doi:10.1093/nar/gkt544

Novel approach for selecting the best predictor for identifying the binding sites in DNA binding proteins

R. Nagarajan[1], Shandar Ahmad[2] and M. Michael Gromiha[1,*]

[1]Department of Biotechnology, Indian Institute of Technology Madras, Chennai 600036, India and [2]National Institute of Biomedical Innovation, Osaka, Japan

Received April 17, 2013; Revised May 20, 2013; Accepted May 25, 2013

ABSTRACT

Protein–DNA complexes play vital roles in many cellular processes by the interactions of amino acids with DNA. Several computational methods have been developed for predicting the interacting residues in DNA-binding proteins using sequence and/or structural information. These methods showed different levels of accuracies, which may depend on the choice of data sets used in training, the feature sets selected for developing a predictive model, the ability of the models to capture information useful for prediction or a combination of these factors. In many cases, different methods are likely to produce similar results, whereas in others, the predictors may return contradictory predictions. In this situation, *a priori* estimates of prediction performance applicable to the system being investigated would be helpful for biologists to choose the best method for designing their experiments. In this work, we have constructed unbiased, stringent and diverse data sets for DNA-binding proteins based on various biologically relevant considerations: (i) seven structural classes, (ii) 86 folds, (iii) 106 superfamilies, (iv) 194 families, (v) 15 binding motifs, (vi) single/double-stranded DNA, (vii) DNA conformation (A, B, Z, etc.), (viii) three functions and (ix) disordered regions. These data sets were culled as non-redundant with sequence identities of 25 and 40% and used to evaluate the performance of 11 different methods in which online services or standalone programs are available. We observed that the best performing methods for each of the data sets showed significant biases toward the data sets selected for their benchmark. Our analysis revealed important data set features, which could be used to estimate these context-specific biases and hence suggest the best method to be used for a given problem. We have developed a web server, which considers these features on demand and displays the best method that the investigator should use. The web server is freely available at http://www.biotech.iitm.ac.in/DNA-protein/. Further, we have grouped the methods based on their complexity and analyzed the performance. The information gained in this work could be effectively used to select the best method for designing experiments.

INTRODUCTION

Protein–DNA interactions play vital roles in several biological processes, including gene regulation, DNA repair, DNA replication and DNA packaging. The knowledge about DNA-binding residues and binding specificity would help to understand the recognition mechanism of protein–DNA complexes. The availability of experimental data on binding specificity (1) and 3D structures of protein–DNA complexes (2) encouraged researchers to reveal important factors for understanding protein–DNA recognition. The analysis has been focused on different directions such as amino acid properties, conservation of residues, contribution of non-covalent interactions and conformational changes of DNA (3–26). The importance of hydrogen bonds, electrostatic, hydrophobic and van der Waals interactions along with weak interactions including cation-π has been stressed by several investigators in the field (12,14,19,24,27–32). The contributions of energetic terms along with physical and chemical features have been used to understand the recognition mechanism of protein–DNA complexes. Furthermore, knowledge-based statistical potentials have been derived using atomic contacts between protein and

*To whom correspondence should be addressed. Tel: +91 44 2257 4138; Fax: +91 44 2257 4102; Email: gromiha@iitm.ac.in

DNA, and these potentials have been used to predict the binding specificity of protein–DNA complexes (33,34). Gromiha *et al.* (2004) combined both inter and intra-molecular interactions for understanding the recognition mechanism.

On the other hand, owing to the exponential increase in the gap between the available sequences and structures of DNA-binding proteins in Uniprot (35) and Protein Data Bank (2), several methods have been proposed to identify the binding site residues just from amino acid sequences. These methods are based on amino acid frequency, evolutionary profile, sequence conservation, predicted secondary structure and solvent accessibility, electrostatic potential, hydrophobicity, position-specific scoring matrix by using various machine learning methods such as support vector machine, neural network, Naïve Bayes classifier and random forest (33,36–51). Careful inspection of these methods revealed the fact that they are applicable to specific types of proteins, and the performance of each method varies drastically in the range of 20–90%. This situation leads confusions to the biologists for selecting the best method to identify the binding sites for designing their experiments. Hence, it is essential and important to reveal the applications and predictive ability of existing predictors to specific data sets based on various properties of the query protein.

In this work, we have systematically categorized the protein–DNA complexes into several groups based on the structure of the protein, structure of the DNA, binding motif and function. The complexes in each category have been divided into several sub-categories using known annotations in structural and functional databases. On the other hand, we have collected all the prediction servers, which have either online services or available standalone programs. We have developed necessary in-house programs to analyze the results obtained with each method using nine types of data sets. We noticed that no method is uniformly predicting the binding sites at high accuracy in all the data sets. This is applicable to the most recently developed methods with tuned parameters, efficient techniques and large data set as well as the earliest methods reported in the literature. We have related the performance of each method with different data sets and revealed the correspondence between them. These results would help the biologists to select the best method to design their experiments rather than choosing any specific method arbitrarily or a combination of methods. In addition, the present study explores the necessity of refining/developing bioinformatics tools to improve the performance in specific categories of DNA-binding proteins. Specific examples for the best and worst performance of methods in selected categories of data sets will be discussed.

MATERIALS AND METHODS

Data sets

We have collected all the protein–DNA complexes (2317 entries) deposited in Protein Data Bank (last accessed on 16 May 2012). These complexes were classified into four broader categories based on (i) protein structure, (ii) DNA structure, (iii) binding motif and (iv) protein function as described later in the text. All the data sets have been culled with the sequence identities of <25 and 40%. We obtained similar results, and the data with the cutoff of <25% sequence identity are presented in this article.

