Databases and ontologies

footprintDB: a database of transcription factors with annotated cis elements and binding interfaces

Alvaro Sebastian[1,*] and Bruno Contreras-Moreira[1,2*]

[1]Laboratory of Computational Biology, Estación Experimental de Aula Dei/CSIC, Av. Montañana 1005, Zaragoza, Spain (http://www.eead.csic.es/compbio)

[2]Fundación ARAID, Paseo María Agustín 36, Zaragoza, Spain

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Traditional and high throughput techniques for determining transcription factor binding specificities are generating large volumes of data of uneven quality, which are scattered across individual databases.

Results: FootprintDB integrates some of the most comprehensive freely available libraries of curated DNA binding sites (DBSs), and systematically annotates the binding interfaces of the corresponding transcription factors (TFs). The first release contains 2422 unique TF sequences, 10112 DBSs and 3662 DNA motifs. A survey of the included data sources, organisms and TF families was performed together with proprietary database TRANSFAC, finding that footprintDB has a similar coverage of multicellular organisms, while also containing bacterial regulatory data. A search engine has been designed that drives the prediction of DNA motifs for input TFs, or conversely of TF sequences that might recognize input regulatory sequences, by comparison with database entries. Such predictions can also be extended to a single proteome chosen by the user, and results are ranked in terms of interface similarity. Benchmark experiments with bacterial, plant and human data were performed to measure the predictive power of footprintDB searches, which were able to correctly recover 10%, 55% and 90% of the tested sequences, respectively. Correctly predicted TFs had a higher interface similarity than the average, confirming its diagnostic value.

Availability: Website implemented in PHP, Perl, MySQL and Apache. Freely available from http://floresta.eead.csic.es/footprintdb

Contact: bioquimicas@yahoo.es or bcontreras@eead.csic.es

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Transcription is one of the most important processes in gene expression, and it is modulated primarily by the binding of regulatory proteins called transcription factors (TFs) to short DNA sequences, called *cis* elements or DNA binding sites (DBSs). The DBS recognition mechanism is generally degenerate, as one TF can usually bind to a collection of similar but different cis elements, which can be grouped together to define a DNA motif. Motifs are most frequently represented as position-specific scoring matrices (PSSM) that capture the occurrence of nucleotides in aligned positions of the underlying DBSs (Stormo, 2000). Furthermore, motifs are frequently plotted as sequence logos, which graphically summarize the binding specificities and/or affinities of TFs (Schneider and Stephens, 1990) (see Supplementary Fig. S1).

Traditional experimental methods to identify DBSs are technically challenging and have been frequently limited to determining *cis*-regulatory sites for one TF at a time. These methods, such as DNA footprinting or electrophoretic mobility shift assays (Galas and Schmitz, 1978; Garner and Revzin, 1981; O'Neill and Turner, 1996), yield high quality data and have been the primary source of data for expert-curated databases such as RegulonDB (Salgado, et al., 2013). However, these approaches do not scale well and are currently being replaced by protocols that allow high throughput discovery of DBSs, such as protein binding microarrays, ChIP-chip or ChIP-Seq experiments (Berger and Bulyk, 2006; Johnson, et al., 2007; Ren, et al., 2000). These procedures produce large volumes of raw sequence data, which must be preprocessed and filtered in order to derive DNA motifs. Databases such as JASPAR (Portales-Casamar, et al., 2010) and TRANSFAC (Matys, et al., 2006) are increasingly annotating DBSs produced by such protocols, fueled by papers that report experimentally-derived DBSs and motifs for large repertoires of TFs (Down, et al., 2007; Jolma, et al., 2013; Noyes, et al., 2008)

On other side there are experimental approaches for characterizing the interface residues of TFs, those in charge of recognizing the nucleotide bases of target *cis* elements. Beyond site-directed mutagenesis (O'Neill, et al., 1998; Shortle, et al., 1981), the most accurate methods are X-ray crystallography and NMR studies of protein-DNA complexes. The resulting structures are maintained and published at the Protein Data Bank (PDB) (Berman, et al., 2000). By further digesting these structural models, the 3D-fooprint database routinely annotates the interfaces of all DNA-binding proteins contained therein, following simple geometrical criteria: interface residues must form hydrogen bonds

*To whom correspondence should be addressed.

or hydrophobic contacts with nitrogen bases or else locate heavy atoms within 4.5 Å of any nitrogen base (Contreras-Moreira, 2010).

Here we present footprintDB, a meta-database which integrates the most comprehensive freely available libraries of curated DBSs and systematically annotates, for the first time, the binding interfaces of the corresponding TFs. Furthermore, we survey the redundancy of all included databases and compare them to TRANSFAC, a subscription-based, commercial alternative. Besides allowing users to compare DNA sequences/motifs to records in the database, as most included repositories do, footprintDB can also interrogate complete proteomes in order to identify which TFs are likely to recognize input *cis* elements. Annotated interfaces are particularly valuable for the second type of query, as our benchmarks indicate that TFs with similar interface residues are more likely to bind to similar DBSs. The three unique features of footprintDB are: i) the possibility to search against multiple curated databases at the same time or to add custom databases; ii) the annotation of interface residues within DNA-binding protein domains; and iii) the support for browsing TFs within user-provided proteomes which are most likely to bind a DBS of interest. This resource is available at http://floresta.eead.csic.es/footprintdb.

2 METHODS

2.1 Data sources

FootprintDB is by design a meta-database of TFs attached to their experimentally determined DNA binding preferences (PSSMs and/or DBSs). Therefore it does not incorporate other databases which contain only TF, DBS or predicted regulatory sequences. The first building block is 3D-footprint (Contreras-Moreira, 2010), a database for the structural analysis of protein-DNA complexes, for two reasons: i) it is to our knowledge the only up-to-date source of annotated binding interfaces of TFs; and ii) it contains structure-based PSSMs, motifs inferred from *cis* elements captured in X-ray and NMR complexes, that have been independently validated (AlQuraishi and McAdams, 2011; Lin and Chen, 2013). The remaining databases and repositories currently integrated in footprintDB are:

(1) JASPAR CORE (2009 version, all species redundant set): a high-quality collection of transcription factor DNA-binding preferences, modeled as PSSMs (Portales-Casamar, et al., 2010).

(2) UniPROBE (Universal PBM Resource for Oligonucleotide Binding Evaluation, Sep 2012 version): contains in vitro DNA binding specificities of proteins measured with universal protein binding microarrays (Robasky and Bulyk, 2011).

(3) "HumanTF": sequence-specific binding preferences of human TFs obtained by high-throughput SELEX and ChIP sequencing. It includes a total of 830 binding profiles, describing 239 distinctly different binding specificities (Jolma, et al., 2013).

(4) Athamap: genome-wide map of potential transcription factor binding sites (TFBS) in Arabidopsis thaliana (Bulow, et al., 2009).

(5) RegulonDB (7.5 version): contains curated data of the transcriptional regulatory network of Escherichia coli K12, including PSSMs and DBSs for many TFs (Salgado, et al., 2013).

(6) DBTBS (Database of transcriptional regulation in Bacillus subtilis): A database of transcriptional regulation in Bacillus subtilis (Sierro, et al., 2008).

(7) "DrosophilaTF": Motifs for 56 Drosophila melanogaster transcription factors built from in vitro binding site selection experiments and compiled genomic binding site sequences (Down, et al., 2007).

In addition to these freely available data sources, we also tested TRANSFAC (2012.1 version), a subscription database with transcription factors, their experimentally-proven binding sites and the corresponding PSSMs (Matys, et al., 2006). TRANSFAC is a popular resource in this community and was thus included in our benchmarks. All these data sets were retrieved, curated, completed when necessary (for instance by searching for TF sequences for GenBank/Uniprot identifiers) and imported into our meta-database using custom Perl scripts. To standardize these tasks we created the footprintDB data format, which bundles together TF and DBS sequences, motifs and links to supporting literature and original sources. This format is a blending of 'matrix.dat', 'factor.dat' and 'site.dat' TRANSFAC files in a single file, as shown in Supplementary Fig. S2. By adopting these formats, a friendly web interface allows users to update and insert data for their own private applications, or rather make them available to the community.

