BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 2 2010, pages 287–289
doi:10.1093/bioinformatics/btp631

Databases and ontologies

hPDI: a database of experimental human protein–DNA interactions

Zhi Xie[1], Shaohui Hu[2,3], Seth Blackshaw[1,3,4,5], Heng Zhu[2,3] and Jiang Qian[1,*]

[1]Department of Ophthalmology, [2]Department of Pharmacology and Molecular Sciences, [3]The Center for High-Throughput Biology, [4]Institute for Cell Engineering and [5]Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, MD 21231, USA

Received on August 26, 2009; revised and accepted on November 3, 2009
Advance Access publication November 9, 2009
Associate Editor: Martin Bishop

ABSTRACT

Summary: The human protein DNA Interactome (hPDI) database holds experimental protein–DNA interaction data for humans identified by protein microarray assays. The unique characteristics of hPDI are that it contains consensus DNA-binding sequences not only for nearly 500 human transcription factors but also for >500 unconventional DNA-binding proteins, which are completely uncharacterized previously. Users can browse, search and download a subset or the entire data via a web interface. This database is freely accessible for any academic purposes.

Availability: http://bioinfo.wilmer.jhu.edu/PDI/

Contact: jiang.qian@jhmi.edu

1 INTRODUCTION

Protein–DNA interactions (PDIs) mediate a large range of functions essential for cellular differentiation, development and function. A major class of DNA-binding proteins are the transcription factors (TFs) that regulate gene expression; DNA-binding specificities of TFs have been extensively studied for decades and the results are mainly collected in the TRANSFAC and JASPAR databases (Sandelin *et al.*, 2004; Wingender *et al.*, 1996). In addition, yeast and bacterial one-hybrid techniques (Y1H and B1H, respectively) and the recently developed protein-binding microarray technology also provide an efficient and comprehensive method for identification of specific PDIs. Consequently, PDIs for relatively comprehensive yeast TFs, as well as a few TF subfamilies in *Caenorhabditis elegans*, *Drosophila* and mouse TFs, have been characterized (Badis *et al.*, 2008; Berger *et al.*, 2008; Deplancke *et al.*, 2006; Grove *et al.*, 2009; Newburger and Bulyk 2009; Noyes *et al.*, 2008; Zhu *et al.*, 2009). Despite the long history of studies and recent advances in this field, PDIs of the vast majority of human TFs remain uncharacterized, which comprise a total of ∼1400 proteins (Messina *et al.*, 2004). Furthermore, sequence-specific PDIs of the larger universe of unconventional DNA-binding proteins (uDBPs), aside from TFs, have not been extensively explored, although a few recent studies have suggested that uDBPs, such as protein kinases and metabolic enzymes, do in fact possess property of sequence-specific PDIs, as reviewed in Hu *et al.* (2009).

Our recent PDI study, in using an unbiased protein microarray assays probed by 460 sequence-diverse DNA motifs, has made significant progress towards the goal of identifying PDIs in humans (Hu *et al.*, 2009). Preferred target sites for 493 human TFs have been identified. Comparison of significant consensus sequences (consensus logos) between our study and TRANSFAC has shown considerable agreement. Furthermore, we found that >500 proteins not predicted to act as TFs unexpectedly showed sequence-specific DNA-binding activity. A number of newly identified PDIs have also been confirmed both *in vitro* and *in vivo*. Here, we present a database hosting the experimentally determined DNA-binding sequences obtained from protein microarray assays for both human TFs and uDBPs. The database is available via a web interface that enables users to browse, query and download any PDIs of interest.

2 DATABASE CONTENT

The human protein DNA Interactome (hPDI) database currently holds a collection of over 17 000 preferable DNA-binding sequences for 493 human TFs and 520 uDBPs. TFs containing known DNA-binding domains (DBDs) cover all the major subfamilies, including zf-C2H2, Homeobox, Nuclear hormone receptor, bHLH, Forkhead, bZIP, Ets, HMG box, RHD, STAT, GATA and IRF. In addition, a number of proteins that do not have known DBDs but are annotated as ‘regulation of gene expression’ by GO database are also annotated as TFs (Ashburner *et al.*, 2000). Consensus logos have been generated for 201 TFs for those binding to at least three and <30 oligonucleotide dsDNA probe sequences. Consensus logos are generated using ‘WebLogo’ (Crooks *et al.*, 2004). Among these logos, 166 novel ones for TFs have no previously known binding sites listed in TRANSFAC. It should be noted that the consensus logos from TRANSFAC are generated from the TRANSFAC SITE database where only DNA-binding sequences bound by human proteins are used.

