the molecular mechanisms of protein–DNA interaction, it is necessary to identify the DNA-binding sites in DNA-binding proteins. In the last decade, computational approaches have been developed to predict protein–DNA-binding sites based solely on protein sequences. In this study, we developed a novel predictor based on support vector machine algorithm coupled with the maximum relevance minimum redundancy method followed by incremental feature selection. We incorporated not only features of physicochemical/biochemical properties, sequence conservation, residual disorder, secondary structure, solvent accessibility, but also five three-dimensional (3D) structural features calculated from PDB data to predict the protein–DNA interaction sites. Feature analysis showed that 3D structural features indeed contributed to the prediction of DNA-binding site and it was demonstrated that the prediction performance was better with 3D structural features than without them. It was also shown via analysis of features from each site that the features of DNA-binding site itself contribute the most to the prediction. Our prediction method may become a useful tool for identifying the DNA-binding sites and the feature analysis described in this paper may provide useful insights for in-depth investigations into the mechanisms of protein–DNA interaction.

Keywords Protein-DNA interactions · Structural features · Random Forest (RF) · Maximum relevance minimum redundancy (mRMR) · Incremental feature selection (IFS)

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-014-0812-x) contains supplementary material, which is available to authorized users.

B.-Q. Li
Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China

B.-Q. Li
Shanghai Center for Bioinformation Technology, Shanghai, People's Republic of China

K.-Y. Feng
Beijing Genomics Institute, Shenzhen Beishan Industrial Zone, Beishan Road, Yantian District, Shenzhen 518083, People's Republic of China

J. Ding (✉)
Schepens Eye Research Institute, Harvard Medical School, 20 Staniford St., Boston, MA 02114, USA
e-mail: juan_ding@meei.harvard.edu

Y.-D. Cai (✉)
Institute of Systems Biology, Shanghai University, Shanghai, People's Republic of China
e-mail: cai_yud@126.com; cai_yud@yahoo.com.cn

Introduction

Protein–DNA interactions play important roles in many biological processes including transcription, DNA replication and repair, viral infection, DNA packing and DNA modifications (Luscombe et al. 2000). Therefore, identification of DNA-binding residues can significantly facilitate our understanding of these biological processes and guide site-directed mutagenesis studies for the functional characterization of DNA-binding proteins. Furthermore, it may also contribute to drug design and discovery, such as aiding

Published online: 22 January 2014

the design of artificial transcription factors (Blancafort et al. 2004).

Conventional experimental approaches such as mutagenesis and binding assays to identify DNA-binding sites are expensive and laborious. With the expansion of protein sequence data, there is an urgent need to develop computational tools that can rapidly and reliably identify DNA-binding sites. Therefore, significant interest has been put into developing computational methods to identify amino acid residues that participate in protein–DNA interactions. Two groups of prediction methods, sequence-based and structure-based methods, are found in the literature to tackle the problem (Ofran et al. 2007). Sequence-based methods have an advantage of not requiring the expensive and time-consuming process of experimentally determining protein structure, and 3D structures are only available for <5 % of all known DNA-binding proteins (Ofran et al. 2007). However, it was suggested that computational models exploiting 3D structures could be used to predict binding residues (Szilagyi and Skolnick 2006). As a matter of fact, such models showed very good performance (Szilagyi and Skolnick 2006). Incorporating more effective features is the most effective way to improve the performance of classifiers. Abundant information that can be extracted from 3D structures and sequences, improvement of computation power and the invention of novel classification methods all have driven the advancement of predicting the protein–DNA interaction sites.

Studies of protein–DNA interfaces suggested that amino acids at the interface possess some properties that distinguish them from the rest of the protein (Lejeune et al. 2005). Thus, it is particularly appealing to use machine learning algorithms to model the DNA-binding patterns of amino acid residues and the complex patterns hidden in the available structural data—the resulting classifier may reliably identify DNA-binding residues. So far, several machine learning methods have been reported for the prediction of DNA-binding residues. Ahmad et al. (2004) analyzed the structural data of representative protein–DNA complexes, and utilized the amino acid sequences in these structures to train an artificial neural network (ANN) for the prediction of DNA-binding sites. Yan et al. (2006) constructed Naive Bayes classifiers using the amino acid characteristics of DNA-binding sites and their sequence neighbors. However, the prediction accuracies were relatively low in these studies (Ahmad et al. 2004; Yan et al. 2006), probably because only features of amino acids were used for the classifier construction. In addition, support vector machines (SVMs) and logistic regression models have been proposed for accurate prediction of DNA-binding residues (Kuznetsov et al. 2006; Hwang et al. 2007). More recently, Li et al. has proposed a tool called PreDNA to predict DNA-binding sites in proteins by integrating sequence and geometric structure information based on SVM (Li et al. 2013b).

In this study, we proposed a novel predictor based on support vector machines (SVMs) algorithm coupled with maximum relevance minimum redundancy (mRMR) method followed by incremental feature selection (IFS). We incorporated both 3D structural features from the PDB data and the features derived from protein sequences to construct a classifier. We have demonstrated that 3D structural features contribute to the prediction of DNA-binding site such that the prediction performance with these features is better than without using them.

Materials and methods

Dataset

For this study, we used the data set described in the work of Ofran et al. (2007). Specifically, we downloaded all protein–DNA complexes in the protein data bank (PDB, http://www.rcsb.org/pdb/home/home.do) (Berman et al. 2000). To reduce the bias of similar sequences, we deleted the homology sequences in the original dataset of 693 PDB chains using a threshold of 30 % identity by CD-HIT (Li and Godzik 2006), resulting in a total of 112 chains. In these complexes, an amino acid is considered to be in contact with a nucleotide if the distance between any atoms of the two molecules is no more than 6 Å. Then, we extracted nine-residue protein segment centered on the annotated protein–DNA interaction residue, with four residues upstream and four residues downstream of the interaction site. For the peptides with length <9 amino acid residues, we complement it with "X". We regarded protein segments centered on the annotated interaction as positive data, while the others as the non-interaction segments. 90 PDB chain sequences were randomly picked to construct the training dataset and the remaining 22 sequences for the testing dataset. In this way, for the training dataset we obtained 2,906 positive samples and randomly selected 5,812 negative samples. As for the testing dataset, we obtained 905 positive samples and 7,044 negative samples since in real situation negative samples are much more than positive samples.