Along the paper we will refer to footprintDB as the sum of the formerly listed data sources, excluding TRANSFAC.

2.2 Database structure and web application

Different aspects were considered when conceiving the database, some of them biological and some relevant for data modeling. Transcription control is a very complex mechanism, still not completely understood, and this must be considered in the design. For instance, a single TF can bind to several possibly degenerate DBSs within the same or different genomic regions, or often the same *cis* element is recognized by several TFs. Other relevant questions are: redundancy among sources, miscellaneous annotation formats, incomplete annotation of entries and availability of data retrieval systems. All these factors made footprintDB have a complex relational schema, shown in Supplementary Fig. S3. The web application is written in PHP and JavaScript, with Perl scripts running the queries. Sequence logos are built with Weblogo (Crooks, et al., 2004). The database runs a MySQL engine on an Apache server. A SOAP Web services interface is available at http://floresta.eead.csic.es/footprintdb/ws.cgi. The online documentation includes examples on how to query it.

2.3 Annotation of transcription factor interfaces and Pfam domains

TF sequences in footprintDB have their DNA-binding interfaces annotated by means of BLASTP alignments (Altschul, et al., 1990) against the 3D-footprint library (http://floresta.eead.csic.es/3dfootprint/download/list_interface2dna.txt) with an E-value threshold of 10. Aligned interface positions from one or more protein-DNA complexes are thus transferred to entries in the database. A benchmark with 127 TFs comparing other machine learning interface-inference tools showed that this straightforward BLAST-based strategy is the most accurate among sequence-based methods, as shown in Supplementary Fig. S4.

Pfam domains (Punta, et al., 2012) were also annotated for all TF sequences using HMMSCAN from the HMMER 3.0 suite (Finn, et al., 2011). Family-specific E-value thresholds were optimized to reduce false positive matches, according to benchmark experiments summarized in Supplementary Fig. S5.

2.4 Analysis of data redundancy

Two kinds of redundancy were defined and measured: internal and external. Internal redundancy is defined as the number of redundant DNA motifs (PSSMs) and TF sequences from the same source, while external redundancy is estimated with respect to other sources. Internal redundancy of DNA motifs was measured aligning all PSSMs from the same data source against each other with STAMP (Mahony and Benos, 2007), taking

the best hit to define nearest neighbor clusters of similar DNA motifs. Two E-value thresholds were tested to define redundancy, 1E-10 and 1E-5; motifs with lower E-values were clustered together and labeled as redundant. Internal redundancy of TF sequences was measured running CD-HIT (Li and Godzik, 2006). Two sequence identity thresholds were tested, 90% and 50%, so that aligned sequences with higher identity percentages were clustered together. External redundancy of either DNA motifs or TFs was estimated comparing PSSMs or protein sequences across data sources. External redundancy values estimate how many data entries from each database have similar values in the other databases. These values are asymmetrical because comparisons can be made in both ways: A vs. B and B vs. A. Redundant data among footprintDB, TRANSFAC and JASPAR CORE were clustered and Euler diagrams drawn with eulerAPE v2.0 (available from http://www.eulerdiagrams.org/eulerAPE), as depicted in Fig. 3. These diagrams illustrate data redundancy among the three main repositories, considering that JASPAR CORE is by design contained in footprintDB.

2.5 Protein sequence and DNA motif search

The search engine of footprintDB relies on a Perl script that implements protein sequence searches with BLASTP (Altschul, et al., 1990) and DNA motif scans with STAMP (Mahony and Benos, 2007). Protein searches take FASTA format sequences as input and by default accept hits with a maximum E-value of 1. Results can be ordered by increasing E-value or by decreasing interface similarity. Interface similarity is calculated using a scoring matrix that gives score 1 to amino acids with similar physicochemical properties and 0 to the rest (Supplementary Fig. S6). Motif searches are carried out by STAMP using as input a PSSM in TRANSFAC format. Other accepted inputs are single or multiple DNA sequences, which will be internally converted to PSSMs. The alignment algorithm is an ungapped Smith-Waterman implementation which extends matched motifs with a maximum E-value of 1. The scoring function is the Pearson Correlation Coefficient (PCC) of aligned matrix columns. Motif similarity is defined as the sum of column PCC values. Results can be ordered by increasing E-value or by decreasing motif similarity. Besides these standard alignment scores, the output table resulting from DNA searches is colored according to twilight thresholds: green matches correspond to reliable motif alignments, while motifs over a red background cannot be guaranteed to be correctly aligned (Sebastian and Contreras-Moreira, 2013).

2.6 External proteome search

A remarkable feature of footprintDB is that it allows extending database searches to external proteomes. By doing this, users can transfer search results, which only consider annotated TFs and DNA motifs, to other species of interest. This extension step requires running BLASTP with a default E-value threshold of 0.01. Note that this parameter can be tuned. When possible, resulting hits have their interface residues predicted, so that they can be sorted with respect to the original annotated TFs in the database.

2.7 Benchmark

Three test datasets and three representative species were chosen in order to benchmark the predictive ability of footprintDB: (1) Arabidopsis thaliana data from Athamap; (2) Escherichia coli data from RegulonDB and (3) a subset of 100 randomly-selected DNA motifs and their associated TFs from "HumanTF". Each dataset consisted of DNA motifs recognized by a single TF, and TFs recognizing a single DNA motif. Benchmark searches were performed by setting aside each test set from the meta-database, first excluding and then including annotated data for the corresponding species. Both protein and DNA searches were evaluated:

(1) The TF benchmark consisted in scanning protein sequences against footprintDB+TRANSFAC and then comparing the predicted motifs to the cognate DNA motif of the query. If one of the matched PSSMs is significantly similar to the cognate motif (STAMP E-value≤1E-5), the result is stored together with its rank, E-value, motif similarity, BLASTP E-value, interface similarity, organism and data source.

(2) The DNA benchmark was done by searching input motifs from all three test datasets looking for putative binding homologous TFs within the corresponding proteomes (versions: A. thaliana TAIR9, E.coli U000096.2 and Homo sapiens GRCh37.58). Hits were compared with the TF associated to the query, and defined as correct with a % sequence identity higher than 90.

Fig. 5 illustrates how the benchmark was done, and the sets of obtained predictions are provided as Supplementary Data.

3 RESULTS

3.1 Database contents

The first release of footprintDB contains 2422 unique transcription factor sequences, 3662 PSSMs and 10112 DBSs. As we added the contents of TRANSFAC version 2012.1 to the analysis, these numbers increased significantly to 4923, 5349 and 21988, respectively (see Table 1). In the next section redundancy analyses are performed in order to fairly evaluate the richness of each data source.

Table 1. Number of unique TF sequences, DNA motifs and DBSs in footprintDB sources. TRANSFAC data are included as a reference.

	Source	TFs	Motifs	Sites
	footprintDB	2422	3662	10112
	TRANSFAC	2919	2163	11949
footprintDB	JASPAR CORE	715	1312	2388
footprintDB	3D-footprint	605	802	722
footprintDB	HumanTF	528	818	0
footprintDB	UniPROBE	401	415	2963
footprintDB	RegulonDB	82	82	1862
footprintDB	Athamap	74	84	84
footprintDB	DBTBS	70	88	1234
footprintDB	DrosophilaTF	57	61	863
	Total unique	4923	5349	21988

The most populated species among footprintDB sources, as compared to TRANSFAC and JASPAR CORE, are shown in Fig. 1A, together with the corresponding number of TF sequences annotated in each data source (full statistics are reported as Supplementary Data). We notice that the TF binding preferences of a few organisms are widely covered, such as human, mouse, yeast, fly or *E.coli*. This coverage could already be sufficient for many applications in the case of human or mouse, considering current estimations of the repertoire of TFs in the human genome (Vaquerizas, et al., 2009). However, other species such as A.thaliana, maize or rice (among plants) or mammals like cow, pig

and sheep have a very shallow coverage. Moreover, the analysis unveils that some organisms are better covered by TRANSFAC (for instance *G. gallus*, *R. norvegicus* or *X. laevis*), while others, such as bacteria, are only considered by open-access libraries such as DBTBS. In the benchmark section we assess to what extent well-explored species can help make predictions on the remaining.

