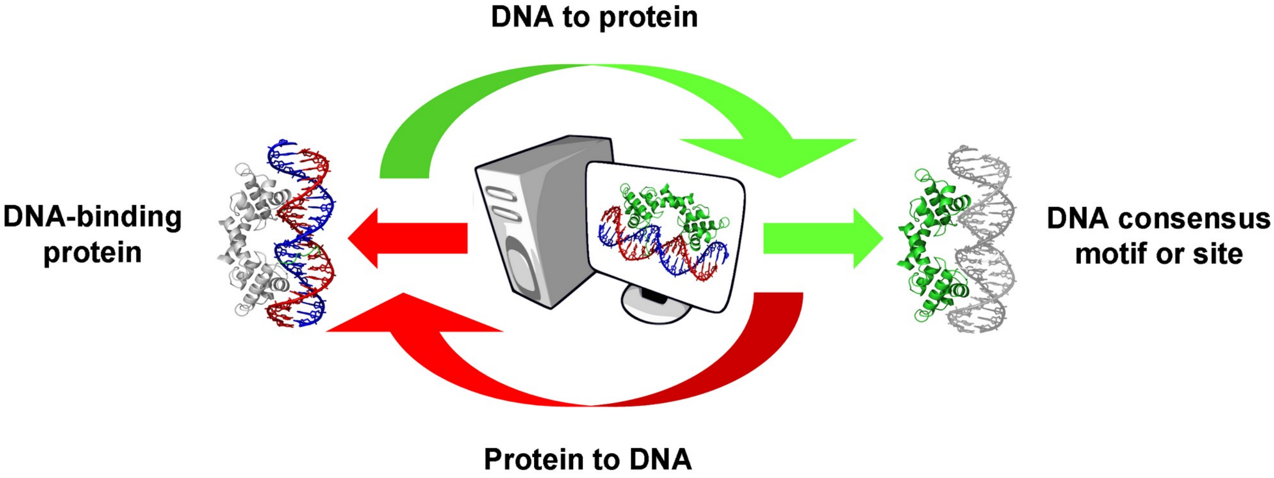
footprintDB



User Manual

[updated Mar 2023]

Alvaro Sebastian Yagüe & Bruno Contreras-Moreira
Computational & Structural Biology Group
Estación Experimental de Aula Dei-CSIC
Av. Montañana 1.005, 50059 Zaragoza (SPAIN)

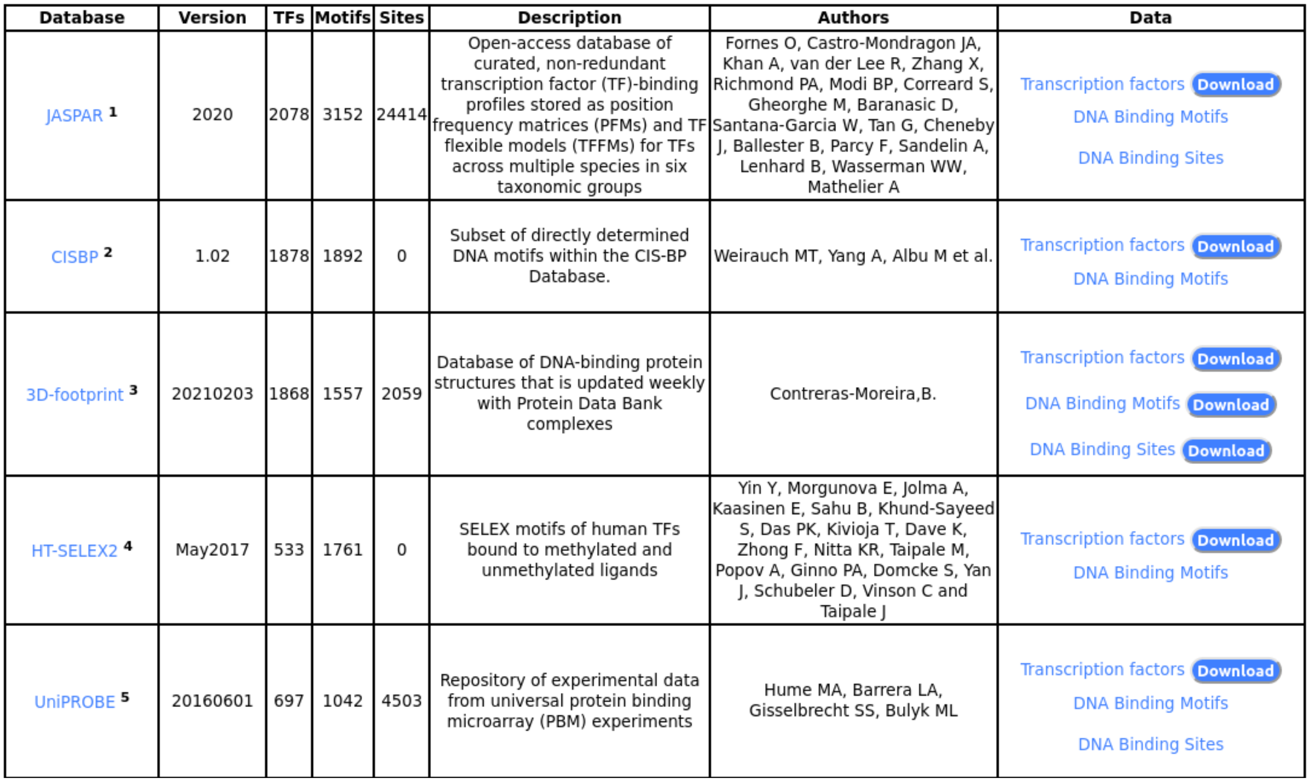
Introduction

footprintDB is a web server for assigning putative cis DNA motifs to input transcription factors (TFs) and conversely for predicting which TFs that might recognize input DNA motifs.

footprintDB consists in a core of **curated and annotated DNA binding data**. These data have been obtained from literature and public repositories and it is stored in a database. Among these data are the protein sequences of the TFs, their DNA binding sites (DBSs) and their Position-Specific Scoring Matrices (PSSM) that summarize the binding preferences, together with their Pfam protein domains, literature references and the set of annotated DNA binding protein interface residues. The main **footprintDB** features are described in detail in the following sections.

1. footprintDB is a repository of databases

footprintDB is a meta-database of TFs attached to their experimentally determined DNA binding preferences. Therefore it does not incorporate other databases which contain only TF sequences or sites. As of Mar 2022, **footprintDB** contains **7169 unique DNA-binding proteins** (mostlyTFs) and **7763 PSSMs)** (1) extracted from the literature and other repositories, including structure-based 3D-footprints, which are also used to annotate interface residues by homology (Contreras-Moreira 2010). Please check https://floresta.eead.csic.es/footprintdb/index.php?databases to see the current contents of footprintDB, including links to the original sources and the relevant citations:



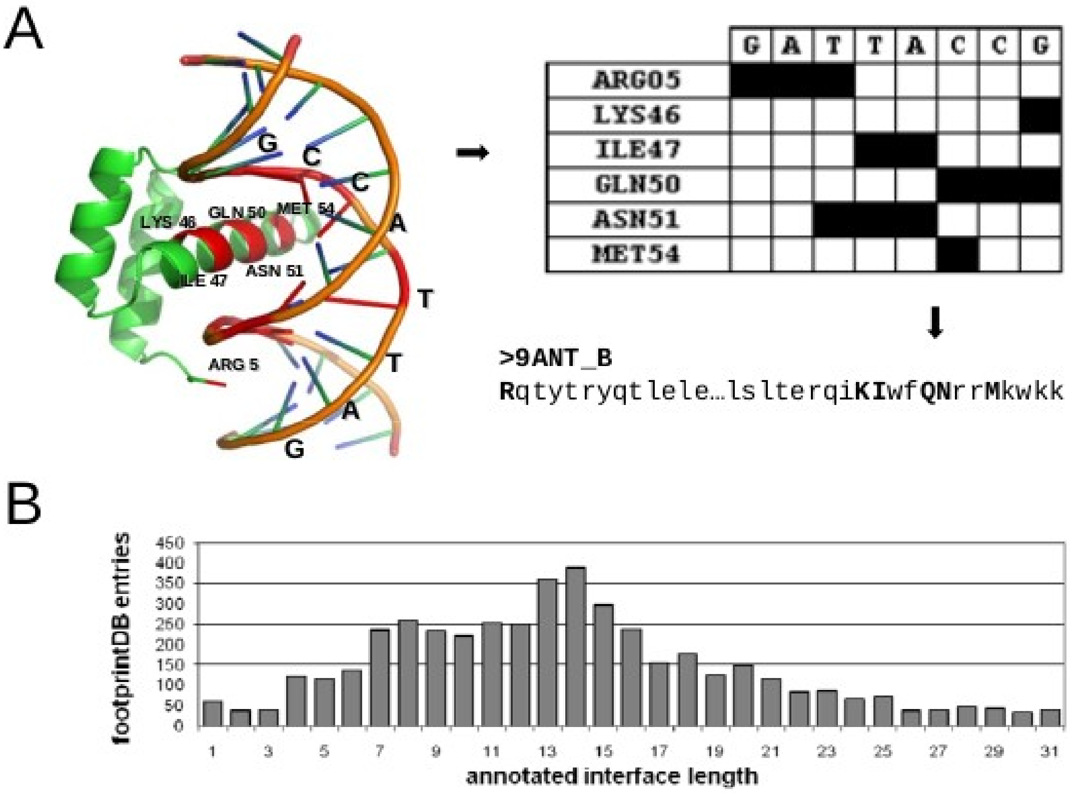
Database	Version	TFs	Motifs	Sites	Description	Authors	Data
JASPAR [1]	2020	2078	3152	24414	Open-access database of curated, non-redundant transcription factor (TF)-binding profiles stored as position frequency matrices (PFMs) and TF flexible models (TFFMs) for TFs across multiple species in six taxonomic groups	Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranasic D, Santana-Garcia W, Tan G, Cheneby J, Ballester B, Parcy F, Sandelin A, Lenhard B, Wasserman WW, Mathelier A	Transcription factors Download DNA Binding Motifs DNA Binding Sites
CISBP [2]	1.02	1878	1892	0	Subset of directly determined DNA motifs within the CIS-BP Database.	Weirauch MT, Yang A, Albu M et al.	Transcription factors Download DNA Binding Motifs
3D-footprint [3]	20210203	1868	1557	2059	Database of DNA-binding protein structures that is updated weekly with Protein Data Bank complexes	Contreras-Moreira,B.	Transcription factors Download DNA Binding Motifs Download DNA Binding Sites Download
HT-SELEX2 [4]	May2017	533	1761	0	SELEX motifs of human TFs bound to methylated and unmethylated ligands	Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, Das PK, Kivioja T, Dave K, Zhong F, Nitta KR, Taipale M, Popov A, Ginno PA, Domcke S, Yan J, Schubeler D, Vinson C and Taipale J	Transcription factors Download DNA Binding Motifs
UniPROBE [5]	20160601	697	1042	4503	Repository of experimental data from universal protein binding microarray (PBM) experiments	Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML	Transcription factors Download DNA Binding Motifs DNA Binding Sites