Feature construction

The features of PSSM conservation scores

Evolutionary conservation plays an important role in biological analysis. A more conserved residue within a protein sequence may indicate an important role by surviving selective pressure. We used position specific iterative

BLAST (PSI BLAST) (Altschul et al. 1997) to measure the conservation status for a specific residue. A 20-dimensional vector was used to denote probabilities of conservation against mutations to 20 different amino acids for a specific residue. For a given peptide, all such 20-dimensional vectors for all residues composed of a matrix called position specific scoring matrix (PSSM). In this study, we used PSSM conservation score to quantify the conservation status of each amino acid in a protein sequence.

The features of amino acid factors

Since each of the 20 amino acids has different and specific properties, the composition of these properties of different residues within a protein can influence the specificity and diversity of the protein structure and function. AAIndex (Kawashima and Kanehisa 2000) is a database containing various physicochemical and biochemical properties of amino acids. Atchley et al. (2005) performed multivariate statistical analyses on AAIndex and transformed AAIndex to five multidimensional and highly interpretable numeric patterns of attribute covariation reflecting polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge. We used these five numerical pattern scores (denoted as "amino acid factors") to represent the respective properties of each amino acid in a given protein.

The features of disorder score

Protein segments lacking fixed three-dimensional structures under physiological conditions play important roles in biological functions (Wright and Dyson 1999; Dunker et al. 2002). The disordered regions of proteins allow for more modification sites and interaction partners and always contain PTM sites, sorting signals, and protein ligands, therefore, are important for protein structuring and functioning (Wright and Dyson 1999; Liu et al. 2002; Tompa 2002). In this study, VSL2 (Peng et al. 2006), which can accurately predict both long and short disordered regions in proteins, was used to calculate disorder score that denotes the disorder status of each amino acid in a given protein sequence.

The features of secondary structure

The post-translational modification of specific residues may be influenced by the solvent accessibility of the relevant residues. In this study, we used the structural features including secondary structure to encode the peptides. The secondary structure was predicted by the predictor SSpro4 (Cheng et al. 2005). SSpro4 predicts secondary structural property of each amino acid to be 'helix', 'strand', or 'other', encoded by 100, 010 and 001, respectively.

The features of protrusion index, depth index, solvent accessible surface area and relative solvent accessible surface area

It has been shown that geometrical properties of the protein surface can influence protein–protein interactions (Jones and Thornton 1997). Thus, supposedly it could influence protein–DNA interactions. In our study, 3D structure features including protrusion index (CX), depth index (DPX), solvent accessible surface area (SAS) and relative solvent accessible surface area (RASA) were used to encode the peptides. These features were predicted by the protein structure and interaction analyzer (PSAIA) from PDB data. PSAIA was developed to compute geometric parameters for large sets of protein structures to predict and investigate protein–protein interaction sites (Mihel et al. 2008).

The feature of B-factor

The B-factors of protein crystal structures reflect the fluctuation of atoms about their average positions and provide important information about protein dynamics. The thermal motion is useful for analyzing the dynamic properties of proteins (Yuan et al. 2005). Therefore, in our study we extracted the B-factor from the PDB data directly to encode the peptides.

The feature space

For each residue of a protein segment, we incorporated 36 features, including 20 features of PSSM conservation score, 1 disorder feature, 5 features of AAFactor, 3 features of the predicted secondary structure, 2 features of solvent accessibility and 5 3D structure features from PDB data. When 3D structures were considered, the predicted solvent accessibility features were excluded. Overall for the nine-residue peptide, there are a total of $34 \times 9 = 306$ features. For nine-residue peptides complemented with "X" residues, all features of these "X" residues are denoted as 0. To determine whether 3D structure features can improve the prediction performance, we also constructed a dataset excluding the five 3D structural features. There are a total of $31 \times 9 = 279$ features for this dataset.

mRMR method

We used maximum relevance minimum redundancy (mRMR) method to rank the importance of all the features (Peng et al. 2005), which has been successfully used in handling various biological prediction problems (Li et al. 2012b, c; Zhang et al. 2012; Gao et al. 2013). mRMR method could rank features based on both their relevance to the target and the redundancy between features. A

smaller index of a feature denotes that it has a better trade-off between maximum relevance to target and minimum redundancy.

Both relevance and redundancy were quantified by mutual information (MI), which estimates how much one vector is related to another. The MI equation was defined as below:

$$I(x,y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \mathrm{d}x\mathrm{d}y \tag{1}$$

In Eq. (1), x, y are vectors, $p(x,y)$ is their joint probabilistic density, and $p(x)$ and $p(y)$ are the marginal probabilistic densities.

Let Ω denote the whole feature set, Ω_s the already-selected feature set containing m features and Ω_t the to-be-selected feature set containing n features. The relevance D between the feature f in Ω_t and the target c can be calculated by:

$$D = I(f,c) \tag{2}$$

The redundancy R between the feature f in Ω_t and all the features in Ω_s can be calculated by:

$$R = \frac{1}{m} \sum_{f_i \in \Omega_s} I(f,f_i) \tag{3}$$

To get the feature f_j in Ω_t with maximum relevance and minimum redundancy, the mRMR function combined Eqs. (2) and (3) and is defined as below:

$$\max_{f_j \in \Omega_t} \left[I(f_j,c) - \frac{1}{m} \sum_{f_i \in \Omega_s} I(f_j,f_i) \right] \quad (j = 1,2,\ldots,n) \tag{4}$$

The mRMR feature evaluation would continue N rounds when given a feature set with N ($N = m + n$) features. After the mRMR evaluation rounds, we get a feature set$S$:

$$S = \left\{ f_1^{'}, f_2^{'}, \ldots, f_h^{'}, \ldots, f_N^{'} \right\} \tag{5}$$

In this feature set S, the index h of each feature indicates at which round that the feature is selected. The smaller the index h is, the earlier the feature satisfies Eq. (4) and the better the feature is.