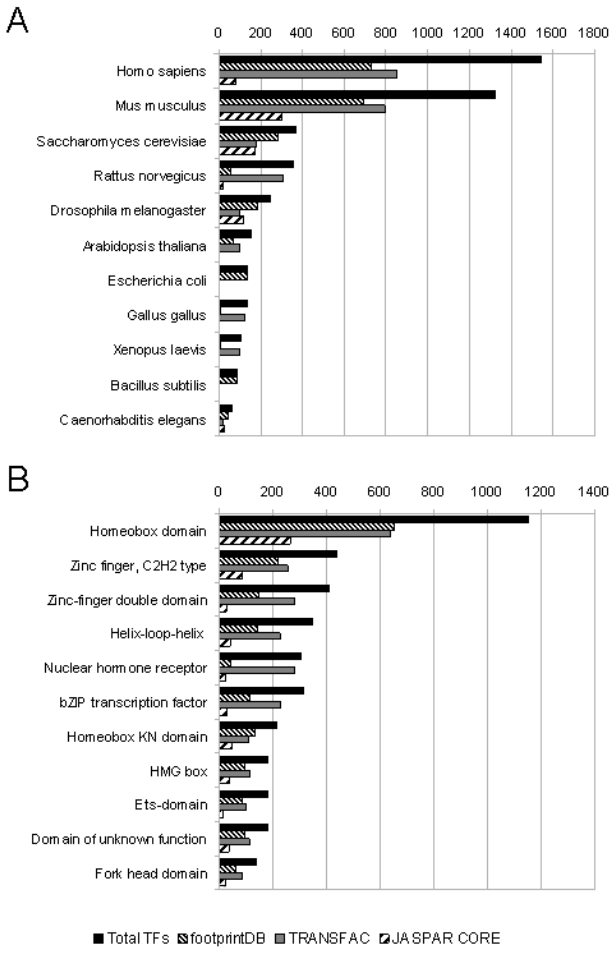


Fig. 1. Species and domain composition of footprintDB entries, compared to TRANSFAC and JASPAR CORE. (A) Most represented species among database entries. (B) Most frequent DNA-binding domains.

The most frequent DNA-binding domains found in the analyzed databases are also summarized in Fig. 1B (see also Supplementary Data). It can be seen that Homeobox and Zinc fingers, widely studied in the literature, are overrepresented. Furthermore, we note that prokaryotic RegulonDB and DBTBS databases do not contain such proteins; instead they are enriched in typical bacterial regulatory proteins, such as helix-turn-helix TFs. One of the unique features of footprintDB is the annotation of interface residues of transcription factors. This annotation relies on the 3D-footprint database, which routinely dissects the interfaces of protein-DNA complexes deposited at the PDB, as explained in Fig. 2A. Alignments between amino acid sequences of TFs and homologous protein-DNA complexes thus drive the transfer of experimentally determined interface residues to equivalent residues of footprintDB entries. Overall, 97% of the total TFs are annotated with this procedure, with most resulting interfaces comprising 7 to 16 amino acid residues (Fig. 2B), in agreement with previous values calculated for Homeobox and Zinc finger families (Contreras-Moreira, et al., 2009) (full interface length statistics are reported as Supplementary Data). As we show later, annotated interfaces are a good filter to decide whether two TFs might be recognizing the same DBS, as required when extending footprintDB searches to external proteomes.

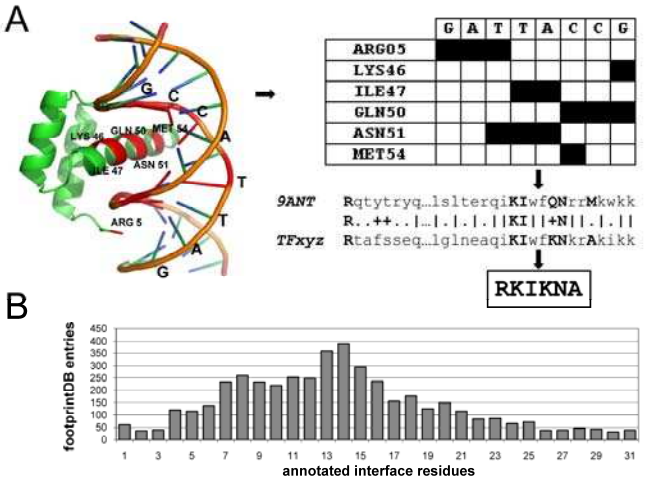


Fig. 2. Annotation of interface residues. (A) 3D-footprint interface of PDB entry 9ANT, which corresponds to Homebox protein Antennapedia in complex with a *cis* element. First, inter-atomic distances are calculated among atoms of amino acid side chains and nitrogen bases. Second, a matrix of interacting residues and their target bases is generated. Third, interface residues are marked as upper-case letters in the sequence, and are further transferred to homologous sequences by means of BLASTP alignments. Note that several PDB complexes can often be used to annotate a single footprintDB entry. (B) Histogram of length of predicted interfaces in footprintDB. Large interfaces usually correspond to proteins with several DNA-binding domains.

3.2 Survey of data redundancy

To estimate the degree of redundancy of the data sources integrated in footprintDB we compared data entries within and between databases, as explained in Material and Methods. When checking internal redundancy, it turns out that some of the largest repositories such as TRANSFAC, JASPAR CORE, HumanTF, UniPROBE and the whole footprintDB contain from 20% to 40% redundant DNA motifs, depending on the similarity cut-off E-values employed to compare PSSMs (1E-5 and 1E-10, respectively, see Supplementary Table S3). For TF sequences we report an even higher redundancy: 40-70% within footprintDB, TRANSFAC and 3D-footprint entries and slightly lower percentages in HumanTF (24-53%), due to proteins with %sequence identity higher than the 90% and 50% cutoffs, respectively (see Supplementary Table S4). JASPAR motif redundancy was anticipated as we analyzed on purpose the all-species redundant set for completeness.

External redundancy is even more relevant to evaluate the added value of each data source. The data in Table 2 summarize the performed comparisons in terms of DNA motifs, indicating that all eukaryotic data sources contain a large fraction of redundant entries among them. This is expected, as databases such as JASPAR and TRANSFAC are built, at least in part, by curating the same literature, and therefore overlap substantially. In fact, TRANSFAC contains 438 DNA motifs identical to JASPAR entries (see Supplementary Table S5). This number increases to 1332 if we consider TRANSFAC PSSMs that align to JASPAR entries with STAMP E-value ≤ 1E-10, which we considered as a cutoff for nearly identical motifs. Besides, these analyses confirm the minimum redundancy between bacterial sources (RegulonDB and DBTBS) and the rest. Note that external redundancy estimates are asymmetrical, as different results are obtained depending on the direction of the comparison.

The picture arising from the comparison of TF sequences across data sources reveals their analogous levels of redundancy (see Supplementary Table S6). For instance, 368 TRANSFAC TFs have identical amino acid sequences in JASPAR; this figure increases to 1062 if we apply a 90% sequence identity redundancy cutoff.

After reviewing the full set of comparisons in Supplementary Tables S5 and S6, it can be concluded that there is an 'eclipse effect': as we relax the redundancy thresholds, eukaryotic datasets progressively overlap till most of their contents turn to be shared. This behavior is shown in the Euler diagrams in Fig, 3 for the three main databases footprintDB, TRANSFAC and JASPAR CORE (which is included in footprintDB).