Classification based on protein structure

We have used the SCOP database (52) for structural classification of proteins based on their structural classes, folding types, superfamilies and families. Our final data set contains 260 protein chains from seven classes, 86 folds, 106 superfamilies and 194 families with the sequence identity of <25%.

Further, we have identified the disordered regions by comparing the structures of proteins in free and complex forms and analyzed the performance of different methods in disordered regions.

Classification based on DNA structure

We have classified the protein–DNA complexes based on DNA structure on two aspects: (i) DNA conformation such as A, B, Z, RH and U and (ii) type of DNA (single stranded, double stranded and palindrome and double stranded and non-palindrome). The conformation of DNA has been obtained from Nucleic acid database (NDB) (53). The databases PDB, NDB and PDIdb (54) have been used to get the information on double/single-stranded DNA and palindrome/non-palindrome DNA. The final data set contains 283 and 301 protein chains based on DNA conformation and type, respectively.

Motif-based classification

The binding motif is considered to be an important factor for identifying the binding sites (55). Hence, we classified the protein–DNA complexes based on their binding motifs, and the major ones are helix-turn-helix, β-barrel and β-ribbon. We obtained the motif information from different databases such as ProNuc (56), PDIdb (54) and Biomolecules gallery (http://gibk26.bio.kyutech.ac.jp/jouhou/image/dna-protein/all/all.html). The final data set contains 69 chains from 15 motifs. We noticed that several complexes are listed under enzymes, which are considered in the classification based on functions.

Functional classification of protein–DNA complexes

We have classified the protein–DNA complexes based on their functions such as enzymes, regulatory proteins and structural proteins. The functional information has been obtained from NDB. The final data set contains 126 enzymes, 149 regulatory proteins and 19 structural proteins with the sequence identity of <25%.

Methods for predicting the binding sites in DNA-binding proteins

We have collected all the available methods for predicting the binding sites in DNA-binding proteins from amino acid sequence, which have either online services or available standalone programs (57). The methods are BindN (39), BindN+ (47), BindN-RF (46), DBS-Pred (37), DBS-PSSM (38), DNABindR (49), DP-Bind with three categories, binary, BLOSUM and PSSM encoding (48), metaDBSite (51) and NAPS (50). The details about the name, features, technique, reference and link for the methods used in the present work are listed in Supplementary Table S1. These methods used different data sets and accuracies reported by the authors are in the range of 70–80%.

Identification of DNA-binding residues

Several criteria have been proposed to identify the DNA-binding sites such as the distance between contacting atoms in protein and DNA (37), reduction in solvent accessibility on binding (58) and interaction energy between protein and DNA (21). Most of the prediction methods analyzed in this work used the distance based criteria for identifying the binding sites. In this approach, a residue in a DNA-binding protein is identified as binding if the distance between any of its heavy atoms and a heavy atom in DNA is ≤3.5 Å. We have identified the binding sites using the same conditions in all the considered protein–DNA complexes.

Assessing the performance of prediction methods

We have assessed the performance of different methods using the measures, sensitivity, specificity, accuracy and Matthews correlation coefficient (MCC). Sensitivity shows the correct prediction of DNA-binding residues, specificity reveals the ability of excluding non-binding residues and accuracy provides the overall performance (59).

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN}) \qquad \textbf{(1)}$$

$$\text{Specificity} = \text{TN}/(\text{TN}+\text{FP}) \qquad \textbf{(2)}$$

$$\text{Accuracy1} = (\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN}) \qquad \textbf{(3)}$$

$$\text{Accuracy2} = (\text{sensitivity}+\text{specificity})/2 \qquad \textbf{(4)}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \qquad \textbf{(5)}$$

In these equations, TP (binding residues predicted as binding), TN (non-binding residues predicted as non binding), FP (non-binding residues predicted as binding) and FN (binding residues predicted as non binding) represent, true positives, true negatives, false positives and false negatives, respectively.

RESULTS AND DISCUSSIONS

We have assessed the performance of all the available methods in different sets of data as described in the 'Materials and Methods' section.

Structural classes

Protein–DNA complexes have been classified into seven structural classes such as all-α, all-β, $\alpha+\beta$, α/β, multi-domain, coiled coil and small proteins. The accuracies obtained with all the 11 considered methods in these data sets are presented in Table 1. From Table 1, we noticed that the performance of a method depends on the structural class. Most of the methods predict well in all-α proteins where as the performance is poor in all-β class of proteins. This trend is similar to protein secondary structure prediction that all-α class proteins are better predicted than all-β class proteins (60). The binding sites in coiled coil proteins are predicted well in most of the methods. The comparison of different methods showed that BindN-RF has the best performance in all-α and all-β proteins. The sensitivity of BindN-RF and DP-Bind_BLOSUM is <60%, although the overall accuracies of these methods are more than metaDBSite. Further, none of the method showed the sensitivity of more than 59% in multi-domain proteins. This might be due to the size of the protein, and the binding site residues are <2%. Hence, we have separated domains in these proteins using SBASE (61) and predicted the binding sites in DNA-binding domain. We observed that the methods DP-Bind_PSSM and NAPS could predict the binding sites with >60% sensitivity, specificity and accuracy. Further, we have evaluated the performance of different methods using MCC, and the results are presented in Supplementary Table S2. We noticed that the trend is similar to that reported using the measure, accuracy.

