D124–D128 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 29 October 2010*
doi:10.1093/nar/gkq992

UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein–DNA interactions

Kimberly Robasky[1,2] and Martha L. Bulyk[1,3,4,*]

[1]Department of Medicine, Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, [2]Bioinformatics Program, Boston University, Boston, MA 02215, [3]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School and [4]Harvard-MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA 02115, USA

Received September 13, 2010; Accepted October 5, 2010

ABSTRACT

The Universal PBM Resource for Oligonucleotide-Binding Evaluation (UniPROBE) database is a centralized repository of information on the DNA-binding preferences of proteins as determined by universal protein-binding microarray (PBM) technology. Each entry for a protein (or protein complex) in UniPROBE provides the quantitative preferences for all possible nucleotide sequence variants ('words') of length *k* ('*k*-mers'), as well as position weight matrix (PWM) and graphical sequence logo representations of the *k*-mer data. In this update, we describe >130% expansion of the database content, incorporation of a protein BLAST (blastp) tool for finding protein sequence matches in UniPROBE, the introduction of UniPROBE accession numbers and additional database enhancements. The UniPROBE database is available at http://uniprobe.org.

INTRODUCTION

A comprehensive understanding of gene-expression regulation requires the thorough characterization of transcription factor (TF)–DNA binding properties. TFs play central roles in transcriptional regulatory networks by binding specific DNA sequences and activating or repressing gene expression. Consequently, TF–DNA-binding specificities have broad impact on cell physiology and development and in evolution (1,2).

Advances in DNA microarray synthesis and the development of protein-binding microarray (PBM) technology (3,4) led to the development of universal PBMs (5), which allow high-throughput measurement of comprehensive data on protein–DNA binding specificities, resulting in large data sets requiring curation and searchability. The Universal PBM Resource for Oligonucleotide-Binding Evaluation (UniPROBE) (6) database was created to satisfy these requirements. Please refer to the original UniPROBE publication (6) for a description of major differences between UniPROBE and the JASPAR (7), TRANSFAC (8) and PAZAR (9) databases. The original UniPROBE publication (6) also provides a detailed description of PBM technology and data types.

Since its inception 2 years ago, the UniPROBE database has continued to expand in size, utility and user base. UniPROBE previously housed data for 177 non-redundant proteins (6). That number has recently grown to over 400 non-redundant proteins or protein complexes, with additional, unpublished PBM data sets already planned for future deposition. Currently, the UniPROBE database averages 933 unique visitors per month (classified by IP address) from over 40 different countries and 3558 page views per month. UniPROBE is the standard for curating universal PBM data, and we invite other researchers generating universal PBM data to contact us about depositing their data in UniPROBE.

DATABASE ADDITIONS

UniPROBE has more than doubled in size since its introduction in January 2009 (6) (Table 1). As of this writing, in addition to the data deposited from the initial set of four publications (5,10–12), PBM data are included from six newer publications (13–18) with additional published (19) and soon to be published data currently in planning for deposition. The new additions include data on TFs

*To whom correspondence should be addressed. Tel: +1 617 525 4725; Fax: +1 617 525 4705; Email: mlbulyk@receptor.med.harvard.edu

Table 1. UniPROBE database contents, with indication of additions in PBM data sets since its introduction in 2009

Reference	Number of proteins or protein complexes	Species
Berger *et al.* (5)	5	*Saccharomyces cerevisiae, Homo sapiens, Mus musculus, Caenorhabditis elegans*
Berger *et al.* (10)	168	*Mus musculus*
Pompeani *et al.* (11)	1	*Vibrio harveyi*
De Silva *et al.* (12)	3	*Plasmodium falciparum, Cryptosporidium parvum*
Grove *et al.* 2009 (13)[a]	21	*Caenorhabditis elegans*
Scharer *et al.* (14)[a]	1	*Homo sapiens*
Lesch *et al.* (15)[a]	1	*Caenorhabditis elegans*
Zhu *et al.* (16)[a]	89	*Saccharomyces cerevisiae*
Badis *et al.* (17)[a]	104	*Mus musculus*
Wei *et al.* (18)[a]	22	*Mus musculus*
Total number:	415	
Non-redundant proteins or protein complexes:	404	
Total, last described (6):	177	
Total added:	238	
Percent increase:	134%	

[a]Indicates data sets that have been added since the last published description of UniPROBE.

