UniPROBE: An Online Database of Protein Binding Microarray Data on Protein–DNA Interactions

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1093/nar/gkn660</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:4878080">http://nrs.harvard.edu/urn-3:HUL.InstRepos:4878080</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
UniPROBE: an online database of protein binding microarray data on protein–DNA interactions

Daniel E. Newburger¹ and Martha L. Bulyk¹,²,³,*

¹Division of Genetics, Department of Medicine, ²Department of Pathology, Brigham and Women’s Hospital and Harvard Medical School and ³Harvard-MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA 02115, USA

Received August 15, 2008; Revised September 18, 2008; Accepted September 21, 2008

ABSTRACT

The UniPROBE (Universal PBM Resource for Oligonucleotide Binding Evaluation) database hosts data generated by universal protein binding microarray (PBM) technology on the in vitro DNA-binding specificities of proteins. This initial release of the UniPROBE database provides a centralized resource for accessing comprehensive PBM data on the preferences of proteins for all possible 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 total, the database hosts DNA-binding data for over 175 nonredundant proteins from a diverse collection of organisms, including the prokaryote Vibrio harveyi, the eukaryotic malarial parasite Plasmodium falciparum, the parasitic Apicomplexan Cryptosporidium parvum, the yeast Saccharomyces cerevisiae, the worm Caenorhabditis elegans, mouse and human. Current web tools include a text-based search, a function for assessing motif similarity between user-entered data and database PWMs, and a function for locating putative binding sites along user-entered nucleotide sequences. The UniPROBE database is available at http://thebrain.bwh.harvard.edu/uniprobe/.

INTRODUCTION

The characterization of transcription factors’ (TFs’) DNA-binding specificities represents a critical step towards understanding the regulation of gene expression and elucidating the biophysical properties governing protein–DNA interactions. The study of DNA-binding specificities therefore has profound implications for the analysis and prediction of the regulatory networks that govern intracellular function, responses to external stimuli, differentiation and development in an organism. Despite recent advances in this field, the vast majority of TFs in most major model organisms and pathogens remain either uncharacterized or poorly described (1).

The development of universal (2) protein binding microarray (PBM) technology (3) (Figure 1) offers a new avenue for the exploration of protein–DNA binding specificity. Universal PBMs provide an efficient and comprehensive method for in vitro interrogation of DNA-binding preferences. PBM technology complements other currently available technologies, such as chromatin immunoprecipitation coupled with either microarray readout (4–7) or high-throughput sequencing (8–10) that identify genomic regions bound in vivo.

Universal PBMs achieve comprehensive, high-resolution determination of proteins’ DNA-binding preferences by measuring the binding preferences of a protein over all possible k-mers of a given length (2,11). Currently employed custom array designs contain a set of 60-bp DNA probes that encompass all possible permutations of either 9 (Bulyk Lab, unpublished data) or 10 bp (12), depending upon the microarray design (2,12). In addition to covering all contiguous 9-mers or 10-mers, these array designs also offer an extensive set of gapped permutations that provide coverage of binding sites of greater length. Together, these data can be synthesized to produce high-confidence measurements of the relative preferences of a protein for all possible sequence variants belonging to a wide range of k-mer patterns typically found in TF binding site motifs (2,12). PBM enrichment scores from the PBM signal intensity data are typically calculated for each of the more than 2.3 million 8-mers (i.e. binding site ‘words’ with eight informative nucleotide positions, including all contiguous 8-mers and a large collection of gapped 8-mers). These 8-mers encompass the full affinity range of DNA binding preferences, from the most preferentially bound k-mers to low-affinity k-mers and nonspecifically associated sequences (2).

The TRANSFAC (13) and JASPAR (14) databases contain hundreds of matrices constructed from DNA binding
Number of binding site sequences in a JASPAR SELEX dataset is just 28, while the median number of nonredundant binding site sequences is just 11. In contrast, the DNA binding profiles obtained from our PBM experiments provide information on the direct binding preferences of a given protein over all $k$-mer DNA binding sequence variants, measured in vitro using the universal PBM technology; the number of sequence variants examined is limited only by the number of features on the microarray. In addition to the $k$-mer binding profiles, these procedures also provide DNA binding sequence PWMs derived from the $k$-mer data using our Seed-and-Wobble algorithm.

The UniPROBE database hosts the high-resolution DNA binding profiles obtained from PBM experiments on known and predicted TFs. The database currently contains DNA binding profiles for many proteins not included in similar databases such as JASPAR, and TRANSFAC, and it offers several tools for searching the database and analyzing user-defined binding profiles or DNA sequences. The resources and analysis tools offered by the UniPROBE database promise to facilitate previously untenable, downstream genomic analyses, and we anticipate that it will represent an important genomic resource as additional PBM data are compiled.

**DATABASE CONSTRUCTION**

The UniPROBE database is managed by a MySQL relational database that provides the back-end for user queries and facilitates the data retrieval necessary for the site's analysis tools. All HTML pages are dynamically generated by PHP scripts hosted on an Apache server, and several JavaScript libraries provide interactive interfaces that facilitate site navigation and form accessibility. The Apache server also hosts all downloadable data, available by HTTP connection.

