Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw315
Advance Access Publication Date: 3 June 2016
Applications Note

OXFORD

Structural bioinformatics

DBSI server: DNA binding site identifier

Shravan Sukumar[1], Xiaolei Zhu[2], Spencer S. Ericksen[3] and Julie C. Mitchell[1,4,*]

[1]Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706, USA, [2]School of Life Sciences, Anhui University, Hefei, Anhui Province 230601, China, [3]Small Molecule Screening Facility and [4]Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706, USA

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on February 22, 2016; revised on April 22, 2016; accepted on May 15, 2016

Abstract

Summary: Protein–nucleic acid interactions are among the most important intermolecular interactions in the regulation of cellular events. Identifying residues involved in these interactions from protein structure alone is an important challenge. Here we introduce the webserver interface to DNA Binding Site Identifier (DBSI), a powerful structure-based SVM model for the prediction and visualization of DNA binding sites on protein structures. DBSI has been shown to be a top-performing model to predict DNA binding sites on the surface of a protein or peptide and shows promise in predicting RNA binding sites.
Availability and Implementation: Server is available at http://dbsi.mitchell-lab.org
Contact: jcmitchell@wisc.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein–nucleic acid interactions are crucial to cellular events in the regulation of transcription, DNA replication and repair; and RNA splicing and post-transcriptional regulation. Identifying residues involved in nucleic acid binding is crucial to understand function. Cocrystal structures are often unavailable, and DNA/RNA binding sites may not be easily discernable from the protein sequence or unbound structure.

Here we present the web interface to DBSI (http://dbsi.mitchell-lab.org), a machine learning approach to classify surface residues as binders or non-binders of DNA. DBSI employs sequence- and structure-based features encompassing a range of physical, chemical, geometric and evolutionary properties of the protein surface (Zhu *et al.*, 2013a). DBSI also implements microenvironment features that allow for small-scale structural perturbation and the role of non-local cooperative effects.

The webserver greatly simplifies the use of DBSI, eliminating the need for downloading, installing and configuring third-party software and requiring the upload of only one file. The DBSI server facilitates online visualization of results, and it ensures privacy of user jobs.

2 Input and output

To use the DBSI server, the user first generates an electrostatic map of their structure (after stripping off heteroatoms) using the CHARMM-GUI (Jo *et al.*, 2008). The resulting PBEQ archive (a tgz file) comprises the sole input to our simple and intuitive user interface. In order to ensure the most accurate predictions, three values in step 2 of the CHARMM-GUI submission should be defined differently from the default values (epsP = 2.0, Dcel_c = 1.0, Dcel_f = 0.5).

The DBSI algorithm computes a set of sequence- and structure-based features and generates DNA binding predictions for each surface residue and a score that quantifies confidence that the residue will bind DNA, using a distance cutoff of 5.0 Å between the binding partners. The interface allows users to track the status of their jobs in the queue and provides a private link to results. Links to results can also be sent to the (optional) email address provided in the job submission screen. Upon job completion, users may visualize DBSI predictions within the browser using JSmol (Fig. 1D) (Hanson *et al.*, 2013). Users may also download a PDB file with the DBSI scores built into the B-factor column, allowing for visualization using PyMol or other molecular viewers(Fig. 1E). A text-based results file

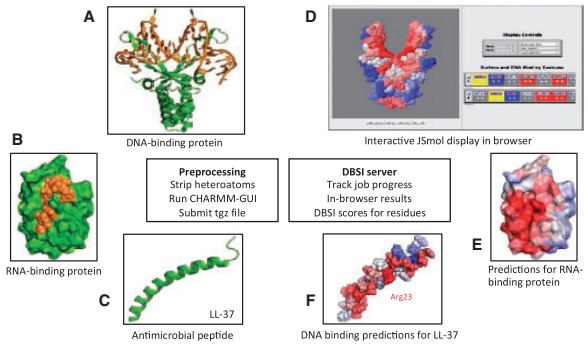

Fig. 1. DBSI server inputs and outputs. Inputs **A–C**: (A) DNA binding protein HU (PDB: 1P51) with DNA. (B) RNA binding protein YTHDC-1 (PDB: 4R3I) with RNA. C) Antimicrobial peptide LL-37 (PDB: 2K6O). Outputs **D–F**: Red coloring indicates predicted binding residues; Blue indicates predicted non-binding residues. (D) Interactive display with DBSI results for protein HU. (E) RNA binding prediction of YTHDC-1. (F) Binding predictions for LL-37 highlight the role Arg23 in binding DNA

with the DBSI classification and score for each surface residue is provided as well. Sample PBEQ archives and a user manual with directions for generating the tgz files, using the server and interpreting results are provided on the DBSI Server website. The DBSI server typically takes around 12 minutes for a 200 residue-sized protein, and is faster on machines with a solid state drive (Supplementary Table S5).

3 Results and performance

DBSI has been developed through a process of rigorous feature selection, model parameter optimization and internal 5-fold cross-validation. DBSI has been tested on an independent dataset comprising 206 high-resolution structures of DNA-binding proteins, showing it to outperform other available models. The AUC of 0.88 obtained by DBSI meets or exceeds that of other recently tested models, with our AUC being computed as an absolute number across the entire dataset rather than the 'per protein average AUC' used in Miao and Westhof (2015), which does not lead to a well-defined classification cutoff. Testing on a dataset with structures of complexed and free DNA binders (Andrabi *et al.*, 2014) has demonstrated that DBSI is very robust to conformational changes induced by binding and can predict binding sites on disordered proteins (Supplementary Tables S3 and S4). Details on testing, statistical analyses and comparisons with other models are available (Zhu *et al.*, 2013a).

