doi:10.1016/j.jmb.2009.02.023 J. Mol. Biol. (2009) 387, 1040–1053

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Identification of DNA-binding Proteins Using Structural, Electrostatic and Evolutionary Features

Guy Nimrod[1], András Szilágyi[2], Christina Leslie[3] and Nir Ben-Tal[1]*

[1]*Department of Biochemistry, The George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel*

[2]*Institute of Enzymology, Hungarian Academy of Sciences, H-1113 Budapest, Hungary*

[3]*Computational Biology Program, Memorial Sloan-Kettering Cancer Center, NY 10065, USA*

Received 8 December 2008; received in revised form 12 February 2009; accepted 12 February 2009
Available online 20 February 2009

Edited by B. Honig

DNA-binding proteins (DBPs) participate in various crucial processes in the life-cycle of the cells, and the identification and characterization of these proteins is of great importance. We present here a random forests classifier for identifying DBPs among proteins with known 3D structures. First, clusters of evolutionarily conserved regions (patches) on the surface of proteins were detected using the PatchFinder algorithm; earlier studies showed that these regions are typically the functionally important regions of proteins. Next, we trained a classifier using features like the electrostatic potential, cluster-based amino acid conservation patterns and the secondary structure content of the patches, as well as features of the whole protein, including its dipole moment. Using 10-fold cross-validation on a dataset of 138 DBPs and 110 proteins that do not bind DNA, the classifier achieved a sensitivity and a specificity of 0.90, which is overall better than the performance of published methods. Furthermore, when we tested five different methods on 11 new DBPs that did not appear in the original dataset, only our method annotated all correctly.

The resulting classifier was applied to a collection of 757 proteins of known structure and unknown function. Of these proteins, 218 were predicted to bind DNA, and we anticipate that some of them interact with DNA using new structural motifs. The use of complementary computational tools supports the notion that at least some of them do bind DNA.

Keywords: DNA-binding proteins; random forests; DNA-binding sites; PatchFinder; structural genomics

Introduction

DNA-binding proteins (DBPs) are involved in processes like DNA transcription, maintenance, replication and the regulation of gene expression, and hence many of these proteins are essential for the viability and proliferation of cells.[1]

As a result of the structural genomics initiatives, there is a growing number of proteins with known structure whose functions are unknown.[2] Presumably, some of these proteins are novel DBPs that are yet to be characterized. Therefore, it is desirable to develop an accurate method for the classification of DBPs from their 3D structure.

Some of the methods for the identification of DBPs have been based on searching for common structural motifs in DNA-binding sites; for example, the helix-turn-helix motif.[3,4] While such methods are successful at identifying proteins with these motifs, they might overlook binding motifs that are yet to be characterized. The observation that the DNA-binding site is usually positively charged, compensating for the negative charges on the DNA backbone, is also commonly used.[3,5–9] Alternative approaches examine evolutionary conservation patterns and the amino acid composition of the protein in order to annotate DBPs.[1,8,10]

Stawiski *et al*. examined positively charged patches on the surface of DBPs in comparison with proteins that do not bind DNA (nDBPs).[7] They trained a neural network (NN) for the identification of DBPs using 12 features, including the patch size,

*Corresponding author. E-mail address: NirB@tauex.tau.ac.il.

Abbreviations used: DBP, DNA-binding protein; nDBPs, proteins that do not bind DNA; NN, neural network; MCC, Matthews correlation coefficient; MSA, multiple sequence alignment; ML-patch, maximum likelihood patch; dsDNA, double-stranded DNA; PSSM, position-specific scoring matrix; ROC, receiver operating characteristic; AUC, area under the curve; PR, precision-recall; RBP, RNA-binding protein.

hydrogen bonding potential, the fraction of evolutionarily conserved positively charged residues and other properties of the protein. The classifier was tested on a dataset of structures of 54 DBPs and 250 nDBPs. They used the Matthews correlation coefficient[11] (MCC; see Materials and Methods) to measure the correlation between the predicted and observed classes and reported an MCC value of 0.74.

Ahmad and Sarai[9] based their NN classifier on the net charge, and the electric dipole and quadrupole moments of the protein. They used a dataset of 78 structures of DBPs and a negative dataset of 110 nDBPs. The algorithm achieved an MCC of 0.68 on this dataset.

Bhardwaj *et al*.[12] examined the sizes of positively charged patches on the surface of DBPs. They used the overall charge of the protein as well as its overall and surface amino acid composition to train a support vector machine classifier. The classifier had sensitivity of 67.4% and specificity of 94.9% using fivefold cross-validation. The analysis was conducted on a non-redundant set (<20% identity between each pair of sequences) of the DBPs gathered from earlier studies[7,10,13] and the dataset of nDBPs used by Stawiski *et al*.[7]

Szilágyi & Skolnick recently developed a logistic regression classifier based on the amino acid composition, the asymmetry of the spatial distribution of specific residues and the dipole moment of the protein.[8] They used a dataset of 138 DBPs that were co-crystallized with DNA and 110 nDBPs, and reported an MCC of 0.74.

Both the methods of Szilágyi & Skolnick and Ahmad & Sarai are particularly noteworthy because of their low sensitivity to the accuracy of the structure, suggesting that the methods may be useful with low-resolution structures or possibly even model structures.

Here, we present a classifier for the detection of DBPs based on the identification and feature representation of functional regions on the surface of proteins. The identification relies on the observation that functional regions in proteins are usually evolutionarily conserved and preserve the functionality of the protein.[14,15]

PatchFinder is an algorithm that we developed recently for identifying such conserved functional regions.[16,17] Briefly, PatchFinder uses as input the 3D structure of the query protein,[18] and a multiple sequence alignment (MSA) of the query protein and its sequence homologues. First, each amino acid position in the protein is assigned an evolutionary conservation score calculated on the basis of the phylogenetic relations among the homologues using Rate4Site.[19] Second, PatchFinder finds the most significant continuous cluster of conserved residues on the protein surface. This cluster is predicted to be the main functional region of the protein, and we refer to it as the maximum likelihood patch (ML-patch).

We present here a PatchFinder analysis of the dataset of DBPs used by Szilágyi & Skolnick. The analysis showed that the regions predicted by PatchFinder are usually DNA-binding sites. Furthermore, the amino acid conservation patterns of the predicted functional regions, their electrostatic potential and other properties were found to be distinctive between the DBPs and the dataset of nDBPs. We utilized these properties along with the features used by Szilágyi & Skolnick in a random forests classifier[20] and achieved an MCC of 0.80, which is better than previously achieved in other studies.

Except for Stawiski *et al*.[7] and Bhardwaj *et al*.,[12] who examined positively charged patches, most previous methods use global properties of the protein as features rather than local properties, which we find to be informative; the PatchFinder approach is more general than these methods, since conservation-defined patches can be used for other functional classes too. This property of the classifier is particularly important in characterizing proteins of novel folds coming out of high-throughput structural genomics.

We used the classifier to predict DBPs in the N-Func database of structures of proteins with unknown function.[17] Our analysis suggests that 218 of the 757 entries of N-Func may bind DNA. We also demonstrated, on the basis of a literature survey and other computational tools, that some of these proteins are likely to bind DNA. Some of the potential false-positives may bind other polynucleotides.

Results

PatchFinder consistently finds the core of the DNA-binding site

In our analysis, we used the dataset of DBPs established by Szilágyi & Skolnick.[8] This dataset is a non-redundant set of 138 structures of proteins bound to a double-stranded DNA (dsDNA). The PatchFinder algorithm uses the conservation analysis as computed by Rate4Site[19] in order to predict the functionally important regions. When fewer than four sequence homologues are available for a query protein, the analysis may be inaccurate.[19] Consequently, PatchFinder predicted the ML-patches for 121 out of the 138 DBPs.