Folds, superfamilies and families

The classification of protein–DNA complexes based on their structures showed that they are distributed in 86 different folding types, 106 superfamilies and 194 families. We

Table 1. Prediction accuracy of binding sites in different classes

Methods	Average Accuracy	all-α	all-β	$\alpha+\beta$	α/β	Coiled coil	Multidomain	Small proteins
BindN	64.2 (74.9)	66.2 (76.3)	62.1 (74.6)	60.3 (74.8)	62.2 (79.6)	75.1 (73.2)	61.3 (79.7)	62.4 (66.5)
BindN+	71.1 (82.8)	76.2 (83.8)	66.0 (83.7)	67.9 (81.9)	66.5 (85.8)	**88.4** (86.9)	65.5 (87.3)	66.9 (70.2)
BindN-RF	71.9 (82.3)	**76.4** (83.7)	**68.0** (82.7)	68.4 (82.8)	67.8 (84.8)	88.1 (86.5)	65.8 (86.6)	**68.5** (69.2)
DBS-Pred	64.3 (72.6)	64.2 (73.0)	62.6 (71.6)	62.0 (71.7)	63.1 (75.4)	74.4 (73.6)	59.8 (76.4)	63.6 (66.5)
DBS-PSSM	70.2 (78.5)	73.2 (80.2)	65.5 (78.2)	65.8 (76.5)	67.1 (83.3)	87.4 (81.6)	65.3 (87.0)	67.3 (62.3)
DP-Bind_Binary	66.9 (68.0)	68.1 68.6)	63.5 (65.8)	63.1 (67.2)	66.3 (70.6)	79.8 (70.4)	62.0 (70.7)	65.4 (62.9)
DP-Bind_BLOSUM	66.1 (67.8)	69.2 (69.5)	63.2 (66.3)	62.8 (67.8)	66.3 (71.5)	75.6 (66.6)	61.1 (70.5)	65.0 (62.4)
DP-Bind_PSSM	72.1 (76.4)	73.7 (78.4)	67.9 (75.3)	**69.6** (76.6)	70.4 (80.5)	88.1 (84.6)	**69.9** (79.4)	64.8 (56.8)
DNABindR	68.0 (71.9)	70.1 (72.9)	62.6 (68.1)	65.2 (71.0)	66.2 (75.2)	82.9 (77.3)	64.2 (77.1)	64.4 (61.7)
metaDBSite	69.9 (72.3)	72.0 (74.1)	66.9 (70.2)	67.2 (71.5)	**69.2** (76.6)	82.0 (74.0)	65.4 (76.6)	66.5 (62.9)
NAPS	63.6 (65.1)	64.6 (64.8)	58.8 (61.3)	59.4 (62.5)	57.6 (66.9)	80.6 (75.0)	62.5 (67.9)	61.6 (57.6)

Accuracies obtained with Equation (3) are given in parentheses. The highest accuracy in each class is shown in bold.

have analyzed the performance of all the 11 prediction methods in all folds, superfamilies and families, and the summarized results are presented in Figure 1 and Supplementary Table S3. DP-Bind_PSSM showed the best performance in >20% of the folds/superfamilies. However, the accuracy of this method is <60% in 13 of the 86 considered folds. BindN-RF scored the highest rank in the classification of families. Methods such as DBS-PSSM and BindN+ predicted the binding sites with topmost accuracy in 10–20% of the considered 186 DNA-binding proteins. Interestingly, one of the earliest prediction methods DBS-Pred (37) also showed the best performance in four folds, four superfamiles and three families. These results showed that the prediction methods are complimenting each other in different types of DNA-binding proteins. It is essential to reveal the best method in specific type of proteins for practical applications.

We have systematically analyzed the correspondence between the structure of the complex and prediction performance, and the methods showing highest and lowest accuracies for identifying the binding sites in 86 folds, 106 superfamilies and 194 families are listed in Supplementary Table S4. Few typical examples for the best and worst predicted folds along with their performances are presented in Table 2. BindN+, DBS-Pred and DP-Bind_PSSM showed the best performance in profilin-like, tetracyclin repressor-like and transcription factor IIA types of folds, respectively. The predicted accuracies are >90% based on the average between sensitivity and specificity. On the other hand, other methods showed a poor performance in these folds with the accuracy in the range of 50–70%. Further, the accuracies of several folds are <70%, and three typical examples are listed in Table 2. The best method showed the accuracy of 57% in Retrovirus zinc finger-like domain fold. The sensitivity and specificity are 70.2 and 43.0%, respectively. In addition, PUA domain-like and HLH-like folds showed the accuracy of 61.1 and 61.2%, respectively. These results indicate the requirement of methods to be applicable to folds in which the binding sites are poorly predicted.

The best predicted superfamilies and their performance are included in Table 2. We observed that pheromone binding and dimeric $\alpha + \beta$ barrel are predicted with the accuracy of >90% where as the lowest accuracies are 65 and 47%, respectively. The binding sites in eukaryotic transcription factors are predicted well with all the methods, and the highest and lowest accuracies are 94 and 72%, respectively. The worst predicted superfamiles are chromo-domain-like, immunoglobulin and RNase A-line with the highest accuracy of ~60% (Table 2). Interestingly, the binding site residues in chromo-domain superfamily are predicted with high specificity, whereas other two superfamilies identify the binding residues with high sensitivity. This suggests that the interface residues in these domains may consist of a small number of residues with strong binding signal, which remain unchanged across the family, whereas there are other residues, which show diversity, and their binding is not directly predicted from sequence features alone.

We observed similar tendency in the classification of families. BindN-RF predicted the binding sites in AraC type transcriptional activator with the accuracy of 99.5%; the sensitivity and specificity are 100 and 99%, respectively. The binding sites in CopG and Z-DNA-binding domain are predicted with >90% accuracy by BindN and DBS-PSSM, respectively.

This analysis revealed that although newly developed methods included several features, fine tuning of parameters and large data set, which showed excellent performance over other methods, simpler methods reported earlier than others may outperform more complex methods on some systems, and hence their availability should be made use of predictions.

