Alex Kern

GENE211

3/18/18

TFDB: A Unified Database for Examining the DNA Binding Affinity of Transcription Factors

Abstract: In this project, I describe a set of programs to combine data from the JASPAR and UniPROBE transcription factor-DNA binding databases into a single SQLite database. This database stores the amino acid sequence of each transcription factor, as well as its name, class, family, position frequency matrix (PFM), the type of experiment used to generate the data, the sequences of its DNA binding domains, and the Uniprot IDs associated with it. I used the SQLAlchemy package in Python to access and query these databases, as well as Biopython, and used SQLAlchemy to construct the database. Neither UniPROBE nor JASPAR store all of the information above, so I developed a set of tools using the Proteins API from EMBL-EBI to extract sequences, DNA binding domains (or others, in principal), and species using Uniprot IDs. In addition, since UniPROBE does not store Uniprot IDs, and some JASPAR Uniprot IDs are not actually Uniprot IDs, I also made a function using the Uniprot RESTful API to find Uniprot IDs by querying Uniprot. Using these tools, I have created the first version of the database, TFDB, containing over 1,900 proteins with over 3,000 DNA binding domains and nearly 1900 position frequency or position probability matrices from 49 species.

Introduction: Transcription factors are some of the most important regulatory proteins in a cell.1 Through binding to DNA, they can modulate the transcription of genes, thus changing their expression levels. Many transcription factors are essential genes, and transcription factors can be deregulated or mutated in various diseases including cancer.2,3 Knowing the affinity of a transcription factor for various DNA sequences is essential for being able to determine where it can bind and thus which genes it can regulate. The field has long recognized the importance of transcription factor-DNA binding, and there have been many techniques developed or used to study these interactions, including bacterial-one-hybrid4, systematic evolution of ligands by exponential enrichment (SELEX)5, ChIP-seq/chip6, Protein-Binding Microarrays (PBMs)7, mechanically induced trapping of molecular interactions (MITOMI)8, and many more. These techniques have allowed the binding specificity of a wide array of transcription factors to be characterized. However, these techniques have limitations and biases in terms of throughput, ability to get accurate binding energies, the amount of information generated, and biological relevance.9 The throughput limitations due to the exponential nature of examining all possible DNA sequences as well as mutations in the transcription factors necessitate the development of computational methods to predict the DNA binding affinity of transcription factors to predict new functions and the effects of mutations.9

In addition to generating transcription factor-DNA binding data, there is also the need to store and represent it. One way to store this information is in the form of consensus sequences, which represent the most common nucleotide at each position in a set of aligned binding sites for a given transcription factor. However, this eliminates a lot of data, since some nucleotides affect binding specificity more than others, many transcription factors bind multiple sites that do not match this consensus, making it difficult to identify binding sites using these sequences.10 To represent the relative importance of residues within a DNA binding motif, the Stormo lab developed the position weight matrix (PWM), which represents the self-information for each nucleotide at each position relative to some background frequency.11 This allows for representation of DNA binding motifs as sequence logos, which show the likelihood of each symbol at each position and the relative importance of each position to binding specificity. However, even PWMs do not represent epistatic relationships between nucleotides, since they treat each nucleotide separately.7 A statistical model to map DNA binding domain sequence to DNA affinity would be useful for predicting alternate binding sites and the effects of mutations on transcription factors, but to be able to train an accurate model, there needs to be a large amount of data.12

Several databases have been built to store the results of transcription factor DNA binding experiments, including JASPAR13, Transfac14, CIS-BP15, and UniPROBE16, among others. Some of these databases, including UniPROBE, store data from only one type of experiment, while others compile data from various types of experiments (e.g. JASPAR). However, none of these databases have all the available data on their own, and many are lacking data which would be useful, such as DNA binding domain sequences, external database IDs, and other information. This makes it difficult to compile a complete set of information to map the sequence and potentially structure of transcription factors to their DNA binding specificities. In this project, I combine and augment two of these databases, JASPAR and UniPROBE, and build a set of tools which can be used to gather necessary information about transcription factors for subsequent database additions. This database will serve as a convenient way to access data on transcription factor-DNA binding and will be used in the Fordyce lab for helping to train a neural network to predict DNA binding.

Methods:

Data Acquisition: I downloaded the SQL tables from JASPAR and UniPROBE, as well as the UniPROBE PWMS, from jaspar.genereg.net and the\_brain.bwh.harvard.edu/uniprobe.

Database Tools: I built MySQL databases using MariaDB: <https://mariadb.com>.

Non-standard Python Packages: SQLAlchemy, Biopython, Requests, BeautifulSoup, PyMySQL

APIs: Proteins API EMBL-EBI17, Uniprot RESTful API

Parsing JASPAR: In the jaspar\_db\_to\_dict script, I used the JASPAR5 module in Biopython to extract the information from JASPAR after connecting to the database. I then used the get\_species function from jaspar\_species\_finder, which uses the Uniprot RESTful API to get the species names from the species taxonomic IDs. Next, I used my get\_sequences and get\_domains functions from the uniport\_domains\_sequences file, which use the Proteins API to get DNA sequences and DNA binding domain information from Uniprot using Uniprot IDs. Since some of the entries in the Uniprot IDs field in were not Uniprot IDs, I then used my get\_uniprot\_id function from search\_uniprot\_ids, which uses the Uniprot RESTful API to query for IDs, and then got domains and sequences for these entries. I then dumped the resulting dictionary to a JSON file.