HOCOMOCO [6]	10	535	1066	0	Homo sapiens comprehensive model collection containing carefully hand-curated TFBS models constructed by integration of binding sequences obtained by both low and high-throughput methods	Kulakovskiy I.V., Vorontsov I.E., Yevshin I.S., Soboleva A.V., Kasianov A.S., Ashoor H., Ba-Alawi W., Bajic V.B., Medvedeva Y.A., Kolpakov F.A., Makeev V.J.	Transcription factors Download DNA Binding Motifs
AthalianaCistrome [7]	v4_May2016	532	861	0	Raw and unmethylated genomic DNA DAP-seq motifs for Arabidopsis thaliana in-vitro-expressed TFs.	OMalley RC, Huang SC, Song L et al.	Transcription factors Download DNA Binding Motifs
HumanTF [8]	1.0	532	818	0	Sequence-specific binding preferences of human TFs obtained by high-throughput SELEX and ChIP sequencing. A total of 830 binding profiles were obtained, describing 239 distinctly different binding specificities	Jolma A., Yan J., Whitington T., Toivonen J., Nitta K.R., Rastas P., Morgunova E., Enge M., Taipale M., Wei G., Palin K., Vaquerizas J.M., Vincentelli R., Luscombe N.M., Hughes T.R., Lemaire P., Ukkonen E., Kivioja T., Taipale J.	Transcription factors Download DNA Binding Motifs
HumanTF2 [9]	1.0	139	663	0	Sequence-specific binding preferences of human TF pairs that bind cooperatively to DNA obtained by CAP-SELEX analysis of 9,400 TF-TF-DNA interactions. This analysis revealed 315 TF-TF interactions recognizing 618 heterodimeric motifs, most of which have n	Jolma A., Yin Y., Nitta K. R., Dave K., Popov A., Taipale M., Enge M., Kivioja T., Morgunova E., Taipale J.	Transcription factors Download DNA Binding Motifs
AthaMYB [10]	1.0	116	224	16	Analysis of the DNA-Binding Activities of the Arabidopsis R2R3-MYB Transcription Factor Family by One-Hybrid Experiments in Yeast.	Kelemen Z., Sebastian A., Xu W., Grain D., Salsac F., Avon A., Berger N., Tran J., Dubrecq B., Lurin C., Lepiniec L., Contreras-Moreira B., Dubos C.	Transcription factors Download DNA Binding Motifs Download DNA Binding Sites Download
FlyZincFinger [11]	1.0	69	141	0	Global analysis of Drosophila Cys2-His2 zinc finger proteins	Enuameh MS et al (Wolfe Lab)	Transcription factors Download DNA Binding Motifs
SMILE-seq [12]	1.0	71	103	0	Validated DNA-binding data from a set of full-length human, mouse and Drosophila TFs, including some heterodimers and distinct structural classes. This data has been obtained by semiautomated protein-DNA interaction characterization technology (SMiLEseq)	Isakova A., Groux R., Imbeault M., Rainer P., Alpern D., Dainese R., Ambrosini G., Trono D., Bucher P., Deplancke B.	Transcription factors Download DNA Binding Motifs
ArabidopsisPBM [13]	20140210	64	108	0	Collection of 63 plant TFs representing 25 families, using protein-binding microarrays. Almost half of these TFs recognize secondary motifs, which in some cases were completely unrelated to the primary element. Analyses of coregulated genes and transcript	Franco-Zorrilla JM Lopez-Vidriero I Carrasco JL Godoy M Vera P Solano R	Transcription factors Download DNA Binding Motifs
Athamap [14]	20091028	74	91	144	AthaMap provides a genome-wide map of potential transcription factor binding sites (TFBS) in Arabidopsis thaliana	Bulow L, Engelmann, S., Schindler, M., Hehl R.	Transcription factors Download DNA Binding Motifs DNA Binding Sites
DBTBS [15]	1.0	71	90	1261	A database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information	Sierro N., Makita Y., de Hoon M.J.L. and Nakai K.	Transcription factors Download DNA Binding Motifs DNA Binding Sites

RegulonDB [16]	7.5	82	82	1862	Curated data of the transcriptional regulatory network of Escherichia coli K12	Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muniz-Rascado L, Martinez-Flores I, Salgado H, Bonavides-Martinez C, Abreu-Goodger C, Rodriguez-Penagos C, Miranda-Rios J, Morett E, Merino E, Huerta AM, Trevino-Quintanilla L, Collado-Vides J.	Transcription factors Download DNA Binding Motifs DNA Binding Sites
DrosophilaTF [17]	1.1	60	62	886	Motif models reported in 51 primary references in the form of position PWMs for 56 Drosophila melanogaster transcription factors	Down TA, Bergman CM, Su J, Hubbard TJ	Transcription factors Download DNA Binding Motifs DNA Binding Sites
EEADannot [18]	2021-11-23	34	46	19	Selected plant motifs and sites manually curated at EEAD-CSIC	Contreras-Moreira B, Sebastian A.	Transcription factors Download DNA Binding Motifs DNA Binding Sites

Table 1: Databases included in footprintDB database and number of Tfs, PWMs and sites retrieved from each one.

TF sequences in footprintDB have their **DNA-binding interfaces residues** annotated by means of BLASTP alignments to protein complexes annotated in the 3D-footprint library (http://floresta.eead.csic.es/3dfootprint/download/list_interface2dna.txt). Aligned interface positions from one or more protein-DNA complexes are thus transferred to entries in the database as explained in the Figure:



	G	A	T	T	A	C	C	G
ARG05	■	■	■					
LYS46								■
ILE47				■	■			
GLN50						■	■	■
ASN51			■	■	■			
MET54						■		

>9ANT_B
RqtytryqtIele…lslterqiKIwfQNrrMkwkk

Figure 1: Annotation of interface residues applying the geometrical rules of 3D-footprint. (A) Interface of PDB entry 9ANT, which corresponds to Homebox protein Antennapedia in complex with a cis element. First, inter-atomic distances are calculated among heavy atoms of both amino acid side chains and nitrogen bases. Second, a matrix of interface contacts is compiled. Third, interface residues are marked as upper-case letters in the sequence. (B) Histogram of predicted interfaces in footprintDB, transferred from 3D-footprint entries through BLASTP alignments.

2. footprintDB is a search engine

footprintDB search engine is designed primarily to receive two types of queries:

1. INPUT: a DNA motif or site

   OUTPUT: a list of DNA-binding proteins (mainly TFs) known to bind a similar DNA motif

2. INPUT: a protein sequence of a putative DNA-binding protein

   OUTPUT: a list of possibly recognized DNA motifs

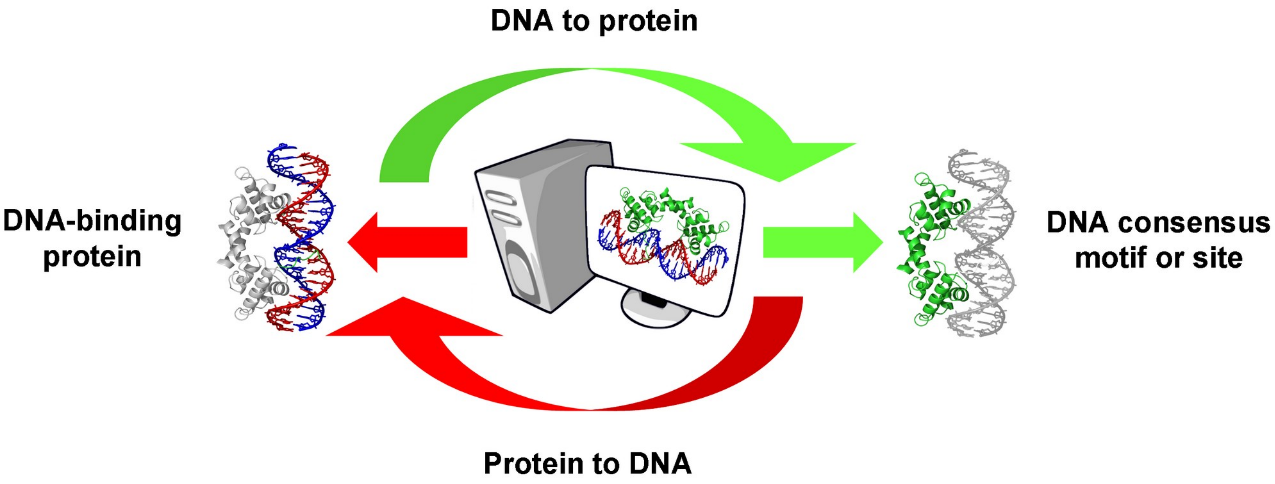
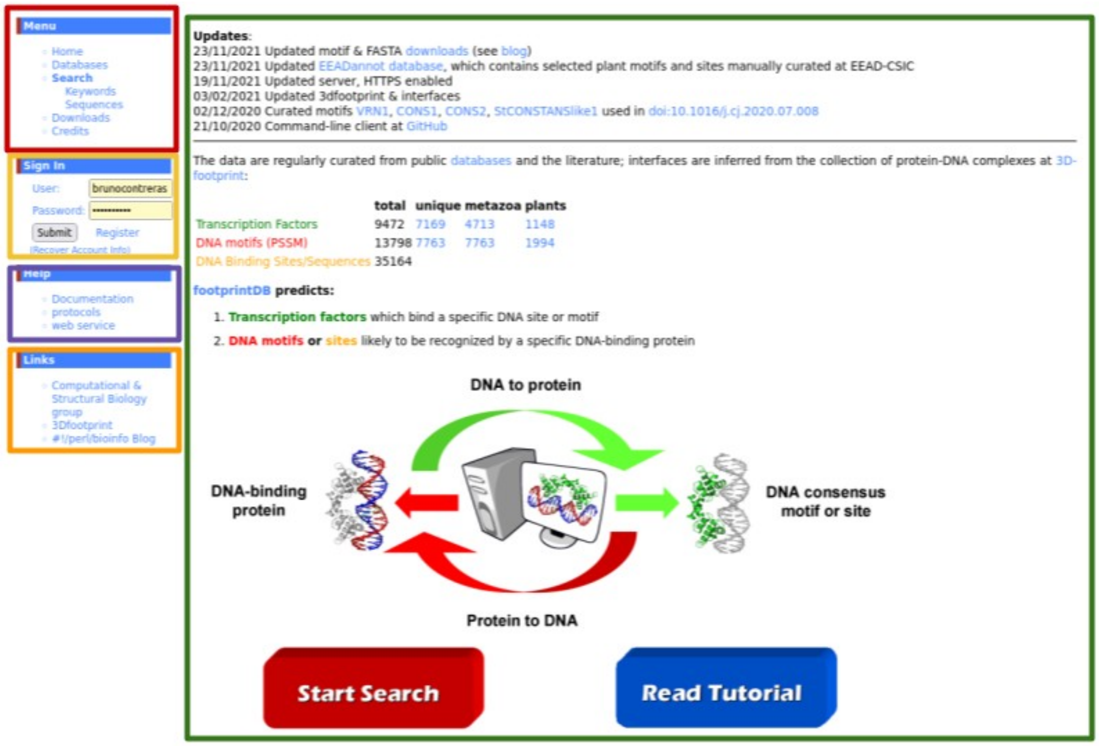


Figure 2: Working schema of footprintDB search engine

Website navigation

1. Website sections



Updates:
23/11/2021 Updated motif & FASTA downloads (see blog)
23/11/2021 Updated EEADannot database, which contains selected plant motifs and sites manually curated at EEAD-CSIC
19/11/2021 Updated server, HTTPS enabled
03/02/2021 Updated 3dfootprint & interfaces
02/12/2020 Curated motifs VRN1, CONS1, CONS2, StCONSTANSlike1 used in doi:10.1016/j.cj.2020.07.008
21/10/2020 Command-line client at GitHub

The data are regularly curated from public databases and the literature; interfaces are inferred from the collection of protein-DNA complexes at 3D-footprint:

	total	unique	metazoa	plants
Transcription Factors	9472	7169	4713	1148
DNA motifs (PSSM)	13798	7763	7763	1994
DNA Binding Sites/Sequences	35164			

footprintDB predicts:

1. **Transcription factors** which bind a specific DNA site or motif
2. **DNA motifs or sites** likely to be recognized by a specific DNA-binding protein

From top to bottom and from left to right:

- **Main menu**: Home, Database listing, Search formulary, Downloads and Credits sections.
- **Sign In Menu**: Authentication form and Registration form link.
- **Help menu**: link to Documentation Section.
- **Links menu**: links to recommended pages.
- **Content**: displays the content of the Sections.
- **Footer**: Shows license and contact info.