Disordered regions

We have analyzed the performance of different methods in disordered regions of 73 protein chains. The results are presented in Table 3. We observed that the methods, BindN-RF and DP-Bind_PSSM, which showed high accuracy in different structures classes (Table 1), have

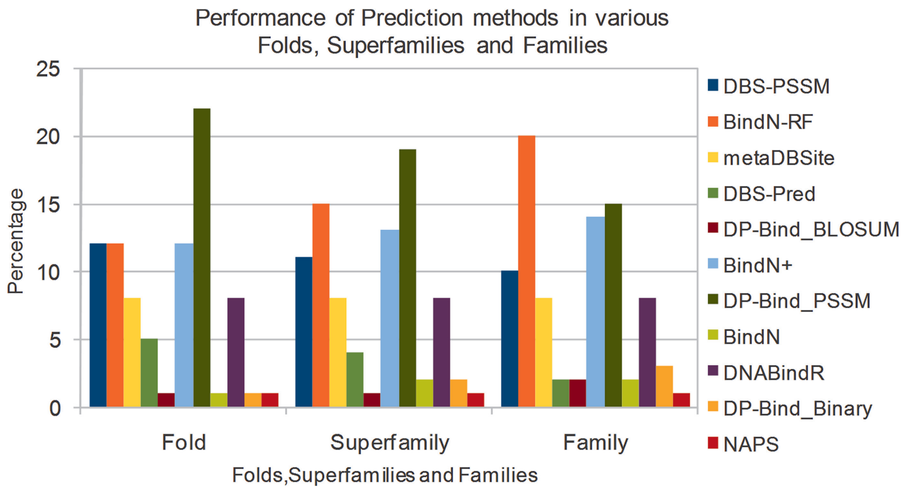

Figure 1. Performance of DNA-binding site prediction methods in various folds, superfamilies and families.

Table 2. Typical examples of best and worst predicted folds, superfamilies and families

Fold/Superfamily/Family	Method	Sensitivity	Specificity	Accuracy1	Accuracy2	MCC	Lowest Accuracy	MCC
Fold								
Profilin-like (1)	BindN+	100.0	96.4	96.6	98.2	0.32	64.5 (DP-Bind_BLOSUM)	0.20
Tetracyclin repressor-like, C terminal domain (2)	DP-Bind_PSSM	96.2	89.6	89.8	92.9	0.28	51.1 (DP-Bind_BLOSUM)	0.20
Transcription factor IIA(TFIIA), beta-barrel domain (2)	DBS-Pred	100.0	80.4	82.0	90.2	0.16	67.9 (NAPS)	0.13
HLH-like (1)	*BindN+*	*40.0*	*82.4*	*72.7*	*61.2*	*0.38*	*49.6 (NAPS)*	*0.16*
PUA domain-like (1)	*DNABindR*	*75.0*	*47.3*	*51.0*	*61.1*	*0.21*	*54.4 (NAPS)*	*0.13*
Retrovirus zinc finger -like domains (2)	*DP-Bind_PSSM*	*70.2*	*43.0*	*54.7*	*56.6*	*0.29*	*46.5 (DNABindR)*	*0.22*
Superfamily								
Pheromone-binding, quourm-sensing transcription factors (1)	BindN+	100.0	96.4	96.6	98.2	0.31	64.5 (DP-Bind_BLOSUM)	0.20
Dimeric alpha + beta barrel (1)	BindN-RF	87.5	96.4	95.9	92.0	0.34	47.3 (DBS-Pred)	0.17
DNA-binding domain-eukaryotic transcription factors (1)	DBS-PSSM	100.0	88.5	90.5	94.3	0.28	72.3 (DBS-Pred)	0.20
Chromo domain-like (1)	*DBS-Pred*	*27.8*	*73.9*	*60.9*	*50.8*	*0.20*	*42.2 (NAPS)*	*0.14*
Immunoglobin (3)	*DBS-PSSM*	*77.8*	*60.0*	*60.4*	*68.9*	*0.29*	*37.4 (NAPS)*	*0.13*
RNase A-like (1)	*DP-Bind_PSSM*	*71.4*	*47.0*	*48.4*	*59.2*	*0.29*	*34.9 (BindN)*	*0.18*
Family								
AraC type transcriptional activator (1)	BindN-RF	100.0	99.0	99.1	99.5	0.32	65.9 (DBS-Pred)	0.19
CopG-like (1)	BindN	100.0	81.1	83.7	90.5	0.22	78.1 (BindN-RF)	0.20
Z-DNA binding domain (1)	DBS-PSSM	100.0	81.1	82.5	90.6	0.26	47.4 (DP-Bind_Binary)	0.19
T7 RNA polymerase (1)	*DP-Bind_PSSM*	*50.0*	*88.0*	*86.2*	*69.0*	*0.28*	*58.5 (NAPS)*	*0.13*
RecA protein-like (ATPase-domain) (1)	*BindN-RF*	*33.3*	*87.7*	*86.5*	*60.5*	*0.33*	*44.1 (DNABindR)*	*0.23*
SRA domain-like (1)	*DNABindR*	*75.0*	*47.3*	*51.0*	*61.1*	*0.23*	*54.4 (NAPS)*	*0.13*

The worst predicted folds/superfamilies/families are shown in italics.