Support vector machines (SVMs)

In this study, the support vector machines (SVMs) approach was introduced and its fast learning algorithm called sequential minimal optimization (SMO) was also implemented. Support vector machine (SVM) is a kind of learning machines based on the statistical learning theory and has been successfully and widely used in dealing with various biological prediction problems (Bock and Gough 2001; Ding and Dubchak 2001; Hua and Sun 2001). The basic idea of applying SVM to feature selection can be outlined as follows. First, map the input vectors into one feature space (possible with a higher dimension), either linearly or nonlinearly, which is relevant to the selection of the kernel function. Then, within the feature space from the first step, seek an optimized linear division, i.e., construct a hyper-plane that separates the samples into two classes (this can be extended to multi-classes). SVM training always seeks a global optimized solution and avoids over-fitting, so it has the ability to deal with a large number of features. However, the biggest problem for training a SVM is the very large quadratic programming (QP) optimization problem. Unlike previous SVM learning algorithms that use numerical QP as an inner loop, SMO uses an analytic QP step to solve this problem without any extra matrix storage, improving its scaling and computation time significantly. Since first proposed by John C. Platt in 1998 (Platt 1998), SMO has been tested on both real-world and artificial problems (Shevade et al. 2000; Knebel et al. 2008; Takahashi et al. 2008).

Tenfold cross-validation method

Tenfold cross-validation was used to evaluate the performance of a classifier (Kohavi 1995). In tenfold cross-validation the data are first divided into ten equally sized folds. Subsequently, ten iterations of training and validation are performed such that in each iteration a different fold of the data is left for validation while the remaining ninefolds are used for training. Let TP denotes true positive. TN denotes true negative. FP denotes false positive and FN denotes false negative. To evaluate the performance of the predictor, the prediction accuracy, specificity, sensitivity and MCC (Matthews's correlation coefficient) were calculated as below:

$$\begin{cases} \text{accuracy} = \dfrac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\ \text{sensitivity} = \dfrac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{specificity} = \dfrac{\text{TN}}{\text{TN} + \text{FP}} \\ \text{MCC} = \dfrac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \end{cases} \tag{6}$$

Incremental feature selection (IFS)

Based on the ranked features according to their importance after mRMR evaluation, we used incremental feature selection (IFS) (Li et al. 2012a, d, 2013a) to determine the optimal number of features.

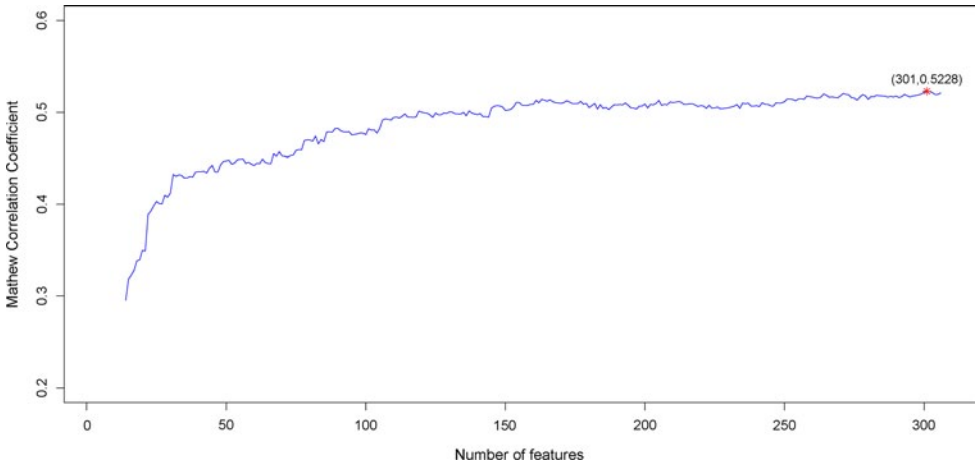

Fig. 1 A plot to show the change of the MCC values versus the feature numbers. The IFS curves were drawn based on the data in Online Supporting Information S2. The MCC first exceeded 0.5140 when 163 features as given in Supporting Information S3 were used The 163 features thus obtained were used to form the optimal feature set for the DNA-binding site predictor

During IFS procedure, features in the ranked feature set are added one by one from higher to lower rank. A new feature set is composed when one feature is added. Thus, N feature sets would be composed when given N ranked features. The i-th feature set is:

$$S_i = \{f_1, f_2, \ldots, f_i\} \ (1 \leq i \leq N) \tag{7}$$

For each of the N feature sets, an SVM predictor was constructed and tested using tenfold cross-validation test. With N prediction accuracies, sensitivities, specificities and MCCs calculated, we obtain an IFS table with one column being the index i and the other columns to be the prediction accuracy, sensitivity, specificity and MCC. We then could get the optimal feature set (S_{optimal}), using which the predictor achieves the best prediction performance.

Results

The mRMR result

Listed in the Online Supporting Information S1 are two kinds of outcomes obtained by running the mRMR software: one is called MaxRel feature table that ranks the 306 features according to their relevance to the class of the samples; and the other is called mRMR feature table that lists the ranked 306 features using mRMR criteria. In the mRMR feature table, a feature with a smaller index implies that it is more important for DNA-binding site prediction. Such list of ranked features was to be used in the following IFS procedure for the optimal feature set selection.

