Predicting DNA-Binding Sites by Exploring the Distribution of Atom Groups around the Surface

Jing Hu[1] and Changhui Yan[2]
[1] Department of Mathematics and Computer Science, Franklin & Marshall College, Lancaster, PA, USA
[2] Department of Computer Science, North Dakota State University, Fargo, North Dakota, USA

Abstract - *DNA-binding proteins perform various functions in the cells. Determining the structures of protein-DNA complexes using experimental methods are hindered by many obstacles. Thus, computational methods for predicting DNA-binding sites on protein structures are needed to elucidate the mechanism of protein-DNA interactions. In this study, we divided atoms of amino acid residues into 14 groups and used a vector consisting of the distribution of these atom groups to describe the characteristics of protein surface around an amino acid. We then trained a Random Forest method to predict DNA-binding sites on protein surface. The predictions were then refined using a post-processing procedure based on the clustering of DNA-binding residues on the surface. The method achieved an accuracy of 80.8% when evaluated using 10-fold cross-validation. The results show that the distribution of different types of atoms around the surface provides sufficient structural information for predicting DNA-binding sites on protein structures.*

Keywords: Random Forest, DNA-binding, prediction, features

1 Introduction

Structural genomics projects are yielding an increasingly large number of protein structures with unknown function. As a result, computational methods for predicting functional sites on these structures are in urgent demand. There has been significant interest in developing computational methods for identifying amino acid residues that participate in protein-DNA interactions based on combinations of sequence, structure, evolutionary information, and chemical or physical properties. Some methods predict DNA-binding sites using protein sequence-derived information as input [1-3]. Compared to methods that make prediction based on protein structures, these methods have the advantage that they can be applied to proteins whose high-resolution structures are unavailable. However, they also suffer relatively low predicting performance. Thus, methods that can explore structural features to detect DNA-binding sites are also needed. For example, Jones et al. [4] analyzed residue patches on the surface of DNA-binding proteins and used electrostatic potentials of residues to predict DNA-binding sites. Later, they extended that method by including DNA-binding structural motifs [5]. In related studies, Tsuchiya et al. [6] used a structure-based method to identify protein-DNA binding sites based on electrostatic potentials and surface shape, and Keil et al. [7] trained a neural network classifier to identify patches likely to be DNA-binding sites based on physical and chemical properties of the patches. Neural network classifiers have also been used to identify protein-DNA interface residues based on a combination of sequence and structural information [8, 9]. Many recent studies have also been published [10-13].

Bagley and Altman [14] developed a FEATURE method to investigate the radial distributions of properties around protein sites like binding sites for calcium, the milieu of disulfide bridges, and the serine protease active site. Later, the method was also used to detect zinc-binding sites [15], phosphorylation sites [16], and peptide binding sites [17]. Using a similar approach, in this study, we investigated the distribution of atomic groups around the DNA-binding sites and trained a random forest method to predict DNA-binding sites on protein structures.

2 Materials and methods

2.1 Datasets

139 protein-DNA complexes were extracted from the PDB [18]. All the structures had resolution better than 3.0 Å and R factor less than 0.3. Each protein in this set had at least 40 amino acid residues and the mutual sequence similarity between the proteins in this set was less than 30%.

2.2 Definition of binding-site residues

Binding-site residues were defined based on atom distance [19]. A protein residue was defined to be a DNA-binding residue if the distance from any of its atoms to any atom of the interacting DNA was less than 5 Å. The 139 proteins had 26,862 residues in total and 5,932 of them were DNA-binding residues. A residue was defined to be a surface residue if its relative accessibility is at least 5% as calculated using NACCESS [20].

2.3 Microenvironmental features of DNA-binding sites on protein surface

We calculated the distance from nucleotides to protein surface. The average distance is 6 Å. Thus, for every surface residue, we define a sphere such that the center of the sphere is 6 Å from the protein surface and the line connecting the sphere center and the most exposed atom of the residue was perpendicular to the protein surface. Then we counted the number of different types of atoms from amino acids that fall into the sphere. The atoms were divided into 14 types as described in [17], namely: C3 (alphatic carbons; sp3), C= (carbonyl carbon; sp2), O= (carbonyl oxygen; sp2), N2H (nitrogen of amides; sp2; also sp2 neutral nitrogen of side chains), Car (aromatic carbon; sp2; general), O2- (negatively charged oxygens (-1/2) in carboxylates; sp2), SH (sulphur in thiols; sp3), OH (hydroxyl group; sp3), NarH (aromatic nitrogen with a hydrogen; sp2), NarH+ (aromatic nitrogen with a hydrogen and a postive charge; sp2), Set (sulphur in thioethers; sp3), C+ (carbon of carbocations; sp2), N3H+ (sp3 nitrogen with a hydrogen and a positive charge), N2H+ (sp2 nitrogen with a hydrogen and a positive charge). Thus, for each surface amino acid residue, a vector of 14 features was obtained. These vectors describe the structural characteristics on the protein surface centering at each surface amino acid. We used these vectors to train a classifier to classify surface residues into DNA-binding and non-DNA-binding classes based on these structural characteristics. Different radius values of the sphere were tested and the best result was achieved when the radius was 20 Å.