Table 3. Prediction performance of binding sites in disordered regions

Method	Sensitivity	Specificity	Accuracy1	Accuracy2	MCC
DBS-Pred	61.3	60.7	60.8	61.0	0.17
BindN	55.5	67.5	65.2	61.5	0.19
BindN+	61.3	64.6	64.0	63.0	0.21
BindN-RF	55.5	68.3	65.9	61.9	0.19
DP-Bind_Binary	78.1	48.4	54.0	63.3	0.21
DP-Bind_BLOSUM	73.0	50.3	54.5	61.6	0.18
DP-Bind_PSSM	65.7	56.4	60.6	61.0	0.20
NAPS	59.1	58.9	58.9	59.0	0.14
DNABindR	75.9	51.9	56.4	63.9	0.22
metaDBSite	73.0	56.0	59.2	64.5	0.23
DBS-PSSM	65.0	61.1	61.8	63.0	0.20

less sensitivity and specificity, respectively in disordered regions. The overall accuracy also reduced to 62%. On the other hand, DBS-Pred maintained the accuracy of 61% for disordered regions. The accuracy obtained with different methods given in Table 3 showed the necessity of developing new methods for predicting the binding sites in disordered regions.

Motifs

We have grouped the protein–DNA complexes into 15 different motifs, which have the representation of 1–30 complexes. The best performance of each method in all the motifs is shown in Supplementary Table S5. In this Table, we have also included the number of motifs, sensitivity, specificity and accuracy. We noticed that BindN+ performed the best in alpha/beta, beta sheet and helix-loop-helix motifs. On the other hand, the performance is poor in Zalpha motif. BindN-RF showed the best performance in 9 of the 15 considered motifs. DBS-PSSM is ranked as the first in the ribbon-helix-helix and Zalpha motifs.

We have analyzed the performance of each method in all these motifs with the condition that the sensitivity and specificity are >60%, and the results are shown in Figure 2. We observed that all the methods performed well at least 2 of the 15 considered motifs. BindN-RF showed the best performances in 12 of 15 motifs followed by BindN+ (10/15). DBS-PSSM, DNABindR and metaDBSite showed the sensitivity and specificity of >70% in 5–8 motifs.

Type of DNA

We have classified the protein–DNA complexes based on three types of DNA such as single-stranded, double-stranded and palindrome, and double-stranded and non-palindrome DNA. We observed that the performance is poor for all the methods to predict the binding sites

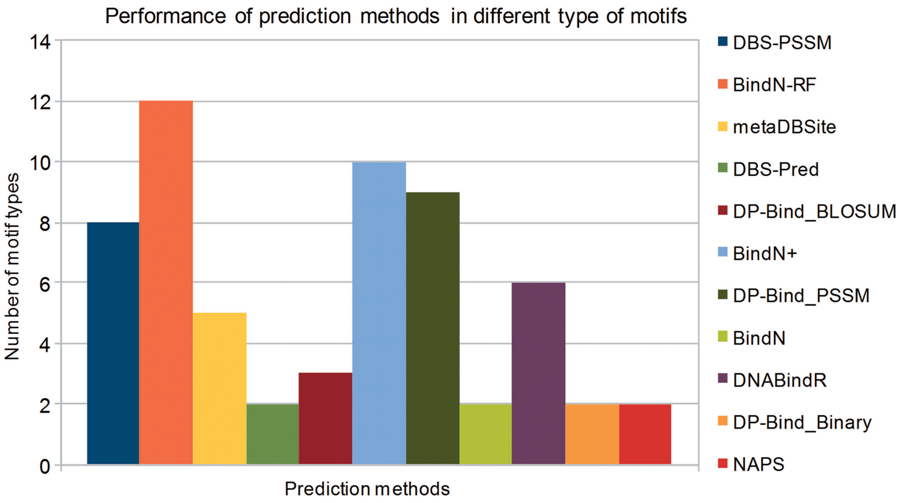

Figure 2. Performance of prediction methods in 15 different types of DNA binding motifs. Number of motifs, which are predicted with the sensitivity and specificity of >60% each in all considered methods are shown.

when the DNA is of single strand. The highest accuracy is 61.5% with the sensitivity and specificity of 49.9 and 73%, respectively, obtained for the method DP-Bind_PSSM. The binding sites with double-stranded DNA are predicted with >70% accuracy in both palindrome and non-palindrome cases. Further, the performance with double-stranded palindrome DNA–protein complexes is better than that with non-palindrome DNA. The accuracies are 71 and 76%, respectively. This result is understandable because many more double-stranded DNA-binding proteins have been solved and hence included in training sets than those binding to single-strand DNA. For example, the first published method for predicting DNA-binding sites (DBS-Pred) used only dsDNA-binding proteins for training the model.

DNA conformation

We have collected the DNA conformation details from NDB and accordingly classified the considered protein–DNA complexes. Majority of the DNA have the conformation of B-type. The prediction method, DP-Bind_PSSM showed the highest accuracy of 71% to predict the binding sites. The RH and Z-DNA types are predicted with the accuracy of 71%. Supplementary Table S6 shows the performance in the complexes with different types of DNA.

Functional classification of protein–DNA complexes

We have classified the protein–DNA complexes based on their functions and are mainly under three categories, namely, enzymes, regulatory and structural proteins.

The enzymes are classified into 17 groups, which have one to 52 protein–DNA complexes. The sensitivity, specificity and accuracy of the best methods in each group of enzymes are presented in Supplementary Table S7. We noticed that none of the prediction methods worked well in 13 of the 17 groups. Only four groups of enzymes, kinase, phosphatase, recombinase invertase and recombinase resolvase are predicted well with the accuracy of >80%. DNA endonuclease is a major group of enzymes with 52 complexes, and the prediction accuracy is 71% with the sensitivity of 63% and specificity of 78%. For the class of rare enzymes with only one complex, the accuracy varies from poor to good. The excellent performance of several methods in these enzymes might be due to the presence of these proteins in the training set of their respective methods. In contrast, DNA reverse transcriptase has two proteins, and the performance is poor in all the methods; the highest accuracy is 59.4% with the sensitivity of 39.4%. DNA polymerase with 17 samples is predicted poorly with the accuracy of 66%.