Parsing UniPROBE: Using SQLAlchemy, I used multiple queries to get the information I could from UniPROBE. I then used my get\_uniprot\_id and get\_domains functions to obtain these data. In addition, UniPROBE stores internal publication IDs instead of pubmed IDs, so I switched the IDs to pubmed IDs. I also used the uniprobe\_pwm\_parser and os to search through the downloaded PWM files to get probability matrices and then dumped to a JSON.

Building the database: In database\_builder, I build the database using the JSON files and the schema defined in DB\_setup which is described using SQLAlchemy’s Object Relational Model.

Results: Figure 1 depicts the Entity Relationship Model of the current iteration of this database. Currently, proteins and experiments have a one-to-one relationship, meaning there is one experiment per protein and one protein per experiment. However, there are many DNA Binding domains which are contained in the same protein, and potentially multiple proteins with the same DNA binding domain, due to complexes, mutants, and partial constructs. There is also one set of protein information per protein sequence, and one protein per set of protein information. Finally, each protein has multiple lines of PFMs corresponding to it, since the PFM rows represent one position in the motif. Since the protein name is stored in every table, it is easy to query to get all the information relating to a protein

Figure 2 depicts some interesting summary statistics about the database. Figure 2A shows the number of DNA binding domains belonging to different broad classes, and Figure 2B shows the number of proteins from each species and the fraction they make up of the database. Perhaps unsurprisingly, human, *Arabidopsis* and mouse proteins make up over 60% of the proteins in the database.

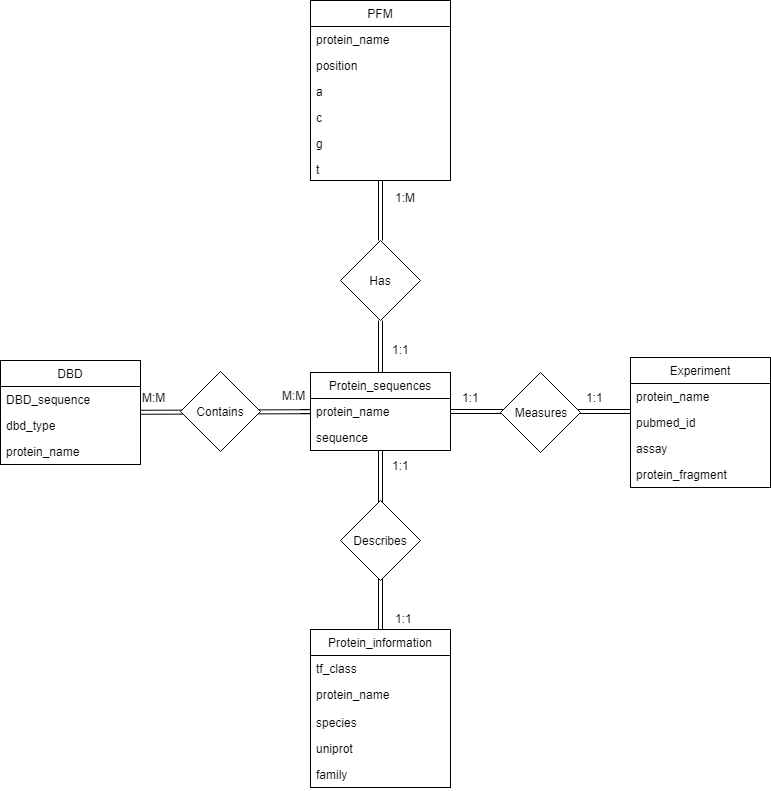


Figure 1: Entity-Relationship Model of TFDB Version 1.0. Relationships in diamonds, each box is a table, and the columns are in the boxes. The cardinalities of the relationships are depicted along the lines connecting the boxes.

Figure 2A

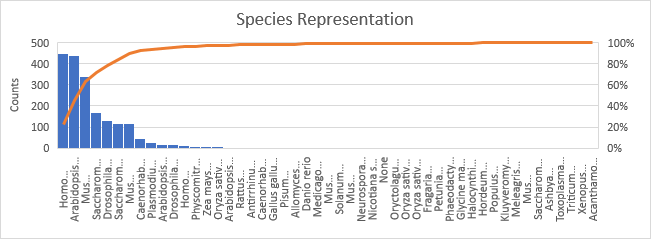
B

Figure 2: Summary statistics about the proteins in TFDB 1.0: A) Counts of DNA binding domains belonging to various broad classes. B) Counts and percentages of proteins belonging to each species.