IFS result

By adding the ranked features one by one, we built 306 individual predictors for the 306 sub-feature sets to predict the DNA-binding sites. We then tested the prediction performance for each of the 306 predictors and obtained the IFS results (Supporting Information S2). Shown in Fig. 1 is the IFS curve plotted based on the data of Supporting Information S2. As we can see from the figure, the MCC first exceeded 0.5140 when 163 features as given in Supporting Information S3 were used and finally achieved the maximum of 0.5228 with 301 features after a period of fluctuation. Based on these 301 features, the predictive sensitivity, specificity and accuracy were 0.6215, 0.8802 and 0.7940, respectively (Table 1). To avoid introduction of irrelevant features, our further analysis will be focused on the 163 features.

Table 1 The predicted results with and without 3D structural features

Dataset	3D structures	Sn	Sp	Ac	MCC
Training dataset	Yes	0.6215	0.8802	0.7940	0.5228
	No	0.5499	0.8780	0.7686	0.4573
Testing dataset	Yes	0.4718	0.8661	0.8212	0.2842
	No	0.3978	0.8718	0.8178	0.2342

Sn sensitivity, *Sp* specificity, *Ac* accuracy, *MCC* Matthews's correlation coefficient

Comparison of prediction performance with and without 3D structure features

To determine whether 3D structural features contribute to the prediction of DNA-binding sites, we constructed training and testing datasets without 3D structural features. Listed in the Online Supporting Information S4 are the prediction accuracy, specificity, sensitivity and MCC based on the training dataset without 3D structural features. As we can see in Table 1, the prediction accuracy and MCC for training dataset were better when using 3D structural features (accuracy 0.7940, MCC 0.5228) than without

Table 2 Comparison with other DNA-binding prediction methods

Method	Sn	Sp	Ac	MCC
Ours (SVM with 3D structure feature)	0.4718	0.8661	0.8212	0.2842
PreDNA	0.3478	0.8886	0.8268	0.2177
BindN	0.3845	0.8014	0.7539	0.1426
BindN_RF	0.4110	0.8234	0.7764	0.1850

Sn sensitivity, *Sp* specificity, *Ac* accuracy, *MCC* Matthews's correlation coefficient

(accuracy 0.7686, MCC 0.4573). For the independent testing dataset, the performance of the predictor with 3D structural features (accuracy 0.8212, MCC 0.2842) is also better than without (accuracy 0.8178, MCC 0.2342). These comparisons suggested that the 3D structural features indeed contributed to the prediction of DNA-binding sites and could improve the prediction performance.

Comparisons with other methods

We compared the prediction performance of our method with the three other DNA-binding prediction methods PreDNA (Li et al. 2013b), BindN (Wang and Brown 2006) and BindN-RF (Wang et al. 2009) on the same independent testing dataset. As shown in Table 2, our method considering 3D structural features has the best performance (ac 0.8212, MCC 0.2842) compared with BindN (ac 0.7539, MCC 0.1426) and BindN-RF (ac 0.7764, MCC 0.1850). For PreDNA (ac 0.8268, MCC 0.2177), though the accuracy of our method was slightly worse, our MCC was much better than that of PreDNA. In summary, our method achieved the highest MCC, and outperformed three other methods on an independent testing dataset.

Discussion

Feature analysis

The distribution of the number of each type of features in the final optimal feature set was investigated and shown in Fig. 2a. Of the 163 features, 85 were from PSSM conservation scores, 28 from the amino acid factors, 3 from the disorder scores, 13 from the predicted secondary structural propensities and 34 from the 3D structure. Five kinds of features made contributions to the prediction of DNA-binding sites. It was revealed by the site-specific distribution of the optimal feature set (see Fig. 2b) that site 5 played the most important roles in determining the DNA-binding sites, indicating that the features from the DNA-binding site itself play the most important role in the prediction. In addition, the features of site 1 and site 9 also contributed significantly on the prediction of DNA-binding sites.

Feature analysis of PSSM conservation score

As mentioned above, among the 163 features, 85 belonged to the PSSM conservation features. It can be clearly seen from Fig. 3a that each of the 20 different amino acid types contributed differently to the prediction of DNA-binding sites in term of the PSSM conservation score. In this regard, the amino acid cysteine (C) and threonine (T) contributed

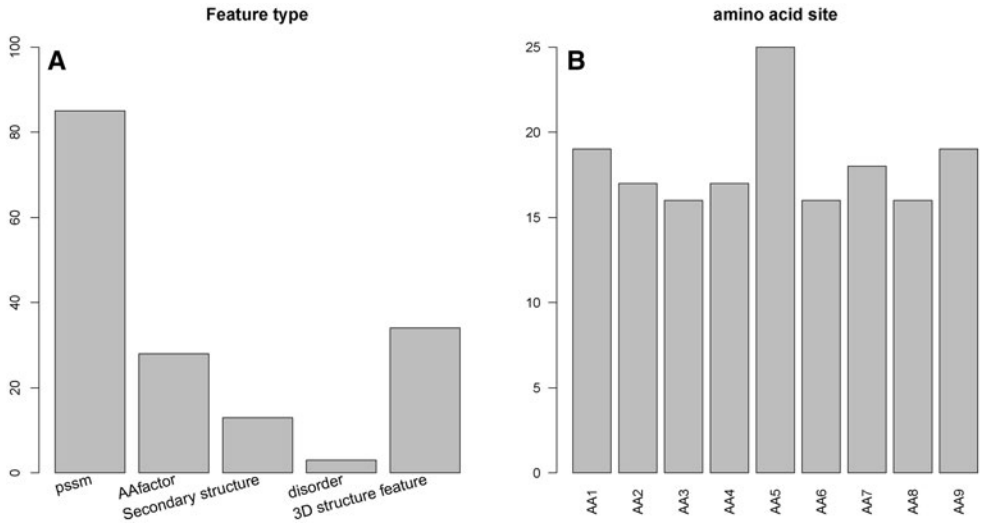

Fig. 2 A two-dimensional histogram to characterize the optimal feature set. The contributions to the DNA-binding site prediction from **a** the six different feature types, and **b** each of the nine subsites. See Sect. "Feature analysis" for further explanation