Regulatory proteins are classified into 13 groups with 149 chains and the accuracy of different methods lies in the range of 60–80% (Supplementary Table S7). Further inspection of Supplementary Table S7 shows that few classes of regulatory proteins such as DNA repair repressor, transcription factor co-activator and transcription factor termination have poor performance to identify the binding site residues with high sensitivity; the sensitivity is 33–46%.

Considering the structural proteins, the average accuracy is in the range of 65–78% for the 19 DNA-binding proteins in this data set. In this group of proteins, we noticed a balance between sensitivity and specificity in most of the methods. Further, one of the poorly performed methods, NAPS showed the best performance in viral coat protein.

General trends on different prediction methods

In addition, we have evaluated the performance of different prediction methods using two independent sets of test data: (i) using the protein–DNA complex structures deposited recently (since June 2011) and (ii) the structures, which were not used in individual methods for developing the respective algorithm. The results obtained with these two sets of data are presented in Table 4. We observed that the balance between sensitivity and specificity lies in the range of 60–70% in most of the methods for both the data sets. However, the accuracy is >75% in

Table 4. Prediction performance of different methods in two independent data sets

Method	Data set 1			Data set 2		
	Accuracy1	Accuracy2	MCC	Accuracy1	Accuracy2	MCC
BindN	76.1	63.1	0.17	76.4	61.4	0.14
BindN+	80.2	69.2	0.28	79.6	68.7	0.26
BindN-RF	78.0	69.5	0.28	75.3	68.7	0.24
DBS-Pred	72.6	62.4	0.16	72.8	62.2	0.14
DBS-PSSM	78.3	66.5	0.25	78.4	69.7	0.23
NAPS	63.5	60.2	0.13	64.8	60.3	0.12
DNABindR	71.6	66.3	0.21	72.1	66.7	0.20
metaDBSite	74.7	68.7	0.24	78.2	66.2	0.22
DP-Bind_Binary	67.9	65.9	0.19	68.6	67.7	0.19
DP-Bind_BLOSUM	68.4	66.1	0.19	67.3	65.4	0.17
DP-Bind_PSSM	75.9	70.3	0.27	77.7	70.0	0.25

Data set 1: List of DNA–protein complexes analyzed in this work and not used in the respective methods.
Data set 2: List of DNA–protein complexes published from June 2011, after the publication of all the analyzed methods.

several methods, when the accuracy was evaluated using Equation (3), which shows the ability of different methods for either correctly predicting the binding sites or excluding non-binding sites. The data presented in this work based on different categories of data sets would be a valuable resource for the biologists to select the best method for their target DNA–binding protein.

Comparison between the best predicted method and combination of methods

The method, metaDBSite, combined six different methods and developed a prediction system for identifying the binding sites in DNA-binding proteins. We have compared the performance of metaDBSite with the best predicted method in different groups of DNA-binding proteins and the results are presented in Supplementary Table S8. We noticed that among 86 folds, metaDBSite performed the best only in six folds. Similar trend is observed in all the nine classification of data sets. This analysis emphasizes the importance of the present method over combination of different methods. In addition, we have estimated the difference in accuracy between the best method and metaDBSite, and we noticed an improved accuracy of up to 54% in all the DNA-binding proteins and the average accuracy is 9.6%. We have also carried out an ensemble-based prediction based on the majority of voting of the 11 methods used in this work, and we observed a similar trend that we obtained with metaDBSite predictor.

Grouping of methods based on their complexities

We have combined the methods into three groups based on their complexities such as (i) additive feature models (models which treat each input feature independent of the other), (ii) complex feature models (which use non-additive combination of features) without using PSSM and (iii) complex feature models using PSSM. The performance of these three groups of models was analyzed in all the considered data sets, and the results are presented in Supplementary Table S9. The results showed that the performance of additive feature models is similar to complex feature models without using PSSM. The complex feature models, which use PSSM, showed better performance for identifying the binding sites in most of the classes. However, the performance of these models to identify the binding sites of disordered regions was poor.

Applications

The insights obtained in the present work have several applications, and some of them are discussed later in the text. (i) For a protein with known structure and without the information of the complex, one can get all the structural information such as class, family, superfamily and so forth. In this case, depending on the type of the protein-specific method can be used to identify the binding sites, and the results will be reliable for designing experiments. (ii) Currently, protein secondary structure prediction are reported to show the accuracy of close to 85%, and structural class can be predicted with the accuracy of >95%. On a large scale analysis, it is possible to predict the structural class and apply suitable method to identify the binding sites. For example, the correct prediction of structural classes would predict the binding sites with the higher accuracy than the average accuracy of best methods reported in the literature. (iii) For a specific protein, it is possible to obtain the structural information using homology modeling or *ab initio* structure-prediction methods with reasonable accuracy. For selecting the best prediction method, the modeled structure would be sufficient to obtain the necessary structural information. The binding sites can be predicted by selecting the respective method based on structural information, which will be reliable for designing experiments. In addition, other information reported in this work can also be combined to get the desired information.

The data presented in Table 1 suggested that BindN+, BindN-RF and DP-Bind_PSSM are the best methods for identifying the binding sites in DNA-binding proteins. However, inspection of these methods showed a wide range of accuracies. For example, BindN+ showed the worst performance in predicting the binding sites in HMG-D protein (1QRV), and the average accuracy is

35%. On the other hand, BindN-RF showed the best performance with an accuracy of 87% in this protein. BindN-RF showed an accuracy of 67% in T4 phage beta-glucosyltransferase (1M5R), whereas DNABindR performed well with an accuracy of 85%. The accuracy is 31% in centromere-binding protein using DP-Bind_PSSM, and BindN-RF could predict with the highest accuracy of 55%, which requires further improvement. These data demonstrated the necessity of selecting methods for efficient prediction and the requirement of improvements in specific proteins.