Perhaps the most interesting case is the overlap between footprintDB and TRANSFAC (Fig. 3, Table 2, Supplementary Tables S5 and S6). They share 22% of motifs and 14% of TRANSFAC TFs (Fig. 3A,D). If nearly identical DNA motifs (E-value≤1E-10) and TFs (with %sequence identity≥90) are considered, these percentages increase to 71% of motifs and 56% of TFs (Fig. 3B,E). Further relaxing these thresholds to E-value≤1E-5 (short motifs, or motifs that share a common pattern but not the whole motif) and sequence identity>=50% (proteins with common domain architecture) then both databases are almost equivalent, as they seem to share 95% of motifs and 98% of TFs (Fig. 3C,F). Such data overlap is also very noticeable for JASPAR.

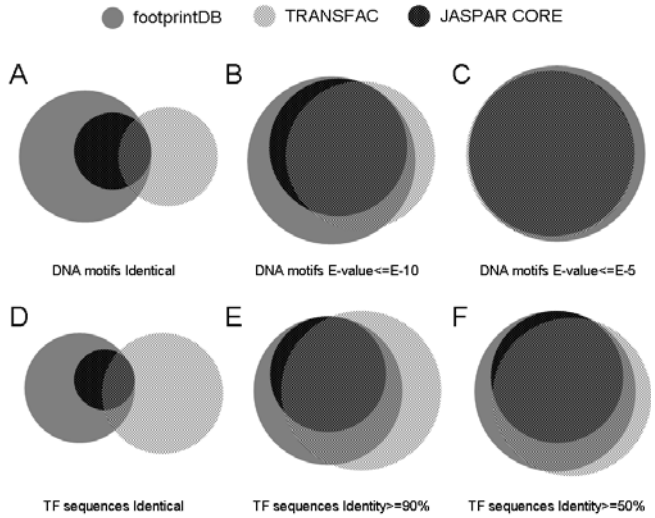


Fig. 3. Euler diagrams representing redundancy for DNA motifs (A, B and C) and TF sequences (D, E and F) annotated in footprintDB, TRANSFAC and JASPAR CORE databases.

3.3 Website

The web interface of footprintDB displays the main menu on the left side, which provides access to the current list of publicly available data sources and to the search engine, in addition to the documentation (Supplementary Fig. S7). By default anonymous users can perform any of these tasks:

(1) Listing the included repositories, their versions, references and authors, with links to the original websites. From this table it is possible to browse transcription factors (TFs), DNA-binding motifs (PSSMs) and DNA-binding sites (DBSs) curated in each individual data source.

(2) Accessing a single entry (TF, PSSM or DBS) to display all the available information for that record, including references to primary literature and experimental evidence, and download them in TRANSFAC format (Example in Supplementary Fig. S8).

(3) Keyword search of TF, PSSM or DBS accessions. The form supports filtering by database, organism or related TF domain and results can be downloaded in TRANSFAC format.

(4) Sequence search, to scan protein or DNA sequences and PSSMs. The search process is explained in the next section.

In addition, registered users have access to the following extra features:

(1) Storing and reusing previous searches.

(2) Inserting/removing their own databases in TRANSFAC or footprintDB formats.

3.4 Search engines and external proteomes

The main purpose of footprintDB is to support searching for annotated TFs and/or regulatory sequences, as depicted in Fig. 4. The search engine is designed primarily to receive two types of queries: (1) a DNA consensus motif, PSSM or site; and (2) a protein sequence of a putative DNA-binding protein. Therefore, two kinds of output will be produced, respectively: (1) a list of DNA-binding proteins predicted to bind a similar DNA motif; and (2) a list of DNA motifs recognized by similar proteins annotated

in any of the included data sources. Search results can be sorted by E-value, motif similarity or interface similarity.

Moreover, the user can also look for homologues in a third party species, by simply specifying an appropriate proteome in the formulary dropdown list or by uploading a custom proteome in FASTA format. Together with the standard search results, this option produces a list of homologous proteins with their interfaces annotated. Interface predictions can then be used to filter out BLASTP hits displaying a significantly different set of DNA binding residues. This kind of search is useful in order to scan TFs within a particular organism of interest, for instance to design laboratory experiments. We further illustrate this kind of search in the next section.

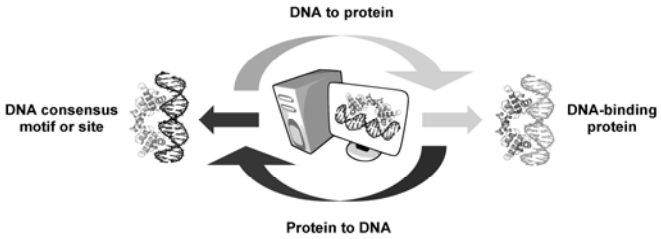


Fig. 4. Main search types supported by footprintDB. Light arrow (top): if the input data is a DNA sequence or motif, the search is powered by STAMP, and the output are proteins likely to bind sequence similar to the input. These proteins might be primary entries in footprintDB or rather endogenous TFs of a proteome of choice, after a secondary call to BLASTP. Dark arrow (bottom): when the input is a protein sequence, a BLASTP search is performed instead, and the user gets a list of putative DNA target sites for it.

Table 2. External redundancy of DNA motifs across data sources integrated in footprintDB. The main diagonal shows the total number of PSSMs in each source. Motifs aligned with STAMP E-values < 1E-10 were called redundant. The top row corresponds to the union of all individual data sources that contribute to footprintDB. Note that a large fraction of 3D-footprint entries are not called redundant within footprintDB because their short DNA motifs fail to produce alignments with E-values < 1E-10. (*) A redundant version of JASPAR was tested.

		footprintDB	TRANSFAC	JASPAR*	3D-footprint	HumanTF	UniPROBE	RegulonDB	Athamap	DBTBS	DrosophilaTF
	footprintDB	**3662**	1531	1295	446	818	412	81	84	76	57
	TRANSFAC	2111	**2163**	963	78	577	401	8	49	5	30
footprintDB	**JASPAR** *	2299	1332	**1312**	60	536	359	5	21	3	20
	3D-footprint	672	154	89	**802**	96	22	4	5	3	7
	HumanTF	1453	651	386	45	**818**	174	4	11	1	14
	UniPROBE	1086	628	386	7	265	**415**	3	5	2	6
	RegulonDB	116	13	10	4	8	8	**82**	0	5	0
	Athamap	130	108	22	6	14	3	0	**84**	1	0
	DBTBS	94	6	4	6	1	3	3	1	**88**	0
	DrosophilaTF	142	64	31	6	34	14	0	0	0	**61**

3.5 Example of footprintDB search

Imagine that we have obtained a set of *cis* elements regulated by a bZIP TF in the genome of *Antirrhinum majus*. Take for instance motif bZIP910, annotated in JASPAR and AthaMap, shown in Supplementary table S8. Now, we have just found out that some of these DBSs are also conserved in *Arabidopsis thaliana*, and want to identify the endogenous TFs that might recognize them, so that we can test them in the lab. To perform such a query, we first paste the input DNA sequences in the search formulary to obtain a list of similar motifs in the database. Among the top four results are bZIP910, XBP1 and TGA1, annotated in different sources and species (snapdragon, human, thale cress and tobacco). All of them are motifs bound by TFs of the bZIP family (basic leucine zipper domain), employing a similar binding interface. However, as we go down the list, motifs start to diverge and hence have associated higher E-values.

If we extend this search by scanning proteins within the *Arabidopsis thaliana* TAIR9 proteome, we find 30 proteins with interfaces identical to that of the query (bZIP910, with interface signature RNRSASR), which should be the best candidates to be tested in the lab for binding. In addition, these results could guide site-directed mutagenesis experiments targeting interface residues. The second hit (human XBP1) produces a list of 21 *A.thaliana* TFs, but an inspection of their interfaces (RKNRAAARK) shows clearly that these are a different subfamily of TFs. Furthermore, these interfaces are in all cases similar but not identical to that of the corresponding human TF. For these reasons, the second row of results should be handled with care. Finally, the third and fourth hits, TGA1 from *A.thaliana* and tobacco, link to up to 10 endogenous proteins with identical interfaces (RQNRAASR), which are similar to the first 30 candidate TFs, and therefore should also be considered for further analyses.

