BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i347–i353
doi:10.1093/bioinformatics/btm174

Prediction of DNA-binding residues from sequence

Yanay Ofran[1,2,*], Venkatesh Mysore[1,3] and Burkhard Rost[1,2,4]

[1]Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, [2]Columbia University Center for Computational Biology and Bioinformatics (C2B2), 1130 St Nicholas Ave. Rm. 802, [3]D.E.Shaw Research, 120 West Forty Fifth Street (current affiliation) and [4]NorthEast Structural Genomics Consortium (NESG), Columbia University, 1130 St Nicholas Ave. Rm. 802, New York, NY 10032, USA

ABSTRACT

Motivation: Thousands of proteins are known to bind to DNA; for most of them the mechanism of action and the residues that bind to DNA, i.e. the binding sites, are yet unknown. Experimental identification of binding sites requires expensive and laborious methods such as mutagenesis and binding essays. Hence, such studies are not applicable on a large scale. If the 3D structure of a protein is known, it is often possible to predict DNA-binding sites *in silico*. However, for most proteins, such knowledge is not available.

Results: It has been shown that DNA-binding residues have distinct biophysical characteristics. Here we demonstrate that these characteristics are so distinct that they enable accurate prediction of the residues that bind DNA directly from amino acid sequence, without requiring any additional experimental or structural information. In a cross-validation based on the largest non-redundant dataset of high-resolution protein–DNA complexes available today, we found that 89% of our predictions are confirmed by experimental data. Thus, it is now possible to identify DNA-binding sites on a proteomic scale even in the absence of any experimental data or 3D-structural information.

Availability: http://cubic.bioc.columbia.edu/services/disis

Contact: yo135@columbia.edu

1 INTRODUCTION

1.1 Protein–DNA interfaces are important but not easy to identify experimentally

Interactions between DNA and proteins are at the heart of many biological processes including transcription and transcriptional regulation, recombination, replication, DNA repair, viral infection, DNA packing and DNA modifications. However, the biophysical underpinnings of these interactions are not entirely clear. Studies of the molecular mechanisms of protein–DNA interaction often focus on protein–DNA interfaces, i.e. the surface residues that bind DNA. Such residues can be identified from 3D structures of protein–DNA complexes (Siggers and Honig, 2007). Unfortunately, 3D structures of such complexes are available for less than 5% of all known DNA-binding proteins. In the absence of a 3D structure of the complex, various biochemical approaches are employed to identify binding residues. There is no standard high-throughput protocol for the identification of DNA-binding sites. Methods such as protein-binding microarrays (Bulyk, 2006) identify DNA-binding proteins on a large scale. However, they do not reveal which residues actually bind the DNA in a straightforward manner.

1.2 Interface residues can be predicted from 3D structures

Studies of protein–DNA interfaces suggest that the amino acids at the interface possess characteristics that distinguish them from the rest of the protein (Lejeune *et al*., 2005; Mandel-Gutfreund and Margalit, 1998; Mandel-Gutfreund *et al*., 1995; Nadassy *et al*., 1999; Pabo and Sauer, 1984). Jones *et al*., 2003 did not only point out electrostatic differences between DNA-binding patches and the rest of the protein surface, but have also demonstrated that these differences may suffice for the prediction of interaction sites from the coordinates of the 3D structure of a protein (Shanahan *et al*., 2004). Other studies have demonstrated that different combinations of electrostatic, biophysical and structural features can predict DNA-binding sites, given the structures of the unbound protein (Ahmad *et al*., 2004; Ahmad and Sarai, 2004; Keil *et al*., 2004; Kuznetsov *et al*., 2006; Tsuchiya *et al*., 2004; Tsuchiya *et al*., 2005); such predictions exploit details about protein surfaces that are not available from the sequence alone, but do not require a structure of the protein-DNA complex. These ground-breaking studies have opened the door for the detailed experimental study of DNA-binding proteins for which unbound 3D structures are available but structures of the protein–DNA complexes are not. Since there are 3D structures for only a small fraction of the known DNA-binding proteins (bound or unbound), it was suggested that computational models of 3D structures could be used to predict binding residues; in fact, such models can succeed very well (Szilagyi and Skolnick, 2006). Yet, for over 40% of known DNA-binding proteins, there are no known homolog that allow for reliable models. Methods that predict binding sites for unbound structures have another potential shortcoming: they are typically trained (and tested) on interfaces obtained from experimentally determined protein–DNA complexes. Then, unbound structures are searched for surface patches that are similar to those observed in the protein–DNA complexes. However, in many cases, proteins undergo substantial changes upon binding DNA. Hence, the unbound and the bound structures often differ. Thus, structure-based prediction methods that are trained on known complexes might fail to identify the binding sites in many unbound structures.