The DBSI model has been tested on a range of proteins and peptides that bind nucleic acids (including large protein complexes such as histones, transcription factors, restriction enzymes, higher-order oligomers, RNA-binding proteins) with promising results (Table 1). Figure 1 shows DBSI results for DNA- and RNA-binding proteins as well as a short peptide. DBSI makes accurate binding site predictions for DNA- (Fig. 1A and D) and RNA- (Fig. 1B and E) binding molecules of varying size and function, from small peptides to large complexes like the nucleosome. We find that for a set of 19 RNA-binding protein structures from the Nucleic Acid Database (Berman *et al.*, 1992), the DBSI server correctly identifies binding sites with high accuracy and specificity, but has a lower sensitivity than observed for DNA-binding proteins. This suggests that sites predicted by DBSI are very likely to bind RNA, but the current model may misclassify weaker RNA binding sites. We are currently performing more thorough testing for a larger set of RNA-specific examples and will extend the DBSI model in the future for optimized performance on RNA-binding proteins. A list of the RNA binding examples and performance metrics is given in Supplementary Table S1.

Table 1. DBSI performance metrics for DNA-binding and RNA-binding proteins

Partner	PDB	AA	AC	SP	SE	PR	F1
DNA	206	38666	0.83	0.85	0.74	0.49	0.59
RNA	19	3473	0.77	0.85	0.43	0.39	0.40

The table shows number of PDB structures (PDB), total number of amino acids (AA), predictive accuracy (AC), specificity (SP), sensitivity (SE), precision (PR) and F1 Score (F1).

In the case of short antimicrobial peptides (Supplementary Table S2), where DNA-binding is thought to be important to activity (Guilhelmelli *et al.*, 2013), DBSI is able to correctly predict residues believed to bind DNA. While it is difficult to verify the DNA-binding residues without structures of the peptide–DNA complexes, molecular dynamics simulations (Jana *et al.*, 2013) and other studies (Wang *et al.*, 2014) suggest that DBSI correctly identifies key binding residues. For example, DBSI identifies residue Arg23 as an important binder of DNA on LL-37 (Fig. 1F).

Predictions from DBSI strongly reinforced the observation that, in contrast to its homologs, the important mitochondrial protein COQ9 was unlikely to bind DNA (Lohman *et al.*, 2014). This is significant, as COQ9, a member of the TetR family of regulators, possesses the characteristic conserved Helix-Turn-Helix DNA binding domain; COQ9 is the only known protein with a TetR transcription factor fold that appears to have lost DNA binding ability, something readily supported by the difference in DBSI predictions between COQ9 and the transcription factor FadR. DBSI has also be used to predict interaction sites for chemically similar molecules to DNA, such as heparin, as evidenced by its strong performance in CAPRI challenge 57 (Zhu *et al.*, 2013b).

With DBSI's proven performance in predicting nucleic acid binding sites, we believe that the ease of use of the webserver coupled with the simple and powerful visualization tools for analysis will make the DBSI server a widely used tool for the identification of key residues in protein–nucleic acid interfaces, thereby aiding the study of interactions between proteins with DNA/RNA and other chemically similar molecules.

Funding

This work was supported by the National Science Foundation [NSF DMS 1160360]

Conflict of Interest: none declared.

References

Andrabi,M. *et al.* (2014) Conformational changes in DNA-binding proteins: relationships with precomplex features and contributions to specificity and stability. *Proteins Struct. Funct. Bioinforma*, **82**, 841–857.

Berman,H.M. *et al.* (1992) A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.

Guilhelmelli,F. *et al.* (2013) Antibiotic development challenges: the various mechanisms of action of antimicrobial peptides and of bacterial resistance. *Front. Microbiol.*, **4**, 1–12.

Hanson,R.M. *et al.* (2013) JSmol and the next-generation web-based representation of 3D molecular structure as applied to proteopedia. *Isr. J. Chem*, **53**, 207–216.

Jana,J. *et al*. (2013) Human cathelicidin peptide LL37 binds telomeric G-quadruplex. *Mol. Biosyst*., **9**, 1833–1836.

Jo,S. *et al*. (2008) CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem*., **29**, 1859–1865.

Lohman,D.C. *et al*. (2014) Mitochondrial COQ9 is a lipid-binding protein that associates with COQ7 to enable coenzyme Q biosynthesis. *Proc. Natl. Acad. Sci. U. S. A*., **111**, E4697–E4705.

Miao,Z. and Westhof,E. *et al*. (2015) A Large-Scale Assessment of Nucleic Acids Binding Site Prediction Programs. *PLoS. Comput. Biol*., **11**, e1004639.

Wang,G. *et al*. (2014) High-quality 3D structures shine light on antibacterial, anti-biofilm and antiviral activities of human cathelicidin LL-37 and its fragments. *Biochim. Biophys. Acta – Biomembr*, **1838**, 2160–2172.

Zhu,X. *et al*. (2013a) DBSI: DNA-binding site identifier. *Nucleic Acids Res*., **41**, 160.

Zhu,X. *et al*. (2013b) Data-driven models for protein interaction and design. *Proteins*, **81**, 2221–2228.