First, we wanted to find out whether the ML-patches correspond to the DNA-binding sites. The ML-patches had an average of 19 residues. In 118/121 (98%) of the ML-patches, at least one of the residues in the patch was in contact with the DNA (see Materials and Methods). In 91/121 (74%) of the cases, at least half of the residues in the ML-patch were in contact with the DNA. On the other hand, the ML-patch included at least half of the residues that are in contact with the DNA in only 21/121 (17%) of the proteins. Figure 1a shows the distribution of precision and sensitivity amongst the proteins in the dataset. According to these data, while most of the residues found by PatchFinder indeed bind DNA, it overlooked a considerable part of the interface. Therefore, we concluded that

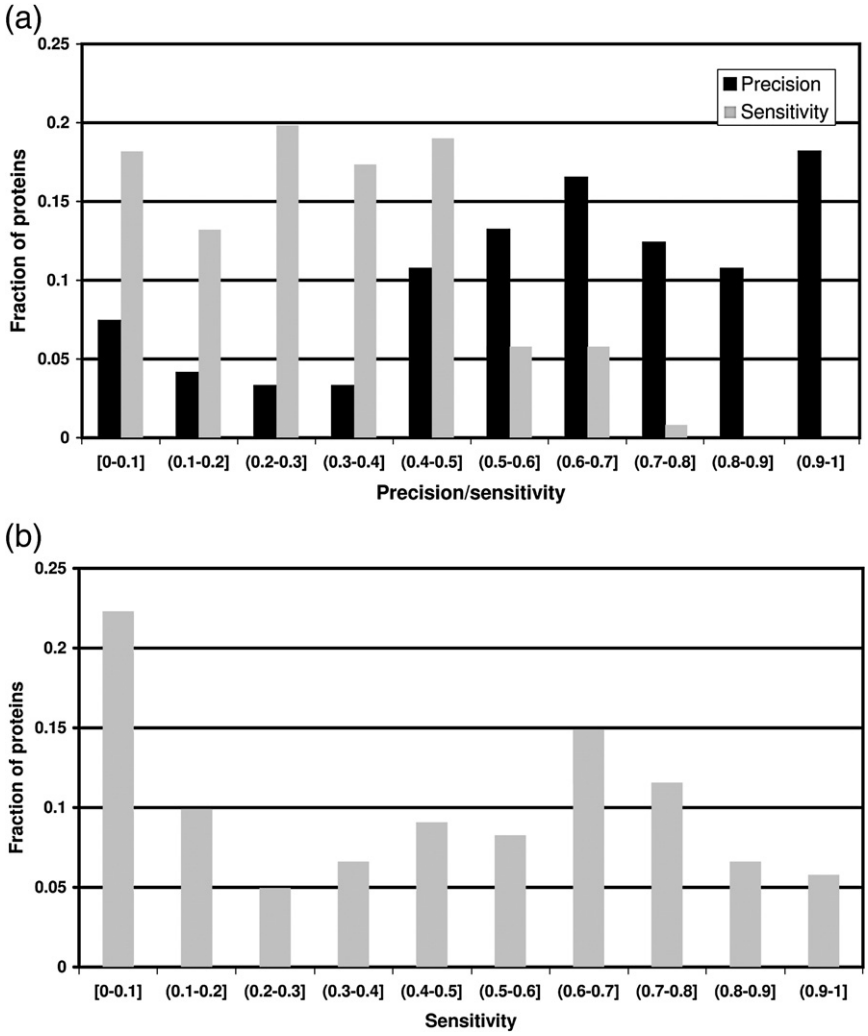

Fig. 1. The performance of PatchFinder in the identification of DNA-binding regions. (a) The fraction of ML-patches within each precision/sensitivity bin. Surface residues within 6 Å of the DNA are considered DNA-binders. Black bars represent the fraction of ML-patches in each precision bin. Precision measures the fraction of the patch residues that bind DNA (Eq. (3)). Grey bars represent the fraction of ML-patches in each sensitivity bin. Sensitivity measures the fraction of DNA-binding residues identified by PatchFinder (Eq. (4)). Sensitivity is low in comparison with the high precision. (b) The fraction of the protein–DNA hydrogen bonds identified by PatchFinder. Grey bars represent the fraction of patches in each sensitivity bin. In this figure, sensitivity was calculated as the fraction of hydrogen bond donors/acceptors within the protein-DNA interface[21] identified by PatchFinder. PatchFinder identified a considerable fraction of the hydrogen bonds between the protein and the DNA in most of the DBPs in the dataset. However, in 23% of the patches, less than 1/10 of the hydrogen bonds were identified.

only part of the DNA-binding site is highly conserved in DBPs.

We used the NUCPLOT program[21] in order to examine the sensitivity of PatchFinder regarding hydrogen bonds between the protein and the DNA. Here, the results were substantially different (Fig. 1b). PatchFinder identified at least half of the hydrogen bonds in 61/121 (50%) of the proteins. Furthermore, the plot revealed two distinct peaks in the sensitivity. The first represents proteins in which PatchFinder found up to 10% of the hydrogen bonds, and the second represents proteins in which PatchFinder found 60–70% of the hydrogen bonds.

PDB id 1dfm represents the structure of one of the DBPs in the dataset. This is a crystal structure of the restriction nuclease BglII from *Bacillus subtilis* along with its recognition DNA sequence.[22] The protein binds as a homodimer to the palindromic DNA sequence AGATCT and cleaves the DNA after the first adenine, with a magnesium ion as co-factor.[23] Figure 2 represents the structure of one unit of the dimer along with the bound DNA molecule. There are 56 residues at the protein–DNA interface. PatchFinder found a patch of 17 residues (red), 16 of which are in contact with the DNA. The ML-patch comprises 29% of the residues in contact with the DNA but 43% of the hydrogen bonds with the DNA. PatchFinder also identified three out of four residues that comprise the active site.[22] In addition, as Fig. 2 shows, the patch corresponds well to the residues that bind the recognition sequence of the DNA (blue).

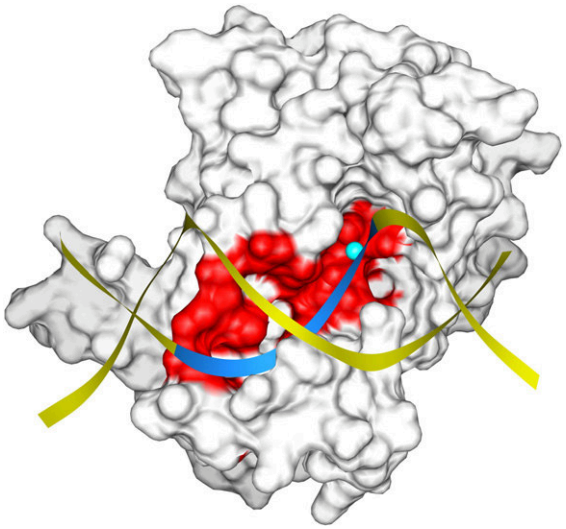

Fig. 2. Restriction nuclease BglII from *Bacillus subtilis*. The crystal structure of restriction nuclease BglII along with its DNA-recognition sequence.[22] The protein is a homodimer and the picture shows a surface representation of one of the subunits. The ML-patch is in red and the rest of the protein is grey. The DNA fragment is shown as strands, with the DNA-recognition sequence coloured blue, and the rest of the molecule is shown in yellow. A calcium ion, located in the active site, is shown in cyan. The figure shows that the ML-patch corresponds well to the interface of the protein with its recognition sequence.

While the structures of the DBPs in the data we used here were determined in complex with a dsDNA, we are more interested in the identification of binding sites of proteins in their unbound state. Differences between the bound- and the apo-conformations of DBPs may occur in their tertiary structure as well as domain organization and even disorder-to-order transitions.[24,25] Obviously, cases in which the apo form is unstructured and the 3D structure of the protein is not available are beyond the scope of this study.