2. Navigation

Left menus must be used to navigate across the website.

The **'Main menu'** is composed by the following sections:

- **Home**: Access to home page with general information.
- **Databases**: Information about the databases included in footprintDB.
- **Search**: Access to search formulary.
- **Downloads**: Access to flat files for bulk download.
- **Credits**: Information about footprintDB creators, citing, data sources and other resources.

The **'Sign In menu'** is composed by an authentication form and a couple of links to:

- Register: Access to a registration form for new users.
- Recover Account Info: registered users can recover their account data.

The 'User menu' is only visible for authenticated users and is composed by the following sections:

- Stored results: Access to a historical record of searches performed by the user.
- Insert database: Insertion of user data collections.
- Manage databases: Manage user data inserted previously.
- Modify account: A form to modify user account data.
- Delete account: An option to remove an account.
- Log out

The '**Help menu**' provides links to extensive footprintDB documentation, including:

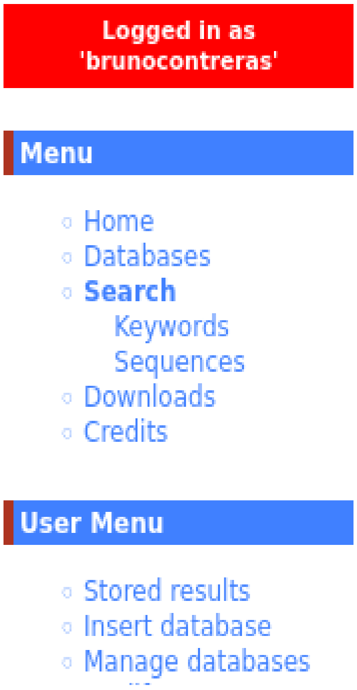
- This manual
- A step-by-step protocol: https://digital.csic.es/handle/10261/136262

3. Registration

Click on the link 'Register' in the 'Sign In' menu on the left or go directly to http://floresta.eead.csic.es/footprintdb/index.php?user_register . Fill in the registration form, required fields are marked with asterisk, and push the 'Register' button to submit the data. You will see the following message if registration was successful: "User successfully registered, you will shortly receive an email with your account information" .

4. Log In

Enter 'User' and 'Password' in the 'Sign In' menu on the left and push the 'Submit' button. If successful, a message will be shown, your user name will be shown in red in the top of the left menus and the 'User Menu' will be now visible:



5. Stored results

Registered users can access to a list of stored searches. Log in and click on the 'Stored results' link in the 'User menu' on the left side. A list of the performed searches will be shown.

Recent search results can be accessed by clicking on the 'view results' link. Old searches are deleted from the server; if you want to repeat one of these searches, click on the 'reuse search' link and the search formulary will be filled with the data of the old search.

6. Log out

Click on the 'Log out' option in the 'User menu' on the left side. You will see the following message: “You have successfully logged out, thank you or using footprintDB” and 'User Menu' will hide (unless automatic cache is activated in your browser; in this case the menu will be visible until any other item is clicked).

7. Recover account info

Click on the 'Recover account info' link in the 'Log In' menu on the left side. Enter your email address and push the 'Recover' button. If any account is associated to that address, you will receive your account data by email and a new auto-generated password.

8. Modify account info

Log in and click on the 'Modify account' link in the 'User menu' on the left side. Modify your data in the formulary, required fields are marked with asterisk, and push the 'Modify' button to submit the data. You will see the following message if registration was successful: “User account successfully modified, you will shortly receive an email with your new account information.” and you will receive an email to remember your account data.

9. Delete account

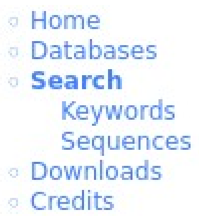
Log in and click on the 'Delete account' link in the 'User menu' on the left side. Please confirm that you want to delete the account by pushing the 'Delete' button.

Searching

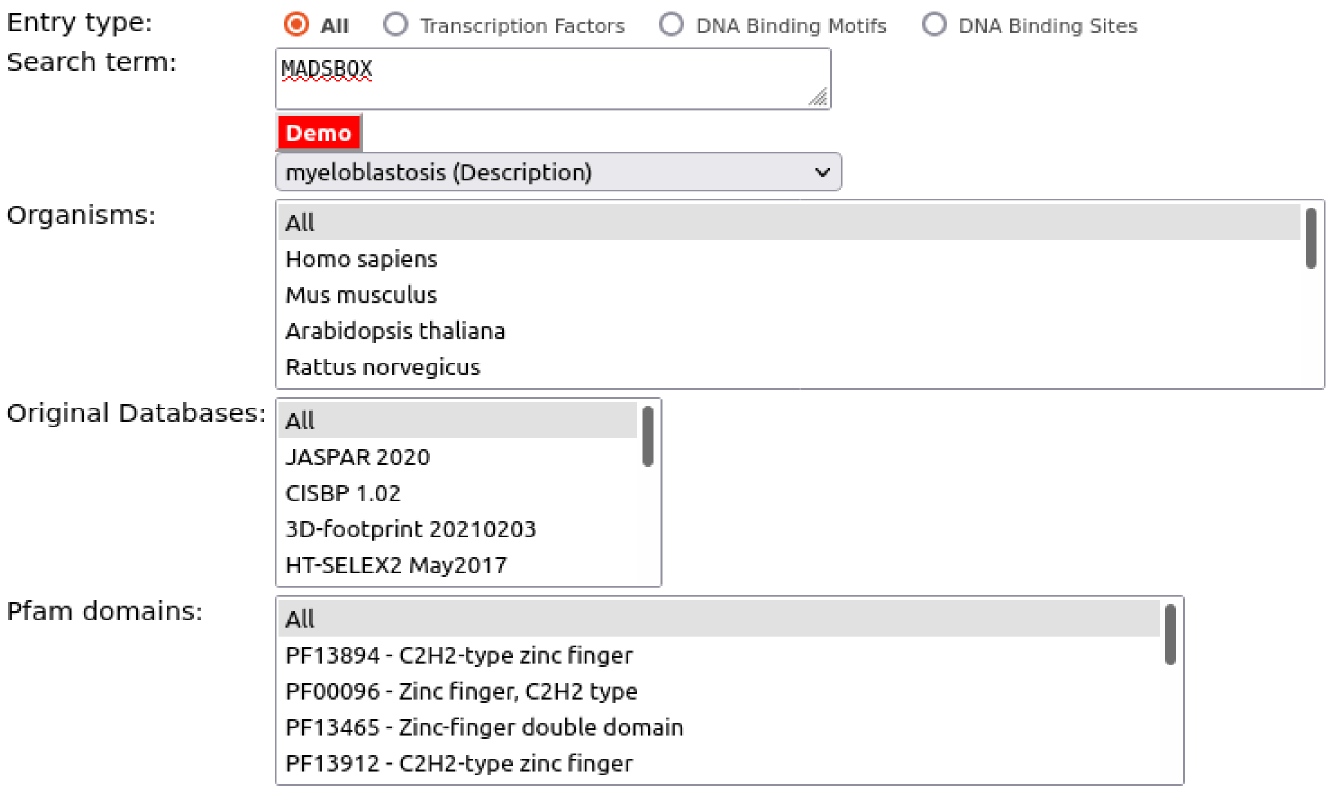
1. Search keywords

If you have a footprintDB account, log in first into the website to store your searches and reuse them.

Click in the ‘Search Keywords’ option of the ‘Main menu’ or go directly to the url: http://floresta.eead.csic.es/footprintdb/index.php?search_entries



The search form looks like this:



Entry type: ● All ○ Transcription Factors ○ DNA Binding Motifs ○ DNA Binding Sites

Search term: MADSBOX

Demo

myeloblastosis (Description)

Organisms:
- All
- Homo sapiens
- Mus musculus
- Arabidopsis thaliana
- Rattus norvegicus

Original Databases:
- All
- JASPAR 2020
- CISBP 1.02
- 3D-footprint 20210203
- HT-SELEX2 May2017

Pfam domains:
- All
- PF13894 - C2H2-type zinc finger
- PF00096 - Zinc finger, C2H2 type
- PF13465 - Zinc-finger double domain
- PF13912 - C2H2-type zinc finger

The search form has the following fields and options:

Entry type: To restrict search to ‘Transcription Factors’, ‘DNA Binding Motifs’ or ‘DNA Binding Sites’

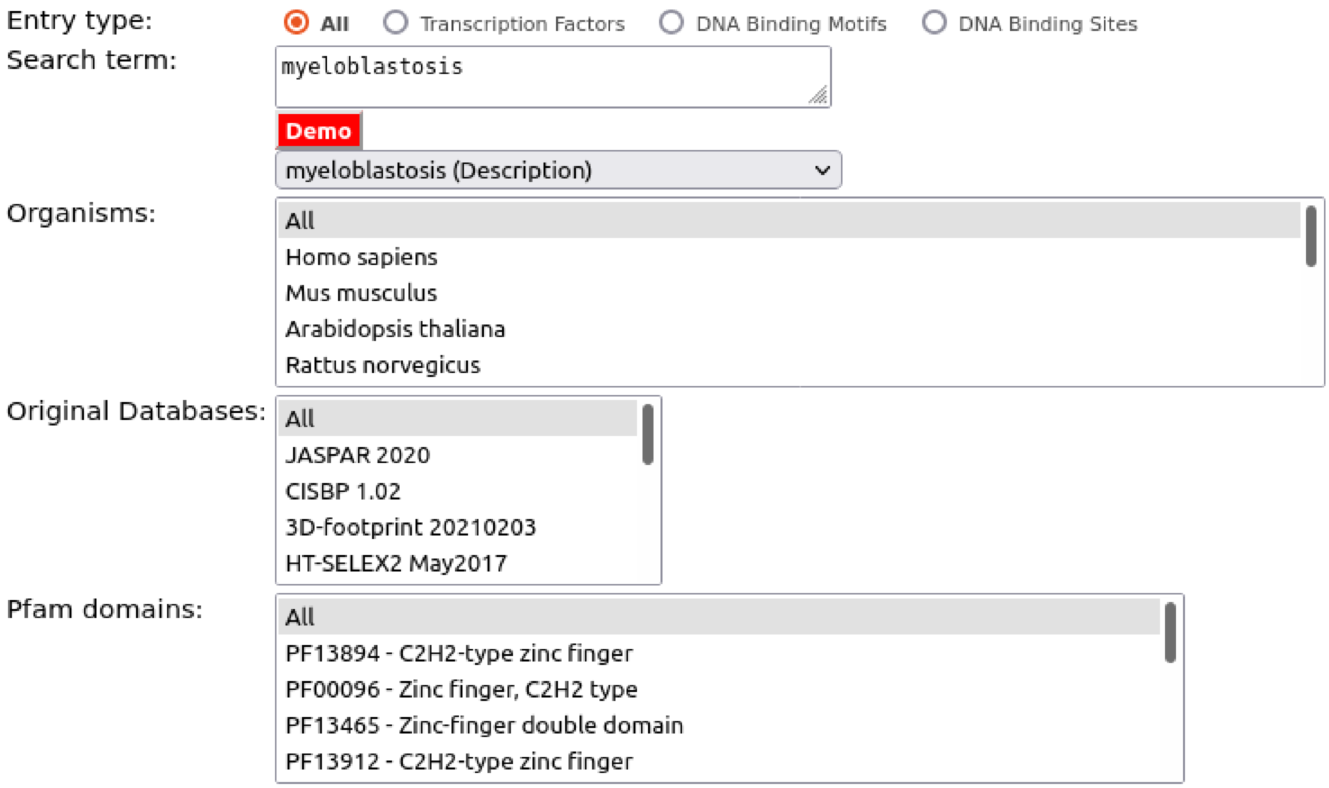
Search term: Text to search, it can be any descriptive word, a transcription factor protein or gene name, UniProt or PDB identifier, original source accession name or DNA site sequence.