Table 1: Summary Statistics

|  |  |
| --- | --- |
| Proteins | 1972 |
| DNA Binding Domains | 3334 |
| Uniprot IDs | 1691 |
| PFMs | 1872 |
| Species | 49 |
| Transcription Factor Classes | 54 |
| Transcription Factor Families | 101 |

Discussion: TFDB 1.0 will provide a useful dataset for examining the relationship between DNA binding domain sequence, protein sequence, and position frequency matrix. It will also allow for examination of the biases of various types of experiments used to measure affinity. While it is not currently the largest database of transcription factor binding profiles, it has all of the information needed for various statistical analyses. The tools I developed for accessing information using Uniprot IDs and searching for Uniprot IDs will also make adding additional data more streamlined.

There are many changes I plan to make to the database for its next iteration. The most significant is that I will change the schema to allow for one-to-many relationships between experiments and protein sequences in case a protein has been measured by multiple methods or labs. I also plan to explicitly include the relationship between experiments and position frequency matrices, to better allow for comparison between methods and even labs. In addition, it would be useful to add more integrity constraints and true foreign keys between the database relationships to make it more robust to adding new information.

CIS-BP is a larger database than either JASPAR or UniPROBE, and so it will be useful to add its data to the database for a more complete set. I already have much of the data organized to add CIS-BP, but since it is less curated than JASPAR or UniPROBE, it will require the schema and integrity changes above. Finally, adding PDB IDs to proteins which have structures will be particularly useful for adding structural information to models, and will be relatively straight-forward by extending the tools I have already built.

Outside of constructing a more complete and robust database, I can begin asking questions about how similarity at the sequence or structural level correlates to similarity in DNA binding affinity, and even more interestingly, which changes give the more fine-tuned specificity changes between transcription factors within a family. With species information, evolutionary conservation can also be used explicitly to examine similarities between transcription factors. Finally, the TFDB will be used to add more sequence and PFM data to the neural network that Tyler Shimko in the Fordyce lab is building.

References:

1. Filtz, Theresa M., Walter K. Vogel, and Mark Leid. "Regulation of transcription factor activity by interconnected post-translational modifications." *Trends in pharmacological sciences* 35.2 (2014): 76-85.
2. Seki, Masafumi, et al. "Recurrent SPI1 (PU. 1) fusions in high-risk pediatric T cell acute lymphoblastic leukemia." *Nature genetics* 49.8 (2017): 1274.
3. Hermosilla, Viviana E., et al. "Developmental SALL2 transcription factor: a new player in cancer." *Carcinogenesis*38.7 (2017): 680-690.
4. Zhu, Lihua Julie, et al. "FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system." *Nucleic acids research*39.suppl\_1 (2010): D111-D117.
5. Stoltenburg, Regina, Christine Reinemann, and Beate Strehlitz. "SELEX—a (r) evolutionary method to generate high-affinity nucleic acid ligands." *Biomolecular engineering*24.4 (2007): 381-403.
6. Valouev, Anton, et al. "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data." *Nature methods*5.9 (2008): 829.
7. Berger, Michael F., and Martha L. Bulyk. "Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors." *Nature protocols* 4.3 (2009): 393.
8. Maerkl, Sebastian J., and Stephen R. Quake. "A systems approach to measuring the binding energy landscapes of transcription factors." *Science* 315.5809 (2007): 233-237.
9. de Boer, Carl., Sadeh, Ronen., Friedman, Nir., Regev, Aviv., “Deciphering cis-regulatory logic with 100 million synthetic promoters.” https://www.biorxiv.org/content/early/2017/11/25/224907
10. Schneider, Thomas D. "Consensus sequence zen." *Applied bioinformatics* 1.3 (2002): 111.
11. Stormo, Gary D. "DNA binding sites: representation and discovery." *Bioinformatics* 16.1 (2000): 16-23.
12. Halevy, Alon, Peter Norvig, and Fernando Pereira. "The unreasonable effectiveness of data." *IEEE Intelligent Systems*24.2 (2009): 8-12.
13. Sandelin, Albin, et al. "JASPAR: an open‐access database for eukaryotic transcription factor binding profiles." *Nucleic acids research* 32.suppl\_1 (2004): D91-D94.Matys, Vea, et al. "TRANSFAC®: transcriptional regulation, from patterns to profiles." *Nucleic acids research* 31.1 (2003): 374-378.
14. Matys, Vea, et al. "TRANSFAC®: transcriptional regulation, from patterns to profiles." *Nucleic acids research* 31.1 (2003): 374-378.
15. Weirauch, Matthew T., et al. "Determination and inference of eukaryotic transcription factor sequence specificity." *Cell* 158.6 (2014): 1431-1443.
16. Newburger, Daniel E., and Martha L. Bulyk. "UniPROBE: an online database of protein binding microarray data on protein–DNA interactions." *Nucleic acids research* 37.suppl\_1 (2008): D77-D82.
17. Nightingale, Andrew, et al. "The Proteins API: accessing key integrated protein and genome information." *Nucleic acids research* 45.W1 (2017): W539-W544.