*To whom correspondence should be addressed.

1.3 First successes in predicting protein–DNA interfaces from sequence alone

Similar problems and challenges exacerbate the attempt to predict other types of interfaces from structure (Jones and Thornton, 2004). For example, pioneering methods for predicting protein–protein interaction sites relied on structural information (Chung *et al*., 2006; Fariselli *et al*., 2002; Fernandez-Recio *et al*., 2004; Jones and Thornton, 1997b; Neuvirth *et al*., 2004). However, several methods have demonstrated that protein–protein interaction sites could be predicted directly from sequence (Koike and Takagi, 2004; Ofran and Rost, 2006; Res *et al*., 2005; Wang *et al*., 2005). This is because interface residues have very unique traits (Ofran and Rost, 2003a, b) and are often organized in groups along the sequence (Ofran and Rost, 2003a). The predictability of interface residues is particularly surprising given the diversity of protein–protein interfaces in structure, size, electrostatic and other physicochemical characteristics (Jones and Thornton, 1997; Lo Conte *et al*., 1999; Ofran and Rost, 2003; Sheinerman *et al*., 2000). Protein–DNA interfaces, on the other hand, may be less diverse in their features since the different DNA segments are more similar to each other than different surface patches on proteins. Thus, it may be possible to identify DNA-binding residues from sequence alone with even greater accuracy than predictions of protein–protein interfaces. Ahmad *et al*., 2004 have used (NN) to predict DNA-binding residues based on their sequence environment and their solvent accessibility, derived from experimentally determined 3D structures. Although the method still relies on experimental 3D structures, its success has demonstrated that some of the characteristics of DNA-binding residues can be identified from sequence alone. In a subsequent study the same group has shown that position specific scoring matrices (PSSMs) alone can predict binding residues with some accuracy. Similarly, Yan *et al*. (2006) have trained naïve Bayes classifiers using sequence neighborhood and evolutionary conservation; they report their method to perform substantially better than simple PSSMs. Despite this success, the performance is still substantially worse than that of methods that benefit from 3D information.

Here, we introduce a novel method that uses only protein sequence information to predict whether or not and with which residues a protein binds DNA. The method relies on sequence environment, evolutionary profiles and predicted structural features (secondary structure, solvent accessibility, globularity). These features were combined through machine learning algorithms, namely through NN and SVM. The algorithms were trained to distinguish between residues that are in contact with DNA and those that are not. Figure 1 sketches the different analyses that were integrated to yield the final prediction. The method that we present here is based on a similar approach to the approach we implemented in our method ISIS (Interaction Sites Identified from Sequence) for predicting protein–protein binding sites (Ofran and Rost, 2006). Thus, we called this new method DISIS—DNA interaction sites identified from sequence.

