BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 5 2010, pages 692–693
doi:10.1093/bioinformatics/btq019

Structural bioinformatics

Advance Access Publication January 19, 2010

iDBPs: a web server for the identification of DNA binding proteins

Guy Nimrod[1,2], Maya Schushan[1], András Szilágyi[3], Christina Leslie[4] 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]The Mina & Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan 52900, Israel, [3]Institute of Enzymology, Hungarian Academy of Sciences, H-1113 Budapest, Hungary and [4]Computational Biology Program, Memorial Sloan-Kettering Cancer Center, NY 10065, USA

Associate Editor: Burkhard Rost

ABSTRACT

Summary: The iDBPs server uses the three-dimensional (3D) structure of a query protein to predict whether it binds DNA. First, the algorithm predicts the functional region of the protein based on its evolutionary profile; the assumption is that large clusters of conserved residues are good markers of functional regions. Next, various characteristics of the predicted functional region as well as global features of the protein are calculated, such as the average surface electrostatic potential, the dipole moment and cluster-based amino acid conservation patterns. Finally, a random forests classifier is used to predict whether the query protein is likely to bind DNA and to estimate the prediction confidence. We have trained and tested the classifier on various datasets and shown that it outperformed related methods. On a dataset that reflects the fraction of DNA binding proteins (DBPs) in a proteome, the area under the ROC curve was 0.90. The application of the server to an updated version of the N-Func database, which contains proteins of unknown function with solved 3D-structure, suggested new putative DBPs for experimental studies.

Availability: http://idbps.tau.ac.il/

Contact: NirB@tauex.tau.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 3, 2009; revised on December 21, 2009; accepted on January 12, 2010

1 INTRODUCTION

DNA binding proteins (DBPs) compose a considerable part of the proteomes of the various organisms (Nimrod *et al.*, 2009), and take part in various processes, such as DNA transcription, replication and packing. There are a number of approaches for the identification of DBPs. Some methods look for direct similarity between the query protein and DBPs (e.g. Gao and Skolnick, 2008; Shanahan *et al.*, 2004). When the DNA binding domain is novel, methods that do not rely directly on previous data may be advantageous. Such methods often rely on electrostatic features of the proteins. DNA is negatively charged, and the DNA binding region of the protein is often positively charged. Therefore, features of positively charged patches on the proteins' surfaces have been examined in order to identify DBPs (Bhardwaj *et al.*, 2005; Stawiski *et al.*, 2003). Other features that represent the distribution of charges within the protein structure have also been used (Ahmad and Sarai, 2004; Szilágyi and Skolnick, 2006), as well as secondary structure content (Stawiski *et al.*, 2003) and the amino acid composition (Szilágyi and Skolnick, 2006).

We recently developed a method for the prediction of DBPs based on the identification of the functional region within the query protein (Nimrod *et al.*, 2009). We showed that patches of highly conserved amino acids, detected by PatchFinder (Nimrod *et al.*, 2008), often delineate the functional regions in proteins in general, and the core of DNA binding regions within DBPs in particular (Nimrod *et al.*, 2009).

Using features of the predicted functional regions and additional global features, we trained a random forests classifier (Breiman, 2001) on a dataset of 138 DBPs and 110 proteins that do not bind DNA (Szilágyi and Skolnick, 2006).

We examined the classifier on a realistic dataset that reflects the fraction of DBPs in proteomes. We evaluated this fraction to be 14% and extended the original dataset by 733 additional proteins that do not bind DNA. The sensitivity and the precision on this dataset were 0.90 and 0.35, respectively, with the default prediction score cutoff. The area under the ROC curve (AUC) was 0.90. We also showed that the performance of the classifier was superior to related methods (Nimrod *et al.*, 2009).

Here, we present the iDBPs web server, which implements the classifier. The server is freely available at http://idbps.tau.ac.il/. It is easy to use and only requires the PDB file (or PDB id) and the chain identifier of the protein of interest.

2 RESULTS

The N-Func database is a collection, which we recently established, of proteins of known three-dimensional (3D)-structure that lack functional annotation (Nimrod *et al.*, 2008). The functional region of each of the proteins in N-Func was predicted using PatchFinder as a first step toward the annotation of these proteins. Here, we present an updated version of the database, which includes 973 PDB entries and their predicted functional regions.