Nadassy *et al*.[25] examined the structures of proteins that were crystallized both in their bound and unbound (apo) states. We analyzed 13 pairs of these protein structures to examine the effect of the conformational changes associated with DNA binding on the identification of the functional sites by PatchFinder. In all cases, the ML-patch of one form overlapped the ML-patch of the other by at least 50%, and in 11 cases, the overlap was 80% or more. The HhaI DNA methyltransferase from *Haemophilus haemolyticus* is an example of one of the proteins examined (Fig. 3). The association of the protein with its recognition DNA sequence invokes base flipping of a target cytosine (red) out of the DNA helix.[26] A methyl group is then transferred from an AdoMet molecule (yellow) to the C5 position of that cytosine. Nadassy *et al*. measured in the DNA-binding site an RMSD of 7.9 Å between the bound and apo forms of the protein.[25] Figure 3 shows a superimposition of the protein in its apo form[27] (green) and the protein (blue) bound to dsDNA[26] (orange). As can be seen in the figure, the association of the protein with the DNA is accompanied by a substantial movement of the active site loop towards the major groove of the dsDNA.[28] Cys81 within this loop is a key residue in the catalytic reaction.[29] PatchFinder found an ML-patch of 15 residues in the bound form of HhaI DNA methyltransferase. This patch comprises most of the AdoMet and the nucleotide-binding pockets. The ML-patch found in the apo form included 12 residues, all of which are in the ML-patch of the bound form, including Cys81.This result and the analysis of 12 additional proteins suggest that even though substantial conformational changes between the bound and apo forms may occur, the effect on the identification of the ML-patch is limited.

Identification of DBPs

Since the region identified by PatchFinder is presumably important for the function of the protein, its physicochemical properties should facilitate functional annotation at some level.

The positive electrostatic charge is usually the most prominent property of the binding site, since the surface of dsDNA is negatively charged due to the backbone phosphates. Hydrogen bonds were also pointed out as important in protein–DNA interactions, especially in recognition.[30-33] The fraction of residues in helical conformation is another noticeable feature found in DNA-binding sites,[7] particularly due to the common helix-related DNA-binding motifs (helix-turn-helix, helix-hairpin-helix and helix-loop-helix).[3] Consequently, we measured

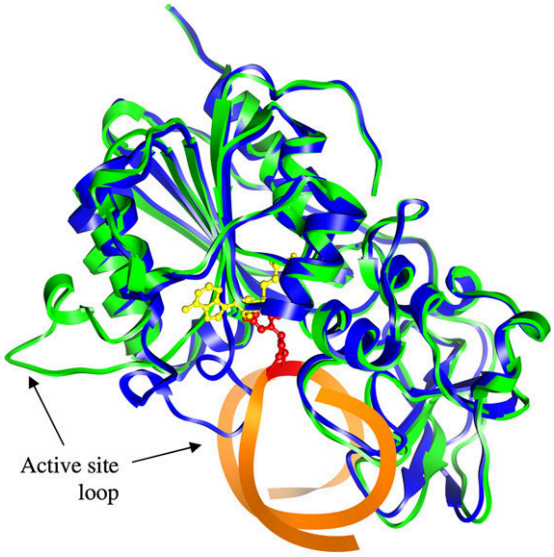

Fig. 3. Conformational changes between the bound and the apo forms of the HhaI DNA methyltrasferase. The crystal structure of the apo form of the protein[27] (green ribbons) is superimposed on the crystal structure of the protein (blue ribbons) bound to a dsDNA (orange strands).[26] The target cytosine, which flips out of the dsDNA, and the AdoMet ligand are colored red and yellow, respectively. The figure demonstrates a considerable shift of the active site loop towards the major groove of the dsDNA.

these properties for the ML-patches of the DBPs and the nDBPs in the datasets. We found that DBPs differ significantly from the rest of the proteins in the features examined (Kolmogorov-Smirnov test; $p<0.001$), even though by itself, none of the features is sufficient for reliable distinction between these two groups.

Another feature that could be helpful in protein annotation is the amino acid composition of the protein and, in particular, that of the DNA-binding region.[25] We used a position-specific scoring matrix (PSSM) representation of the MSAs in order to characterize the amino acid conservation patterns of the ML-patches. Then we defined features that represent the similarity of the mean conservation pattern of the ML-patch to a set of prototype conservation patterns obtained via clustering (see Materials and Methods for details). The features listed above were added to the 10 features developed by Szilágyi & Skolnick.[8] The features are: the dipole moment of the molecule, the spatial asymmetry of Arg, Gly, Asn and Ser, and the percentage of Arg, Ala, Gly, Lys and Asp in the query protein.

Our classifier was based on two separately trained classifiers, depending on the availability of the ML-patch. The first included 16 features that did not require the identification of the ML-patch. This classifier is suitable for cases in which too few sequence homologues were found for the query protein for detecting the ML-patch. There were 28 such cases in our dataset. The second classifier included all 33 features that were calculated for the proteins, including those that require the ML-patch. We trained random forests classifiers[20] with the resulting vectors that represented the proteins in the dataset. The performance of the classifier was evaluated using 10-fold cross validation runs. The resulting classifier had a sensitivity and a specificity of 0.90, and an MCC of 0.80.

Figure 4a represents a receiver operating characteristic (ROC) curve of the classifier that plots the sensitivity *versus* the false-positive detection rate (i.e., 1 – specificity) at various prediction thresholds (grey line). The area under the curve (AUC) is a measure of the quality of the separation between the examined protein classes (i.e., DBPs *versus* nDBPs). An AUC of 0.5 represents a classification that corresponds to a randomly generated prediction, while an area of 1 corresponds to a perfect classifier. The classifier had a high AUC of 0.96. This value is higher than the AUC of 0.93 achieved by Szilágyi And Skolnick on the same dataset.[8]

We also examined the classifier on the datasets used by Bhardwaj *et al.*[12] and Stawiski *et al.*[7] Our classifier was significantly better on these datasets than the methods of the respective authors (see Supplementary Data S1).

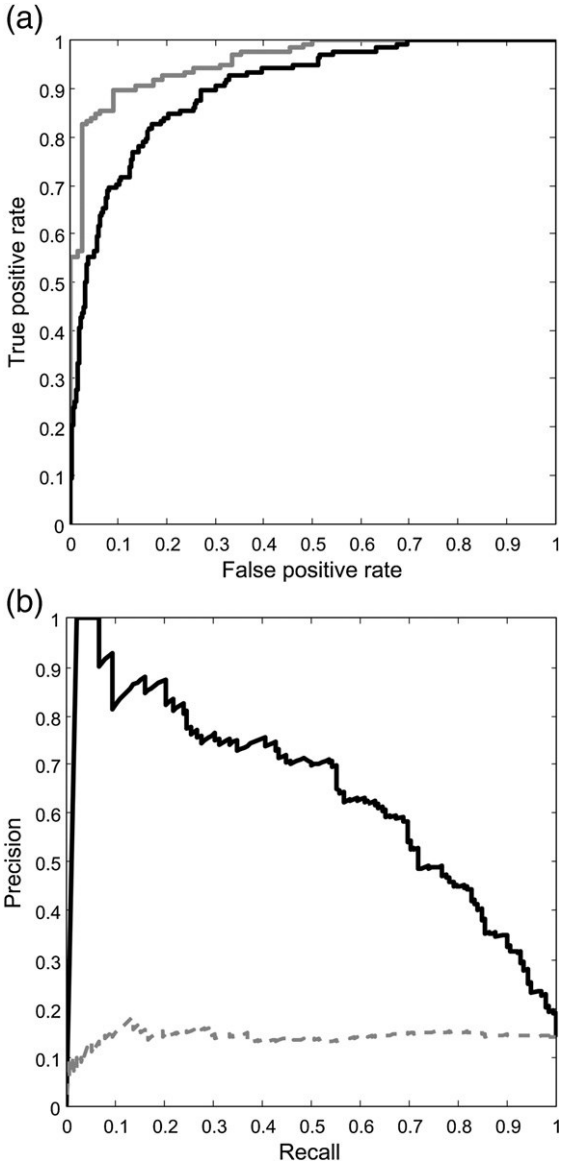

Fig. 4. ROC and PR curves of the classifier. (a) The ROC curve of the classifier. The results were obtained using a dataset of 138 DBPs and 110 nDBPs (grey line) and on the extended dataset of 138 DBPs and 843 nDBPs (black line), which reflect the anticipated proportion of DBPs in reality. The areas under the curves (AUC) are 0.96 and 0.90, respectively. (b) The precision-recall (PR) curve of the classifier, on the extended dataset (black line) in comparison with a random classification (broken grey line) generated by shuffling the classes.