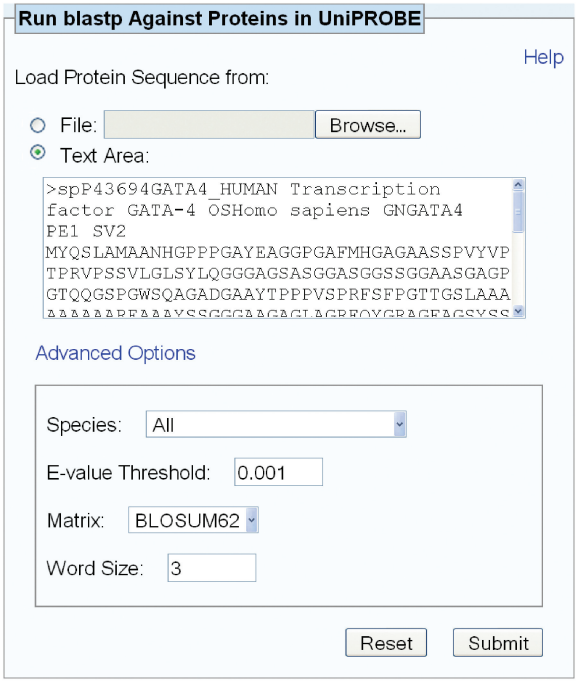

Figure 1. Blastp search of UniPROBE with human GATA4 protein sequence and default parameter settings for the advanced search options.

from *Caenorhabditis elegans, Saccharomyces cerevisiae*, *Mus musculus* and *Homo sapiens*. The UniPROBE database now houses PBM data for 415 individual proteins or protein complexes, nearly all of which are TFs, corresponding to 404 non-redundant proteins or protein complexes.

NEW BLASTP SEARCH FEATURE

In the latest version of UniPROBE, the available online search features have been augmented with a new search tool that permits a user to perform a blastp (20) search of a protein sequence of interest (the 'query protein') against all protein sequences in the UniPROBE database (the 'subject proteins'). This feature incorporates NCBI's Protein–Protein BLAST tool (21), blastp v.2.2.23+, for accurate and efficient alignments. This blastp tool returns a list of links to the Details page for each subject protein that either exactly matches or is similar to the query protein(s) according to user-specified search parameter settings. Links from the Details pages allow further exploration and links to download the PBM data for the matching proteins.

Query protein sequences may be entered manually into a web-page form or uploaded as a text file. The sequence is parsed using fail-safe rules to interpret the format. Multiple sequences can be processed in batch either by specifying one sequence per line, or by entry of FASTA-formatted sequences, which may cross multiple lines but are separated by header lines. Numbers and unnecessary white-space are stripped from the sequence prior to performing the search.

For the subject proteins, the blastp search tool uses a database comprising all the clone insert sequences corresponding to all the PBM experiments with data curated in UniPROBE. For example, consider a search for the human TF GATA4, which is not currently in UniPROBE. Running the blastp tool on the human GATA4 sequence with default parameter settings (Figure 1) results in eight hits, four from yeast and four from mouse, all with the GATA DNA-binding domain (Figure 2). Among the hits, the tool correctly retrieves two hits to Gata3, which is represented in the database by two proteins: the full-length TF and just the DNA-binding domain. The blastp search parameter settings (*E*-value threshold, species, substitution matrix and word size) are passed directly to a local instance of NCBI's blastp executable.

Results are output with the sequence matches within matching subject proteins rendered with yellow highlighting on all the residues within the confines of the alignment. Also provided is the offset of the first aligned residue of the query protein. As defined by blastp, the score provided is a measure of similarity, and the *E*-value is the number of expected matches if the subject protein sequences were generated randomly.