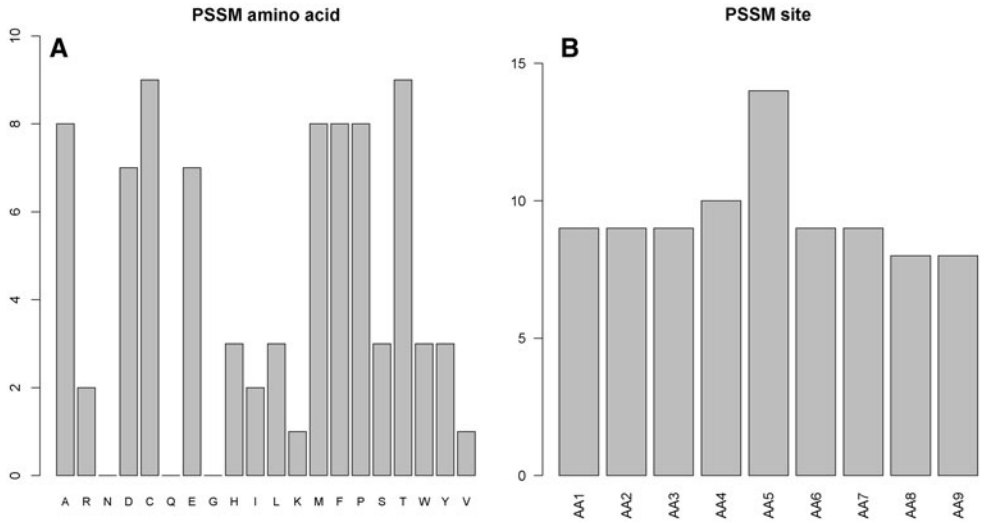

Fig. 3 A two-dimensional histogram to characterize the PSSM features in the optimal features set. **a** The contributions on the DNA-binding site prediction from mutation to each of the 20 amino acid types. **b** The evolutionary conservation status for each of the nine subsites. See Sect. "Feature analysis of PSSM conservation score" for further explanation

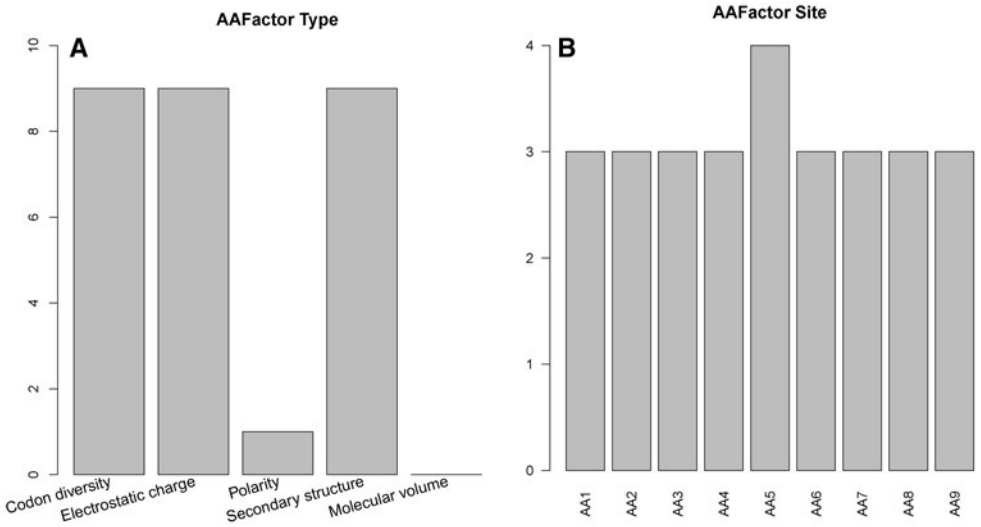

Fig. 4 A two-dimensional histogram to characterize the amino acid factor types in the optimal features set. The contributions on the DNA-binding site prediction from **a** the five different amino acid types, and **b** each of the nine subsites. See Sect. "Feature analysis of amino acid factor" for further explanation

most, followed by alanine (A), methionine (M), phenylalanine (F) and proline (P). It has been reported that there are conserved cysteine residues in the DNA-binding domain of the bovine papillomavirus type 1 E2 protein and a motif consisting of a reactive cysteine residue in a basic region of the DNA-binding domain is a feature common to a number of transcriptional regulatory proteins (McBride et al. 1992). Furthermore, four conserved cysteine residues are required for the DNA-binding activity of nuclear factor I (Novak et al. 1992). Meanwhile, as shown in Fig. 3b, the conservation status at the subsite 5 played the most important role in predicting the DNA-binding sites. Moreover, the conservation status against residue E (glutamic acid) at subsite 5 ("AA5_pssm_7") has an index of 1 in the final optimal feature set, which implied that the conservation status of the interaction site itself played critical roles in the determination of DNA-binding sites. In addition, the top ten optimal features contain four other PSSM conservation features: the conservation status against residue E at site 3 and 4 (index 4, "AA3_pssm_7" and index 3, "AA4_pssm_7"), the conservation status against residue D at site 2 (index 8, "AA2_pssm_4") and the conservation status against residue C at site 7 (index 10, "AA7_pssm_5").