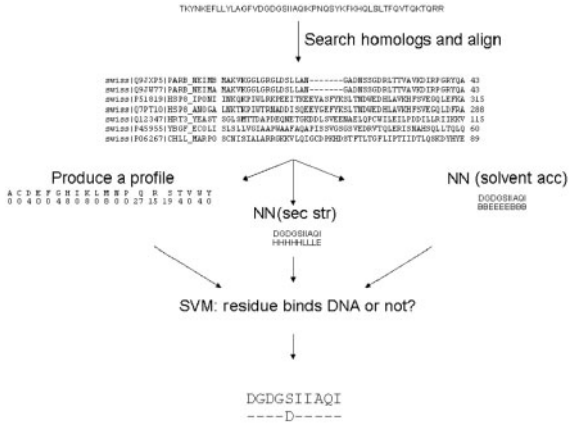

Fig. 1. Schematic description of DNA interaction sites identified from sequence (DISIS) predictions. Given a query protein sequence, DISIS performs the following procedures. First, a standard PSI-BLAST is used to find all proteins related to the query. Then, MaxHom is used to align all the sequences that were found and the alignments is sent to the PROF server, which uses neural networks (NN) to predict the secondary structure and the solvent accessibility of each residue. In addition, for each residue the evolutionary profile and evolutionary conservation are calculated (using MaxHom). Finally, all these features are fed to support vector machines (SVM) to determine for each residue whether it binds DNA or not.

2 METHODS

2.1 Dataset: definition of protein–DNA interfaces

For training and testing we used a non-redundant subset (below) of all protein–DNA complexes in the PDB (Berman *et al*., 2000). For each complex, we defined the protein–DNA interface as all the residues on the protein that were in contact with the DNA. Amino and nucleic acids were considered in contact if any of their atoms were closer than 6 Å. Previous studies used distances from 4 Å to 12 Å between *C*-alpha or between *C*-beta atoms. However, the variations in the size of side chains might result in an under-representation of large residues in the data, as their side chain themselves can extend over several Ångstrøms. Hence, we defined contacts based on the distance between the closest pair of atoms. While this definition is not biased by amino acid size, it is slightly more permissive than some other definitions, i.e. it tends to define more residues as DNA binding. Thus, rather then biasing the data towards some residues, our permissive definition introduced some white noise (Ofran and Rost, 2003).

2.2 Non-redundant subsets

In order to reduce bias from very similar sequences in the database, we built sequence-unique subsets for all types of proteins under consideration. Using the HSSP-value as a measure of sequence similarity (Mika and Rost, 2003; Rost, 1999; Sander and Schneider, 1991), we built three sets of sequences such that no two proteins from different sets had HSSP-values >0. For alignments over 250 residues, this translated to less than 20% pair-wise sequence identity (PIDE), i.e. pairs with >20% PIDE were not included. Altogether, we had 127 064 residues in our dataset, 23 862 were in contact with DNA and 103 202 were not (the list of PDB files we used is available form our website). We used these three sets for training (optimizing connections in NN/SVMs), cross-training (optimizing additional parameters such as when to stop training) and testing. We then rotated through the sets such that ultimately each

protein had been used for testing exactly once. We only reported performance estimates for the test set.

2.3 Input features

We used the principle of sliding windows to capture the sequence environment, i.e. when predicting whether or not residue k binds DNA, we included the information of w consecutive residues: $k-(w-1)/2, \ldots, k, \ldots, k+(w-1)/2$. The features that were used for training and testing included the evolutionary profile of each residue and its neighbors (four on either side, i.e. $w=9$), the level of conservation of the residue and its neighbors (one on each side, i.e. $w=3$ for this feature), the predicted secondary structure of the residue and the predicted solvent accessibility of the residue and its neighbors (one on each side). We tried various numbers of neighboring residues for each of these parameters and found the system to be rather robust under such parameter changes. The combination mentioned earlier yielded the best performance on a cross-training set (see in the following text; note that no parameter was optimized to yield best performance on the final test set).

2.4 Evolutionary profiles

To obtain evolutionary profiles, we first aligned each protein in our dataset against a filtered version of all currently known sequences using PSI-BLAST (Altschul *et al.*, 1997) with three iterations (Przybylski and Rost, 2002) (cut-off at 10^{-3}). We then realigned the final set of proteins suggested by PSI-BLAST with the dynamic programming algorithm MaxHom (Sander and Schneider, 1991; Schneider and Sander, 1996); MaxHom profiles were used as input both into the PROFphd series of methods predicting secondary structure and solvent accessibility (Rost *et al.*, 2004) and to the method described here that predicted residues in protein–DNA interfaces.

2.5 Machine learning algorithms

We used standard feed-forward NN as described in detail elsewhere (Rost and Sander, 1993; Rost, 1996). We implemented SVM (Vapnik, 1995) by using the SVM-light package (Joachims, 1999) with the radial kernel function.