UNIPROBE ACCESSION NUMBERS

A significant new feature is the addition of UniPROBE accession numbers. Each TF PBM data set now has its own UniPROBE accession number, regardless of whether or not its protein is unique in the database. Accession

Query Sequence for **spP43694GATA4_H ... GNGATA4 PE1 SV2** :

MYQSLAMAANHGPPPGAYEAGGPGAFMHGAGAASSPVYVPTPRVPSSVLGLSYLQGGGAG
SASGGASGGSSGGAASGAGPGTQQGSPGWSQAGADGAAYTPPPVSPRFSFPGTTGSLAAA
AAAAAAREAAAYSSGGGAAGAGLAGREQYGRAGFAGSYSSPYPAYMADVGASWAAAAAAS
AGPFDSPVLHSLPGRANPAARHPNLDMFDDFSEGRECVNCGAMSTPLWRRDGTGHYLCNA
CGLYHKMNGINRPLIKPQRRLSASRRVGLSCANCQTTTTTLWRRNAEGEPVCNACGLYMK
LHGVPRPLAMRKEGIQTRKRKPKNLNKSKTPAAPSGSESLPPASGASSNSSNATTSSSEE
MRPIKTEPGLSSHYGHSSSVSQTFSVSAMSGHGPSIHPVLSALKLSPQGYASPVSQSPQT
SSKQDSWNSLVLADSHGDIITA

Protein Match	Species	UniPROBE Accession Number	DNA Binding Domain	Match Offset in Search Protein	Score	E value	Cloned Protein Sequence	Publication	Save	View
Gata6	Mus musculus	UP00100	GATA	198	628	1e-67	PLPVPRGPSTDLLEDLSESRECVNCGSIQTPLWRRDGTGHYLCNACGLYS KMNGLSRPLIKPQKRVPSSRRLGLSCANCHTTTTTLWRRNAEGEPVCNAC GLYMKLHGVPRPLAMKKEGIQTRKRKPKNINKSKA	Badis et al., Science 2009		
Gata5	Mus musculus	UP00080	GATA	206	579	5e-62	PGRRTSFVPDFLEEFPGEGRECVNCGALSTPLWRRDGTGHYLCNACGLYH KMNGVNRPLVRPQKRLSSSRRSGLCCSNCHTATTTLWRRNSEGEPVCNAC GLYMKLHGVPRPLAMKKESIQTRKRKPENPAKIKG	Badis et al., Science 2009		
Gata3	Mus musculus	UP00032	GATA	212	549	1e-58	MEVTADQPRWVSHHHPAVLNGQHPDTHHPGLSHSYMDAAQYPLPEEVDVL FNIDGQGNHVPPYYGNSVRATVQRYPPTHHGSQVCRPPLLHGSLPWLDGG KALGSHHTASPWNLSPFSKTSIHHGSPGPLSVYPPASSSSLSGGHASPHL FTFPPTPPKDVSPDPSLSTPGSAGSARQDEKECLKYQVPLPDSMKLESSH SRGSMTALGGASSSTHHPITTYPPYVPEYSSGLFPPSSLLGGSPTGFGCK SRPKARSSTEGRECVNCGATSTPLWRRDGTGHYLCNACGLYHKMNGQNRP LIKPKRRLSAARRAGTSCANCQTTTTTLWRRNANGDPVCNACGLYYKLHN INRPLTMKKEGIQTRNRKMSSKSKKCKKVHDSLEDFPKNSSFNPAALSRH MSSLSHISPFSHSSHMLTTPTPMHPPSSLSFGPHHPSSMVTAMG	Badis et al., Science 2009		
Gata3	Mus musculus	UP00032	GATA	212	533	1e-56	SPTGFGCKSRPKARSSTEGRECVNCGATSTPLWRRDGTGHYLCNACGLYH KMNGQNRPLIKPKRRLSAARRAGTSCANCQTTTTTLWRRNANGDPVCNAC GLYYKLHNINRPLTMKKEGIQTRNRKMSSKSKKCK	Badis et al., Science 2009		
Gat1	Saccharomyces cerevisiae	UP00287	GATA	216	143	2e-11	NPSPSIVKPGSRRNSSVRKKKPALKKIKSSTSVQSSATPPSNTSSNPDIK CSNCTTSTTPLWRKDPKGLPLCNACGLFLKLHGVTRPLSLKTDIIKKRQR SSTKINNNITPPPSSSLNPGAAGKKKNYTASVAAS	Zhu et al., Genome Res 2009		
Gzf3	Saccharomyces cerevisiae	UP00347	GATA	217	135	2e-10	MASQATTLRGYNIRKRDNVFEPKSSENLNSLNQSEEEGHIGRWPPLGYEA VSAEQKSAVQLRESQAGASISNNMNFKANDKSFSTSTAGRMSPDTNSLHH ILPKNQVKNNGQTMDANCNNNVSNDANVPVCKNCLTSTTPLWRRDEHGAM LCNACGLFLKLHGKPRPISLKTDVIKSRNRKSNTNHAHNLDNFRNQTLIA ELKGDCNIESSGRKANRVTSEDKKKKSSQLLMGTSSTAKISKKPKTESKE RSDSHLSATKLEVLMSGDCSRPNLKPKLPKQDTAIYQEKLLTFPSYTDVK EYSNSAHQSAFIKERSQFNAASFPLNASHSVTSKTGADSPQLPHLSMLLG SLSSTSISNNGSEIVSNCNNGIASTAATLAPTSSRTTDSNPSEVPNQIRS TMSSPDIISAKRNDPAPLSFHMASINDMLETRDRAISNVKTETTPPHFIP FLQSSKAPCISKANSQSISNSVSSSDVSGRKFENHPAKDLGDQLSTKLHK EEEIIKLKTRINELELVTDLYRRHINELDGKCRALEERLQRTVKQEGNKG G	Zhu et al., Genome Res 2009		
Gln3	Saccharomyces cerevisiae	UP00318	GATA	216	145	1e-11	SSNTTNSVRKNSLIKPMSSTSLANFKRAASVSSSISNMEPSGQNKKPLIQ CFNCKTFKTPLWRRSPEGNTLCNACGLFQKLHGTMRPLSLKSDVIKKRIS KKRAKQTDPNIAQNTPSAPATASTSVTTTNAKPIR	Zhu et al., Genome Res 2009		
Gat3	Saccharomyces cerevisiae	UP00319	GATA	258	84	1e-04	MNIKTLCHPEYKRISVESLLNPVEETIDCEKPHSQTKINTAKPISASLYV TTNNTAVVQHNVQKRKGVTRRCPQCAVIKTSPQWREGPDGEVTLCNACGL FYRKIFLVFGKDLAKRYFNEIKGVSVKRKVPKSLYGVTRTR	Zhu et al., Genome Res 2009		