3.6 Search benchmark

The first benchmark consisted in scanning *A.thaliana* TF sequences and DNA motifs from Athamap against footprintDB+TRANSFAC, after excluding all *A.thaliana entries*. Figures 5A and 5B summarize both experiments, and can be extended to the second and third benchmarks explained below.

Overall, 31 out of the 56 tested TFs (55%) were recovered in the TF search, and 27 out of 48 DNA motifs (56%). Among recovered TFs and motifs, 24/31 and 13/27 were first hits, respectively. It is remarkable that most hits were annotated in TRANSFAC, mainly from other plants, human or mouse. In both experiments, the average interface similarities of correctly recovered TFs were 80 and 91%, compared to overall values of 58% and 78%, respectively, as shown in Fig. 5C. When all *A.thaliana* records were put back in footprintDB (still excluding Athamap), the percentages of recovered TFs and motifs increased to 70% and 83%, respectively. Again most results were derived from TRANSFAC, suggesting that, together with Athamap, it is the most comprehensive source of plant regulatory data.

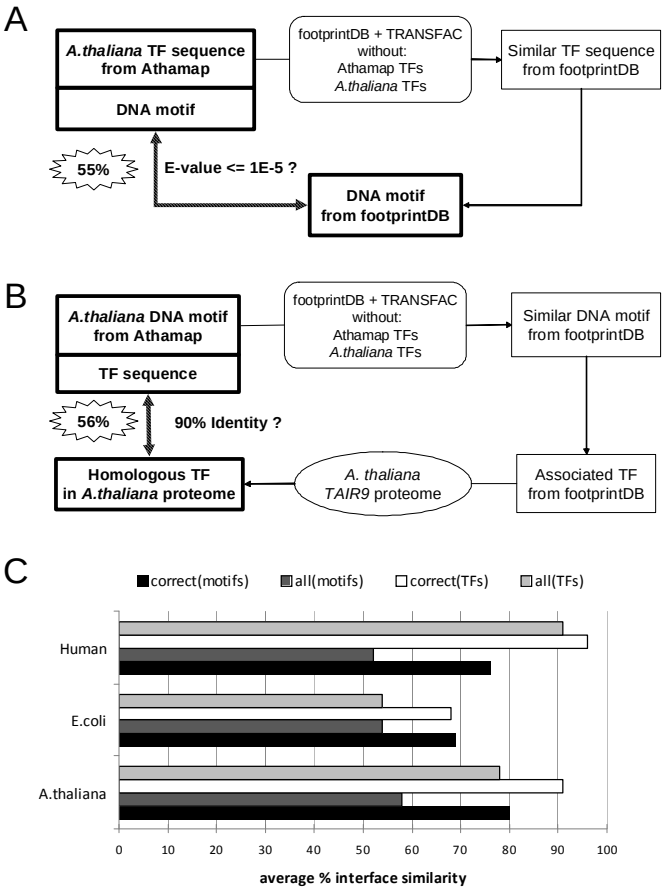


Fig. 5. Benchmark of footprintDB performance using A.thaliana data annotated in Athamap. (A) One TF sequence from Athamap is searched against footprintDB. (B) A DNA motif from Athamap is scanned against footprintDB. In both cases, red arrows represent the comparison of predictions to the cognate sequences, which are taken to be correct or false in terms of STAMP E-value (A) and % sequence identity (B), as explained in Materials and Methods. In the TF experiment (A), 31 out of 56 Athamap sequences (55%) were successfully recovered. In the corresponding experiment with DNA motifs (B), the footprintDB pipeline recovered 27 out of 48 (56%) Athamap motifs. (C) Interface similarity of correct hits and of all predictions among A.thaliana, E.coli and human test sets used during the benchmark.

The second benchmark consisted in scanning *E.coli* sequences from RegulonDB against footprintDB+TRANSFAC, after excluding all *E.coli* records. In total, only 9 out of 82 tested TFs (12%) were successfully retrieved in the TF search, and 8 out of 82 DNA motifs (10%). Among these, 6/9 and 5/8 were first hits, respectively. Most matches were from *B.subtilis* entries annotated in DBTBS and 3D-footprint. In both cases, average interface similarities of recovered TFs were ca. 69%, compared to 54% among all predictions. When *E.coli* records were included back in footprintDB, the percentages of recovered TFs and motifs increased to 18% and 20% respectively.

The third benchmark consisted in scanning TF sequences and DNA motifs from "HumanTF" against footprintDB+TRANSFAC, after excluding all human records. In total, 100 out 100 tested TFs, and 90 out of 100 DNA motifs, were recovered. Overall, 87/100 and 69/90 were first hits, respectively. Matched records were from model multicellular organisms, mostly mouse, but also fly, worm or frog, generally annotated in TRANSFAC or JASPAR. In both experiments, average interface similarities of recovered TFs were 86% and 96%, compared to 52% and 91% for all predictions. When human records were added back to footprintDB, the percentages of recovered TFs and motifs remain the same, but now some best hits were human.

4 DISCUSSION

FootprintDB is an effort to group and unify the most important, diverse and well annotated open access databases of experimentally obtained TF binding preferences. Although a few other resources have a similar philosophy (Portales-Casamar, et al., 2010; Riva, 2012), footprintDB goes one step further by systematically annotating interface residues, those that capture the binding specificity of DNA-binding proteins. This allows linking motif similarity with TF similarity and supports scanning TFs with conserved interfaces in external proteomes. The observed high values of interface similarity among correctly recovered TFs in our benchmarks, compared to the average values among all predictions, confirm their value as a quality control when transferring regulatory annotations by homology.

The search engine is perhaps the most remarkable feature of footprintDB. It can drive the prediction of DNA binding motifs for unknown TF sequences, and also the opposite search, assigning putative TFs to input DNA motifs, which might have been found during in silico promoter analyses. Our benchmarks suggest that footprintDB indeed has predictive power, as it was able to correctly recover TFs and motifs from *E.coli*, *A.thaliana* and *H. sapiens*. However, the performance was better with eukaryotes than with bacteria, as the tested data sources are evidently more redundant for multicellular organisms. This observation exposes that the predictive ability of footprintDB is proportional to the richness of its data sources. In our experience, correct results are obtained most frequently when TF sequences or DNA motifs from phylogenetically related organisms are available in the database. For instance, in our *A.thaliana* TF benchmark, first hits captured the correct TF in 24 out of 31 cases. Overall, 21 of these 24 matches were to plant sequences, including species such as *Zea mays*, *Helianthus annuus*, *Nicotiana tabacum*, *Brassica napus*, *Antirrhinum majus*, *Hordeum vulgare*, *Solanum lycopersicum*, *Daucus carota* and *Oryza sativa*. In contrast, the *E.coli* test had very limited success, as the only available reference organism was *Bacillus subtilis*, which in fact is only remotely related. Beyond

these benchmark experiments, footprintDB has already been profitably applied for identifying endogenous rice TFs (OsEREBP1 and OsEREBP2) that bind specifically to a target sequence within the OsRMC promoter (Serra, et al., 2013). Furthermore, footprintDB has also been extensively tested during the in silico identification of drought stress regulatory proteins in *A.thaliana*, which have been later validated with yeast one-hybrid experiments. Preliminary results further confirm that correct predictions are provided by phylogenetically related entries which are annotated in the database, otherwise results are not reliable (data not shown). Another important result of this unpublished work, which is relevant in this context, is that DNA searches seem to be more sensitive with single *cis* elements as input than with PSSMs.