2.6 Filter of SVM output and default prediction threshold

The SVM assigns a score to each residue central in the sliding input window. Our default threshold for translating an SVM score into a prediction of DNA binding was defined as follows: if this score was >0.35, we predicted the residue to be DNA binding (at this level, 83% of the predictions were confirmed experimentally); if the raw score was between 0.35 and −0.3, we marked the residue as putative DNA binding. For each putative DNA-binding residue, we then scanned its eight sequence neighbors (four on each side), and counted the number of residues that were predicted to be putative or positive DNA binders. If there were five or more such residues we annotated the putative residue as DNA binding; otherwise, we marked it as non-DNA binding. This in-between mapping of predictions effectively corresponded to filtering the output. All other residues were predicted as non-DNA binding. Again, these parameter choices were found to be about right with a broad stability when analyzing the cross-training performance on a single data set split, i.e. these parameters were set before we monitored the performance on the final test set.

2.7 Alternative decision thresholds

Changing the threshold used to translate from the SVM output into DNA binding/non-binding enables dialing through different points in a ROC-like curve (Fig. 2). Effectively this dial lets users of the method focus on extreme ends of the tradeoff between accuracy and coverage: they may focus on few very reliable predictions, or on many less reliable ones. The same threshold can also be used to define a reliability index that predicts the accuracy of a prediction.

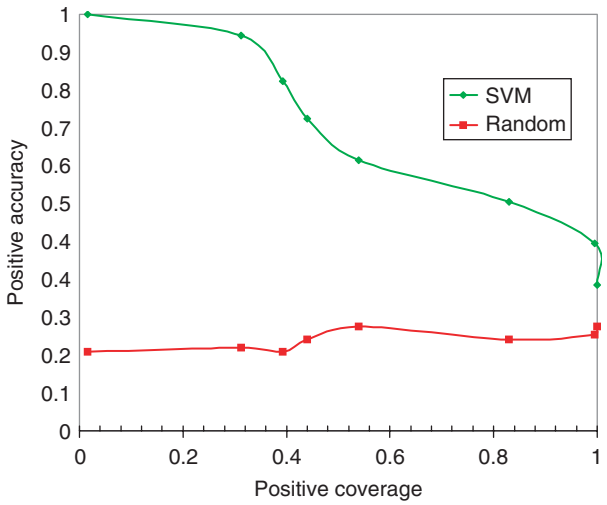

Fig. 2. Accuracy versus coverage: DISIS (green) and a random assignment (red) using PDB interfaces as gold standard—the data was compiled for a set of proteins that was not used for developing the method. The stronger the confidence in our prediction, the higher the accuracy and the lower the coverage, i.e. when we select the strongest predictions, most of these are right. At accuracy of 0.7, DISIS correctly predicted at least one residue in all the proteins in our data set.

2.8 Performance measures

As a single overall measure for performance, we used the two-state per residue accuracy defined as follows:

$$Q_2 = 100 \times \frac{TN+TP}{TN+TP+FN+FP} \quad (1)$$

where TP are the true positives (residues correctly predicted to bind DNA), TN the true negatives (residues correctly predicted not to bind DNA), FN the false negatives (predicted not to bind, observed to bind) and FP the false positives (predicted to bind but not observed).

Since our dataset contained more non-binding residues (81%) than binding residues, methods that over-predict non-binding residues would reach high values of Q_2. In order to capture such over-predictions, we also measured the positive accuracy (ACC; often referred to as specificity or precision) and the positive coverage (COV; often referred to as sensitivity) for the inference ('prediction') of interacting residues by the standard formulae:

$$ACC = \frac{TP}{TP+FP}; \quad COV = \frac{TP}{TP+FN} \quad (2)$$

3 RESULTS

3.1 Assessment on comprehensive non-redundant high-resolution data

The search of protein–DNA complexes in the PDB yielded, after reducing redundancy, 274 complexes with 693 chains and 127 064 residues. 23 862 of these residues (19%) were involved in contacts (closest atom ≤6 Å) between amino and nucleic