Organisms: Select any organism(s) to restrict the search. Multiple species can be selected pushing the Control key on your keyboard. (Use with caution, as many TFs are not associated to an specific organism).

Orignal databases: Select databases or sources to restrict the search. Multiple databases can be selected pushing the Control key.

Pfam domains: Select protein Pfam domains to restrict the search. Multiple domains can be selected pushing the Control key.

To start the search please click the ‘Demo’ button and then the ‘Search’button:



Entry type: ● All ○ Transcription Factors ○ DNA Binding Motifs ○ DNA Binding Sites

Search term: myeloblastosis

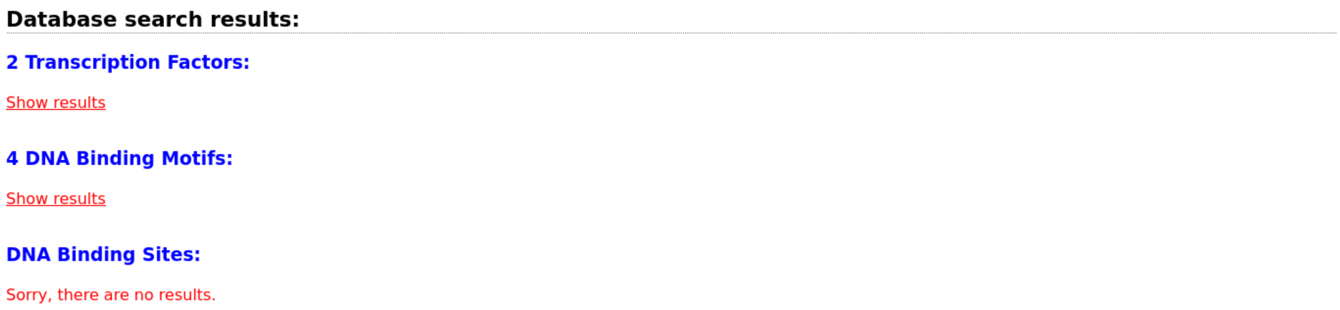
myeloblastosis (Description)

Organisms:
- All
- Homo sapiens
- Mus musculus
- Arabidopsis thaliana
- Rattus norvegicus

Original Databases:
- All
- JASPAR 2020
- CISBP 1.02
- 3D-footprint 20210203
- HT-SELEX2 May2017

Pfam domains:
- All
- PF13894 - C2H2-type zinc finger
- PF00096 - Zinc finger, C2H2 type
- PF13465 - Zinc-finger double domain
- PF13912 - C2H2-type zinc finger

The former search will look for the word ‘myeloblastosis’ in the database, obtaining multiple results that we can expand clicking on ‘Show results’:



Database search results:

2 Transcription Factors:

Show results

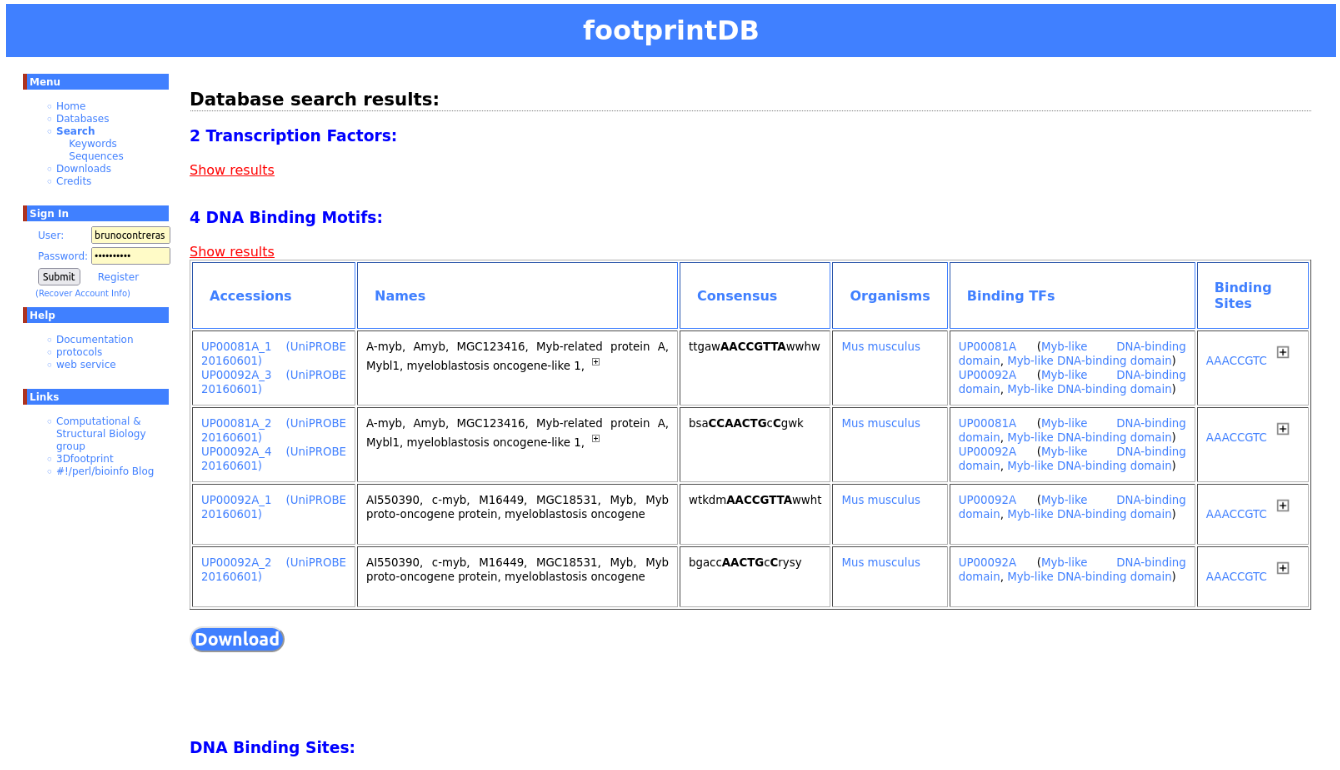
4 DNA Binding Motifs:

Show results

DNA Binding Sites:

Sorry, there are no results.

A full list of the results will be shown, with a short summary of them and the option to access them individually or download them:



Database search results:

2 Transcription Factors:

Show results

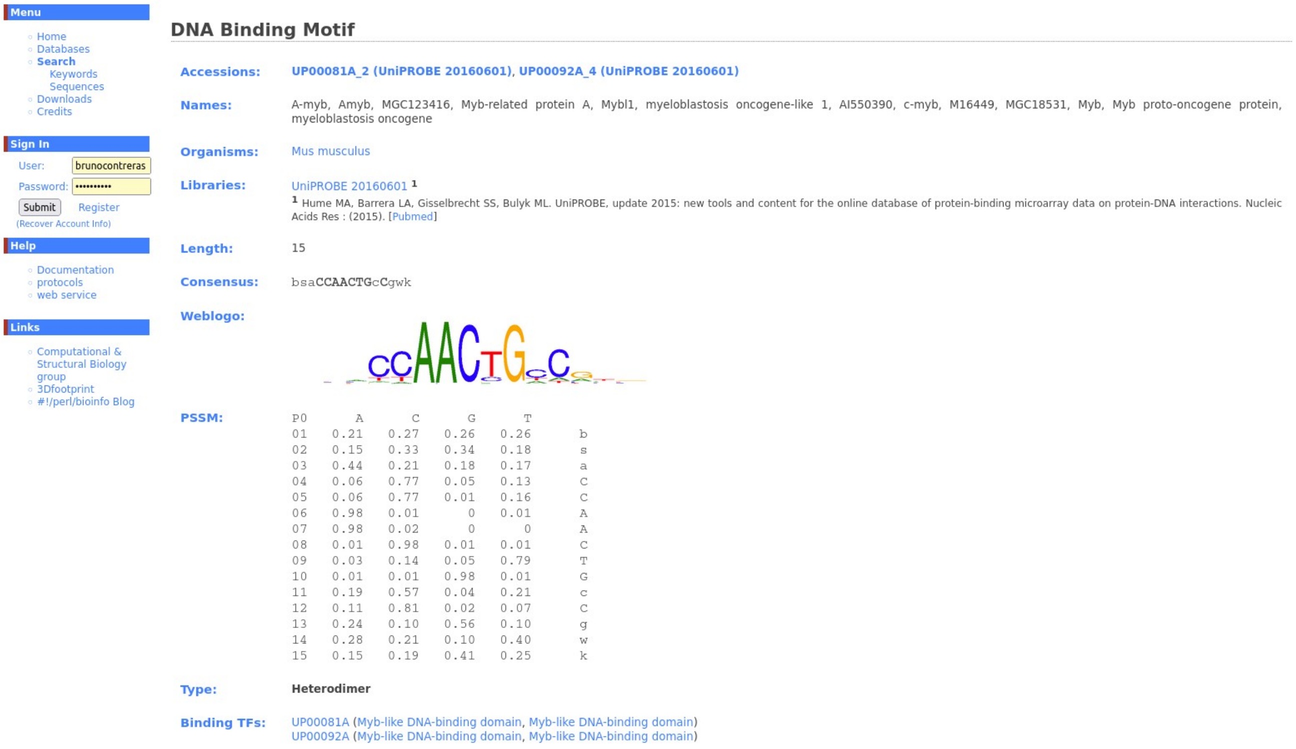
4 DNA Binding Motifs:

Show results

Accessions	Names	Consensus	Organisms	Binding TFs	Binding Sites
UP00081A_1 (UniPROBE 20160601) UP00092A_3 (UniPROBE 20160601)	A-myb, Amyb, MGC123416, Myb-related protein A, Mybl1, myeloblastosis oncogene-like 1, ⊞	ttgaw**AACCGTTA**wwhw	Mus musculus	UP00081A (Myb-like DNA-binding domain, Myb-like DNA-binding domain) UP00092A (Myb-like DNA-binding domain, Myb-like DNA-binding domain)	AAACCGTC ⊞
UP00081A_2 (UniPROBE 20160601) UP00092A_4 (UniPROBE 20160601)	A-myb, Amyb, MGC123416, Myb-related protein A, Mybl1, myeloblastosis oncogene-like 1, ⊞	bsa**CCAACTG**c**C**gwk	Mus musculus	UP00081A (Myb-like DNA-binding domain, Myb-like DNA-binding domain) UP00092A (Myb-like DNA-binding domain, Myb-like DNA-binding domain)	AAACCGTC ⊞
UP00092A_1 (UniPROBE 20160601)	AI550390, c-myb, M16449, MGC18531, Myb, Myb proto-oncogene protein, myeloblastosis oncogene	wtkdm**AACCGTTA**wwht	Mus musculus	UP00092A (Myb-like DNA-binding domain, Myb-like DNA-binding domain)	AAACCGTC ⊞
UP00092A_2 (UniPROBE 20160601)	AI550390, c-myb, M16449, MGC18531, Myb, Myb proto-oncogene protein, myeloblastosis oncogene	bgacc**AACTG**c**C**rysy	Mus musculus	UP00092A (Myb-like DNA-binding domain, Myb-like DNA-binding domain)	AAACCGTC ⊞