Online tool for the correspondence between protein/DNA type and the best method

We have developed a web server to provide the best method for any type of protein/DNA-based on its class, fold, family, superfamily, motif, function, single/double-stranded DNA and DNA conformation. It takes the structural/function information of protein/DNA and displays the best method in the output. The web server is freely available at http://www.biotech.iitm.ac.in/DNA-protein/.

CONCLUSIONS

Selecting the best method for identifying the binding sites in DNA-binding proteins is one of the immediate requirements for biologists to design experiments. We have addressed this problem by carefully analyzing the available prediction methods using nine different types of data sets based on structural information, motifs, DNA types and functional information. The one-to-one correspondence between the subclass of DNA-binding proteins and best/worst prediction method are given for all the studied data sets. These information would be highly valuable to select the best method for understanding the recognition mechanism for specific proteins as well as massive analysis with large data sets.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–9.

ACKNOWLEDGEMENTS

The authors thank the reviewers for their constructive comments. The authors wish to thank Professor Sandor Pongor for helpful discussions. R.N. and M.M.G. thank the Bioinformatics facility of Department of Biotechnology and Indian Institute of Technology Madras for computational facilities. We thank Oxford University Press for partially waiving the publication charges.

FUNDING

DST grant, Government of India [SR/SO/BB-0036/2011] and an ICGEB short-term visiting fellowship (to M.M.G.). Funding for open access charge: Department of Science and Technology, Government of India research grant (partial); Oxford University Press (partial waiver).

Conflict of interest statement. None declared.

REFERENCES

1. Prabakaran,P., An,J., Gromiha,M.M., Selvaraj,S., Uedaira,H., Kono,H. and Sarai,A. (2001) Thermodynamic database for protein-nucleic acid interactions (ProNIT). *Bioinformatics*, **17**, 1027–1034.
2. Berman,H.M., Kleywegt,G.J., Nakamura,H. and Markley,J.L. (2012) The protein data bank at 40: reflecting on the past to prepare for the future. *Structure*, **20**, 391–396.
3. Sarai,A. and Kono,H. (2005) Protein-DNA recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.*, **34**, 379–398.
4. Hogan,M.E. and Austin,R.H. (1987) Importance of DNA stiffness in protein-DNA binding specificity. *Nature*, **329**, 263–266.
5. Gromiha,M.M., Munteanu,M.G., Simon,I. and Pongor,S. (1997) The role of DNA bending in Cro protein-DNA interactions. *Biophys. Chem.*, **69**, 153–160.
6. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
7. Gromiha,M.M. (2005) Influence of DNA stiffness in protein-DNA recognition. *J. Biotechnol.*, **117**, 137–145.
8. Mandel-Gutfreund,Y. and Margalit,H. (1998) Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic Acids Res.*, **26**, 2306–2312.
9. Mandel-Gutfreund,Y., Margalit,H., Jernigan,R.L. and Zhurkin,V.B. (1998) A role for CH•O interactions in protein-DNA recognition. *J. Mol. Biol.*, **277**, 1129–1140.
10. Nadassy,K., Wodak,S.J. and Janin,J. (1999) Structural features of protein-nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
11. Jones,S., van Heyningen,P., Berman,H.M. and Thornton,J.M. (1999) Protein-DNA interactions: a structural analysis. *J. Mol. Biol.*, **287**, 877–896.
12. Jayaram,B., McConnell,K., Dixit,S.B., Das,A. and Beveridge,D.L. (2002) Free-energy component analysis of 40 protein-DNA complexes: a consensus view on the thermodynamics of binding at the molecular level. *J. Comput. Chem.*, **23**, 1–14.
13. Gromiha,M.M., Siebers,J.G., Selvaraj,S., Kono,H. and Sarai,A. (2004) Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *J. Mol. Biol.*, **337**, 285–294.
14. Lejeune,D., Delsaux,N., Charloteaux,B., Thomas,A. and Brasseur,R. (2005) Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins*, **61**, 258–271.
15. Yamasaki,S., Terada,T., Kono,H., Shimizu,K. and Sarai,A. (2012) A new method for evaluating the specificity of indirect readout in protein-DNA recognition. *Nucleic Acids Res.*, **40**, e129.
16. Bouvier,B., Zakrzewska,K. and Lavery,R. (2011) Protein-DNA recognition triggered by a DNA conformational switch. *Angew. Chem. Int. Ed. Engl.*, **50**, 6516–6518.
17. Fuxreiter,M., Simon,I. and Bondos,S. (2011) Dynamic protein-DNA recognition: beyond what can be seen. *Trends Biochem. Sci.*, **36**, 415–423.
18. Kolomeisky,A.B. (2011) Physics of protein-DNA interactions: mechanisms of facilitated target search. *Phys. Chem. Chem. Phys.*, **13**, 2088–2095.
19. Zou,X., Ma,W., Solov'yov,I.A., Chipot,C. and Schulten,K. (2011) Recognition of methylated DNA through methyl-CpG binding domain proteins. *Nucleic Acids Res.*, **40**, 2747–2758.
20. Zahran,M., Daidone,I., Smith,J.C. and Imhof,P. (2010) Mechanism of DNA recognition by the restriction enzyme EcoRV. *J. Mol. Biol.*, **401**, 415–432.
21. Gromiha,M.M. and Fukui,K. (2011) Scoring function based approach for locating binding sites and understanding the recognition mechanism of protein-DNA complexes. *J. Chem. Inf. Model.*, **51**, 721–729.