The expected performance on "real" user data

The fraction of DBPs in a proteome is much smaller than it is in the training data. We expect that the user data would be similar to the distribution of proteins within proteomes. We used the gene ontology[34] (GO) database and the annotations available for the genomes of *Homo sapiens*, *Saccharomyces cerevisiae* and *Escherichia coli*, in order to evaluate the fraction of DBPs in a set of randomly selected proteins. Consequently, we estimated the fraction of DBPs to be 14% (see Materials and Methods).

On the basis of this estimate, we examined the classifier on an extended non-redundant dataset with 733 additional nDBPs. Along with the original

dataset of 138 DBPs and 110 nDBPs,[8] we had a non-redundant set of 981 proteins in which 14% are DBPs. On this dataset, the specificity of the classifier dropped from 0.90 to 0.72, while the sensitivity stayed at 0.90. However, by applying different score cutoffs, one can select a suitable cutoff for specific needs. For example, at a sensitivity of 0.85 the specificity is 0.82. The AUC, which integrates the various true-positive (sensitivity) and false-positive detection rate (1–specificity) values at different cutoffs, was 0.90 (Fig. 4a, black curve).

A precision-recall (PR) curve is a plot of the precision *versus* the recall (sensitivity) of the classifier at various prediction score cutoffs. When the data are highly skewed, as in our case, a PR curve is considered better than ROC curve for the analysis of a classifier's performance.[35] The PR curve of the classifier (Fig. 4b, continuous line) is significantly better than that obtained at random (Fig. 4b, broken line). Supplementary Data Table S1 gives the specificity and recall values along with the corresponding classification score thresholds.

Ranking the features by their contribution

We examined the contribution of the various features to the overall performance of the classifier on the 220 proteins for which an ML-patch was predicted. The features were divided into seven categories as follows: electrostatic potential, hydrogen bond donors/acceptors, secondary structure, the amino acid conservation patterns of the ML-patch, amino acid asymmetry, amino acid content in the proteins and the dipole moment. The last three categories are of the features developed by Szilágyi & Skolnick.[8] We trained seven different classifiers. In each classifier, one group of features was omitted. The unique contribution of each category was measured as the change in misclassification rate of the dataset using 10-fold cross-validation (i.e., the change in the total number of misclassified proteins divided by the size of the dataset). According to this analysis (Fig. 5), the features of the electrostatic potential, the amino acid conservation patterns of the ML-patch and the secondary structure had the largest contribution to the classification.

Analysis of false predictions

Using 10-fold cross-validation, the classifier misclassified 25 proteins: 14 false-negatives and 11 false-positives. For most of these proteins, more than 40% of the decision trees classified the protein correctly. It is worthwhile to note that in almost all of the false-positive predictions, the average electrostatic potential of the ML-patch was positive, a marked property of most of the DBPs. These proteins bind molecules with a mostly negatively charged surface like heme and NADPH.

RNA, in particular, has physicochemical properties similar to those of DNA as a polynucleotide. Consequently, the interactions of proteins with DNA and RNA are similar as well.[36] Shazman & Mandel-Gutfreund have recently developed a classifier for the identification of RNA-binding proteins.[37] They showed that although the classifier identified RNA-binding proteins (RBPs) well, it could not distinguish between RBPs and DBPs. We examined our classifier on the non-redundant set of 76 RBPs used by Shazman & Mandel-Gutfreund. The classifier misclassified 51 (67%) of the proteins, predicting them as DBPs. Even though this rate is better than expected at random (i.e. false-positive rate of 90%), we concluded that our classifier does not distinguish well between RBPs and DBPs, at least when trained on its current training set.

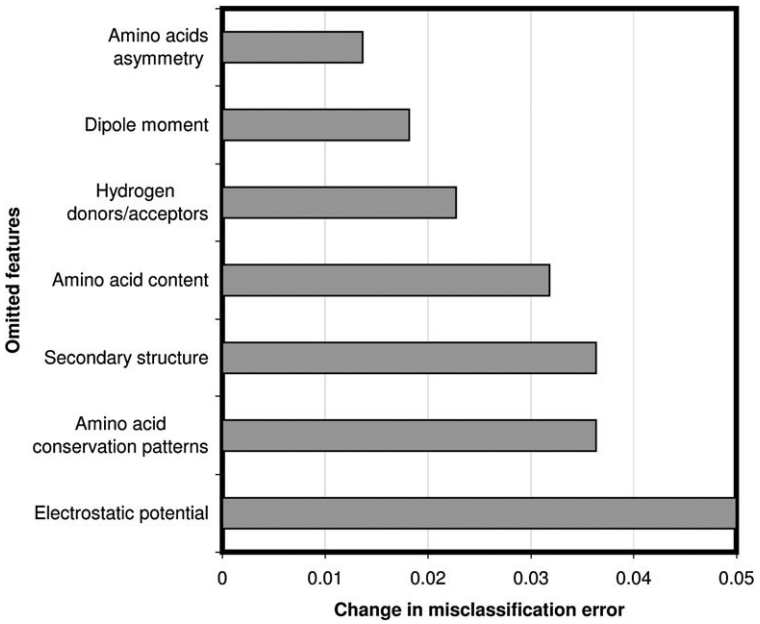

Fig. 5. Determinants of the classifier. The graph represents the change in misclassification rate upon excluding each feature category from the input vectors. A high rate represents a big contribution of the category to the overall performance of the final classifier. According to this analysis, the highest contribution comes from the electrostatic potential, the amino acid conservation patterns of the ML-patch and the secondary structure.

A small but independent dataset

Since Szilágyi & Skolnick gathered the dataset of DBPs we used here, new structures of DBPs have been deposited in the PDB. Applying the same filtering criteria as used for the assembly of the original set, we found 11 additional crystal structures of DBPs bound to dsDNA. These proteins share 35% or less sequence identity with the entries in the dataset used by Szilágyi & Skolnick. Our classifier correctly predicted all of these proteins as DBPs. The lowest score assigned to a protein in this set was 0.65. At this score cutoff, the specificity on the extended dataset is 0.87. We tested three published structure-based methods[6,8,9] and a sequence-based method[10] on this set, and found that between two and eight of the DBPs were misclassified by them (see Supplementary Data Table S4). While this additional test set is too small to provide a reliable comparison, it suggests that our method is more sensitive and identifies DBPs better than related methods.

Identification of DBPs in the N-Func database

Structural genomics efforts like the Protein Structure Initiative (PSI) have determined the structures of hundreds of proteins as part of an effort to map the protein fold space. Many of these proteins are of unknown function, and are referred to as hypothetical proteins.[38] Some of these are presumably DBPs.

We recently established the N-Func database: a collection of 757 hypothetical proteins of known 3D-structure.[17] N-Func includes the PatchFinder prediction of the functional region for each protein, based on MSAs taken from the HSSP database.[39]

Applying the classifier on N-Func, 218 proteins were predicted as DBPs at a score cutoff of 0.5. The list is available as Supplementary Data Table S2 along with the fraction of trees in the forest that classified each protein as DBP, which corresponds to the confidence of the classification (see Materials and Methods). Additionally, we supply links to a detailed analysis of the ML-patch of each protein, including 3D visualization of the patch (using FirstGlance in Jmol).

The score cutoff of 0.5 is very permissive, having an expected precision of only 35% at a sensitivity of 90% (see Fig. 4b). Nevertheless, by applying different classification score cutoffs on the list (see Supplementary Data Table S1), sensitivity can be lowered in order to improve precision or *vice versa*.