**DATABASE CONTENT**

The UniPROBE database hosts the results of PBM experiments, subsequent computational analyses performed on these data, and protein annotations. The site currently hosts PBM data for over 175 nonredundant proteins from a wide range of organisms, including the prokaryote *Vibrio harveyi*, the eukaryotic malarial parasite *Plasmodium falciparum*, the parasitic Apicomplexan *Cryptosporidium parvum*, the yeast *Saccharomyces cerevisiae*, the worm *Caenorhabditis elegans*, mouse and human. These data already encompass the majority of mouse homeodomain TFs and will soon include more than 100 additional mouse proteins (labs of Bulyk, M.L. and Hughes, T.R., unpublished results), nearly 90 additional *S. cerevisiae* proteins (labs of Bulyk, M.L. and LaBaer, J., unpublished results), and over 20 additional *C. elegans* proteins (labs of Bulyk, M.L. and Walhout, A.J., unpublished results). The UniPROBE database will additionally host data for *Drosophila melanogaster* TFs from ongoing projects in the Bulyk laboratory, and we anticipate the addition of several datasets from other laboratories using this microarray technology.
For each DNA-binding protein, the server holds several different data types, including: (i) unprocessed 60-mer probe signal intensity data; (ii) normalized probe intensities; (iii) TF-binding DNA profile representations and (iv) publication-specific data. The unprocessed (or ‘raw’) Agilent array data include information on probe position, 60-mer probe sequence, and Cy3 (DNA, from incorporated Cy3-dUTP) and Alexa 488 (protein, from Alexa 488 conjugated anti-GST antibody) signal intensities, the latter of which are necessary for accurately assessing relative DNA binding (2). The normalized probe intensities are derived from the raw data after adjusting for relative DNA concentrations at each spot and for spatial nonuniformities within the microarray. The in-house software used for normalization and subsequent binding profile generation will soon be available for download as the Universal PBM Analysis Suite (11).

The database offers several different binding profile representations for assessing TF specificity and affinity. First, we use our Seed-and-Wobble algorithm (2) to generate PWM motifs, which represent the observed probability of finding a given nucleotide in a given position within a TF’s DNA target site. PWMs currently serve as one of the primary methods for quantitative representation of DNA binding site motifs (13,14,22), and they are useful for creating graphical sequence logos (23,24) and for performing sequence analysis using any of several published software tools (22,25). Graphical sequence logos of the Seed-and-Wobble motifs, generated using the algorithms defined by enoLOGOS (24), are also present in the UniPROBE database.

Second, we provide two k-mer-based DNA binding profiles for each TF. The universal PBM designs facilitate k-mer binding profile construction because they allow for full coverage of all 8-mers of width 12 or less. The k-mer profiles have several advantages over the traditional PWM model. For example, comprehensive coverage of ungapped and gapped 8-bp sequence variants can provide insight into nucleotide interdependence within DNA binding site sequences; whereas, mononucleotide independence is implicit in traditional PWMs (26,27). The database’s first k-mer-based binding profile consists of the median signal intensities and PBM enrichment scores associated with each contiguous 8-mer, where enrichment scores are calculated using a variant of the Wilcoxon–Mann–Whitney statistic and range from 0.5 for the most favored k-mers to −0.5 for the most disfavored k-mers. The second k-mer-based binding profile includes the top-scoring gapped 8-mer patterns (up to 10 positions) as determined by a 0.25 enrichment score cutoff; we used this threshold to avoid excessive file size for the gapped pattern profile and note that the Universal PBM Analysis Suite (11) can be used to generate full profiles for all gapped 8-mer patterns up to 12-nt positions in length.

In addition to these PBM data files, the UniPROBE database also provides relevant experimental information and factor-specific annotations. Experimental features include protein expression method, sequence and construct information (i.e. full-length protein or DNA-binding domain only). Factor annotations include functional descriptions and links to a variety of protein and gene reference databases. The website structure and interface section subsequently discusses these external annotations in further detail.

WEBSITE STRUCTURE AND INTERFACE

The Browse page of the UniPROBE database site presents a table containing each hosted TF, that protein’s structural class, and the publication with which the protein’s PBM data are associated. The entries are accompanied by brief descriptions of protein function retrieved from IHOP (28) or from a species-specific database such as SGD (29) or WormBase (30). The table then presents a link to a zipped file containing all factor-associated PBM data and a link to a Details page containing further factor annotations and a display of relevant features from the PBM experiments.

The Details page (Figure 2) described above has several components. The first section (Figure 2A) provides additional annotations for the factor of interest, including unique gene or protein accession numbers (if available as provided by the species-specific database), gene synonyms, DNA-binding domain amino acid sequence (if available) and links to databases such as IHOP (28), RefSeq (31), UniProt (32) and JASPAR (14). The second section (Figure 2B) displays the PWM and the matrix motif logo derived from Seed-and-Wobble analysis of PBM k-mer data. Although it does not directly display k-mer data, links to the data files containing complete k-mer data, normalized 60-mer probe data, and raw probe data are provided below the motif logo. The final section of the Details page (Figure 2C) displays a table that presents the experimental conditions and protein sequences used to produce each PBM dataset for the TF of interest.