22. Ahmad,S., Keskin,O., Sarai,A. and Nussinov,R. (2008) Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Res.*, **36**, 5922–5932.
23. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248–1253.
24. Zhou,P., Tian,P., Ren,Y., Zou,J. and Shang,Z. (2010) Systemic classification and analysis of themes in protein-DNA recognition. *J. Chem. Inf. Model.*, **50**, 1476–1488.
25. Pabo,C.O. and Nekludova,L. (2000) Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.*, **301**, 597–624.
26. Prabakaran,P., Siebers,J.G., Ahmad,S., Gromiha,M.M., Singarayan,M.G. and Sarai,A. (2006) Classification of protein-DNA complexes based on structural descriptors. *Structure*, **14**, 1355–1367.
27. Cherstvy,A.G. (2011) Electrostatic interactions in biological DNA-related systems. *Phys. Chem. Chem. Phys.*, **13**, 9942–9968.
28. Mirny,L.A. and Gelfand,M.S. (2002) Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Res.*, **30**, 1704–1711.
29. Oda,M. and Nakamura,H. (2000) Thermodynamic and kinetic analyses for understanding sequence-specific DNA recognition. *Genes Cell*, **5**, 319–326.
30. Wintjens,R., Lievin,J., Rooman,M. and Buisine,E. (2000) Contribution of cation-pi interactions to the stability of protein-DNA complexes. *J. Mol. Biol.*, **302**, 395–410.
31. Rooman,M., Lievin,J., Buisine,E. and Wintjens,R. (2002) Cation-pi/H-bond stair motifs at protein-DNA interfaces. *J. Mol. Biol.*, **319**, 67–76.
32. Gromiha,M.M., Santhosh,C. and Suwa,W. (2004) Influence of Cation-pi Interactions in Protein-DNA Complexes. *Polymer*, **45**, 633–639.
33. Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, **35**, 114–131.
34. Donald,J.E., Chen,W.W. and Shakhnovich,E.I. (2007) Energetics of protein-DNA interactions. *Nucleic Acids Res.*, **35**, 1039–1047.
35. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
36. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
37. Ahmad,S., Gromiha,M.M. and Sarai,A. (2004) Analysis and Prediction of DNA-binding proteins and their binding residues based on composition, sequence and structure information. *Bioinformatics*, **20**, 477–486.
38. Ahmad,S. and Sarai,A. (2005) PSSM based prediction of DNA-binding sites in proteins. *BMC Bioinformatics*, **6**, 33.
39. 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.
40. Kuznetsov,I.B., Gou,Z., Li,R. and Hwang,S. (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins*, **64**, 19–27.
41. Ofran,Y., Mysore,V. and Rost,B. (2007) Prediction of DNA-binding residues from sequence. *Bioinformatics*, **23**, i347–i353.
42. Ho,S.Y., Yu,F.C., Chang,C.Y. and Huang,H.L. (2007) Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM-PSSM method. *Biosystems*, **90**, 234–241.
43. Bhardwaj,N. and Lu,H. (2007) Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions. *FEBS Lett.*, **581**, 1058–1066.
44. Wu,J., Liu,H., Duan,X., Ding,Y., Wu,H., Bai,Y. and Sun,X. (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.
45. Xu,B., Yang,Y., Liang,H. and Zhou,Y. (2009) An all-atom knowledge-based energy function for protein-DNA threading, docking decoy, discrimination, and prediction of transcription-factor binding profiles. *Proteins*, **76**, 718–730.
46. Wang,L., Yang,M.Q. and Yang,J.Y. (2009) Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genomics*, **10**, S1.
47. Wang,L., Huang,C., Yang,M.Q. and Yang,J.Y. (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Systems Biol.*, **4**, S3.
48. Hwang,S., Gou,Z. and Kuznetsov,I.B. (2007) DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics*, **23**, 634–636.
49. Yan,C., Terribilini,M., Wu,F., Jernigan,R., Dobbs,D. and Honavar,V. (2006) Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, **7**, 262.
50. Carson,M.B., Langlois,R. and Lu,H. (2010) NAPS: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Res.*, **38**, W431–W435.
51. Si,J., Zhang,Z., Lin,B., Schroeder,M. and Huang,B. (2011) metaDBSite: a meta approach to improve protein DNA-binding site prediction. *BMC Syst. Biol.*, **5**, S7.
52. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
53. Berman,H.M., Olson,W.K., Beveridge,D.L., Westbrook,J., Gelbin,A., Demeny,T., Hsieh,S.-H., Srinivasan,A.R. and Schneider,B. (1992) The Nucleic Acid Database: A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids. *Biophys. J.*, **63**, 751–759.
54. Norambuena,T. and Melo,F. (2010) The Protein-DNA Interface database. *BMC Bioinformatics*, **11**, 262.
55. Shanahan,H.P., Garcia,M.A., Jones,S. and Thornton,J.M. (2004) Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Res.*, **32**, 4732–4741.
56. Bourne,P.E. and Desai,N. (1987) PRONUC: a software package for the analysis of protein and nucleic acid sequences. *Comput Methods Programs Biomed.*, **24**, 27–36.
57. Gromiha,M.M. and Nagarajan,R. (2013) Computational approaches for predicting the binding sites and understanding the recognition mechanism of protein-DNA complexes. *Adv. Prot. Chem. Str. Biol.*, **91**, 65–99.
58. Tjong,H. and Zhou,H.-X. (2007) DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res.*, **35**, 1465–1477.
59. Gromiha,M.M. (2010) *Protein Bioinformatics: From Sequence to Function*. Elsevier Publishers, New Delhi.
60. Gromiha,M.M. and Selvaraj,S. (1998) Protein secondary structure prediction in different structural classes. *Protein Eng.*, **11**, 249–251.
61. Murvai,J., Vlahovicek,K. and Pongor,S. (2000) A simple probabilistic scoring method for protein domain identification. *Bioinformtics*, **16**, 1155–1156.