We further analyzed the proteins in N-Func that are predicted as DBPs with score >0.78. With this score threshold, the expected specificity is 0.95 and the sensitivity is 0.58. Even though the specificity is high at this cutoff, about one-third of the predicted DBPs are expected to be false-positives. We examined these proteins with two algorithms for the identification of DBPs implemented in the ProFunc server.[40] The first algorithm searches for the DNA-binding helix-turn-helix motif in the query protein.[4] The second algorithm searches for local structural similarity between the query protein and 3D templates of known DNA/RNA-binding proteins.[41] In addition, we examined whether the proteins have folds[42] or sequence motifs[43] that are related to DBPs. The data are summarized in Supplementary Data Table S3. For most of the proteins, we found additional support for a DNA-binding functionality. Some of the proteins, on the other hand, are predicted to bind RNA or other ligands. This may indicate the similarity between DNA and RNA-binding proteins,[36] and the difficulty to distinguish between them as was shown earlier[44] and in the section Analysis of false predictions, above. Furthermore, there are known examples of proteins that show affinity to both RNA and DNA (e.g., the archeal chromatin protein Alba[45]).

Below, we present two detailed examples of predicted DBPs.

2fna: Q97Y08_SULSO from *Sulfolobus solfataricus*

PDB entry 2fna refers to a protein of 356 residues from *Sulfolobus solfataricus*. The 3D structure of the protein was determined as part of the Joint Center for Structural Genomics (JCSG) initiative. The protein bound to an ADP molecule was crystallized and was assigned the Pfam motif Archaeal ATPase.[46] Our classifier predicted that the protein binds DNA. We analyzed the protein with ProFunc, a web server for function prediction of proteins with known 3D structure.[40] The server includes various general tools, e.g. for the identification of sequence[47] or structural[48] similarity, as well as tools that specialize in specific functional classes like DBPs.[3,4,41] ProFunc's tools that specialize in DBPs overlooked the

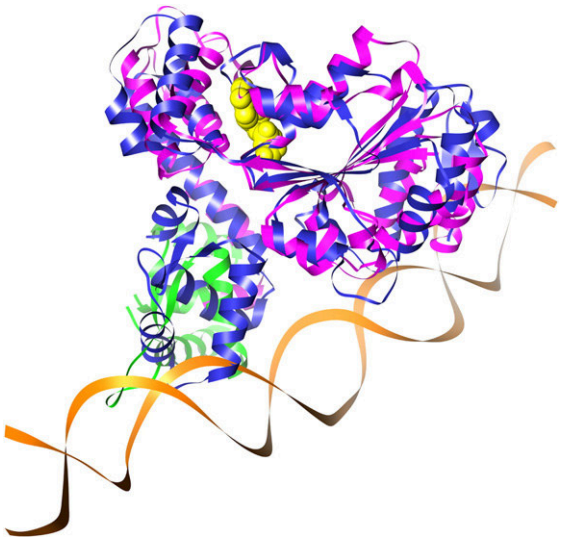

Fig. 6. Structural similarity between 2fna and cdc6 from *Sulfolobus solfataricus*. The structure of the protein of unknown function Q97Y08_SULSO (PDB id 2fna; blue ribbons) was superimposed on the structure of cdc6 (magenta and green) bound to dsDNA (orange strands) and to an ADP molecule (yellow in spacefill representation).[49] The similarity supports our prediction that 2fna also binds DNA.

possibility that 2fna binds DNA. However, other tools in ProFunc identified both structural and sequence similarity between 2fna and archaeal cdc6,[49,50] a protein involved in pre-replication complexes.[50] cdc6 is a DBP that also binds ATP/ADP.[49] Figure 6 presents a superimposition, made with UCSF Chimera,[51] of the structures of 2fna (blue) and the cdc6 from *S. solfataricus* (magenta and green) bound to a dsDNA (orange).[49] Evidently, there is close structural similarity between 2fna and most of the cdc6 from *S. solfataricus* (magenta). Furthermore, the superimpositions of the protein chains placed together the ADP molecules crystallized with each protein. The C-terminal domain of cdc6 (green) is shifted considerably in comparison with the corresponding region in 2fna. However, both domains have a similar winged helix fold. Therefore, we suggest that 2fna, which shares the same structure, also binds DNA.

1t06: Q81BA8_BACCR from *Bacillus cereus*

Another protein that our classifier predicted to bind DNA is Q81BA8_BACCR from *Bacillus cereus*. The structure of the protein was determined as part of the initiative of the Midwest Center for Structural Genomics (MCSG) by Zhang and colleagues (PDB id 1t06). The protein appears in the PDB file as a homodimer and has a large positively charged surface cavity (Fig. 7a). We analyzed the protein using the Skan algorithm for the identification of structural similarity.[52] Skan identified significant structural similarity between 1t06 and the structure of Q82ZI8_ENTFA from *Enterococcus faecalis* (PDB id 2b6c; Fig. 7b). In addition, the proteins are classified within the same SCOP family.[53] However, they share a low level of sequence identity of only 17%.

Q82ZI8_ENTFA is a hypothetical protein that was crystallized by the same center. It has a positively charged surface cavity and, like 1t06, was predicted by our classifier to bind DNA. Q82ZI8_ENTFA belongs to a newly characterized Pfam entry PF08713.[46] PF08713 includes 3-methyladenine DNA glycosylases,[54,55] and is annotated as a family of DNA alkylation repair enzymes. As the protein appears in the PDB file, the interface between its monomers is at a different location within the fold in comparison with 1t06 (data not shown). However, the PQS[56] server for the analysis of protein quaternary structures predicts that this interface is due to crystal packing rather than being physiological.

The two proteins share a low level of sequence similarity but a similar fold, and both are predicted to be DBPs. This suggests that although the proteins are evolutionarily distant, they retained properties related to DNA binding, which were identified by the classifier.

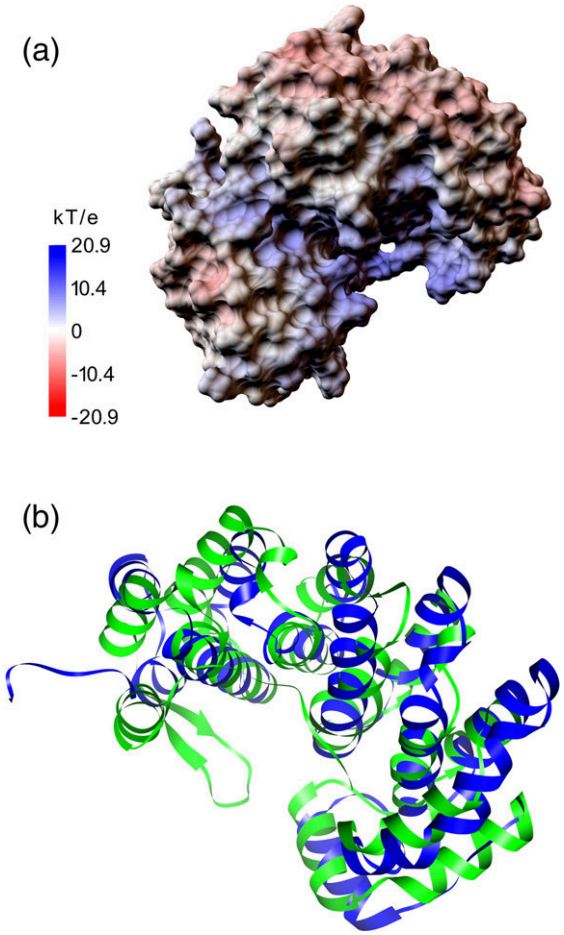

Fig. 7. Hypothetical protein Q81BA8_BACCR from *Bacillus cereus*, a potential DBP. (a) The electrostatic potential of the homodimeric structure of the protein mapped onto the molecular surface, according to the colour scale on the left. The figure reveals a large positively charged surface patch. The figure was produced with the Python Molecular Viewing environment (PMV).[79] (b) The crystal structure of one unit of the homodimer of Q81BA8_BACCR (green ribbons) superimposed on the structure of Q82ZI8_ENTFA (blue ribbons). Q82ZI8_ENTFA belongs to the Pfam entry of predicted DNA alkylation repair enzymes and, like Q81BA8_BACCR, is predicted by the classifier to bind DNA.

Discussion

We introduce here a new approach for the detection of DBPs. The improvement over previous methods is based predominantly on various properties of the functional regions of the proteins. In the following, we discuss some of the implications and limitations of the approach.