Next, we applied the iDBPs server to N-Func in order to identify potential DNA binders. The results, available as Supplementary Table 1, include the prediction score of each protein as well as the corresponding estimated precision and sensitivity.

Using the default prediction threshold, 233 proteins were identified as potential DBPs. At this threshold, the expected precision is only 0.35, while the sensitivity is 0.9. However, one can filter

*To whom correspondence should be addressed.

the results using different thresholds in order to gather predictions with high precision. Supplementary Figure 2 presents an example of predicted DBP from N-Func.

We previously showed that many of the patches cover most of the hydrogen bonds within the protein–DNA interface in DBPs (Nimrod *et al.*, 2009). Here, we also show that they cover most of the interface positions that interact with the DNA bases (Supplementary Material and Supplementary Fig. 1).

3 IMPLEMENTATION

3.1 Prediction of functional regions in the protein

PatchFinder uses as input the protein structure (or a model) and a multiple sequence alignment (MSA) of the query protein and its sequence homologs. The MSA is generated automatically using the procedure implemented in ConSurf-DB (Goldenberg *et al.*, 2009). PatchFinder searches for statistically significant clusters of evolutionarily conserved residues on the protein surface (ML-patches), which often correspond to the functional regions in proteins (Nimrod *et al.*, 2008). When only a few sequence homologs are available for the query protein, the conservation signal cannot be calculated reliably and the functional region is not predicted. In such cases, the iDBPs server uses a classifier that was trained on the global features alone.

3.2 The classifier's input features

The features calculated for the ML-patches are: average surface electrostatic potential, secondary structure content, patch size (number of residues) and cluster-based amino acid conservation patterns (Nimrod *et al.*, 2009).

The global features include the average electrostatic potential, the secondary structure content and the protein size. They also include the protein's dipole moment, its amino acid composition, the spatial asymmetry of residues within the protein structure (Szilágyi and Skolnick, 2006) and the fraction of hydrogen donors/acceptors on the protein surface.

3.3 The web server

The web server requires the user to upload a protein structure in PDB format (or provide the PDB id), indicate the chain identifier of the query proteins and provide an e-mail address (optional). Once the calculations are finished, the results are sent to the user and include the prediction score as well as the expected sensitivity and precision at this score cutoff as calculated on the extended dataset. When available, a link to the PatchFinder results is also supplied. The PatchFinder results include the MSA, the evolutionary rates computed for each position in the protein (Mayrose *et al.*, 2004), the list of residues composing the ML-patch and the confidence of the prediction. In addition, the user can also visualize the ML-patch on the 3D-structure of the protein using the FirstGlance in Jmol applet.

3.4 Update of the N-Func database

The procedure we used to gather the structures in N-Func is described in detail in the original publication with the following modifications: sequence homologs were collected and multiply aligned using the protocol of the ConSurf-DB server (Goldenberg *et al.*, 2009) on the UniProt database (Bairoch *et al.*, 2005).

ACKNOWLEDGEMENTS

We acknowledge Haim Ashkenazy for his help with the establishment and maintenance of the iDBP server.

Funding: BLOOMNET ERA-PG grant; Edmond J. Safra Bioinformatics program at Tel Aviv University (to M.S.); OTKA (grant PD73096 to A.S.).

Conflict of Interest: none declared.

REFERENCES

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

Bairoch,A. *et al.* (2005) The universal protein resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Bhardwaj,N. *et al.* (2005) Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Res.*, **33**, 6486–6493.

Breiman,L. (2001) Random forests. *Machine Learn.*, **45**, 5–32.

Gao,M. and Skolnick, J. (2008) DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. *Nucleic Acids Res.*, **36**, 3978–3992.

Goldenberg,O. *et al.* (2009) The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res.*, **37**, D323–327.

Mayrose,I. *et al.* (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, **21**, 1781–1791.

Nimrod,G. *et al.* (2008) Detection of functionally important regions in 'hypothetical proteins' of known structure. *Structure*, **16**, 1755–1763.

Nimrod,G. *et al.* (2009) Identification of DNA-binding proteins using structural, electrostatic and evolutionary features. *J. Mol. Biol.*, **387**, 1040–1053.

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

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

Szilágyi,A. and Skolnick,J. (2006) Efficient prediction of nucleic acid binding function from low-resolution protein structures. *J. Mol. Biol.*, **358**, 922–933.