2.4 Classifier for predicting DNA-binding residues

We used a Random Forest (RF) method [21] to train a classifier to predict DNA-binding residues. A RF is a method consisting of an ensemble of tree-structured classifiers. It has been applied to solve many bioinformatics problems in recent years. In this study, we used the implementation of RF in WEKA package [22]. Ten fold cross-validations were used to evaluate the performance of the classifier. The proteins in the dataset were randomly split into 10 subsets. In each round of experiments, 9 subsets were used as training set to train a classifier, and the remaining subset was used as test set. This procedure was repeated 10 times with each subset being used as test set once. From a protein in the training set, the feature vectors associated with all binding-site residues were used as positive examples. We noticed that the sphere of a binding-site residue and that of a non-binding surface residue might overlap in space. Thus, to reduce noise in the training set, for the negative examples we only considered the surface residues whose spheres did not overlap with any spheres of binding-site residues. The feature vectors extracted from these residues were used as negative training examples. For a protein from the test set, all surface residues were used as test examples, so that a prediction was made for every surface residue.

2.5 Assessment of prediction performance

Prediction performance was evaluated using sensitivity, precision, accuracy (ACC), and Matthews' correlation coefficient (MCC):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{4}$$

where TP was the number of true positives (i.e., residues predicted to be DNA-binding residues that are in fact DNA-binding residues); TN was the number of true negatives (i.e., residues predicted to be non-DNA-binding residues that are in fact non-DNA-binding residues); FN was the number of false negatives (i.e., residues predicted to be non-DNA-binding residues that are in fact DNA-binding residues) and FP was the number of false positives (i.e., residues predicted to be DNA-binding residues that are in fact not interface residues). Sensitivity is a measure of the percentage of DNA-binding residues that are correctly predicted. Specificity is the fraction of non-DNA-binding residues that are correctly predicted. Accuracy is the percentage of overall predictions that are correct. MCC (Matthews correlation coefficient) measures the correlation between predictions and actual class labels, which is in the range of [-1, 1], with 1 denoting perfect predictions and -1 denoting that every example is incorrectly predicted. In a two-class classification, if the numbers of examples of the two classes are not equal, MCC is a better measure than accuracy [23].

3 Results

3.1 Identification of DNA-binding residues by the Random Forest method

A Random Forest (RF) classifier was trained to predict whether a surface residue is DNA-binding residue based on the feature vector associated with its surrounding sphere. 10-fold cross-validation was used to evaluate the performance of the classifier. Table 1 (column 2) shows that the classifier achieved an overall accuracy of 67.3% with a MCC of 0.2, and 57.9% of DNA-binding residues and 69% of non-DNA-binding residues are correctly identified.

3.2 Post-processing of prediction results

A visualization of the DNA-binding sites revealed that DNA-binding residues clustered on protein surface to form a contiguous patch. Thus, the predicted DNA-binding residues were also expected to form a patch on the surface. However, when we analyzed the prediction results by RF, we found that that some predicted DNA-binding residues were isolated on protein surface, and in some cases, the predicted DNA-binding sites form multiple small patches on the surface. Thus, we designed a post-processing procedure to remove isolated predictions and merge small patches into a large one. For a surface residue that was predicted to be DNA-binding residue, if less than 2 of its neighboring surface residues were predicted to be DNA-binding, then we changed its prediction to non-DNA-binding. For a surface residue that was predicted to be non-DNA-binding, if more than 60% of its neighboring residues were predicted to be DNA binding, then we changed its prediction to DNA binding. After the post-processing (Table 1, column 3), the prediction performance was improved to an overall accuracy of 73.5% with a MCC of 0.26, and 57.2% of DNA-binding residues and 76.0% of non-DNA-binding residues are correctly identified. Compared this with the results without post-processing, we can see that the post-processing improve accuracy, MCC, and precision at only little cost of sensitivity.