Table 1. Accuracy of predictions using different features

Method	Q_2
Sequence only	59
Evolutionary data	67
Evolution + sequence	78
ISIS (protein–protein interaction)	68
DISIS	89

Using the sequence environment of each residue alone, it is possible to predict most residues correctly. Using evolutionary information alone (e.g. PSSM), the performance is even better. The combination of the two, further improves the performance. DISIS uses these two features but also adds predicted secondary structure and predicted solvent accessibility. It is interesting to compare DISIS to ISIS, which uses the same features to predict protein–protein interaction sites. The accuracy of ISIS is much lower, suggesting that the similarities between different protein–DNA interfaces are greater than the similarities between protein–protein interfaces.

acid(s), i.e. bound DNA. We used this set in a 3-fold cross-validation (one-third for training, one-third for cross-training, one-third for testing; full rotation to ascertain that each residue was used for testing exactly once). All results reported the performance on all 274 protein chains in the test set. The SVMs were trained to classify individual residues into two classes: either DNA binding or non-DNA binding.

3.2 Raw SVM: high accuracy at low coverage

Our first observation was that the raw SVM already performed significantly better than random. In particular, the raw output yielded very high positive accuracy (Equation 2 correctly predicted DNA-binding residues/all predicted to bind) but fairly low positive coverage (Equation 2 correctly predicted DNA-binding residues/observed DNA-binding residues). For each residue, the SVM returns a number between −3 and 3. It was possible to find a cutoff score such that 95% of the residues with this score or higher were indeed observed in the complex to be part of the interface. However, for this level of accuracy, the coverage was below 5%, i.e. only 5% of the observed residues were successfully identified at this level. We also observed that accuracy dropped steeply below this cutoff score. For example, raising coverage by only 3% points to 8%, dropped accuracy by 30 points to 65%. Although the accuracy of the raw output kept dropping when increasing coverage further, it remained significantly higher than the level expected at random (observed DNA-binding residues/total number of residue = 0.18). These results indicated that the signal identified by the SVM was very strong for some residues. However, for most of the DNA-binding residues in the dataset the signal was less distinct. Nevertheless, the observation that the accuracy remained considerably above random at all levels of coverage suggested that improving performance through postprocessing of the raw SVM output may be possible.

3.3 Positive and two-state accuracy

Positive accuracy and coverage (Fig. 2) only reflect the performance on DNA-binding residues. Other methods typically also report performance for non-binding residues. Usually, this is accomplished by simply quoting the two-state per residue accuracy (which is biased by the correct prediction of non-binding, Methods, Equation 1). The two-state accuracy is the total number of correctly predicted residues, both positive and negative, over the total number of residues. For our method it is 89% (Table 1).

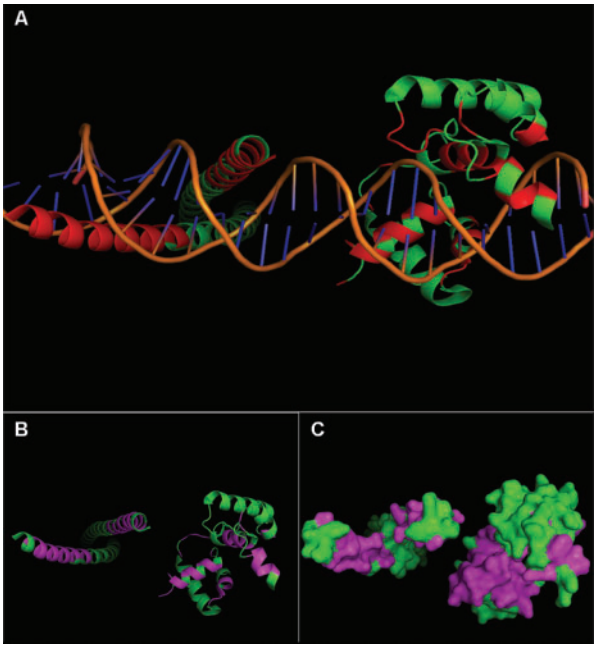

Fig. 3. Prediction of binding residues in *c*-Myb and C/EBPβ, bound to DNA—the ternary complex of the *c*-Myb protein, a regulator of proliferation and differentiation of hematopoietic cells, and the *C*-terminal portion C/EBPβ, a CAAT-enhancer binding protein, both bound to DNA, is used to demonstrate the predictions of DISIS. The predictions were made based only on the sequence of the two proteins and were later mapped to the structure. Note that the residues predicted to bind DNA (purple) create a contiguous patch on the surface of the protein (C). However, in the cartoon representation (B), it is apparent that the predictions are not contiguous in sequence.