PDIs of those uDBPs identified in our recent study (Hu *et al.*, 2009) are archived based on different protein classes in the database, including protein kinases, chromatin-associated proteins, RNA-binding proteins, transcriptional co-regulators, other nucleic acid-binding proteins rather than TFs and RNA-binding proteins, protein associated with DNA repair and replication, mitochondrial proteins and all other categories. Using the same criteria as the TFs,

[*]To whom correspondence should be addressed.

Table 1. Statistics of the hPDI database

TF sub-family	No.	uDBP classes	No.	Overall statistics	No.
Total	493	Total	520	No. of PDIs	17 718
Zf-C2H2	95	RNA-binding proteins	207	No. of DNA-binding proteins	1013
Homeobox	44	All other categories	132	No. of DNA-binding logos	437
Other sub-families	36	Mitochondrial proteins	97	Mean sequences bound per protein	17
HLH	22	Chromatin-associated proteins	73		
Nuclear hormone_receptor	17	Other nucleic acid binding	50		
zf-CCHC	12	DNA repair and replication	50		
Myb	11	Transcriptional co-regulators	43		
HMG_box	11	Protein kinases	14		
Ets	10				
MH	8				
bZIP_1	6				
Forkhead	6				
IRF	6				
TFs without identified DBDs	209				

TFs without identified DBDs are defined as proteins annotated as 'regulation of transcription' at GO database but without known DBDs defined by Pfam database. Some proteins may belong to more than one protein class.

consensus logos for 236 uDBPs have been generated and archived. The class/family coverage of proteins in hPDI is summarized in Table 1.

3 DATABASE ARCHITECHTURE AND DATA RETRIEVAL

We have developed a web interface for the hPDI database. Perl CGI is used to connect the database and dynamically generate user-friendly HTML front-end queries, using Apache web server. Users may perform the following tasks on the web.

(1) Protein view: Users can search a protein of interest. The protein view pages will provide the relevant information of the protein, such as annotation, protein class, DNA-binding sequences/logos and the position weight matrix (PWM).

(2) Class view: Users can browse the DNA-binding profiles for any particular TF subfamily or protein class for a quick overview.

(3) Motif search: The database can be queried with user-defined DNA sequences and the 'best matching' motifs are returned. The server also allows wild cards in DNA sequences, such as N, Y, R. Using this function, users can find the potential binding proteins (both TFs and uDBPs) for a given DNA sequence.

(4) All the DNA sequences, PWMs and their binding proteins, as well as the protein sequences, can be downloaded from our server.

4 DISCUSSION

The hPDI database offers important resources that complement the existing TRANSFAC, JASPAR and UniProt databases in several distinct ways (Newburger and Bulyk, 2009; Sandelin *et al.*, 2004; Wingender *et al.*, 1996). First, it holds a significant number of experimentally obtained consensus DNA-binding sequences of human TFs, where >300 human TFs are unique in our collection. Users can combine our data collection with other databases for better definition of DNA-binding profiles for human TFs, since these datasets are generated using completely independent techniques. Second, the data access is freely available for all the academic users. Finally and most importantly, it is the only database to maintain DNA-binding sequences for uDBPs, which may play much more profound roles in regulation of gene transcription or other cellular functions than previously thought. Indeed, our in-depth study of an uDBP, ERK2, using combined *in vitro* and *in vivo* approaches, has revealed that ERK2 acts as a transcriptional repressor of interferon-gamma response genes in human cells, illustrating the biological relevance of uDBPs identified using this approach. We have also included all the raw binding data files in the database and detailed data processes are referred to the Supplementary Data in our previous publication (Hu *et al.*, 2009). We expect that this database will play a significant role in the field of regulatory genomics and proteomics. We will continuously update the hPDI database as new data are generated.

Funding: National Institutes of Health; W. M. Keck Foundation Distinguished Young Investigator in Medical Research Award; Ruth and Milton Steinbach Fund; generous gift from Mr and Mrs Robert and Clarice Smith.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Badis,G. *et al.* (2008) A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell*, **32**, 878–887.

Berger,M.F. *et al.* (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**, 1266–1276.

Crooks,G.E. *et al.* (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Deplancke,B. *et al.* (2006) A gene-centered C. elegans protein-DNA interaction network. *Cell*, **125**, 1193–1205.

Grove,C.A. *et al.* (2009) A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. *Cell*, **138**, 314–327.

Hu,S. *et al.* (2009) Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. *Cell*, **139**, 610–622.

Messina,D.N. *et al.* (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res.*, **14**, 2041–2047.

Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, D77–D82.

Noyes,M.B. *et al.* (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, **133**, 1277–1289.

Sandelin,A. *et al.* (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.

Wingender,E. *et al.* (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.

Zhu,C. *et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.*, **19**, 556–566.