PatchFinder identifies the core of the DNA-binding site

PatchFinder traces the main conservation signal on the protein surface, which is often the functional region of the protein.[15,16] Our hypothesis was that this region often mediates the interaction with the DNA in DBPs. In support of this hypothesis, we showed that at least some of the residues in the ML-patch are in contact with the DNA in 118 out of

121 of the DBPs in the dataset. In most cases, at least half of the residues in the ML-patch were in contact with the DNA (i.e., high precision), while a considerable part of the DNA-binding region was overlooked (low sensitivity).

It has been reported that DNA-binding sites are often evolutionarily conserved.[1] However, it was concluded also that the observed conservation signal in these regions was not strong enough to enable the identification of binding sites.[1] We demonstrate here that this claim is only partially true: In most cases, evolutionary conservation enables the identification of a core of the interface while the rest of the interface residues are less conserved.

Many of the DBPs have some specificity to the target DNA sequence.[1] According to common models, the interaction with DNA begins with non-specific association of the protein with DNA and proceeds with recognition of the target site on the DNA.[57-59] According to these models, the protein slides along the DNA and hops between close DNA segments, "searching" for its target sequence before the recognition.[57,59] An implicit assumption in these models is that protein–DNA interactions include two components: specific and non-specific. We suggest that, generally speaking, the (evolutionarily conserved) ML-patches, detected by PatchFinder, mediate the specific interactions with the DNA. This hypothesis may explain the low level of sensitivity along with the high precision of the patches. The specific functionality of the protein (e.g., recognition of the DNA sequence or catalytic activity) typically requires a highly conserved region. However, in order to preserve the initial non-specific interactions with DNA, the geometric and chemical constraints on the protein region involved in these non-specific interactions are presumably more permissive. Therefore, this region may be less conserved. The initial association of the protein with the DNA is presumably driven mostly by non-specific Coulomb attraction between positively charged residues on the protein surface and the negatively charged DNA molecule.[60] Indeed, examination of the DNA-binding regions of the DBPs in the dataset showed that the fraction of the positively charged residues arginine and lysine in these regions is higher than in the ML-patches in most (62%) of these proteins.

It is known that hydrogen bonds contribute to the recognition of the target DNA sequence.[30-32] We analyzed the pattern of protein–DNA hydrogen bonds with NUCPLOT.[21] The analysis showed an enrichment of hydrogen bonds in most of the ML-patches in comparison with the rest of the DNA-binding site.

Another interesting result from the same analysis suggested a partitioning of the ML-patches into two groups, according to the fraction of protein–DNA hydrogen bonds identified by each patch (Fig. 1b). A preliminary inspection of 16 protein–DNA complexes with the highest fraction of hydrogen bonds in their ML-patches showed that the vast majority of these proteins are transcription factors, which typically require specific recognition of the DNA. By contrast, inspection of the 16 proteins whose ML-patches were devoid of hydrogen bonds showed that this group is enriched with proteins with catalytic activity, including an endonuclease,[61] a methyltransferase[61] and a DNA polymerase.[62] In these proteins, the main conservation signal came from the catalytic site rather than the recognition site.

It is important to note that PatchFinder is intended for the identification of various functional regions in proteins.[16,17] Thus, methods that specialize in the identification of DNA-binding regions are presumably superior to PatchFinder in their coverage of the region.[6,7,63]

Identification of DBPs

We examined various features of the ML-patches of the proteins in our dataset in order to discriminate between DBPs and nDBPs. The classifiers developed by Stawiski *et al.*[7] and Bhardwaj *et al.*[12] examined similar features of positively charged regions in the proteins. One fundamental difference between these methods and ours lies in the different regions examined. We examined the regions that are most likely to be functional in each protein. We assumed that the differences between functional regions in the DBPs and nDBPs are more pronounced than the differences between positively charged regions of the two classes. This hypothesis is supported by the marked contribution of the amino acid conservation patterns of the ML-patch to the overall performance (Fig. 5).

The expected significance of real data

The classifier presented here outperformed related methods, both on the Szilágyi & Skolnick dataset that was used for training/testing and on the independent set of 11 DBPs, collected later. We reached a sensitivity and a specificity of 0.90 on the Szilágyi & Skolnick dataset. When the fraction of DBPs is closer to its fraction in proteomes (i.e., 14%) the specificity changes significantly (see Fig. 4). These results suggest that there is still room for improvement. This could possibly be achieved by, e.g., developing new features or using specialized classifiers for specific classes of DBPs.

Measure of confidence for the performance analysis

While it is common in the bioinformatics community to perform a bootstrap-style analysis in order to compute the variance of the cross-validation error estimate, recent theoretical and empirical work concludes that this procedure is not statistically justified. For example, a recent empirical study[64] used simulations to show that both cross-validation and bootstrapping can produce unreliable estimates of the true error rate when the sample size (dataset) is small and found no practical procedure to estimate the uncertainty of these error estimates. For this reason, we do not compute a bootstrap estimate of

the variance of our classifier's cross-validation error rate. Moreover, we caution that the dataset used for training and testing is clearly not a representative sample of the underlying distribution of proteins, which may also cause the empirical error estimate to deviate from the true error rate. Having said that, it is encouraging to note that the classifier detected all the 11 DBPs of the new set correctly (see A small but independent dataset section in Results).

Predictions on N-Func

We provide a list of 218 predicted DBPs from the N-Func database of proteins of unknown function. Preliminary analysis of some of these proteins showed that at least some of them are likely to bind DNA. Some of the proteins predicted as DBPs may present new DNA-binding motifs. Therefore, we encourage further investigation of the potential DBPs in N-Func, in particular those with the highest prediction score.

Materials and Methods

Datasets

Our analysis was based on the datasets of DBPs and nDBPs used by Szilágyi & Skolnick.[8] The first is a non-redundant set (up to 35% sequence identity between each sequence pair) of 138 DBPs that were co-crystallized with dsDNA at a resolution of 3 Å or better. The negative dataset of nDBPs is a representative set of 110 proteins that meet the same redundancy criteria.[9,65] The performance of the classifier on the datasets was measured using 10-fold cross-validation.

The extended dataset (with the lower—presumably more realistic—fraction of DBPs) included additional 733 structures of nDBPs, so that it will form a non-redundant set of proteins with the original dataset of 248 structures. The additional structures were gathered using the PISCES server, which produces lists of PDB entries according to a variety of filters.[66] We used a pre-compiled list of PDB entries. Entries in this list include crystal structures with a resolution better than 3.0 Å. The sequence identity between each pair of sequences is less than 25%. From this list, we removed sequences that share more than 35% sequence identity with one of the sequences in the original training set. Finally, we removed PDB entries that had no GO annotation[34] or had a GO annotation containing any of the following strings: 'DNA', 'MOLECULAR_FUNCTION UNKNOWN' or 'PROCESS UNKNOWN'. From that list, we selected 733 entries at random to extend the list of nDBPs.

Prediction of functional regions using PatchFinder

We predicted the functionally important region of each protein with an algorithm that we developed, called PatchFinder.[16,17] The algorithm searched for statistically significant clusters of spatially close and evolutionarily conserved residues on the protein surface. We showed earlier that these clusters often delineate the regions in the proteins that mediate interactions with other molecules.[16] The algorithm receives as input an MSA of the query protein and its sequence homologues, as well as the 3D coordinates of the protein in PDB format.[18]

Homologous sequences for each protein in the set were gathered from the UniProt database[67] by three PSI-BLAST[47] iterations. Next, the sequences (up to 300) were aligned by CLUSTAL W.[68]

Conservation analysis is often inaccurate for proteins with fewer than four sequence homologues.[19] Therefore, 28 proteins were removed from the dataset, leaving 121 DBPs and 99 nDBPs with ML-patches; however, we trained a second classifier using the 16 global features that did not require the identification of the ML-patch.

Defining the DNA-binding site

We needed to define the DNA-binding site in order to examine the performance of our method. To this end, we considered solvent-accessible residues within 6 Å from the DNA as residues that are in contact with DNA.[63] In this context, it is noteworthy that the crystal structures in the dataset included only fragments of DNA. Furthermore, the protein–DNA complex, as it appears in the crystal, may represent only one out of several binding conformations. Hence, we could not exclude the possibility that a few other residues may be in contact with DNA.