3.4 Example for DISIS performance

We illustrated the performance of DISIS for a particular example (Fig. 3), namely that for the ternary complex of the *c*-Myb protein, a regulator of proliferation and differentiation of hematopoietic cells, and the *C*-terminal portion C/EBPβ, a CAAT-enhancer binding protein, both bound to DNA. All the predicted binding residues (purple) fell within the patches that bind DNA, although the predictions were not consecutive in sequence.

4 DISCUSSION

4.1 Combining SVM and NN

DISIS is based on a combination of physicochemical features, evolutionary information and predicted structural features. Correlations between such features and binding are typically so

subtle that we cannot use simple linear statistics to predict them. Therefore, we used a combination of artificial NN and SVMs for this task. These algorithms implicitly, yet reliably, identified common denominators between protein–DNA interfaces that have no sequence similarity. Thereby, we developed a method that predicted residues in protein–DNA interfaces for uncharacterized sequences.

4.2 Performance estimates provided lower bounds

The experimental data that we used was incomplete with respect to the positives: the fact that a particular residue is not observed to bind DNA in a particular complex does not prove that it will not bind to DNA at all. Any 3D complex provides only a snapshot of the interaction at a given moment. Therefore, the treatment of FP is a crucial factor in the assessment. Since it is common for DNA-binding proteins to change their conformation during their interaction with DNA (Richter and Eigen, 1974; von Hippel and Berg, 1989), the missing data problem may not only affect surface but also buried residues. Furthermore, a single protein can have several alternative DNA-binding sites. Thus, when a method predicts a certain residue as DNA binding, it may be correct although the 3D complex does not support this prediction. Nevertheless, we deemed all residues that were not observed in the PDB complexes as negative examples (i.e. not DNA binding) and any prediction that identified any of these residues as DNA binding as incorrect (FP). This solution was conservative in the sense that it clearly under-estimated the true performance, at least for the major score that we reported, namely the accuracy in predicting interaction residues. Note that the fact that a protein may change its conformation upon binding to DNA did not at all influence our prediction method, since we used no information from the 3D complex other than the labels on the residues (binding/non-binding) during training. All the information that we used for testing would have been identical between a bound and an unbound structure, as it was entirely sequence-based.

4.3 DISIS succeeded in the absence of annotations as well as for singletons

Several studies have suggested that evolutionary information could help in predicting DNA-binding residues (Sarai and Kono, 2005; Stawiski *et al*., 2003; Szilagyi and Skolnick, 2006; Yan *et al*., 2006). However, 30–70% of known protein domains have no annotated homologs (Fischer and Eisenberg, 1999). The applicability of methods that exclusively use homology-related information is therefore limited, particularly for analyses on the scale of entire proteomes. Our approach relies on a combination of features and is hence capable of providing reliable predictions even in the absence of evolutionary information. Furthermore, DISIS only used information available for all unannotated sequences. About 10% of the sequences in our test set had less than 10 family members, while 3% had no known family member in publicly available databases. Conservation-based methods would not be able to analyze these sequences. However, DISIS provides predictions for these sequences with two state accuracy of 0.76—substantially higher than the expected at random. The numbers for family sizes for complexes from the PDB clearly over-estimated the situation for entire proteomes for which we may observe 15–40% singletons (Liu and Rost, 2001; Liu and Rost, 2002; Liu *et al*., 2004), i.e. the value of our method is likely significantly higher than what the above estimates suggest.