Number of matches: 8

Figure 2. Results from blastp search of all protein sequences in UniPROBE using the human full-length GATA4 protein sequence as the query.

numbers are five digits prefixed with 'UP' (abbreviation for 'universal PBM'), e.g. UP00350. Accession numbers are returned as part of the search results and are also listed on each protein's Details page. A user can use the 'Quick Search' tool to find TFs by accession number. Accession numbers can be requested prior to publication of new PBM data sets, such as for unpublished PBM data sets in new article submissions.

OTHER NEW FEATURES

New to this version of UniPROBE is the inclusion of PBM data for protein complexes. This functionality was implemented to accommodate homodimer and heterodimer data for bHLH TFs from *C. elegans* (13). This feature allows the Details page to render data sets for the protein of interest and for each of the proteins with which the protein of interest dimerizes.

The UniPROBE statistics cited here were derived with the aid of several minor but useful enhancements. It is now possible to use 'Text Search' to find TFs by publication; TFs can be searched by species using the same tool. The search results now include the total number of TFs returned. To easily distinguish between separate, published PBM data sets for the same protein, a reference to the publication for each separate data set has been added to the bottom of all TF Details pages, along with the array design number(s).

For convenience a new, shorter URL (http://uniprobe.org) has been registered, which redirects to the legacy UniPROBE URL (http://thebrain.bwh.harvard.edu/uniprobe).

FUTURE DIRECTIONS

Future updates planned for UniPROBE include additional user and administrative tools. Currently in development is a negative control sequence generator which, given an *E*-score threshold indicative of DNA-binding preference, will generate random sequence of user-specified length that does not include any 8-mer with scores exceeding the given threshold for user-selected TFs and species in UniPROBE. Another planned feature is the display of sequence alignments resulting from the blastp searches of UniPROBE. Also under development are administrative tools to allow for self-deposition and automated pre-publication UniPROBE accession number requests. The template for the Details page will be generalized to support self-deposition of PBM data for protein complexes. These tools and others will be facilitated, and system performance will generally improve, with the implementation of a newly designed database schema. As always, we continue to encourage user registration and feedback for error reports and feature requests, some of which motivated the development of the new features described here.