Descriptors

Average electrostatic potential

The average electrostatic potential was computed for the surface atoms of the proteins as well as for the surface atoms of the ML-patch using the following procedure.

1. Assignment of the radii, charges and hydrogen atoms to each atom in the protein using PDB2PQR[69] and the CHARMM force field.[70]
2. Calculation of the electrostatic potential at the points of the grid that encloses the protein using the Adaptive Poisson-Boltzmann Solver (APBS).[71]
3. Interpolation of the electrostatic potential values on the surface points of the protein that were generated by SURF.[72]
4. Averaging the electrostatic potential of the points of the entire protein surface and those of the ML-patch.

Hydrogen bond donors/acceptors

For each protein, we calculated the number of unsatisfied hydrogen bond donors per exposed atom on the protein surface. For this task, we used HBPLUS[73] to identify hydrogen bond donors already satisfied by hydrogen bond acceptors within the protein. These donors were subtracted from the total number of hydrogen bond donors in the protein. A similar procedure was performed for hydrogen bond acceptors.

The average number of hydrogen bond donors/acceptors per exposed atom within the ML-patch provides some separation between DBPs and nDBPs. However, these features did not improve the performance of the classifier, and we therefore used these features only as calculated for the whole protein surface.

Secondary structure content

The DSSP program assigns the secondary structure to each residue in a protein with a known 3D structure.[74] We used DSSP to compute the fraction of residues in the

ML-patch that are in α-helical conformation, the fraction of residues that are in a β-strand and these values for the whole protein surface.

Characterization of the ML-patch based on pre-calculated amino acid conservation patterns

This multi-dimensional descriptor is aimed at representing the amino acid composition of the ML-patch of the query protein and the corresponding positions in its sequence homologues. The pre-calculation was conducted on a non-redundant set of 609 proteins from N-Func. This dataset and the dataset on which we test the classifier are disjoint sets.

1. Removal of redundancy from each MSA with CD-HIT.[75] After filtration, each MSA contained sequences that shared, at most, 90% sequence identity.
2. Calculation of a PSSM for each MSA.[76]
3. Extraction of the PSSM positions that correspond to the ML-patches of the proteins in the dataset.
4. *K*-means clustering of the vector representations of PSSM positions in ML-patches gathered from all the DBPs and the nDBPs. The distance d between each pair of vectors $\mathbf{x}$ and $\mathbf{y}$ was measured with a simple Euclidean distance function.

We applied this procedure with K values between 2 and 20 and found that clustering the data into more than 12 clusters did not improve the performance in terms of AUC. Thus, 12 clusters were constructed and their centroids were calculated.

Based on the 12 pre-calculated cluster centroids, we constructed for each protein a 12-dimensional vector whose components corresponded to the average distances of the vectors at PSSM positions corresponding to ML-patch residues to the 12 cluster centroids. Equation (1) represents the calculation of the ith element in the vector of a protein in the dataset. C_i corresponds to the ith K-means centroid and $\mathbf{x}_j$ correspond to the PSSM vector of position j.

$$v_i(x) = \frac{1}{|MLpatch|} \sum_{j \in MLpatch} d(C_i, x_j) \quad (1)$$

$\mathbf{x} = (\mathbf{x}_k \mid k$ is the position in the PSSM corresponding to a residue in the ML-patch).

Additional features

In addition to the features above, we used the number of residues in the proteins, the number of residues in the ML-patch, and the number of solvent-accessible atoms in the protein.

Classification with random forests

The random forests classifier builds an ensemble of decision trees.[20] Each tree is built on the basis of a subset of the training set. The split at each tree-node is based on a feature selected out of a random subset of the input descriptors. Once the forest of trees has been built, new instances are classified according to the decision of the majority of trees.

We used random forests with 50,000 decision trees. Five descriptors were selected randomly at each node split. The performance of the classifier was evaluated using 10-fold cross-validation.

It is noteworthy that we also examined the data with a support vector machine and achieved similar performance.

Integration of the data from Szilágyi & Skolnick

Szilágyi & Skolnick used a logistic-regression classifier and achieved an MCC of 0.74.[8] In order to reproduce this performance with random forests, we transformed the data using principal components analysis (PCA). The transformation was calculated with the Weka software package.[77] The original 10-dimensional vectors were transformed to nine dimensions conserving 95% of the variance in the original space. Thus, the principal components representation seems to improve over the original feature representation, even though there is little dimensionality reduction. The transformed descriptors were then integrated in the final classifier.

The transformation parameters were deduced from the set of 609 proteins in N-Func described above.

Evaluation of the fraction of DBPs in genomes

It has been estimated that the fraction of genes encoding for DBPs is 2–3% of prokaryotic genomes and 6–7% of eukaryotic genomes.[78] However, due to the availability of new functional data and annotation databases since the calculation of these figures, we decided to re-compute these estimates. We used the GO database in order to evaluate the fraction of DBPs in the genomes of *H. sapiens*, *S. cerevisiae* and *E. coli*.[34] For each of these genomes, we looked at the proteins that are assigned with GO numbers and measured the fraction among them with DNA-related GO numbers (i.e., GO numbers that include the word 'DNA' in their title). The fraction of proteins related to DNA was 14%, on average.

Measures for performance evaluation

We examined the performance of the various classifiers using the following measures, where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively:

$$\mathrm{MCC} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TN + FP) \cdot (TN + FN) \cdot (TP + FP) \cdot (TP + FN)}} \quad (2)$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\mathrm{Sensitivity\ (Recall)} = \frac{TP}{TP + FN} \quad (4)$$

$$\mathrm{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

Acknowledgements

We thank Gilad Wainreb, Matan Kalman, Yanay Ofran, Eran Bacharach and Phaedra Agius for helpful discussions. We thank Roman Laskowski

for conducting the ProFunc calculations on the dataset. A.S. was supported by grant PD73096 from the Hungarian Scientific Research Fund. This work was supported by the BLOOMNET ERA-PG grant.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2009.02.023