This study is also an up-to-date comprehensive comparison of TF databases. Fogel et al. made a statistical analysis of an early version of TRANSFAC (Fogel, et al., 2005), while other papers have studied the similarity of motifs annotated in TRANSFAC and JASPAR (Kielbasa, et al., 2005; Schones, et al., 2005). In our study we find significant data redundancy between TRANSFAC and JASPAR databases. However, the most significant overlap found is between footprintDB and TRANSFAC, as summarized in Fig. 3. These analyses suggest that footprintDB and TRANSFAC contain overall almost equivalent data, so footprintDB can be currently used as an open-access alternative, bearing in mind that organism coverage is also an important factor, as already discussed in the *A.thaliana* benchmark. It remains to be seen whether available funding will allow footprintDB (and its integrated datasets) to keep the pace of scheduled updates of commercial alternatives such as TRANSFAC.

Significant internal redundancy is observed among TF sequences of 3D-footprint, "HumanTF" and TRANSFAC, as well as footprintDB. In the first case, this is mostly explained in terms of the intrinsic redundancy of the PDB. With respect to "HumanTF", the observed redundancy is due to the fact that this source includes both complete protein sequences and domains of orthologous TFs from mouse and human. A similar explanation is valid for TRANSFAC, which appears to frequently annotate orthologous TFs from relates species. Indeed, inspection of the resulting clusters suggests that most redundant TFs at 90% of sequence identity are probably inparalogues and orthologues from phylogenetically close organisms. Inspection of relaxed clusters (50% sequence identity cut-off) unveils that they gain more divergent homologous proteins of the same family, which bind to regulatory elements using the same Pfam domains.

While our survey reveals a comprehensive coverage of human and murine TFs, both in TRANSFAC and in open-access repositories, it also shows that prokaryotes are still only served by specialized expert-curated resources such as RegulonDB and DBTBS. In fact these organism-specific repositories are reported to be the least redundant (as also observed for DrosophilaTF). By combining freely available data sources, footprintDB aims to be a reference meta-database covering bacteria, plants and animals, although our benchmark clearly shows that its predictive power is greater for multicellular organisms. Despite the wide coverage of this meta-database, our benchmarks encourage the addition of any other relevant high quality resources/datasets, as we found out with the plant regulatory data. For this reason the Web interface allows users to import their own data collections, which can optionally be shared with other users, and we hope that the adoption of this tool by the community will translate into a richer set of curated data repositories.

ACKNOWLEDGEMENTS

We thank our colleagues from the STREG project (L.Bülow, R.Hehl, G.Huep, B.Weisshaar, C.Dubos, L.Lepiniec and A.Koller) for their feedback and support while developing this resource. We would also like to acknowledge the authors of the databases integrated in footprintDB for their feedback and permissions to use their data: L.Bülow, R.Hehl (Athamap), C. Bergman (DrosophilaTF), J.Taipale (HumanTF), A. Sandelin (JASPAR), J. Collado-Vides (RegulonDB) and M. Bulyk (UniPROBE).

Funding: This work was supported by Programa Euroinvestigación/Plant KBBE 2008 [EUI2008-03612]. Results have been achieved under the framework of the Transnational (Germany, France, Spain) Cooperation within the PLANT-KBBE Initiative, with Funding from Ministerio de Ciencia e Innovación, Agence Nationale de la Recherce (ANR) and BMBF.

REFERENCES

AlQuraishi, M. and McAdams, H.H. (2011) Direct inference of protein-DNA interactions using compressed sensing methods, *Proc Natl Acad Sci U S A*, **108**, 14819-14824.

Altschul, S.F., *et al.* (1990) Basic local alignment search tool, *J Mol Biol*, **215**, 403-410.

Berger, M.F. and Bulyk, M.L. (2006) Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins, *Methods Mol Biol*, **338**, 245-260.

Berman, H.M., *et al.* (2000) The Protein Data Bank, *Nucleic Acids Res*, **28**, 235-242.

Bulow, L., *et al.* (2009) AthaMap, integrating transcriptional and post-transcriptional data, *Nucleic Acids Res*, **37**, D983-986.

Contreras-Moreira, B. (2010) 3D-footprint: a database for the structural analysis of protein-DNA complexes, *Nucleic Acids Res*, **38**, D91-97.

Contreras-Moreira, B., Sancho, J. and Angarica, V.E. (2009) Comparison of DNA binding across protein superfamilies, *Proteins*, **78**, 52-62.

Crooks, G.E., *et al.* (2004) WebLogo: a sequence logo generator, *Genome Res*, **14**, 1188-1190.

Down, T.A., *et al.* (2007) Large-scale discovery of promoter motifs in Drosophila melanogaster, *PLoS Comput Biol*, **3**, e7.

Finn, R.D., Clements, J. and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching, *Nucleic Acids Res*, **39**, W29-37.

Fogel, G.B., *et al.* (2005) A statistical analysis of the TRANSFAC database, *Biosystems*, **81**, 137-154.

Galas, D.J. and Schmitz, A. (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity, *Nucleic Acids Res*, **5**, 3157-3170.

Garner, M.M. and Revzin, A. (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system, *Nucleic Acids Res*, **9**, 3047-3060.

Johnson, D.S., *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions, *Science*, **316**, 1497-1502.

Jolma, A., *et al.* (2013) DNA-Binding Specificities of Human Transcription Factors, *Cell*, **152**, 327-339.

Kielbasa, S.M., Gonze, D. and Herzel, H. (2005) Measuring similarities between transcription factor binding sites, *BMC Bioinformatics*, **6**, 237.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22**, 1658-1659.

Lin, C.K. and Chen, C.Y. (2013) PiDNA: predicting protein-DNA interactions with structural models, *Nucleic Acids Res*.

Mahony, S. and Benos, P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities, *Nucleic Acids Res*, **35**, W253-258.

Matys, V., *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, *Nucleic Acids Res*, **34**, D108-110.

Noyes, M.B., *et al.* (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites, *Cell*, **133**, 1277-1289.

O'Neill, L.P. and Turner, B.M. (1996) Immunoprecipitation of chromatin, *Methods Enzymol*, **274**, 189-197.

O'Neill, M., Dryden, D.T. and Murray, N.E. (1998) Localization of a protein-DNA interface by random mutagenesis, *EMBO J*, **17**, 7118-7127.

Portales-Casamar, E., *et al.* (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles, *Nucleic Acids Res*, **38**, D105-110.

Punta, M., *et al.* (2012) The Pfam protein families database, *Nucleic Acids Res*, **40**, D290-301.

Ren, B., *et al.* (2000) Genome-wide location and function of DNA binding proteins, *Science*, **290**, 2306-2309.

Riva, A. (2012) The MAPPER2 Database: a multi-genome catalog of putative transcription factor binding sites, *Nucleic Acids Res*, **40**, D155-161.

Robasky, K. and Bulyk, M.L. (2011) UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions, *Nucleic Acids Res*, **39**, D124-128.

Salgado, H., *et al.* (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more, *Nucleic Acids Res*, **41**, D203-213.

Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences, *Nucleic Acids Res*, **18**, 6097-6100.

Schones, D.E., Sumazin, P. and Zhang, M.Q. (2005) Similarity of position frequency matrices for transcription factor binding sites, *Bioinformatics*, **21**, 307-313.

Sebastian, A. and Contreras-Moreira, B. (2013) The twilight zone of cis element alignments, *Nucleic Acids Res*, **41**, 1438-1449.

Serra, T.S., *et al.* (2013) OsRMC, a negative regulator of salt stress response in rice, is regulated by two AP2/ERF transcription factors, *Plant Mol Biol*, **82**, 439-455.

