Protein-DNA Structure-Affinity Database

The protein-DNA structure-affinity database is comprised of multiple files. This document provides a detailed description of their contents.

Metadata Database (Database.accdb)

This file is a relational database (MS Access file format) that contains all the metadata information concerning the protein-DNA complexes in the database. It is comprised of several tables whose contents are described below.

Complex Table

This is the central table in the database. Its entries correspond to the unique protein-DNA complexes in the database. Each complex corresponds to a single protein domain that binds DNA. The contents of each entry field are described below:

ID: Unique identifier specific to the table.

Complex Name: Unique name that identifies a complex across all tables in the database. Names obey the following specific format:

pdb {pcn -> {fdscn, bdscn}, {spr, epr} -> {{sbfs, ebfs}, {sbbs, ebbs}}}

pdb: PDB code of the structure containing the complex.
pcn: Protein chain number in the PDB file that contains the DNA-binding protein domain.
fdscn: Forward DNA strand chain number in the PDB file to which the protein binds.
bdscn: Backward DNA strand chain number in the PDB file to which the protein binds.
spr: Starting protein residue of the recognition helix.
epr: Ending protein residue of the recognition helix.
sbfs: Starting DNA base on the forward strand to which protein binds.
ebfs: Ending DNA base on the forward strand to which protein binds.
sbbs: Starting DNA base on the backward strand to which protein binds.
ebbs: Ending DNA base on the backward strand to which protein binds.

PDB Name: PDB code of the structure containing the complex.

Protein Name: Common protein name for the complex.

Gene Name: Common gene name for the complex.

Motif: Protein family type (e.g. HTH).

Submotif: Protein family subtype (e.g. Winged HTH).

Organism: Organism from which protein structure was obtained.

Quantitative: List of databases / original sources that contain quantitative DNA-binding information on the complex.

Binding Site Table

Entries in this database correspond to unique and experimentally verified DNA binding sites.

ID: Unique identifier specific to the table.

Complex Name: Name of the complex which binds this DNA sequence.

Sequence: Base sequence of the DNA binding site.

DB: Database or literature source from which binding site information was obtained.

DB ID: The binding site ID in source database (if applicable).

Quality/Source: Information on the quality or type of experiment used to identify binding site.

Organism: Organism(s) in which binding site was identified.

PWM Table

This table contains information about the PWMs of each complex in the database.

ID: Unique identifier specific to the table.

Complex Name: Unique name that identifies a complex across all tables in the database.

Source: The source(s) used in deriving the PWM. If PWM was directly obtained from an external source, the source name is used. If PWM was constructed from a partial or the complete set of binding sites in the database, the indicators “Some/All binding sites in this database” are used, respectively.

N: The number of binding sites used in deriving the PWM.

DB Table

This table contains a list of all database and literature sources from which DNA binding sites were obtained.

ID: Unique identifier specific to the table.

Title: Name of database or DOI/PMD/PMCID of literature source.

Organism Table

This tables contains a list of all organisms from which DNA binding sites were obtained.

ID: Unique identifier specific to the table.

Species Name: Scientific name of organism.

StructuralMotif Table

This table contains a list of all structural families with representative complexes in the database.

ID: Unique identifier specific to the table.

Classification: Domain family name.

StructuralSubmotif Table

This table contains a list of all structural subfamilies with representative complexes in the database.

ID: Unique identifier specific to the table.

Classification: Domain subfamily name.

Parent Classification: Domain family name.

StructuralOutliers Table

This table contains a list of proteins whose HTH domains each formed a structurally unique cluster.

ID: Unique identifier specific to the table.

Protein Name: Common protein name.

Organism: Host organism of protein.

PDB IDs: PDB structures associated with protein.

PDB Structures (Structures Folder)

PDB files containing the atomic coordinates of all complexes in the database are included in the "Structures" folder. The name of each file corresponds to the unique "Complex Name" in the metadata database. All files have been standardized in the following way to facilitate programmatic use:

- Residues are hydrogenated.
- Water molecules have been stripped.
- Missing atoms have been added.
- Protein chain(s) always precede DNA chains (when applicable).
- DNA chains are always ordered with the forward strand first.
- DNA molecules are always double-stranded.
- Only protein-contacting DNA basepair positions have been retained. As such, the PWM positions can be mapped uniquely and in a one-to-one fashion to the basepair positions in the PDB files. If a basepair position is missing from the PDB file, its corresponding entry in the PWM is marked NA.

Each PDB file is provided in multiple formats, which are placed in separate folders inside the "Structures" folder. They are:

Full: Contains entirety of DNA-binding protein and DNA molecule.

HTH + DNA: Contains only the recognition helix of the protein and the DNA molecule.

HTH Only: Contains only the recognition helix of the protein.

DNA Only: Contains only the DNA molecule.

Position Weight Matrices (PWMs Folder)

A position-weight matrix or PWM was derived for each complex in the database. These PWMs are provided as matrices in text files. The name of each file corresponds to the unique "Complex Name" in the metadata database. Rows corresponds to distinct DNA basepair positions (13 in total). If a complex is missing a basepair position, the contents of the corresponding row are filled with "NA". Columns correspond to different nucleotides, ordered as "A", "T", "C", "G". Entries correspond to the probability of binding a nucleotide at a given position.