Feature analysis of amino acid factor

Illustrated in Fig. 4 are the contributions of different amino acid factors and their subsite locations to the DNA-binding

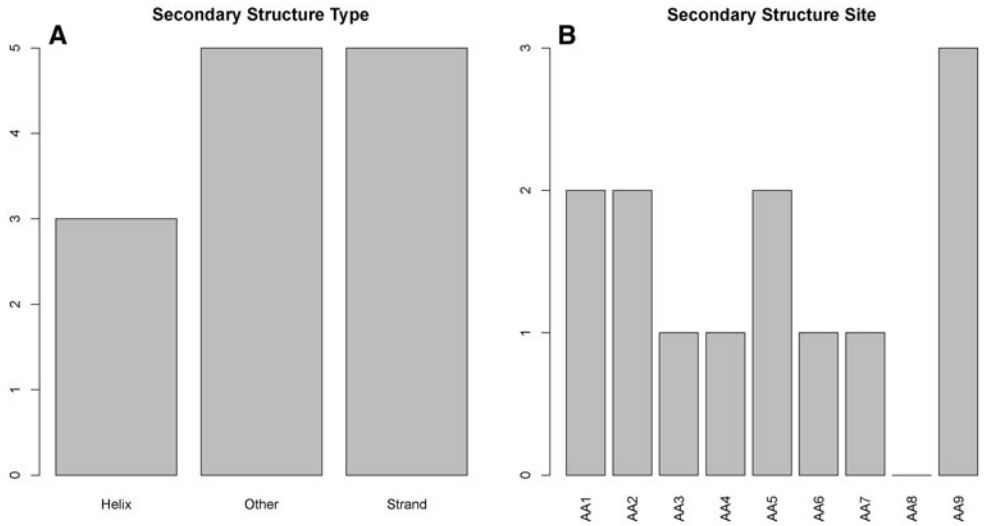

Fig. 5 A 2-dimensional histogram to characterize the secondary structure types in the optimal features set. The contributions on the DNA binding site prediction from **a** the three different types of the secondary structure, and **b** each of the 9 subsites. See Sect. "Secondary structure features analysis" for further explanation

site prediction. It can be seen from Fig. 4a that the codon diversity, electrostatic charge and secondary structure were the most important feature to the DNA-binding site prediction. The codon diversity of subsite 5 has an index of 7 in the optimal feature set, which implied that it is critical for the determination of DNA-binding sites. It has been shown that the Homo sapiens genome had stronger codon bias for DNA-binding transcription factors than the *Saccharomyces cerevisiae* genome (Hudson et al. 2011). The electrostatic charge of subsite 6 has an index of 5 in the optimal feature set, which suggested that electrostatic charges have great influence on the prediction of DNA-binding sites. Jones et al. (2003) had pointed out electrostatic differences between DNA-binding patches and the rest of the protein surface, and had also demonstrated that these differences may suffice for the prediction of interaction sites from the coordinates of the 3D structure of a protein (Shanahan et al. 2004). As shown in Fig. 4b, the amino acid residues at the subsite 5 contributed the most to the DNA-binding sites prediction.

Disorder feature analysis

Within the final optimal feature set, three disordered features were selected. Such two disordered features were from subsites 1, 5 and 8. Particularly, the disorder feature of subsite 8 had the index of 14 in the final optimal feature set, suggesting that it was one of the most important features in the DNA-binding site prediction. The intrinsic disorder within and flanking the DNA-binding domains of human transcription factors has been reported (Guo et al. 2012). Furthermore, intrinsic disorder is more prevalent among transcription factors than other types of proteins (Liu et al. 2006).

Secondary structure features analysis

The feature and site-specific distribution of the secondary structure in the optimal feature set was given in Fig. 5, from which we can see that all the three kinds of secondary features can influence on the prediction of DNA-binding site (panel a). Particularly, it has been shown that helix-turn-helix motif is a common substructure of DNA-binding proteins (Brennan and Matthews 1989). While secondary structures at subsites 9 contributed the most on the DNA-binding site determination (panel b). The secondary structure of subsite 9 has an index of 2 in the optimal feature set indicating it is an important feature for the prediction of DNA-binding sites.

CX, DPX, ASA, RASA and B-factor feature analysis

Shown in Fig. 6 are the five kinds of 3D structural features in the optimal feature set. There are nine ASA features, eight CX features, seven DPX features, seven RASA features and three B-factor features in the optimal feature set. It can be seen from Fig. 6a that all five kinds of features contribute to the prediction of DNA-binding sites, and ASA and CX contribute most, followed by DPX and RASA. Moreover, it can be seen from Fig. 6b that 3D structural features at the subsite 7 and subsite 9 contribute significantly to the DNA-binding site prediction, which suggests that the 3D structural features from the residues away from the binding site play crucial role in the prediction of

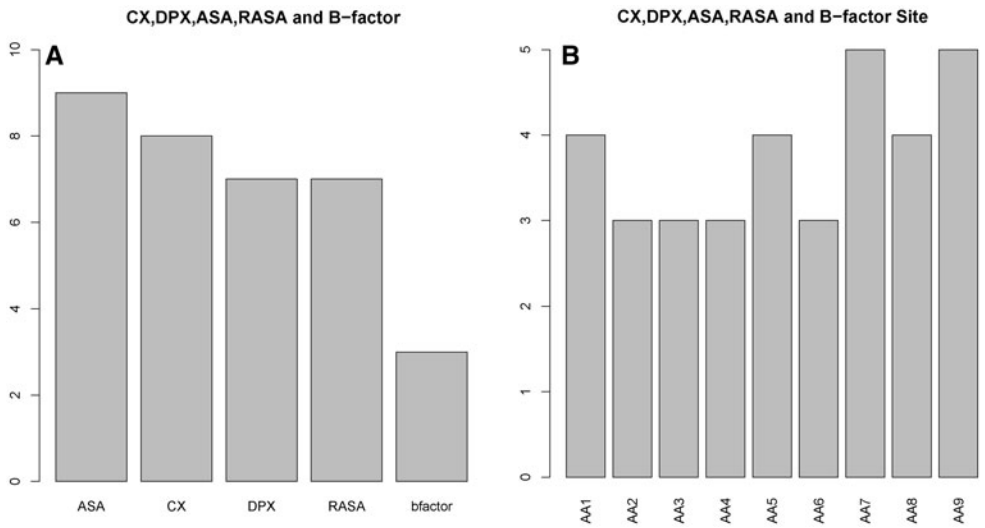

Fig. 6 A two-dimensional histogram to characterize the 3D structure feature types in the optimal features set. The contributions on the DNA-binding site prediction from **a** the protrusion index (CX), depth index (DPX), solvent accessible surface area (ASA), relative solvent accessible surface area (RASA), B-factor, and **b** each of the nine subsites. See Sect. "CX, DPX, ASA, RASA and B-factor feature analysis" for further explanation