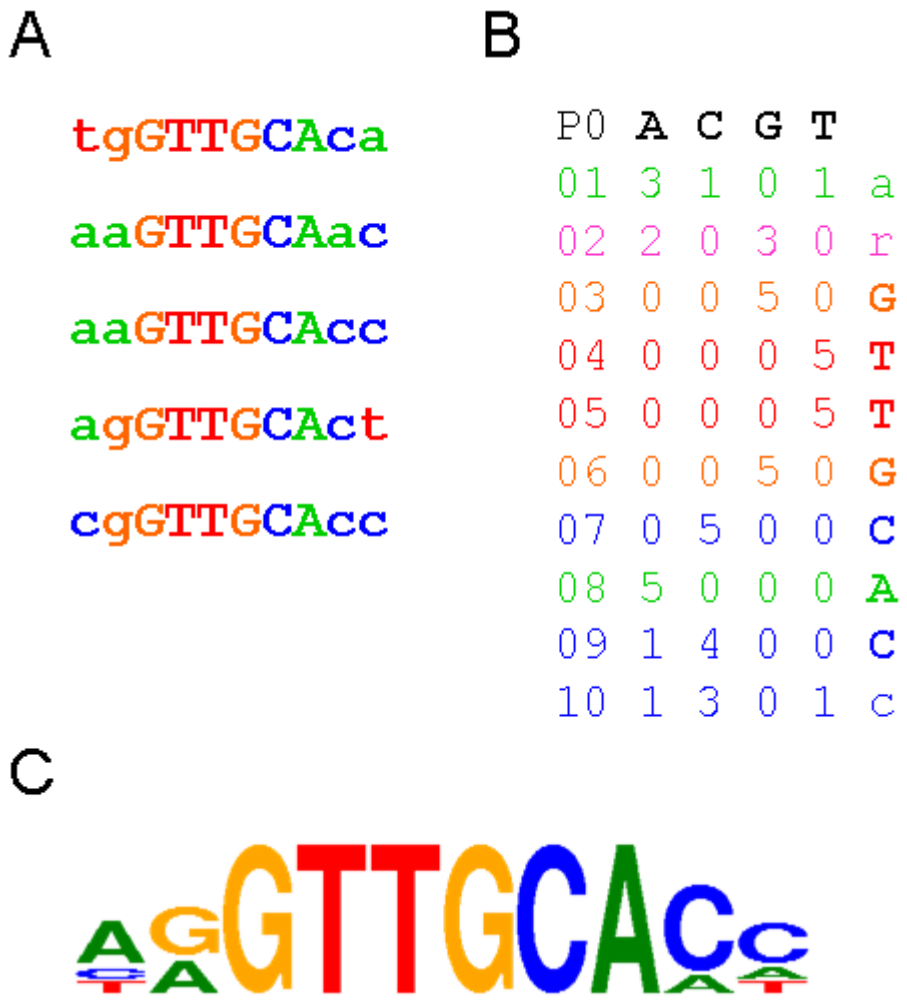
Shortle, D., DiMaio, D. and Nathans, D. (1981) Directed mutagenesis, *Annual review of genetics*, **15**, 265-294.

Sierro, N., *et al.* (2008) DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information, *Nucleic Acids Res*, **36**, D93-96.

Stormo, G.D. (2000) DNA binding sites: representation and discovery, *Bioinformatics*, **16**, 16-23.

Vaquerizas, J.M., *et al.* (2009) A census of human transcription factors: function, expression and evolution, *Nature reviews. Genetics*, **10**, 252-263.

SUPPLEMENTARY FIGURES



Supplementary figure S1. Three typical representations of DNA motifs. (A) Multiple alignment of DNA binding sites recognised by a TF, usually *cis* elements identified in different promoters. (B) Position-specific scoring matrix (PSSM) in TRANSFAC notation. (C) Sequence logo, with base heights proportional to their conservation across sites.

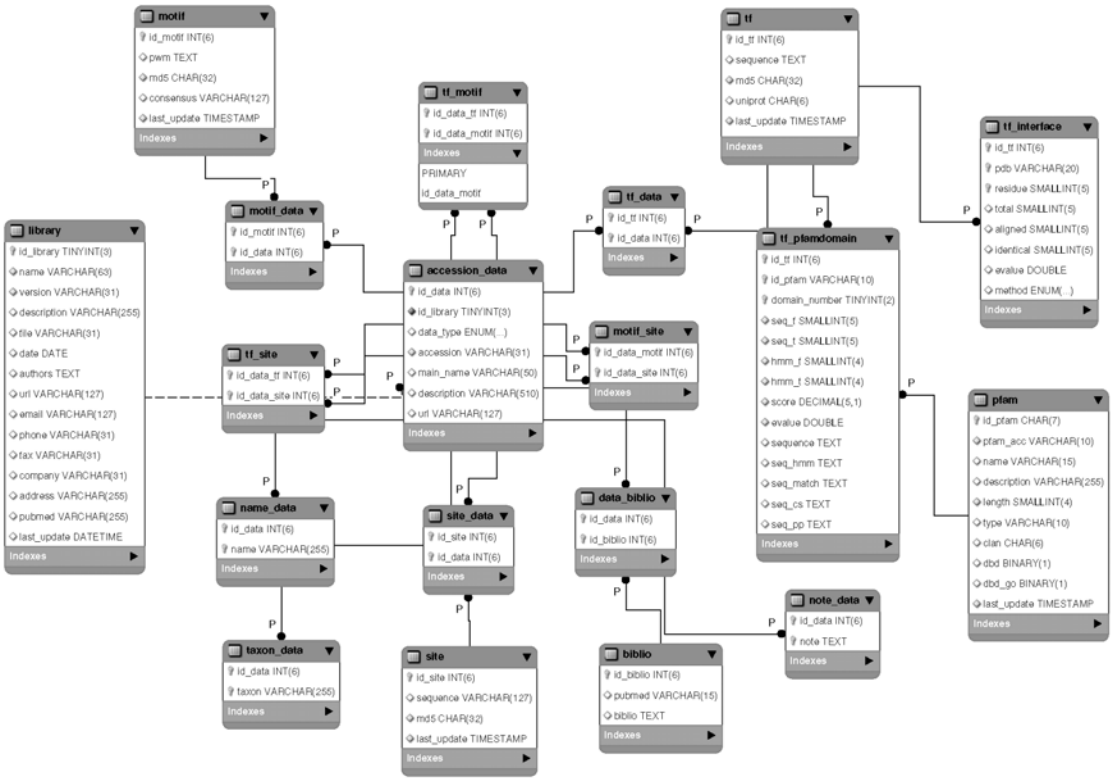
A

```
# FOOTPRINTDB FORMAT SPECIFICATIONS:

VV  Header with library data fields (Separated by ';')
VV  File: ; Name: ; Version: ; Date: ;
VV  Authors: ; Url: ; Email: ; Phone: ; Fax: ; Company: ; Address: ;
VV  Url: ; Pubmed: ; Description: ;
XX  End of section (header, motif, factor and site sections)
//  End of entry

# MOTIF SECTION:
MO  Accession
DE  Description
NA  Names (Separated by ';')
P0  PSSM
01
...
LN  Url
CC  Annotations (Separated by ';')
RX  PUBMED: Pubmed ID
RL  Reference details
XX

# FACTOR SECTION:
FA  Accession
DE  Description
NA  Names (Separated by ';')
SQ  Sequence
IN  (Blast prediction interface) Model: Residues; Total= ; Aligned= ;
IN  Identical= ; %ID= ; e-value= ; method=
SC  Uniprot Uniprot ID
OS  Organisms (Separated by ';')
LN  Url
CC  Annotations (Separated by ';')
RX  PUBMED: Pubmed ID
RL  Reference details
XX

# SITE SECTION:
SI  Accession
DE  Description
NA  Names (Separated by ';')
SQ  Sequence
LN  Url
CC  Annotations (Separated by ';')
RX  PUBMED: Pubmed ID
RL  Reference details
XX

# If the SITE has not Pubmed-Reference data, scripts will retrieve that
data from site's motif.

//
```

