How-to for methods used in:

"*Integrating sequence and gene expression information predicts human genome-wide DNA-binding proteins and suggests a cooperative mechanism*" by Ahmad et al. (2017 under review).

This document is under constant improvement. If you have a question, which is not addressed, please send us an email shandar@jnu.ac.in and we will incorporate answer to your question into this document.

Please note that minor differences between data used in the manuscript and protocols presented here may appear due to constant updating of various public resources from which these information is compiled. The following text is meant to conveniently implement the protocols used in the paper and in some cases a simpler alternative has been included here, which may be somewhat different from the one used in the manuscript. For example, GO terms were originally compiled from TargetMine, but may also be extracted from Uniprot data files, which are easier to download as a full database and are provided here. That is a suggested route for this information (see below).

1. Genome-wide data sets of human/ alternative organism proteins:

UniProtKB is the primary source for this information. Latest version of the complete data set of proteins (for ALL organisms) together with related details can be downloaded from:
ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.dat.gz
After uncompressing the downloaded file, species-wise data can be extracted easily from the downloaded file by parsing the field identified by "OS" (uniprot_sprot.dat).

Following lists the scripts uploaded at gigease.sciwhylab.org to enable extraction of various information's from the *uniprot_sprot.dat*.

2. *Extracting sequence features and keyword data from uniprot_sprot.dat*

All sequence feature field data from the protein master file (*uniprot_sprot.dat*) can be extracted by parsing field labeled “FT”. An example bash script to do so is provided in *gigeasa* website in the directory “*feature-computing-codes/*”, by the name *get-FT.sh*. Similarly, the keyword data can be extracted by parsing “KW” field of the same file, with an example script to do so being provided *get-KW.sh*.

3. *Annotation of DBP:*

Please note that the elaborate annotation of DBP in multiple ways has been implemented here only for human DBPs, even though ideas presented here are general and can be easily extended to other systems.

3.1 DBP as per UniProt Sequence feature (DBP by SeqFT):

DBP annotation by keyword can be extracted as per the second line of get-FT.sh, described above. This simply looks for natural language based parsing of FT field.

3.2 DBP as per UniProt Keyword (DBP by KW):

DBP annotation by keyword can be extracted as per the second line of get-KW.sh, described above. Please note that the KW field has been significantly updated in the Uniprot website since our first use for this work and has become more accurate in describing DBP since then.

3.3 Annotation of DBP by Pfam:

Pfam based annotation requires two steps. First a reference list of Pfam domains, which can be directly annotated as DBP is needed. Next, Pfam domain content of each protein will be needed to determine if a DBP annotated domain is present. The second component of this can be achieved by parsing the uniprot_sprot.dat, described above using get-Pfam.sh script available on this website. For generating a pfam domain reference list of DBP-associated domains, precompiled lists in

Supplementary Table ST2(b) of the manuscript can be used. This list is prepared as follows:

A keyword search for “DNA binding” on Pfam website returned four sets of results based on the occurrence of this term in the Pfam header, Pfam sequence abstracts, matching PDB headers and title records and text abstracts of corresponding InterPro entries. Except for Pfam header list of 224 entries, of which 75 are observed in 352 SP_human proteins, text search contains many domains which are only indirectly involved in DNA-binding or co-occur with another DNA-binding domain. We wished to collect in this group only those domains, which directly participate in DNA-binding and the occurrence of such domains in a protein implies a high confidence that this protein has a DNA-binding function. To achieve this goal, we developed a protocol to create a reference list of DNA-binding Pfam domains, as described in Supplementary Figure SF1. As a result, 121 Pfam domains with high confidence in their DNA-binding behaviour were obtained (Supplementary Table ST2(b)). This list can be used for annotating DBPs by Pfam for any organism.

3.4 Annotation of DBP by sequence similarity with proteins that were observed in protein-DNA complex structures in the Protein Data Bank (PDB):

A protein is annotated as DBP by PDB if it showed a sequence identity of 90% or higher with a protein involved in protein-DNA complexes. All we need to get this annotation is the data set of protein-DNA complexes. Some of the proteins in a complex may not be binding to DNA, so we need to compute all the DNA-binding sites of all residues in the protein (any atom close to 3.5) and removed those proteins, which have no DNA contact. In our manuscript, we collected 3106 protein-DNA complexes from the PDB containing 6670 protein chains, of which 6182 had at least one residue contacting DNA in the complex (any protein atom within 3.5 Å of any DNA atom). (Results specific to human proteins in this regard are available as part of the supplementary material of the manuscript).

3.5. DBP annotation by Gene Ontology (GO).

Complete list of GO terms associated with a given protein can be extracted from uniprot_sprot.dat file using the script (get-GO.sh) provided in gigeasa website here. Gene Ontology (GO) term “DNA-binding” with accession code GO:0003677 includes all proteins involved in DNA-binding. We have observed that many UniProt entries associated with DNA-binding are NOT annotated by this term but a more specific (Child term) such as “Single Stranded DNA-binding” etc. This problem was solved by using TargetMine which could be searched for proteins associated with a GO term whose parent term includes a query GO term. However, TargetMine does not include GO annotations for all species and therefore any extension of the method will require careful examination of all Child terms of GO terms extracted from uniprot_sprot.dat.