References

1. Luscombe, N. M. & Thornton, J. M. (2002). Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.* **320**, 991–1009.
2. Friedberg, I. (2006). Automated protein function prediction–the genomic challenge. *Brief Bioinform.* **7**, 225–242.
3. Shanahan, H. P., Garcia, M. A., Jones, S. & Thornton, J. M. (2004). Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Res.* **32**, 4732–4741.
4. Ferrer-Costa, C., Shanahan, H. P., Jones, S. & Thornton, J. M. (2005). HTHquery: a method for detecting DNA-binding proteins with a helix-turn-helix structural motif. *Bioinformatics*, **21**, 3679–3680.
5. Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, **268**, 1144–1149.
6. Tsuchiya, Y., Kinoshita, K. & Nakamura, H. (2004). Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins: Struct. Funct. Genet.* **55**, 885–894.
7. Stawiski, E. W., Gregoret, L. M. & Mandel-Gutfreund, Y. (2003). Annotating nucleic acid-binding function based on protein structure. *J. Mol. Biol.* **326**, 1065–1079.
8. Szilágyi, A. & Skolnick, J. (2006). Efficient prediction of nucleic acid binding function from low-resolution protein structures. *J. Mol. Biol.* **358**, 922–933.
9. Ahmad, S. & Sarai, A. (2004). Moment-based prediction of DNA-binding proteins. *J. Mol. Biol.* **341**, 65–71.
10. Ahmad, S., Gromiha, M. M. & Sarai, A. (2004). Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.
11. Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
12. Bhardwaj, N., Langlois, R. E., Zhao, G. & Lu, H. (2005). Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Res.* **33**, 6486–6493.
13. Jones, S., Shanahan, H. P., Berman, H. M. & Thornton, J. M. (2003). Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.* **31**, 7189–7198.
14. Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T. & Ben-Tal, N. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* **33**, W299–W302.
15. Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E. & Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154.
16. Nimrod, G., Glaser, F., Steinberg, D., Ben-Tal, N. & Pupko, T. (2005). In silico identification of functional regions in proteins. *Bioinformatics*, **21**, i328–i337.
17. Nimrod, G., Schushan, M., Steinberg, D. M. & Ben-Tal, N. (2008). Detection of functionally important regions in 'hypothetical proteins' of known structure. *Structure*, **16**, 1755–1763.
18. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.
19. Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.* **21**, 1781–1791.
20. Breiman, L. (2001). Random forests. *Mach. Learn.* **45**, 5–32.
21. Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (1997). NUCPLOT: a program to generate schematic diagrams of protein-nucleic acid interactions. *Nucleic Acids Res.* **25**, 4940–4945.
22. Lukacs, C. M., Kucera, R., Schildkraut, I. & Aggarwal, A. K. (2000). Understanding the immutability of restriction enzymes: crystal structure of BglII and its DNA substrate at 1.5 Å resolution. *Nature Struct. Biol.* **7**, 134–140.
23. Pingoud, A., Fuxreiter, M., Pingoud, V. & Wende, W. (2005). Type II restriction endonucleases: structure and mechanism. *Cell Mol. Life Sci.* **62**, 685–707.
24. Dyson, H. J. & Wright, P. E. (2002). Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* **12**, 54–60.
25. Nadassy, K., Wodak, S. J. & Janin, J. (1999). Structural features of protein-nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
26. Klimasauskas, S., Kumar, S., Roberts, R. J. & Cheng, X. (1994). HhaI methyltransferase flips its target base out of the DNA helix. *Cell*, **76**, 357–369.
27. Cheng, X., Kumar, S., Posfai, J., Pflugrath, J. W. & Roberts, R. J. (1993). Crystal structure of the HhaI DNA methyltransferase complexed with S-adenosyl-L-methionine. *Cell*, **74**, 299–307.
28. Roberts, R. J. (1994). An amazing distortion in DNA induced by a methyltransferase. *Biosci. Rep.* **14**, 103–117.
29. O'Gara, M., Klimasauskas, S., Roberts, R. J. & Cheng, X. (1996). Enzymatic C5-cytosine methylation of DNA: mechanistic implications of new crystal structures for HhaL methyltransferase-DNA-AdoHcy complexes. *J. Mol. Biol.* **261**, 634–645.
30. Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. (1999). Protein-DNA interactions: A structural analysis. *J. Mol. Biol.* **287**, 877–896.
31. Mandel-Gutfreund, Y., Schueler, O. & Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J. Mol. Biol.* **253**, 370–382.
32. Pabo, C. O. & Sauer, R. T. (1984). Protein-DNA recognition. *Annu. Rev. Biochem.* **53**, 293–321.
33. Pabo, C. O. & 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.
34. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. *et al.* (2000). Gene ontology:

tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29.
35. Davis, J. & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240, ACM, Pittsburgh, PA.
36. 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: Struct. Funct. Genet.* **61**, 258–271.
37. Shazman, S. & Mandel-Gutfreund, Y. (2008). Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput. Biol.* **4**, e1000146.
38. Lubec, G., Afjehi-Sadat, L., Yang, J. W. & John, J. P. (2005). Searching for hypothetical proteins: theory and practice based upon original data and literature. *Prog. Neurobiol.* **77**, 90–127.
39. Schneider, R. & Sander, C. (1996). The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.* **24**, 201–205.
40. Laskowski, R. A., Watson, J. D. & Thornton, J. M. (2005). ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.* **33**, W89–W93.
41. Laskowski, R. A., Watson, J. D. & Thornton, J. M. (2005). Protein function prediction using local 3D templates. *J. Mol. Biol.* **351**, 614–626.
42. Holm, L. & Park, J. (2000). DaliLite workbench for protein structure comparison. *Bioinformatics*, **16**, 566–567.
43. Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D. *et al.* (2007). New developments in the InterPro database. *Nucleic Acids Res.* **35**, D224–D228.
44. Shazman, S. & Mandel-Gutfreund, Y. (2008). Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput. Biol.* **4**, e1000146.
45. Sandman, K. & Reeve, J. N. (2005). Archaeal chromatin proteins: different structures but common function? *Curr. Opin. Microbiol.* **8**, 656–661.
46. Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T. *et al.* (2006). Pfam: clans, web tools and services. *Nucleic Acids Res.* **34**, D247–D251.
47. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
48. Krissinel, E. & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D*, **60**, 2256–2268.
49. Dueber, E. L., Corn, J. E., Bell, S. D. & Berger, J. M. (2007). Replication origin recognition and deformation by a heterodimeric archaeal Orc1 complex. *Science*, **317**, 1210–1213.
50. Liu, J., Smith, C. L., DeRyckere, D., DeAngelis, K., Martin, G. S. & Berger, J. M. (2000). Structure and function of Cdc6/Cdc18: implications for origin recognition and checkpoint control. *Mol. Cell*, **6**, 637–648.
51. Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C. & Ferrin, T. E. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics*, **7**, 339.
52. Yang, A. S. & Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.* **301**, 665–678.
53. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
54. Alseth, I., Rognes, T., Lindback, T., Solberg, I., Robertsen, K., Kristiansen, K. I. *et al.* (2006). A new protein superfamily includes two novel 3-methyladenine DNA glycosylases from Bacillus cereus, AlkC and AlkD. *Mol. Microbiol.* **59**, 1602–1609.
55. Dalhus, B., Helle, I. H., Backe, P. H., Alseth, I., Rognes, T., Bjoras, M. & Laerdahl, J. K. (2007). Structural insight into repair of alkylated DNA by a new superfamily of DNA glycosylases comprising HEAT-like repeats. *Nucleic Acids Res.* **35**, 2451–2459.
56. Henrick, K. & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358–361.
57. von Hippel, P. H. & Berg, O. G. (1989). Facilitated target location in biological systems. *J. Biol. Chem.* **264**, 675–678.
58. Hu, T., Grosberg, A. Y. & Shklovskii, B. I. (2006). How proteins search for their specific sites on DNA: the role of DNA conformation. *Biophys. J.* **90**, 2731–2744.
59. Slutsky, M. & Mirny, L. A. (2004). Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophys. J.* **87**, 4021–4035.
60. Takeda, Y., Ross, P. D. & Mudd, C. P. (1992). Thermodynamics of Cro protein-DNA interactions. *Proc. Natl Acad. Sci. USA*, **89**, 8180–8184.
61. Flick, K. E., Jurica, M. S., Monnat, R. J., Jr & Stoddard, B. L. (1998). DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-PpoI. *Nature*, **394**, 96–101.
62. Nair, D. T., Johnson, R. E., Prakash, S., Prakash, L. & Aggarwal, A. K. (2004). Replication by human DNA polymerase-iota occurs by Hoogsteen base-pairing. *Nature*, **430**, 377–380.
63. Ofran, Y., Mysore, V. & Rost, B. (2007). Prediction of DNA-binding residues from sequence. *Bioinformatics*, **23**, i347–i353.
64. Isaksson, A., Wallman, M., Göransson, H. & Gustafsson, M. G. (2008). Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recogn. Lett.* **29**, 1960–1965.
65. Rost, B. & Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl Acad. Sci. USA*, **90**, 7558–7562.
66. Wang, G. & Dunbrack, R. L., Jr (2005). PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* **33**, W94–W98.
67. Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S. *et al.* (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159.
68. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
69. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. (2004). PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32**, W665–W667.

70. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. *et al.* (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, **102**, 3586–3616.
71. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA*, **98**, 10037–10041.
72. Varshney, A. & F. P. Brooks, J. (1993). Fast analytical computation of richards's smooth molecular surface. *IEEE Visualization '93*,, 300–307.
73. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793.
74. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
75. Li, W. & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
76. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
77. Witten, I. H. & Frank, E. (2005). In *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edit. Morgan Kaufmann, San Francisco, CA.
78. Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. (2000). An overview of the structures of protein-DNA complexes. *Genome Biol.* **1**; REVIEWS001.1-001.37.
79. Sanner, M. F. (1999). Python: a programming language for software integration and development. *J. Mol. Graph. Model.* **17**, 57–61.