DNA Binding Sites:

If we click in the Accession of any of the results we are shown the individual data for it:



DNA Binding Motif

Accessions: **UP00081A_2 (UniPROBE 20160601), UP00092A_4 (UniPROBE 20160601)**

Names: A-myb, Amyb, MGC123416, Myb-related protein A, Mybl1, myeloblastosis oncogene-like 1, AI550390, c-myb, M16449, MGC18531, Myb, Myb proto-oncogene protein, myeloblastosis oncogene

Organisms: Mus musculus

Libraries: UniPROBE 20160601 [1]

[1] Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res : (2015). [Pubmed]

Length: 15

Consensus: bsa**CCAACTG**c**C**gwk

Weblogo:

PSSM:

P0	A	C	G	T	
01	0.21	0.27	0.26	0.26	b
02	0.15	0.33	0.34	0.18	s
03	0.44	0.21	0.18	0.17	a
04	0.06	0.77	0.05	0.13	C
05	0.06	0.77	0.01	0.16	C
06	0.98	0.01	0	0.01	A
07	0.98	0.02	0	0	A
08	0.01	0.98	0.01	0.01	C
09	0.03	0.14	0.05	0.79	T
10	0.01	0.01	0.98	0.01	G
11	0.19	0.57	0.04	0.21	c
12	0.11	0.81	0.02	0.07	C
13	0.24	0.10	0.56	0.10	g
14	0.28	0.21	0.10	0.40	w
15	0.15	0.19	0.41	0.25	k

Type: **Heterodimer**

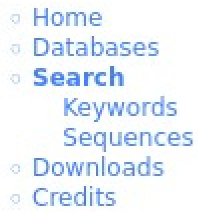
Binding TFs: UP00081A (Myb-like DNA-binding domain, Myb-like DNA-binding domain)
UP00092A (Myb-like DNA-binding domain, Myb-like DNA-binding domain)

2. Search DNA motifs

a. Find transcription factors that bind DNA motifs similar to the query

If you have a footprintDB account, log in first into the website to store your searches and reuse them.

Click on the ‘Start to Search’ button in the Home page, click in the ‘Search sequences’ option of the ‘Main menu’ or go directly to the url: http://floresta.eead.csic.es/footprintdb/index.php?search



The search form has the following fields and options:

Search name: Name a title for the search.

Email: Please type your email if you desire to receive the results by email.

Input type: Please choose ‘DNA sites or motifs’.

Limit number of results per query: Enter the number of desired results per query to speed up.

Order results by: Allows to order results by DNA or TF similarity or by E-value.

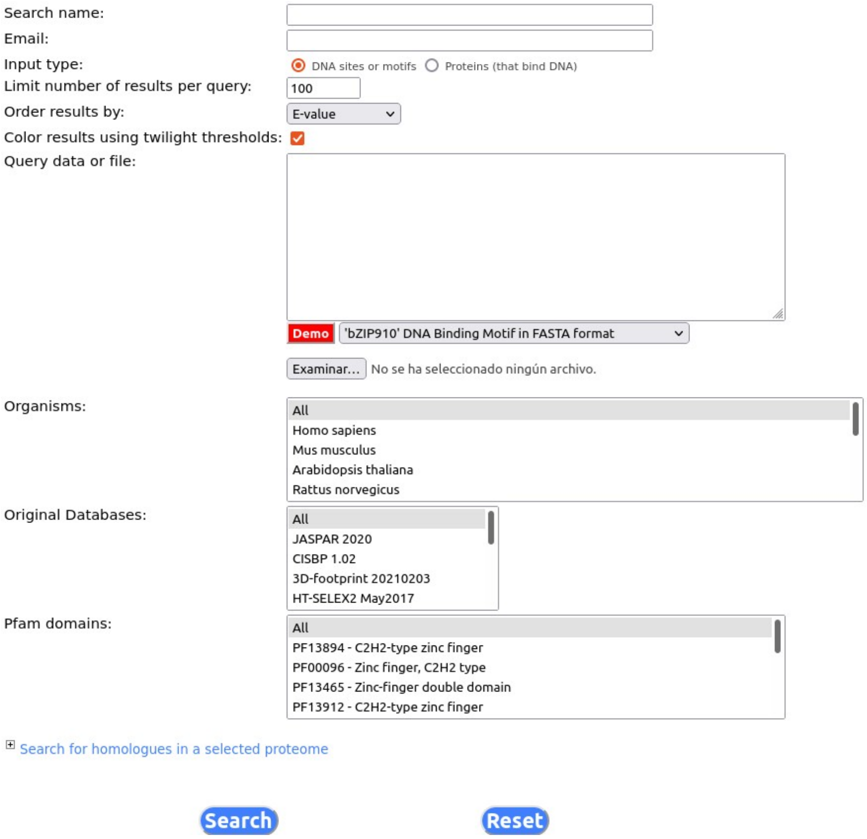
Color results using twilight thresholds: Only available for DNA search, mark in green color results that pass the thresholds defined in our previous article and in red if not (Sebastian and Contreras-Moreira 2013).

Query data or file: Enter your DNA sites or motifs in the text area or upload them from a file. The only valid formats are: FASTA and TRANSFAC. You can also use sample data pushing the ‘Demo’ button.

```
>bZIP910 (JASPAR CORE)
ATGACGT
CTGACGT
ATGACGT
CTGACGT
GTGACGT
GTGACGT
```

```
DE  bZIP transcription factor
P0    A    C    G    T
01    0    0   18    0    G
02   14    0    4    0    A
03    0    0    0   18    T
04    0    0   18    0    G
05   18    0    0    0    A
06    0   18    0    0    C
07    0    0   18    0    G
08    0    0    0   18    T
```

Table 2: Examples of FASTA (left) and TRANSFAC (right) formats for DNA input .



Search name:

Email:

Input type: ● DNA sites or motifs ○ Proteins (that bind DNA)

Limit number of results per query: 100

Order results by: E-value

Color results using twilight thresholds: ☑

Query data or file:

Demo 'bZIP910' DNA Binding Motif in FASTA format

Examinar... No se ha seleccionado ningún archivo.

Organisms:
- All
- Homo sapiens
- Mus musculus
- Arabidopsis thaliana
- Rattus norvegicus

Original Databases:
- All
- JASPAR 2020
- CISBP 1.02
- 3D-footprint 20210203
- HT-SELEX2 May2017

Pfam domains:
- All
- PF13894 - C2H2-type zinc finger
- PF00096 - Zinc finger, C2H2 type
- PF13465 - Zinc-finger double domain
- PF13912 - C2H2-type zinc finger

⊞ Search for homologues in a selected proteome

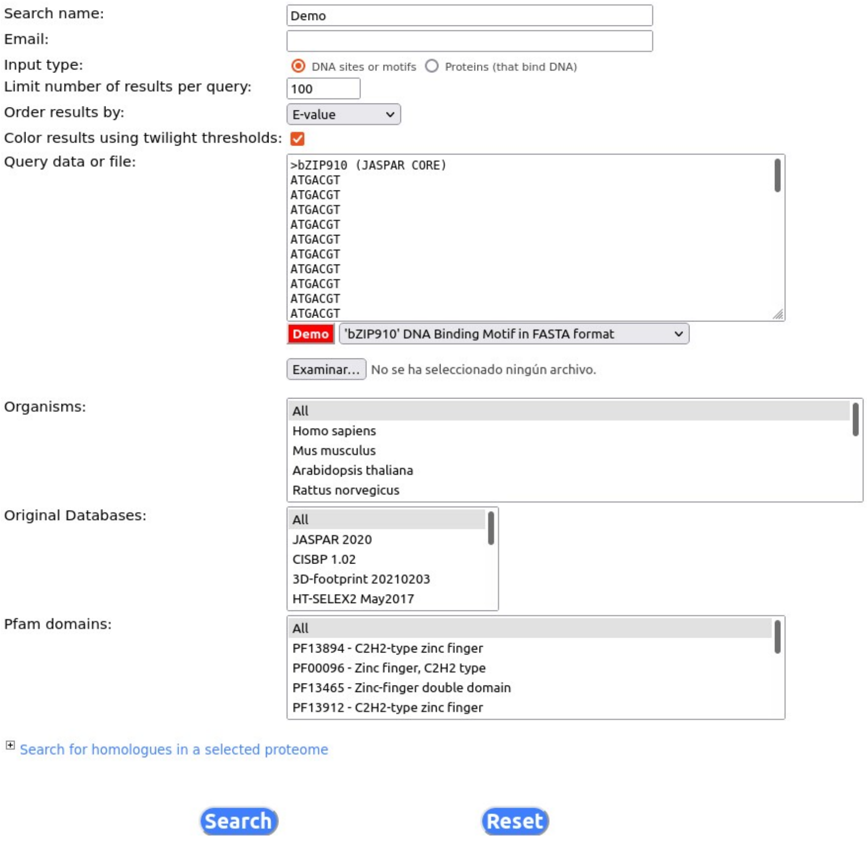
Organisms: Select any organism(s) to restrict the search. Multiple species can be selected pushing the Control key on your keyboard. (Use with caution, as many TFs are not associated to an specific organism).

Databases: Select databases or sources to restrict the search. Multiple databases can be selected pushing the Control key.

Pfam domains: Select protein Pfam domains to restrict the search. Multiple domains can be selected pushing the Control key.

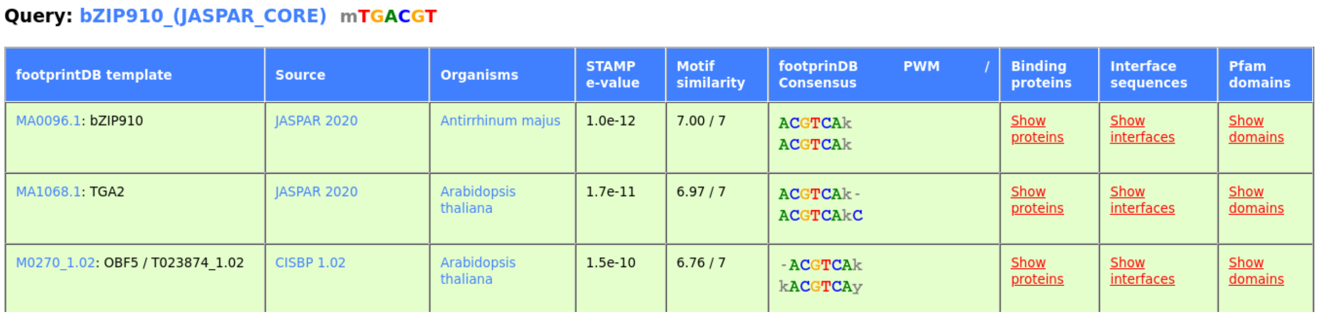
Please not that when search type 'DNA sites or motifs' is selected, the option to automatically 'Search for homologues in a selected proteome' is shown, which will be explained in the next section.