4. Computing sequence features of proteins for analysis and prediction:

This particular step may be a bit difficult to reproduce quickly because we use our previously reported tools for computing some of the features. To overcome the difficulty in reproducing these features, we have taken two steps here. First, we started compiling pre-computed feature sets for this step and making them available to the users. As a start, annotations for mouse and Arabidopsis thaliana are provided. Second, we are providing the source codes for all the tools we have used in our work for computing these features. These tools are still not prepared for sharing purpose and require setting up paths and installation scripts. We are working on that, but provide the raw source which can be used by expert users who wish to understand the working of these tools. The description of these tools is provided below.

In this work, we compute binding sites on proteins for five different types of ligands viz. ATP, Carbohydrates, RNA, DNA and Proteins based on our own published methods. First, we compute binding sites for each protein using these tools and then create a whole protein summary features. The scripts to convert binding site data into summary features are available from gigeasa website and they are defined as follows:

1. The average score for the top 5, top 10, top 25 and third quartile score of predicted DNA-binding sites (described as pDBSs in the manuscript).
2. Features as in (1) based on binding sites of ATP, carbohydrate mRNA and proteins each.

5. Sequence feature based prediction of DBPs

After assigning all DBP annotations as a target class to be predicted and sequence features computed from DBS as described above, model training is performed as described in the manuscript. We have attempted multiple linear regression, support vector machine and random forests for the predictions, of which, as expected MLR is slightly less accurate while the other two give similar performances under various cross-validation and parameterizations. The source code to train models with all three methods (MLR, SVM, RF) are provided in the training-codes/ directory of gigease website.

6. DBP annotation of genes:

First of all, probe-level annotations of the Affymetrix platform are converted to gene names by the dictionary provided in the Gene expression omnibus (GEO) platform file.

In our manuscript DBP coding genes were selected by virtue of their GO annotations, obtained from TargetMine. All levels of GO associations for biological process, molecular function and cellular component were utilized for this purpose. Users may also extract Gene/protein mapping from uniprot_sprot.dat file by parsing the corresponding field.

Mapping genes in Affymetrix to protein name:

The mapping can be done by TargetMine or uniprot_sprot.dat file. In some cases, a gene maps to more than one proteins in the protein list. In such cases, we have arbitrarily used the first occurrence of the protein as mapping.

7. Compiling gene expression profiles:

This step can be done reliably only at an organism level. At the time of starting this work, GEO contains about 1.2 million gene expression profiles, accumulated over some 13,000 different “platforms”. For each organism expression data are available as single “platform” file from GEO. However, multiple platforms are available for the same species. In our work, we have prepared expression profile data for human (as in the manuscript), mouse and Arabidopsis thaliana (as on gigeasa.sciwhylab.org) by using the platform with highest number of transcriptomes (samples) in each case.

Most recent data for human, mouse and Arabidopsis can be downloaded from GEO under platform IDs GPL570, GPL1261 and GPL198 respectively. Platform files contain sample wise expression values from which a simple text parser can extract a matrix in which columns represent a sample name and rows contain the “probe ID”. Probe ID is converted to gene name as per the annotation file available from GEO for each platform. When more than one probe corresponds to a gene, the highest expression value is assigned to the gene.

8. Compiling DBP annotations for genes in Affymetrix microarray chip:

Probe-level annotations of the Affymetrix platform were converted to gene names by the dictionary provided in the Gene expression omnibus (GEO) platform file. DBP coding genes were selected by virtue of their GO annotations, obtained from TargetMine. All levels of GO associations for biological process, molecular function and cellular component were utilized for this purpose, allowing the non-discriminating features to be automatically eliminated by a feature selection and training model.

9. Computing gene expression level (EL) features:

EL histograms with 20 global equal-probability bins are computed by defining equal-probability bins of $M \times N$ EL values, where M is the number of genes and N is the number of EL values for each gene. In this study, M=20,318 and N=72,488 as stated above. For the 20 equal bin values defined from a global pool of genes and samples, individual EL feature profiles for each gene were computed by counting the relative number of occurrences (out of M values) in each of the 20 bins. These 20 values represent our EL feature set for each gene.

10. Computing co-expression level (CEL) features:

To compute the co-expression features for each of the M genes, their EL values in N samples were compared with those for all the other genes. The resulting *M* values (Pearson correlation coefficients) are summarized as co-expression histograms similar to the EL probability features described above. Again, the histogram bins were recomputed by considering all the *N×(N-1)/2* co-expression values from unique pairs of genes and the distributions of co-expression levels in these 20 bins were used as co-expression level (CEL) features of that gene.

11. Computing network gene ontology composition (NGC) features:

Network gene ontology composition (NGC) features were derived by computing a histogram of GO terms' occurrences; given a gene, GO terms for its *T* top co-expressed (positive correlations) and *L* least co-expressed (negative correlations) genes were pooled and counted. In this study, both *T* and *L* were set to 50, resulting in GO histograms based on the annotations of top 100 "co-expressed genes" for each query gene. Overall 138 GO terms were found to be present in at least 200 genes in the entire list (used for filtering) and hence the NGC features were composed of 138-dimensional integer valued vectors.