DNA-binding sites. It has been reported that DNA-binding probability of a residue is significantly higher for residues with higher solvent accessible area (Ahmad et al. 2004) and the difference in the solvent accessible surface area has been used to predict the DNA-binding sites (Zen et al. 2009). Moreover, it has been shown that both the protrusion index (CX) and accessible surface area (ASA) are important features to distinguish hot spots from non-hot spots in protein interfaces (Xia et al. 2010). Furthermore, it has been shown that the average values of the B-factors of 20 types of residues in DNA-binding group were also significantly lower than the non-binding group (Xiong et al. 2011).

Directions for experimental validation

The selected features at different sites may provide clues for researchers to find or validate new determinants of DNA-binding sites. For example, we revealed that amino acid C (Cysteine) plays a pivotal role in the DNA-binding site determination, consistent with previous reports (McBride et al. 1992; Novak et al. 1992). In addition, we highlighted the importance of codon diversity and electrostatic charge, which agrees with previous studies (Jones et al. 2003; Shanahan et al. 2004; Hudson et al. 2011). There were three disorder features in the optimal feature set, indicating that disorder features were important for prediction of DNA-binding site, consistent with previous reports that intrinsic disorder within and flanking the DNA-binding domains of human transcription factors exist (Liu et al. 2006; Guo et al. 2012). It was revealed in our study that secondary structure of helix played an important role in determination of DNA-binding sites, whose role in protein–DNA interaction has been confirmed (Brennan and Matthews 1989). The role of 3D structural features such as RASA, ASA, CX and DPX in the prediction of DNA-binding sites has also been supported by previous studies (Ahmad et al. 2004; Zen et al. 2009; Xia et al. 2010; Xiong et al. 2011). Thus, the remaining features in the optimal feature set were worthy of validation by further research.

In this study, we have developed a prediction model for predicting DNA-binding residues of proteins. A major contribution of this study is to incorporate several new 3D structural features calculated from the PDB data, which has been shown to significantly improve the prediction performance. Our results indicate that the combination of traditional features and 3D structural features is effective in predicting the DNA-binding residues. Our method outperformed three other DNA-binding site prediction methods based on the same testing dataset. Moreover, we have identified an optimal feature list containing the most important features to identify DNA-binding residues. The selected features in this study may shed some light on the mechanism of protein–DNA interaction and provide guidelines for experimental validations.

Acknowledgments This work was supported by grants from National Basic Research Program of China (2011CB510102, 2011CB510101), National Natural Science Foundation of China (31371335), Innovation Program of Shanghai Municipal Education Commission (12ZZ087), and the grant of "The First-class Discipline of Universities in Shanghai".

References

Ahmad S, Gromiha MM, Sarai A (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 20(4):477–486. doi:10.1093/bioinformatics/btg432btg432

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402 (pii:gka562)

Atchley WR, Zhao J, Fernandes AD, Druke T (2005) Solving the protein sequence metric problem. Proc Natl Acad Sci USA 102(18):6395–6400. doi:10.1073/pnas.0408677102

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242 (pii:gkd090)

Blancafort P, Segal DJ, Barbas CF 3rd (2004) Designing transcription factor architectures for drug discovery. Mol Pharmacol 66(6):1361–1371. doi:10.1124/mol.104.002758

Bock JR, Gough DA (2001) Predicting protein–protein interactions from primary structure. Bioinformatics 17(5):455–460

Brennan RG, Matthews BW (1989) The helix-turn-helix DNA binding motif. J Biol Chem 264(4):1903–1906

Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 33:W72–W76. doi:10.1093/Nar/Gki396

Ding CH, Dubchak I (2001) Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17(4):349–358

Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z (2002) Intrinsic disorder and protein function. Biochemistry 41(21):6573–6582

Gao Y-F, Li B-Q, Cai Y-D, Feng K-Y, Li Z-D, Jiang Y (2013) Prediction of active sites of enzymes by maximum relevance minimum redundancy (mRMR) feature selection. Mol Biosyst 9:61–69

Guo X, Bulyk ML, Hartemink AJ (2012) Intrinsic disorder within and flanking the DNA-binding domains of human transcription factors. Pac Symp Biocomput :104–115 (pii:9789814366496_0011)

Hua S, Sun Z (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol 308(2):397–407. doi:10.1006/jmbi.2001.4580

Hudson NJ, Gu Q, Nagaraj SH, Ding Y-S, Dalrymple BP, Reverter A (2011) Eukaryotic evolutionary transitions are associated with extreme codon bias in functionally-related proteins. PLoS One 6(9):e25457. doi:10.1371/journal.pone.0025457

Hwang S, Gou Z, Kuznetsov IB (2007) DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics 23(5):634–636. doi:10.1093/bioinformatics/btl672

Jones S, Thornton JM (1997) Analysis of protein–protein interaction sites using surface patches. J Mol Biol 272(1):121–132. doi:10.1006/jmbi1997.1234

Jones S, Shanahan HP, Berman HM, Thornton JM (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res 31(24):7189–7198

Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. Nucleic Acids Res 28(1):374 (pii:gkd029)

Knebel T, Hochreiter S, Obermayer K (2008) An SMO algorithm for the potential support vector machine. Neural Comput 20(1):271–287. doi:10.1162/neco.2008.20.1.271