To start the search please click on the 'Demo' button and then on 'Search' :



Search name:	Demo
Email:	
Input type:	● DNA sites or motifs ○ Proteins (that bind DNA)
Limit number of results per query:	100
Order results by:	E-value
Color results using twilight thresholds:	☑
Query data or file:	>bZIP910 (JASPAR CORE) ATGACGT ATGACGT ATGACGT ATGACGT ATGACGT ATGACGT ATGACGT ATGACGT ATGACGT ATGACGT
	Demo 'bZIP910' DNA Binding Motif in FASTA format
	Examinar... No se ha seleccionado ningún archivo.
Organisms:	All Homo sapiens Mus musculus Arabidopsis thaliana Rattus norvegicus
Original Databases:	All JASPAR 2020 CISBP 1.02 3D-footprint 20210203 HT-SELEX2 May2017
Pfam domains:	All PF13894 - C2H2-type zinc finger PF00096 - Zinc finger, C2H2 type PF13465 - Zinc-finger double domain PF13912 - C2H2-type zinc finger

In the former search we take as query a DNA motif in FASTA format (a list of DNA binding sites, all of them with the same length). We want to search at most 10 TFs with similar DNA motifs without filtering organisms neither domains in all available databases. We obtain the following results (only the first 3 are shown):



Query: bZIP910_(JASPAR_CORE) mTGACGT

footprintDB template	Source	Organisms	STAMP e-value	Motif similarity	footprinDB Consensus PWM /	Binding proteins	Interface sequences	Pfam domains
MA0096.1: bZIP910	JASPAR 2020	Antirrhinum majus	1.0e-12	7.00 / 7	ACGTCAk ACGTCAk	Show proteins	Show interfaces	Show domains
MA1068.1: TGA2	JASPAR 2020	Arabidopsis thaliana	1.7e-11	6.97 / 7	ACGTCAk- ACGTCAkC	Show proteins	Show interfaces	Show domains
M0270_1.02: OBF5 / T023874_1.02	CISBP 1.02	Arabidopsis thaliana	1.5e-10	6.76 / 7	-ACGTCAk kACGTCAy	Show proteins	Show interfaces	Show domains

We notice that the first result is the query itself (because Demo query is from JASPAR database, included in footprintDB) and the others are similar DNA motifs present in footprintDB. We can click on the links ‘Show proteins’, ‘Show interfaces’ and ‘Show domains’ to retrieve information about proteins that bind the similar DNA domain retrieved in the search (when there are annotated TFs for the DNA motifs, second result has not related TF).

Predicted DNA binding residues for each protein are shown coloured in the interface sequence. Left-clicking on the ‘footprintDB template’ accession name or on the DNA aligned sequence will display the corresponding footprintDB DNA motif information.

DNA Binding Motif

Accessions: MA0096.1 (JASPAR 2020)

Names: bZIP910

Organisms: Antirrhinum majus

Libraries: JASPAR 2020 [1]
[1] O Fornes, JA Castro-Mondragon, A Khan, R van der Lee, X Zhang, PA Richmond, BP Modi, S Correard, M Gheorghe, D Baranasic, W Santana-Garcia, G Tan, J Ceneby, B Ballester, F Parcy, A Sandelin, B Lenhard, WW Wasserman, A Mathelier. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 48(D1):D87-D92 (2020) [Pubmed]

Notes: SELEX

Length: 7

Consensus: mTGACGT

Weblogo:

PSSM:

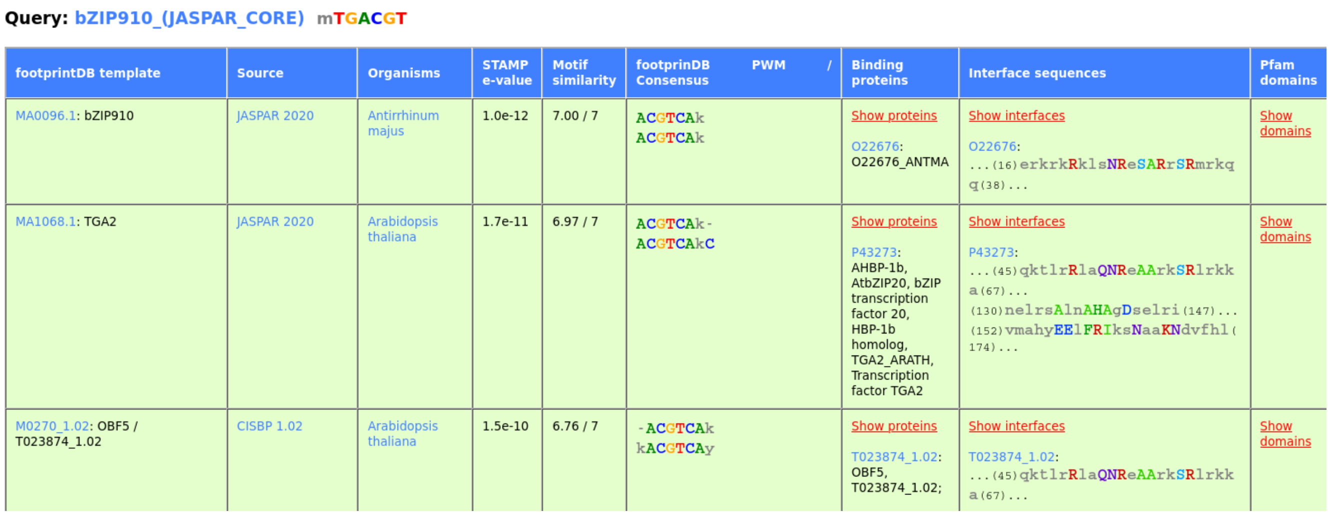
```
P0      A      C      G      T
01     15     15      5      0      m
02      0      0      0     35      T
03      0      0     35      0      G
04     35      0      0      0      A
05      0     35      0      0      C
06      0      0     35      0      G
07      0      0      0     35      T
```

Binding TFs: O22676 (bZIP transcription factor, Basic region leucine zipper)

Binding Sites:
MA0096.1.1 / MA0096.1.10 / MA0096.1.12 / MA0096.1.14 / MA0096.1.3 / MA0096.1.5 / MA0096.1.8
MA0096.1.11
MA0096.1.13 / MA0096.1.2 / MA0096.1.4 / MA0096.1.6 / MA0096.1.7 / MA0096.1.9
MA0096.1.15
MA0096.1.16 / MA0096.1.18
MA0096.1.17
MA0096.1.19
MA0096.1.20

Publications: Martinez-Garcia J. F., Moyano E., Alcocer M. J. C., Martin C. Two bZIP proteins from Antirrhinum flowers preferentially bind a hybrid C-box/G-box motif and help to define a new sub-family of bZIP transcription factors.. Plant J. 13:489-505 (1998). [Pubmed]

In the same way, left-clicking on the TF accession name in ‘Binding proteins’ or ‘Interface sequences’ columns will show the full information page for the TF:



Query: bZIP910_(JASPAR_CORE) mTGACGT

footprintDB template	Source	Organisms	STAMP e-value	Motif similarity	footprinDB Consensus	PWM /	Binding proteins	Interface sequences	Pfam domains
MA0096.1: bZIP910	JASPAR 2020	Antirrhinum majus	1.0e-12	7.00 / 7	ACGTCAk ACGTCAk		Show proteins O22676: O22676_ANTMA	Show interfaces O22676: ...(16)erkrkRklsNReSARrSRmrkqq(38)...	Show domains
MA1068.1: TGA2	JASPAR 2020	Arabidopsis thaliana	1.7e-11	6.97 / 7	ACGTCAk- ACGTCAkC		Show proteins P43273: AHBP-1b, AtbZIP20, bZIP transcription factor 20, HBP-1b homolog, TGA2_ARATH, Transcription factor TGA2	Show interfaces P43273: ...(45)qktlrRlaQNReAArkSRlrkka(67)... (130)nelrsAlnAHAgDselri(147)... (152)vmahyEElFRIksNaaKNdvfhl(174)...	Show domains
M0270_1.02: OBF5 / T023874_1.02	CISBP 1.02	Arabidopsis thaliana	1.5e-10	6.76 / 7	-ACGTCAk kACGTCAy		Show proteins T023874_1.02: OBF5, T023874_1.02;	Show interfaces T023874_1.02: ...(45)qktlrRlaQNReAArkSRlrkka(67)...	Show domains

Transcription Factor

Accessions: **bZIP910 (Athamap 20091028), T028334_1.02 (CISBP 1.02), O22676 (JASPAR 2020)**

Names: bZIP910, O22676_ANTMA, T028334_1.02;

Organisms: Antirrhinum majus

Libraries: Athamap 20091028 [1], CISBP 1.02 [2], JASPAR 2020 [3]

[1] BÃ¼low L, Engelmann S, Schindler M, Hehl R. AthaMap, integrating transcriptional and post-transcriptional data. Nucleic acids research 37:D983-6 (2009). [Pubmed]

[2] Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014 Sep 11;158(6):1431-43. doi: 10.1016/j.cell.2014.08.009. [Pubmed]

[3] O Fornes, JA Castro-Mondragon, A Khan, R van der Lee, X Zhang, PA Richmond, BP Modi, S Correard, M Gheorghe, D Baranasic, W Santana-Garcia, G Tan, J Ceneby, B Ballester, F Parcy, A Sandelin, B Lenhard, WW Wasserman, A Mathelier. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 48(D1):D87-D92 (2020) [Pubmed]

Notes: family:bZIP

Length: 133

Pfam Domains:
13 - 61 Basic region leucine zipper
17 - 75 bZIP transcription factor

Sequence: (in bold interface residues)

```
  1 MASQQRSTSPGIDDDERKRKRKLSNRESARRSRMRKQQRLDELIAQESQMQEDNKKLRDT   60
 61 INGATQLYLNFASDNNVLRAQLAELTDRLHSLNSVLQIASEVSGLVLDIPDIPDALLEPW  120
121 QLPCPIQADIFQC
```

Interface Residues: 25, 26, 28, 29, 32, 33

3D-footprint Homologues: 1dh3_C, 2dgc_A, 5t01_B

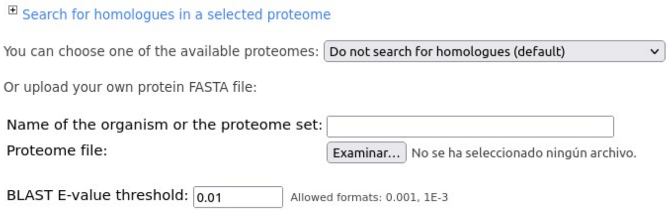
Binding Motifs:
bZIP910(1) GATGACGTGGCm
bZIP910(2) GgrTGCTGACGT
M2679_1.02 GATGACGTGGCm
MA0096.1 mTGACGT

Binding Sites:
MA0096.1.1 / MA0096.1.10 / MA0096.1.12 / MA0096.1.14 / MA0096.1.3 / MA0096.1.5 / MA0096.1.8
MA0096.1.11
MA0096.1.13 / MA0096.1.2 / MA0096.1.4 / MA0096.1.6 / MA0096.1.7 / MA0096.1.9

Other data shown are: the source database, organisms, Pfam domains, the set of interface residues - which are the key residues mediating specific DNA recognition-, STAMP E-value and DNA motif similarity score (sum of the Pearson correlation coefficients of the aligned DNA motif positions).

b. Find in a selected proteome homologous transcription factors that bind DNA motifs similar to the query

Please follow the steps explained in the former section ‘Find transcription factors that bind DNA motifs similar to the query’ until you see the search formulary. The menu ‘Search for homologues in a selected proteome’ will be available at the bottom of the page. Click on the title ‘Search for homologues in a selected proteome’ to expand the homologue search options:



⊞ Search for homologues in a selected proteome

You can choose one of the available proteomes: Do not search for homologues (default)

Or upload your own protein FASTA file:

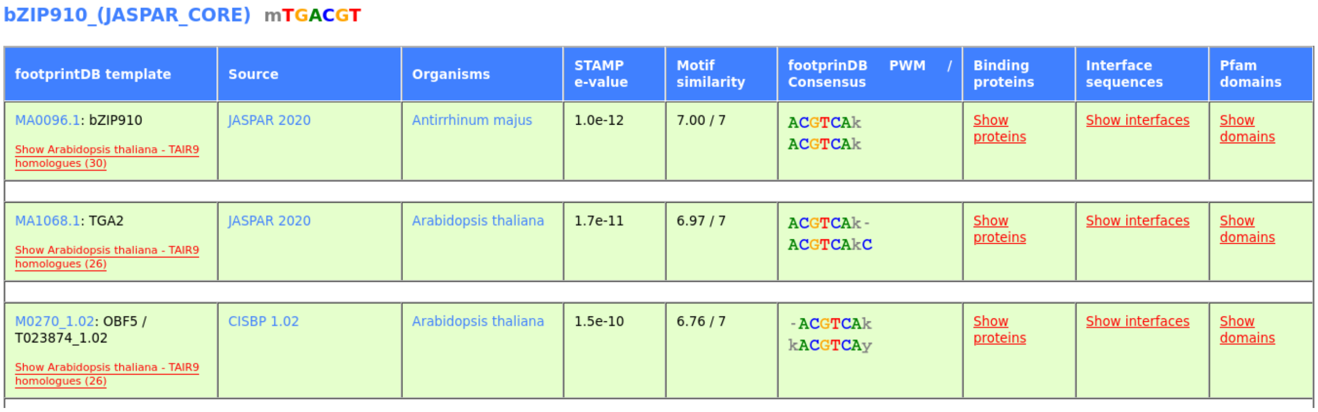
Name of the organism or the proteome set:

Proteome file: No se ha seleccionado ningún archivo.