In addition to the Browse and Details pages, several other pages facilitate the download of specific materials. The Downloads page distributes ZIP files containing all instances of specific data types (i.e. all PWMs or all raw data) in the entire database, along with ZIP files of the data associated with each given publication. The Downloads page also provides links to the core SQL tables used by the database and documentation for these tables. As an alternative method of accessing UniPROBE files, one can use the Apache HTTP Server index to browse for files or use the Explore page to view the directory structure in a Microsoft Explorer style interface.

Most pages in the database website also provide forms for performing text searches on the database and for interrogating PBM-derived binding profiles. The following section describes these tools in detail.

SEARCH AND ANALYSIS TOOLS

Several tools for conducting database searches and performing analyses on TF k-mer binding profiles (Figure 3) enhance the database’s utility. A simple search field in the top right corner of the site’s horizontal navigation bar provides a modified full-text MySQL search across species names, gene names, synonyms and annotations. Under the
Advanced Search navigation bar, a customizable text search allows the user to enter multiple key words within specific database fields for a higher precision query. These terms may be linked by AND or OR Boolean operators by selecting the Match All or Match Any radio buttons, respectively (Figure 3A). The Browse page presents the match results (Figure 3B) for both search methods in the same table format used by the default Browse display, which provides basic annotations for each matching protein and links to download or view the associated PBM data.

The Advanced Search bar also contains two analysis tools available for online use. The first tool, which uses the Tomtom program (33) from the Meta-MEME suite (22), provides a platform for comparing standard motif representations against the PBM-derived PWMs in the database (Figure 3C). The user may enter or upload up to 20 binding site representations in one of the following formats: frequency matrix, count matrix, Meta-MEME 3.x motif or IUPAC motif. As additional options, the user can restrict the query to PWMs from a particular species, specify a maximum similarity threshold, set a minimum matrix overlap or choose the comparison algorithm. Available algorithms include Euclidean distance, Pearson correlation, Kullback–Liebler Divergence and the Sandelin–Wasserman function, which are described in detail in the documentation for the Tomtom program (33). Upon query submission, the tool returns a table of statistics describing the best scoring alignment between a given pair of motifs, a Tomtom E-value (33) that quantifies their similarity and a graphical alignment of the two motifs’ logos (Figure 3D). This table also provides links to each matching factor’s Details page for further investigation of the PBM data.
The second analysis tool uses contiguous 8-mer PBM enrichment score data to scan user-supplied input DNA sequences for putative TF binding sites (Figure 3E). To perform this DNA scan, the user must first upload or enter a FASTA file containing up to 30 DNA sequences, each of a 10-kb maximum permissible length. The user then specifies the species of interest and an enrichment score threshold, and the tool scans the input DNA sequence using an 8-bp sliding window to detect whether any TFs from the species of interest have enrichment scores greater than the user’s threshold for that particular sequence. The website displays the results of the scan both as an HTML table available for plain-text download and as a simple graphic indicating binding site position and TF identity (Figure 3F). This tool may be useful not only for generating hypotheses about putative regulatory interactions but also for minimizing the unintentional creation of new binding sites for unrelated factors when designing site-directed mutagenesis experiments.

**FUTURE DIRECTIONS**

The upcoming publication of several large PBM datasets of yeast, fly and mouse TFs will contribute significantly to the breadth of coverage in the database. We encourage users of the database to register at http://thebrain.bwh.harvard.edu/uniprobe/register.php to receive updates concerning the addition of new datasets and changes to the database interface or analysis tools. We also encourage other labs generating universal PBM data to contact us by email if they wish to add their data to the UniPROBE database following the acceptance of their data for publication. The development of several additional tools may also enhance the website, and they may include a download manager, a local BLASTP function for identifying matches in our database to a user-specified query protein, and a DNA-binding preference prediction tool for user-specified query proteins.

**AVAILABILITY AND LICENSE**

All data hosted by the UniPROBE database are freely available for distribution at the database website. The sequences of the 60-mer DNA probes synthesized on our custom-designed universal arrays are available under the terms of the academic research use license described at http://thebrain.bwh.harvard.edu/uniprobe/academic-license.php. All pages have been tested under Firefox 2.0, Firefox 3.0 and Internet Explorer 7.

**ACKNOWLEDGEMENTS**

We thank A. Philippakis and F. S. He for the survey of SELEX datasets in the 2006 JASPAR database, I. Adzhubei for technical assistance, M. Berger for helpful discussions and S. Gisselbrecht, R. P. McCord, A. Gehrke, and A. Aboukhalil for critical reading of the article.
FUNDING
National Institutes of Health (R01 HG003985 to M.L.B.). Funding for open access charge: National Institutes of Health (R01 HG003985 to M.L.B.).

Conflict of interest statement. None declared.

REFERENCES