AVAILABILITY AND LICENSE

All data hosted by the PBM database are freely available for distribution at the database website. The sequences of the 60-mer DNA probes synthesized on the custom-designed universal arrays are available under the terms of the academic research use license available at http://thebrain.bwh.harvard.edu/uniprobe/academic-license.php.

ACKNOWLEDGEMENTS

The authors thank Ivan Adzhubey for technical assistance and Dan Newburger and Mike Berger for helpful discussions.

FUNDING

Funding for open access charge: National Institutes of Health (grant number R01 HG003985 to M.L.B.).

Conflict of interest statement. None declared.

REFERENCES

1. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
2. Bulyk,M.L. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol.*, **5**, 201.
3. Bulyk,M.L., Huang,X., Choo,Y. and Church,G.M. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl Acad. Sci. USA*, **98**, 7158–7163.
4. Mukherjee,S., Berger,M.F., Jona,G., Wang,X.S., Muzzey,D., Snyder,M., Young,R.A. and Bulyk,M.L. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, **36**, 1331–1339.
5. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
6. 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.
7. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.
8. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
9. Portales-Casamar,E., Arenillas,D., Lim,J., Swanson,M.I., Jiang,S., McCallum,A., Kirov,S. and Wasserman,W.W. (2009) The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res.*, **37**, D54–D60.
10. Berger,M., Badis,G., Gehrke,A., Talukder,S., Philippakis,A., Penacastillo,L., Alleyne,T., Mnaimneh,S., Botvinnik,O. and Chan,E. (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**, 1266–1276.
11. Pompeani,A.J., Irgon,J.J., Berger,M.F., Bulyk,M.L., Wingreen,N.S. and Bassler,B.L. (2008) The Vibrio harveyi master quorum-sensing regulator, LuxR, a TetR-type protein is both an activator and a repressor: DNA recognition and binding specificity at target promoters. *Mol. Microbiol.*, **70**, 76–88.
12. De Silva,E.K., Gehrke,A.R., Olszewski,K., Leon,I., Chahal,J.S., Bulyk,M.L. and Llinas,M. (2008) Specific DNA-binding by Apicomplexan AP2 transcription factors. *Proc. Natl Acad. Sci. USA*, **105**, 8393–8398.
13. Grove,C.A., De Masi,F., Barrasa,M.I., Newburger,D.E., Alkema,M.J., Bulyk,M.L. and Walhout,A.J.M. (2009) A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. *Cell*, **138**, 314–327.
14. Scharer,C.D., McCabe,C.D., Ali-Seyed,M., Berger,M.F., Bulyk,M.L. and Moreno,C.S. (2009) Genome-wide promoter analysis of the SOX4 transcriptional network in prostate cancer cells. *Cancer Res.*, **69**, 709–717.
15. Lesch,B.J., Gehrke,A.R., Bulyk,M.L. and Bargmann,C.I. (2009) Transcriptional regulation and stabilization of left-right neuronal identity in C. elegans. *Genes Dev.*, **23**, 345–358.
16. Zhu,C., Byers,K.J.R.P., McCord,R.P., Shi,Z., Berger,M.F., Newburger,D.E., Saulrieta,K., Smith,Z., Shah,M.V., Radhakrishnan,M. *et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.*, **19**, 556–566.

17. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
18. Wei,G.-H., Badis,G., Berger,M.F., Kivioja,T., Palin,K., Enge,M., Bonke,M., Jolma,A., Varjosalo,M., Gehrke,A.R. *et al.* (2010) Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *The EMBO J.*, **29**, 2147–2160.
19. Alibés,A., Nadra,A.D., De Masi,F., Bulyk,M.L., Serrano,L. and Stricher,F. (2010) Using protein design algorithms to understand the molecular basis of disease caused by protein-DNA interactions: the Pax6 example. *Nucleic Acids Res.*, doi:10.1093/nar/gkq683 [Epub ahead of print, 4 August 2010].
20. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
21. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.