BLAST E-value threshold: 0.01

Now you might select a species to search for homologues in its proteome or either upload a custom file with a proteome file in FASTA format and choose a BLAST E-value threshold for the Blastp search against the proteome (Default 0.01).

Search parameters are the same of the previous example, but in this case we choose to include among the results the subset of ‘Arabidopsis thaliana - TAIR 9’ proteins which are presumably homologous to each of the reported DNA-binding proteins. Indeed we obtain the same previous results but in a slightly different order, with proteins with a higher number of homologues shown first (only the first 3 are discussed for brevity):



bZIP910_(JASPAR_CORE) mTGACGT

footprintDB template	Source	Organisms	STAMP e-value	Motif similarity	footprinDB PWM / Consensus	Binding proteins	Interface sequences	Pfam domains
MA0096.1: bZIP910 Show Arabidopsis thaliana - TAIR9 homologues (30)	JASPAR 2020	Antirrhinum majus	1.0e-12	7.00 / 7	ACGTCAk ACGTCAk	Show proteins	Show interfaces	Show domains
MA1068.1: TGA2 Show Arabidopsis thaliana - TAIR9 homologues (26)	JASPAR 2020	Arabidopsis thaliana	1.7e-11	6.97 / 7	ACGTCAk- ACGTCAkC	Show proteins	Show interfaces	Show domains
M0270_1.02: OBF5 / T023874_1.02 Show Arabidopsis thaliana - TAIR9 homologues (26)	CISBP 1.02	Arabidopsis thaliana	1.5e-10	6.76 / 7	-ACGTCAk kACGTCAy	Show proteins	Show interfaces	Show domains

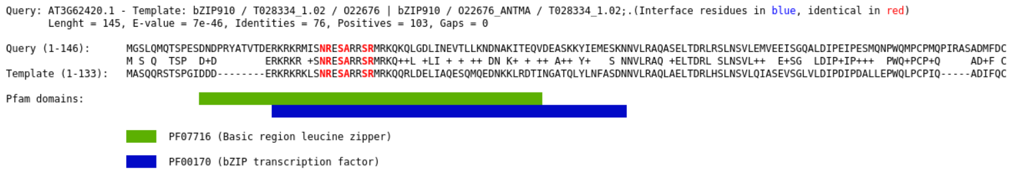
Each row now contains a TF that recognizes a motif similar to the query, and the motif alignment is also shown, as explained in the previous section. However, the provided red link ‘Show Arabidopsis thaliana – TAIR9 homologues’ allows us to display a list of proteins, just beneath the name of each matched footprintDB name:

bZIP910_(JASPAR_CORE) mTGACGT

footprintDB template	Source	Organisms	STAMP e-value	Motif similarity	footprinDB PWM / Consensus	Binding proteins	Interface sequences	Pfam domains
MA0096.1: bZIP910 Show Arabidopsis thaliana - TAIR9 homologues (30)	JASPAR 2020	Antirrhinum majus	1.0e-12	7.00 / 7	ACGTCAk ACGTCAk	Show proteins	Show interfaces	Show domains

Arabidopsis thaliana - TAIR9 protein / gene / interface	footprintDB template	Blast E-value	Interface similarity	Template alignment	Interface sequences	Related results
Protein: AT3G62420 (sequence) **Interface**: krmisNReSArrSRmrkqk	bZIP910 / T028334_1.02 / O22676	7e-46	6 / 6	1-146	krklsNReSArrSRmrkqq	bZIP910_(JASPAR_CORE) (MA1407.1)
Protein: AT1G75390 (sequence) **Interface**: krkqsNReSArrSRmrkqk	bZIP910 / T028334_1.02 / O22676 bZIP910 / T028334_1.02 / O22676	4e-25 3e-17	6 / 6 6 / 6	39-173 39-105	krklsNReSArrSRmrkqq krklsNReSArrSRmrkqq	bZIP910_(JASPAR_CORE) (MA1407.1)
Protein: AT2G18160 (sequence) **Interface**: krmlsNReSArrSRmrkqk	bZIP910 / T028334_1.02 / O22676	1e-23	6 / 6	29-123	krklsNReSArrSRmrkqq	bZIP910_(JASPAR_CORE) (MA1407.1)
Protein: AT4G34590 (sequence) **Interface**: krmlsNReSArrSRmkkqk	bZIP910 / T028334_1.02 / O22676	1e-21	6 / 6	25-108	krklsNReSArrSRmrkqq	bZIP910_(JASPAR_CORE) (MA1407.1)
Protein: AT1G68880 (sequence) **Interface**: rrkvsNReSArrSRmrkqr	bZIP910 / T028334_1.02 / O22676	2e-12	6 / 6	38-111	krklsNReSArrSRmrkqq	bZIP910_(JASPAR_CORE) (MA1335.1 / M0210)
Protein: AT3G30530 (sequence) **Interface**: rrmisNReSArrSRmrkqr	bZIP910 / T028334_1.02 / O22676	2e-12	6 / 6	79-169	krklsNReSArrSRmrkqq	bZIP910_(JASPAR_CORE) (MA1407.1)
Protein: AT5G49450 (sequence) **Interface**: krklsNReSArrSRlkkqk	bZIP910 / T028334_1.02 / O22676	9e-12	6 / 6	1-134	krklsNReSArrSRmrkqq	-

Homologous proteins will be shown under each TF. Each new row contains data of one protein; left-clicking on the protein name will open a new window with protein sequence in FASTA format. Left-clicking on the 'Blast E-value', 'Interface similarity' or 'Template alignment' columns will show the Blast alignment with the corresponding footprintDB protein sequence with coloured protein domains (Pfam version 24.0) highlighting in red the identical interface residues and in blue the rest of the interface. The last column 'Related results' shows other footprintDB TF results which are presumably homologous to the same *Arabidopsis thaliana* protein.



```
Query: AT3G62420.1 - Template: bZIP910 / T028334_1.02 / O22676 | bZIP910 / O22676_ANTMA / T028334_1.02;.(Interface residues in blue, identical in red)
       Lenght = 145, E-value = 7e-46, Identities = 76, Positives = 103, Gaps = 0

Query (1-146):      MGSLQMQTSPESDNDPRYATVTDERKRKRMISNRESARRSRMRKQKQLGDLINEVTLLKNDNAKITEQVDEASKKYIEMESKNNVLRAQASELTDRLRSLNSVLEMVEEISGQALDIPEIPESMQNPWQMPCPMQPIRASADMFDC
                    M S Q  TSP  D+D        ERKRKR +SNRESARRSRMRKQ++L +LI + + ++ DN K+ + ++ A++ Y+   S NNVLRAQ +ELTDRL SLNSVL++  E+SG  LDIP+IP+++  PWQ+PCP+Q     AD+F C
Template (1-133):   MASQQRSTSPGIDDD--------ERKRKRKLSNRESARRSRMRKQQRLDELIAQESQMQEDNKKLRDTINGATQLYLNFASDNNVLRAQLAELTDRLHSLNSVLQIASEVSGLVLDIPDIPDALLEPWQLPCPIQ-----ADIFQC

Pfam domains:

                    🟩 PF07716 (Basic region leucine zipper)

                    🟦 PF00170 (bZIP transcription factor)
```

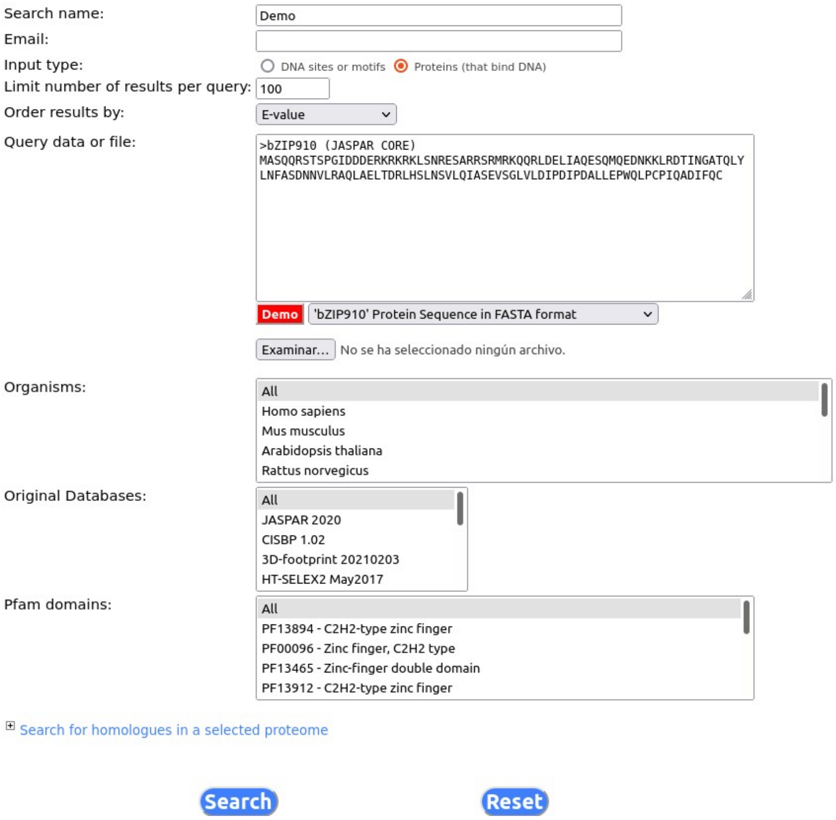
3. Search protein sequences

a. Find transcription factors with similar sequences

Click on the ‘Search Sequences’ button in the Home page, click in the ‘Search’ option of the ‘Main menu’ or go directly to the url: http://floresta.eead.csic.es/footprintdb/index.php?search

After clicking on ‘**Proteins (that bind DNA)**’ you’ll see a search form with fields and options as explained in the previous Section. In this case there is only a noticeable difference with respect to the input format of the sequence to search.

Query data or file: Enter your DNA sites or motifs in the text area or upload them from a file. The only valid format is FASTA . You can use sample data pushing the ‘Demo’ button.