4.4 Comparison to other methods

Prediction methods can only be compared meaningfully using the same datasets and the same standards of measuring performance. Such comparisons are scientifically meaningful only if the methods are similar in their goals and scope. Hence, studies that rely on structure (or on modeled structure) are not comparable to our method. Similarly, methods that classify proteins as DNA binding but do not identify the binding sites are also not comparable to our method. Two studies have suggested that prediction of DNA-binding site directly from sequence may be possible (Sarai and Kono, 2005; Yan *et al*., 2006). Their analysis is based on different datasets; the comparison of published values for performance and those that we estimated had, therefore, very limited validity. To compare the performance of alternative approaches to that of DISIS, we implemented methods that are based on principles used in other studies and tested them on our data set to estimate the importance of particular features (Table 1). Methods that are based on sequence data alone achieve levels of accuracy substantially higher than random (Sarai and Kono, 2005) (Table 1). When other features, such as conservation, are added the accuracy improves substantially (Yan *et al*., 2006). DISIS, which also incorporated predicted secondary structure and predicted solvent accessibility reached levels of accuracy that were substantially higher.

4.5 DNA binding marked by clearer signals than protein–protein binding

ISIS, a method that we have developed to predict interaction sites in transient protein–protein interactions (Ofran and Rost, 2006), which uses input features similar to DISIS, was much less accurate than DISIS. While this gap could be attributed to various factors (e.g. ISIS is based on a system of NN, while DISIS is based on a combination of NN and an SVM), the most probable explanation is that the sequence signal that marks DNA-binding sites is much stronger than that of protein–protein interaction sites. Put differently, these results suggested that the similarity between different DNA-binding sites is greater than the similarity between different protein-binding sites.

5 CONCLUSIONS

We showed that DNA-binding sites in different proteins share common denominators that could be characterized as a combination of their physicochemical features (as manifested by their sequence environment and by their evolutionary profile), their local structure (secondary structure elements and exposure to solvent) and their evolutionary conservation. This fact enabled the accurate prediction of binding sites even

in the absence of any experimental information (in particular, without using 3D structures), in the absence of annotations and even in the absence of evolutionary information. Comparing the performance of our final method DISIS to the performance of related methods that used less information, we concluded that the underlying structure predictions were essential for the success in predicting DNA binding. A major challenge for the postgenomic era is the development of large-scale, automated tools for the functional annotation of proteins (Roberts, 2004). DISIS responds to this challenge by providing an *in silico* tool that can reliably annotate binding sites in DNA-binding proteins.

ACKNOWLEDGEMENTS

We thank Guy Nimrod, Nir Ben-Tal (Tel Aviv University), and Trevor Siggers (Harvard University), for helpful discussions. We thank Jinfeng Liu, Andrew Kernytsky and Michael Honig (Columbia University) for help with computers and databases. This work was supported by the Grants 1-R01-GM64633 from the National Institute of General Medicine (NIGMS) at the National Institutes of Health (NIH) and 2-R01-LM007329 from the National Library of Medicine (NLM). Last, but not the least, we thank all those who deposit their experimental data in public databases, and to those who maintain these databases.

Conflict of Interest: none declared.

REFERENCES

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

Ahmad,S. and Sarai,A. (2004) Moment-based prediction of DNA-binding proteins. *J. Mol. Biol.*, **341**, 65–71.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bulyk,M.L. (2006) Analysis of sequence specificities of DNA-binding proteins with protein binding microarrays. *Methods Enzymol.*, **410**, 279–299.

Chung,J.L. *et al*. (2006) Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins*, **62**, 630–640.

Fariselli,P. *et al*. (2002) Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.*, **269**, 1356–1361.

Fernandez-Recio,J. *et al*. (2004) Identification of protein-protein interaction sites from docking energy landscapes. *J. Mol. Biol.*, **335**, 843–865.

Fischer,D. and Eisenberg,D. (1999) Finding families for genomic ORFans. *Bioinformatics*, **15**, 759–762.

Joachims,T. (1999) Making large-scale SVM learning practical. In Scholkopf,B., Burges,C. and Smola,A. (eds.), *Advances in Kernel Methods - Support Vector Learning*. MIT press, Cambridge Massachusetts.

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

Jones,S. and Thornton,J.M. (1997a) Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.*, **272**, 121–132.