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, pp 1137–1143

Kuznetsov IB, Gou Z, Li R, Hwang S (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins 64(1):19–27. doi:10.1002/prot.20977

Lejeune D, Delsaux N, Charloteaux B, Thomas A, Brasseur R (2005) Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. Proteins 61(2):258–271. doi:10.1002/prot.20607

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659. doi:10.1093/bioinformatics/btl158

Li B-Q, Cai Y-D, Feng K-Y, Zhao G-J (2012a) Prediction of protein cleavage site with feature selection by random forest. PLoS One 7(9):e45854. doi:10.1371/journal.pone.0045854

Li B-Q, Feng K-Y, Chen L, Huang T, Cai Y-D (2012b) Prediction of protein–protein interaction sites by random forest algorithm with mRMR and IFS. PLoS One 7(8):e43927. doi:10.1371/journal.pone.0043927

Li B-Q, Hu L–L, Chen L, Feng K-Y, Cai Y-D, Chou K-C (2012c) Prediction of protein domain with mRMR feature selection and analysis. PLoS One 7(6):e39308. doi:10.1371/journal.pone.0039308

Li BQ, Hu LL, Niu S, Cai YD, Chou KC (2012d) Predict and analyze *S*-nitrosylation modification sites with the mRMR and IFS approaches. J Proteomics 75(5):1654–1665. doi:10.1016/j.jprot.2011.12.003

Li B-Q, Huang T, Zhang J, Zhang N, Huang G-H, Liu L, Cai Y-D (2013a) An ensemble prognostic model for colorectal cancer. PLoS One 8(5):e63494. doi:10.1371/journal.pone.0063494

Li T, Li QZ, Liu S, Fan GL, Zuo YC, Peng Y (2013b) PreDNA: accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information. Bioinformatics 29(6):678–685. doi:10.1093/bioinformatics/btt029btt029

Liu J, Tan H, Rost B (2002) Loopy proteins appear conserved in evolution. J Mol Biol 322(1):53–64 (pii:S0022283602007362)

Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK (2006) Intrinsic disorder in transcription factors. Biochemistry 45(22):6873–6888. doi:10.1021/bi0602718

Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) An overview of the structures of protein-DNA complexes. Genome Biol 1(1):REVIEWS001

McBride AA, Klausner RD, Howley PM (1992) Conserved cysteine residue in the DNA-binding domain of the bovine papillomavirus type 1 E2 protein confers redox regulation of the DNA-binding activity in vitro. Proc Natl Acad Sci USA 89(16):7531–7535

Mihel J, Sikic M, Tomic S, Jeren B, Vlahovicek K (2008) PSAIA—protein structure and interaction analyzer. BMC Struct Biol 8:21. doi:10.1186/1472-6807-8-21

Novak A, Goyal N, Gronostajski RM (1992) Four conserved cysteine residues are required for the DNA binding activity of nuclear factor I. J Biol Chem 267(18):12986–12990

Ofran Y, Mysore V, Rost B (2007) Prediction of DNA-binding residues from sequence. Bioinformatics 23(13):i347–i353. doi:10.1093/bioinformatics/btm174

Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238. doi:10.1109/TPAMI.2005.159

Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z (2006) Length-dependent prediction of protein intrinsic disorder. BMC Bioinf 7:208. doi:10.1186/1471-2105-7-208

Platt JC (1998) Sequential minimal optimization: a fast algorithm for training support vector machines. Microsoft Research. MSR-TR-98–14

Shanahan HP, Garcia MA, Jones S, Thornton JM (2004) Identifying DNA-binding proteins using structural motifs and the

electrostatic potential. Nucleic Acids Res 32(16):4732–4741. doi:10.1093/nar/gkh80332/16/4732

Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KK (2000) Improvements to the SMO algorithm for SVM regression. IEEE Trans Neural Netw 11(5):1188–1193. doi:10.1109/72.870050

Szilagyi A, Skolnick J (2006) Efficient prediction of nucleic acid binding function from low-resolution protein structures. J Mol Biol 358(3):922–933. doi:10.1016/j.jmb.2006.02.053

Takahashi N, Guo J, Nishi T (2008) Global convergence of SMO algorithm for support vector regression. IEEE Trans Neural Netw 19(6):971–982. doi:10.1109/TNN.2007.915116

Tompa P (2002) Intrinsically unstructured proteins. Trends Biochem Sci 27(10):527–533 (pii:S0968000402021692)

Wang L, Brown SJ (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 34(Web Server issue):W243–W248. doi:10.1093/nar/gkl298

Wang L, Yang MQ, Yang JY (2009) Prediction of DNA-binding residues from protein sequence information using random forests. BMC Genom 10(Suppl 1):S1. doi:10.1186/1471-2164-10-S1-S1

Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. J Mol Biol 293(2):321–331. doi:10.1006/jmbi.1999.3110

Xia J-F, Zhao X-M, Song J, Huang D-S (2010) APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinf 11(1):174

Xiong Y, Liu J, Wei DQ (2011) An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins 79(2):509–517. doi:10.1002/prot.22898

Yan C, Terribilini M, Wu F, Jernigan RL, Dobbs D, Honavar V (2006) Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinf 7:262. doi:10.1186/1471-2105-7-262

Yuan Z, Bailey TL, Teasdale RD (2005) Prediction of protein B-factor profiles. Proteins 58(4):905–912. doi:10.1002/prot.20375

Zen A, de Chiara C, Pastore A, Micheletti C (2009) Using dynamics-based comparisons to predict nucleic acid binding sites in proteins: an application to OB-fold domains. Bioinformatics 25(15):1876–1883. doi:10.1093/bioinformatics/btp339

Zhang N, Li B-Q, Gao S, Ruan J-S, Cai Y-D (2012) Computational prediction and analysis of protein (gamma)-carboxylation sites based on a random forest method. Mol Biosyst 8:2946–2955