Search name: Demo

Email:

Input type: ○ DNA sites or motifs ● Proteins (that bind DNA)

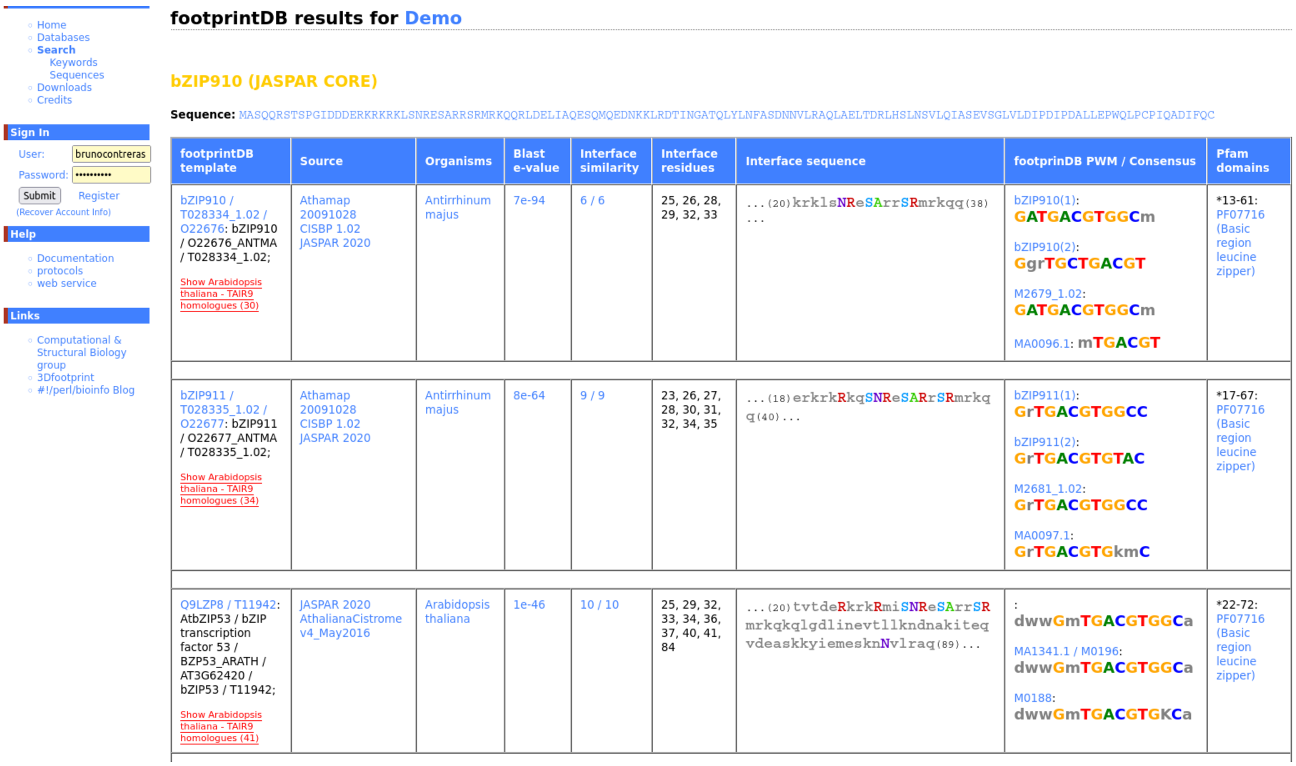
Limit number of results per query: 100

Order results by: E-value

Query data or file:

```
>bZIP910 (JASPAR CORE)
MASQQRSTSPGIDDDERKRKRKLSNRESARRSRMRKQQRLDELIAQESQMQEDNKKLRDTINGATQLY
LNFASDNNVLRAQLAELTDRLHSLNSVLQIASEVSGLVLDIPDIPDALLEPWQLPCPIQADIFQC
```

In this search we query a protein sequence in FASTA format. In particular, we wish to search no more than 10 TFs with similar sequence and their associated DNA binding motifs without filtering organisms nor domains in all available databases. We obtain the following results (only the first 3 are shown):



footprintDB results for Demo

bZIP910 (JASPAR CORE)

Sequence: MASQQRSTSPGIDDDERKRKRKLSNRESARRSRMRKQQRLDELIAQESQMQEDNKKLRDTINGATQLYLNFASDNNVLRAQLAELTDRLHSLNSVLQIASEVSGLVLDIPDIPDALLEPWQLPCPIQADIFQC

footprintDB template	Source	Organisms	Blast e-value	Interface similarity	Interface residues	Interface sequence	footprinDB PWM / Consensus	Pfam domains
bZIP910 / T028334_1.02 / O22676: bZIP910 / O22676_ANTMA / T028334_1.02; Show Arabidopsis thaliana - TAIR9 homologues (30)	Athamap 20091028 CISBP 1.02 JASPAR 2020	Antirrhinum majus	7e-94	6 / 6	25, 26, 28, 29, 32, 33	...(20)krklsNReSArrSRmrkqq(38)...	bZIP910(1): GATGACGTGGCm bZIP910(2): GgrTGCTGACGT M2679_1.02: GATGACGTGGCm MA0096.1: mTGACGT	*13-61: PF07716 (Basic region leucine zipper)
bZIP911 / T028335_1.02 / O22677: bZIP911 / O22677_ANTMA / T028335_1.02; Show Arabidopsis thaliana - TAIR9 homologues (34)	Athamap 20091028 CISBP 1.02 JASPAR 2020	Antirrhinum majus	8e-64	9 / 9	23, 26, 27, 28, 30, 31, 32, 34, 35	...(18)erkrkRkqSNReSARrSRmrkqq(40)...	bZIP911(1): GrTGACGTGGCC bZIP911(2): GrTGACGTGTAC M2681_1.02: GrTGACGTGGCC MA0097.1: GrTGACGTGkmC	*17-67: PF07716 (Basic region leucine zipper)
Q9LZP8 / T11942: AtbZIP53 / bZIP transcription factor 53 / BZP53_ARATH / AT3G62420 / bZIP53 / T11942; Show Arabidopsis thaliana - TAIR9 homologues (41)	JASPAR 2020 AthalianaCistrome v4_May2016	Arabidopsis thaliana	1e-46	10 / 10	25, 29, 32, 33, 34, 36, 37, 40, 41, 84	...(20)tvtdeRkrkRmiSNReSArrSRmrkqkqlgdlinevtllkndnakiteqvdeaskkyiemesknNvlraq(89)...	: dwwGmTGACGTGGCa MA1341.1 / M0196: dwwGmTGACGTGGCa M0188: dwwGmTGACGTGKCa	*22-72: PF07716 (Basic region leucine zipper)

As earlier, we can notice that the first result is the query itself (query is from JASPAR collection and is present in footprintDB) and the other results are transcription factors with similar interface sequences (results are ordered by Blastp E-value) and they have annotated also similar DNA motifs.

Each row contains a TF with similar sequence to the query. Predicted DNA binding residues are shown coloured in the interface sequence and all the DNA motifs annotated for that TF are shown. Left-clicking on the Blast E-value or the interface similarity score will show the alignment of the footprintDB TF sequence with the query. Left-clicking on the ‘footprintDB template’ TF accession name will display the full information about the TF:

User: brunocontreras

Password: ••••••••••

Transcription Factor

Accessions: **bZIP910 (Athamap 20091028), T028334_1.02 (CISBP 1.02), O22676 (JASPAR 2020)**

Names: bZIP910, O22676_ANTMA, T028334_1.02;

Organisms: Antirrhinum majus

Libraries: Athamap 20091028 [1], CISBP 1.02 [2], JASPAR 2020 [3]

[1] BÃ¼low L, Engelmann S, Schindler M, Hehl R. AthaMap, integrating transcriptional and post-transcriptional data. Nucleic acids research 37:D983-6 (2009). [Pubmed]

[2] Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014 Sep 11;158(6):1431-43. doi: 10.1016/j.cell.2014.08.009. [Pubmed]

[3] O Fornes, JA Castro-Mondragon, A Khan, R van der Lee, X Zhang, PA Richmond, BP Modi, S Correard, M Gheorghe, D Baranasic, W Santana-Garcia, G Tan, J Ceneby, B Ballester, F Parcy, A Sandelin, B Lenhard, WW Wasserman, A Mathelier. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 48(D1):D87-D92 (2020) [Pubmed]

Notes: family:bZIP

Length: 133

Pfam Domains:

13 - 61 Basic region leucine zipper
17 - 75 bZIP transcription factor

Sequence:
(in bold interface residues)

```
  1 MASQQRSTSPGIDDDERKRKRKLSNRESARRSRMRKQQRLDELIAQESQMQEDNKKLRDT   60
 61 INGATQLYLNFASDNNVLRAQLAELTDRLHSLNSVLQIASEVSGLVLDIPDIPDALLEPW  120
121 QLPCPIQADIFQC
```

Interface Residues: 25, 26, 28, 29, 32, 33

3D-footprint Homologues: 1dh3_C, 2dgc_A, 5t01_B

Binding Motifs:

bZIP910(1) **GATGACGTGGC**m
bZIP910(2) **G**gr**TGCTGACGT**
M2679_1.02 **GATGACGTGGC**m
MA0096.1 m**TGACGT**

Binding Sites:

MA0096.1.1 / MA0096.1.10 / MA0096.1.12 / MA0096.1.14 / MA0096.1.3 / MA0096.1.5 / MA0096.1.8
MA0096.1.11
MA0096.1.13 / MA0096.1.2 / MA0096.1.4 / MA0096.1.6 / MA0096.1.7 / MA0096.1.9
MA0096.1.15
MA0096.1.16 / MA0096.1.18
MA0096.1.17
MA0096.1.19

Similarly, left-clicking on DNA binding motif ‘footprintDB PWM / Consensus’ accession name will show the DNA binding motif information.

Other data shown are: the source organism(s), Pfam domains, the set of interface residues -which are the key residues mediating specific DNA recognition-, Blastp E-value and interface similarity score.

b. Find in a selected proteome homologous transcription factors that bind TF sequences similar to the query

Please follow the steps explained in the former section ‘Find transcription factors with similar sequences’ until you see the search formulary. The menu ‘Search for homologues in a selected proteome’ will be available at the bottom of the page. Then follow the same procedure explained in the previous section. Homologous protein sequences from the selected genome will be shown and they can be accessed as previously explained.

Bibliography

Main citation:

Sebastian A, Contreras-Moreira B (2014) footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics* 30, 258-65. https://doi.org/10.1093/bioinformatics/btt663

Contreras-Moreira B (2010) 3D-footprint: a database for the structural analysis of protein-DNA complexes. Nucleic Acids Res 38(Database issue): D91-97. https://doi.org/10.1093/nar/gkp781

Sebastian A, Contreras-Moreira B (2013). The twilight zone of cis element alignments. Nucleic Acids Res 41(3): 1438-1449. https://doi.org/10.1093/nar/gks1301

Dubos C, Kelemen Z, Sebastian A *et al.* (2014) Integrating bioinformatic resources to predict transcription factors interacting with *cis*-sequences conserved in co-regulated genes. BMC Genomics 15**,** 317. https://doi.org/10.1186/1471-2164-15-317

Contreras-Moreira B, Sebastian A. FootprintDB: Analysis of Plant Cis-Regulatory Elements, Transcription Factors, and Binding Interfaces. In: Hehl R. (eds) Plant Synthetic Promoters. *Methods in Molecular Biology*, vol 1482. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-6396-6_17 30, 258-65 (2016).

Please check also https://floresta.eead.csic.es/footprintdb/index.php?databases to get the updated citations of the current databases in footprintDB.