Jones,S. and Thornton,J.M. (1997b) Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.*, **272**, 133–143.

Jones,S. and Thornton,J.M. (2004) Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.*, **8**, 3–7.

Keil,M. *et al*. (2004) Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a neural network. *J. Comput. Chem.*, **25**, 779–789.

Koike,A. and Takagi,T. (2004) Prediction of protein-protein interaction sites using support vector machines. *Protein Eng. Des. Sel.*, **17**, 165–173.

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

Lejeune,D. *et al*. (2005) Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins*, **61**, 258–271.

Liu,J. and Rost,B. (2001) Comparing function and structure between entire proteomes. *Protein Sci.*, **10**, 1970–1979.

Liu,J. and Rost,B. (2002) Target space for structural genomics revisited. *Bioinformatics*, **18**, 922–933.

Liu,J. *et al*. (2004) Automatic target selection for structural genomics on eukaryotes. *Proteins: Structure, Function, and Bioinformatics*, **56**, 188–200.

Lo Conte,L. *et al*. (1999) The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.

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.

Mandel-Gutfreund,Y. *et al*. (1995) Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J. Mol. Biol.*, **253**, 370–382.

Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.

Nadassy,K. *et al*. (1999) Structural features of protein-nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.

Neuvirth,H. *et al*. (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.*, **338**, 181–199.

Ofran,Y. and Rost,B. (2003a) Analysing six types of protein-protein interfaces. *J. Mol. Biol.*, **325**, 377–387.

Ofran,Y. and Rost,B. (2003b) Predicted protein-protein interaction sites from local sequence information. *FEBS Lett.*, **544**, 236–239.

Ofran,Y. and Rost,B. (2006) ISIS: Interaction Sites Identified from Sequence. *Bioinformatics*, **23(2)**, e13–e16.

Pabo,C.O. and Sauer,R.T. (1984) Protein-DNA recognition. *Annu. Rev. Biochem.*, **53**, 293–321.

Przybylski,D. and Rost,B. (2002) Alignments grow, secondary structure prediction improves. *Proteins*, **46**, 197–205.

Res,I. *et al*. (2005) An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, **21**, 2496–2501.

Richter,P.H. and Eigen,M. (1974) Diffusion controlled reaction rates in spheroidal geometry. Application to repressor–operator association and membrane bound enzymes. *Biophys. Chem.*, **2**, 255–263.

Roberts,R.J. (2004) Identifying protein function–a call for community action. *PLoS Biol.*, **2**, E42.

Rost,B. (1996) PHD: predicting one-dimensional protein structure by profile based neural networks. *Method. Enzymol.*, **266**, 525–539.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Rost,B. *et al*. (2004) The PredictProtein server. *Nucleic Acids Res.*, **32**, W321–326.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Sarai,A. and Kono,H. (2005) Protein-DNA recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.*, **34**, 379–398.

Schneider,R. and Sander,C. (1996) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **24**, 201–205.

Shanahan,H.P. *et al*. (2004) Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Res.*, **32**, 4732–4741.

Sheinerman,F.B. *et al*. (2000) Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct Biol.*, **10**, 153–159.

Siggers,T.W. and Honig,B. (2007) Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Res.*, **35(4)**, 1085–1097.

Stawiski,E.W. *et al*. (2003) Annotating nucleic acid-binding function based on protein structure. *J. Mol. Biol.*, **326**, 1065–1079.

Szilagyi,A. and Skolnick,J. (2006) Efficient prediction of nucleic acid binding function from low-resolution protein structures. *J. Mol. Biol.*, **358**, 922–933.
Tsuchiya,Y. *et al.* (2004) Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins*, **55**, 885–894.
Tsuchiya,Y. *et al.* (2005) PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces. *Bioinformatics*, **21**, 1721–1723.
Vapnik,V.N. (1995) *The nature of statistical learning theory*. Springer, New York.
von Hippel,P.H. and Berg,O.G. (1989) Facilitated target location in biological systems. *J. Biol. Chem.*, **264**, 675–678.
Wang,B. *et al.* (2005) Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Lett*, **580(2)**, 380–384.
Yan,C. *et al.* (2006) Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformat.*, **7**, 262.