3.3 Relaxation of prediction results after post-processing

In this study, the DNA-binding residues were defined based on their distance to the binding DNA using a cutoff chosen in a previous study [19]. However, different cutoff values had been used in many other studies. In our study, the majority of the false positive predictions were very close to the observed DNA-binding residues (either being the direct neighbor of a DNA-binding residue or separated from the DNA-binding sites by only one residue). Some of these false positive predictions could have been counted as true positives if a different cutoff value was used. To account for the uncertainty in the cutoff value, we re-evaluated the performance by relaxing the criterion of true positive as in [9]. With the relaxed criterion when a surface residue was predicted to be a DNA-binding residue, the prediction is considered a true positive prediction if (1) the surface residue was indeed a DNA-binding residue, or (2) it was a direct neighbor (on the protein surface) of a DNA-binding residue. After the relaxation (Table 1, column 4), the prediction had an accuracy of 80.8%, with 0.50 MCC, 71.5% sensitivity and 80.8% precision.

Table 1. Prediction performances of the proposed method

	Random Forest[1]	Post-processing[2]	Relaxation[3]
Sensitivity (%)	57.9	57.2	71.5
Specificity (%)	69.0	76.0	83.5
ACC (%)	67.3	73.5	80.8
MCC	0.20	0.26	0.50

[1]Predictions by the Random Forest method. [2]Predictions from the Random Forest method were processed using the post-processing procedure. [3]A relaxed criterion of true positive was used to evaluate the performance.

4 Conclusions

In this study, we used vectors consisting of the distribution of atom groups to describe the characteristics of protein surface and used them to train a RF method to predict DNA-binding residues. A post-processing procedure was used to refine the predictions based on the distribution of DNA-binding residues over the protein surface. After the post-processing, the predicted DNA-binding sites form a contiguous path on the protein surface. The accuracy of the method reached 80.8% based on a relaxed criterion. The results confirmed that the distribution of atom groups on the protein surface provided useful structural information for predicting DNA-binding sites.

5 References

[1] Yan, C., et al., Identifying amino acid residues involved in protein-DNA interactions from sequence. BMC Bioinformatics, 2006. **7**: p. 262.

[2] Ahmad, S. and A. Sarai, PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics, 2005. **6**(1): p. 33.

[3] Wang, L. and S.J. Brown, BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucl Acids Res, 2006. **34**: p. W243-W248.

[4] Jones, S., et al., Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucl Acids Res, 2003. **31**(24): p. 7189-7198.

[5] Shanahan, H.P., et al., Identifying DNA-binding proteins using structural motifs and the electrostatic potential. Nucl Acids Res, 2004. **32**(16): p. 4732-4741.

[6] Tsuchiya, Y., K. Kinoshita, and H. Nakamura, Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. Proteins, 2004. **55**(4): p. 885-894.

[7] Keil, M., T. Exner, and J. Brickmann, Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a neural network. J Comput Chem, 2004. **25**(6): p. 779-789.

[8] Ahmad, S., M.M. Gromiha, and A. Sarai, Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics, 2004. **20**(4): p. 477-486.

[9] Tjong, H. and H.-X. Zhou, DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. Nucl. Acids Res., 2007. **35**(5): p. 1465-1477.

[10] Xiong, Y., J. Liu, and D.-Q. Wei, An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins: Structure, Function, and Bioinformatics, 2011. **79**(2): p. 509-517.

[11] Cai, Y., et al., A novel sequence-based method of predicting protein DNA-binding residues, using a machine learning approach. Molecules and Cells, 2010. **30**(2): p. 99-105.

[12] Alibés, A., L. Serrano, and A. Nadra, Structure-based DNA-binding prediction and design, in Methods Mol Biol. 2010. p. 77-88.

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

[14] Bagley, S.C. and R.B. Altman, Characterizing the microenvironment surrounding protein sites. Protein Science, 1995. **4**(4): p. 622-635.

[15] Ebert, J.C. and R.B. Altman, Robust recognition of zinc binding sites in proteins. Protein Science, 2008. **17**(1): p. 54-56.

[16] Fan, S. and X. Zhang, Characterizing the microenvironment surrounding phosphorylated protein sites. Genomics Proteomics Bioinformatics, 2005. **3**(4): p. 213-217.

[17] Petsalaki, E., et al., Accurate Prediction of Peptide Binding Sites on Protein Surfaces. PLoS Comput Biol, 2009. **5**(3): p. e1000335.

[18] Berman, H.M., et al., The Protein Data Bank. Nucl Acids Res, 2000. **28**(1): p. 235-242.

[19] Ofran, Y. and B. Rost, Analysing six types of protein-protein interfaces. J. Mol. Biol., 2003. **325**(2): p. 377-387.

[20] Hubbard, S.J., NACCESS. 1993, Department of Biochemistry and Molecular Biology, University College, London.

[21] Breiman, L. RF/tools: A class of two-eyed algorithms. in SIAM workshop. 2003.

[22] Witten, I.H. and E. Frank, Data mining: practical machine learning tools and techniques with Java implements. 1999, San Mateo, CA: Morgan Kaufmann.

[23] Baldi, P., et al., Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics, 2000. **16**: p. 412-424.