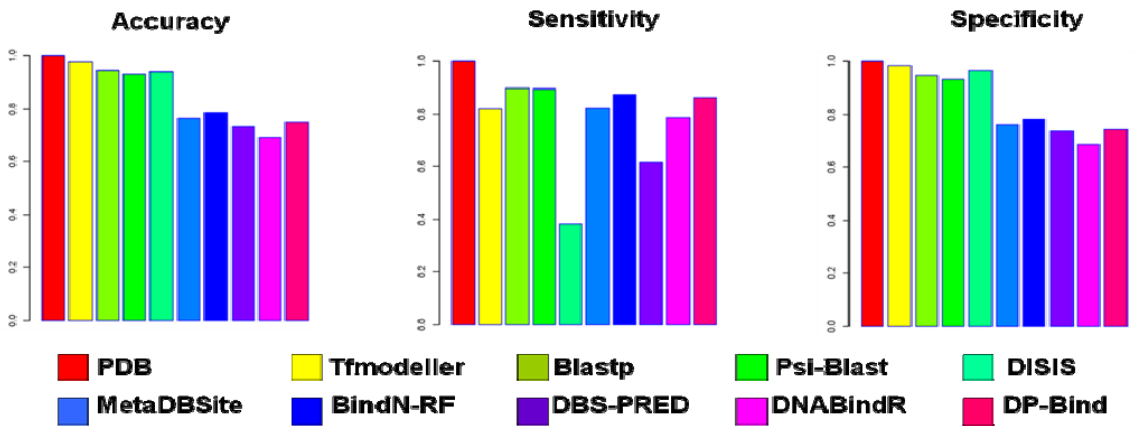
B

```
# TRANSFAC FORMAT SPECIFICATIONS:

VV  Header with library version
XX  End of field
//  End of entry

# MOTIF FILE:
AC  Accession
XX
ID  Identifier
XX
NA  Main name
XX
DE  Description
XX
BF  Binding factor accession; Name; Species: ...
XX
P0  PSSM
01
...
XX
BS  Binding site data sequence; Accession;
XX
CC  Annotation
XX
RN [1] Reference number and Accession
RX  PUBMED: Pubmed ID
RA  Reference Authors
RT  Reference Title
RL  Reference Journal, Number, Issue, Pages (Year)
XX
//

# FACTOR FILE:
AC  Accession
XX
ID  Identifier
XX
FA  Main name
XX
SY  Name synonyms (Separated by ';')
XX
OS  Organisms (Separated by ',')
XX
SQ  Sequence
XX
SC  Uniprot Uniprot ID
XX
FF  Annotation
XX
MX  Motif accession;
XX
BS  Binding site accession;
XX
RN  [1] Reference numer and Accession
RX  PUBMED: Pubmed ID
RA  Reference Authors
RT  Reference Title
RL  Reference Journal, Number, Issue, Pages (Year)
XX
//

# SITE FILE:
AC  Accession
XX
ID  Identifier
XX
DE  Description
XX
OS  Organisms (Separated by ',')
XX
SQ  Sequence
XX
BF  Binding factor accession; Name; Species: ...
XX
MX  Motif accession;
XX
RN [1] Reference numer and Accession
RX  PUBMED: Pubmed ID
RA  Reference Authors
RT  Reference Title
RL  Reference Journal, Number, Issue, Pages (Year)
XX
//
```

Supplementary figure S2. Comparison of the unified footprintDB format file (A) with the equivalent factor, motif and site files in TRANSFAC format (B).

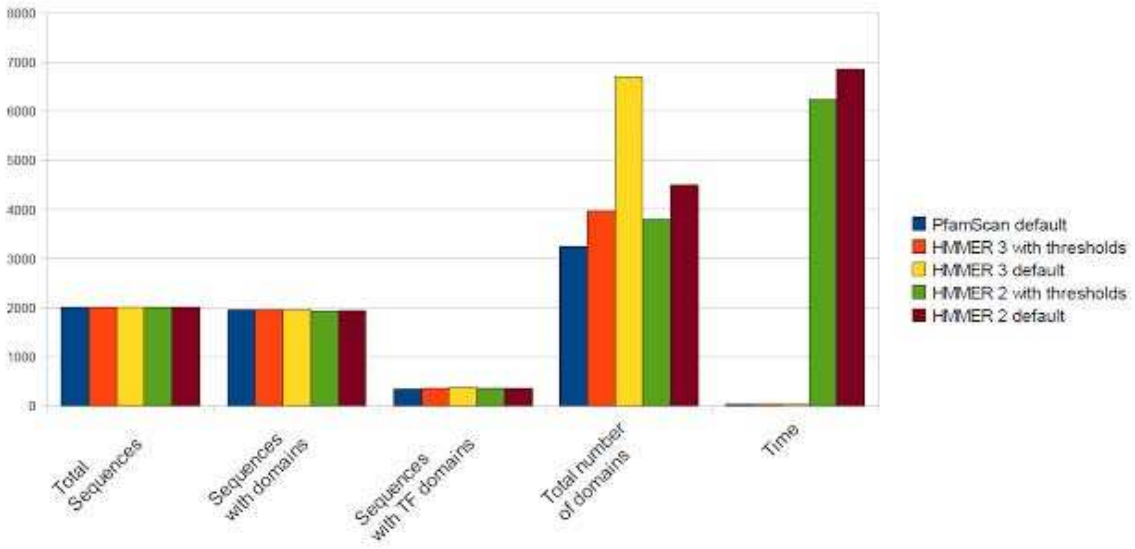


Supplementary figure S3. Table and column relational schema of footprintDB MySQL database. Three tables (*tf, site, motif*) store non redundant information of TFs, DBS and PSSMs, respectively. These are connected through tables *tf_data, motif_data, site_data* to table *accession_data* that stores redundant annotation of entries with links to their source repositories. Table *accesion_data* table is also linked to *library*, which contains repository information: name, version, authors, description, publications, url, etc. Other tables store names, synonyms, comments, taxonomies or bibliography annotations. TFs have two additional tables (*tf_interface* and *tf_pfamannotation*) with interface residues and Pfam domain annotation respectively. Finally, tables *user* and *user_library* store info about users and access privileges.

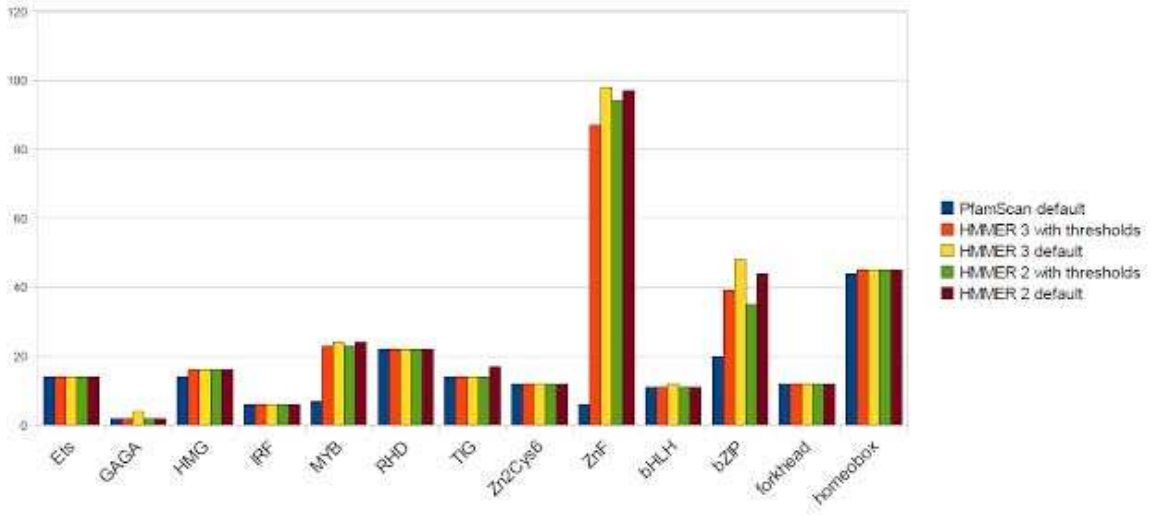


Supplementary figure S4. Leave-one-out cross-validation using a non-redundant set of 127 PDB structures of TF-DNA complexes with interface residues annotated in 3Dfootprint. Blastp and Psi-Blast perform local sequence alignments against the 3D-footprint library of annotated interfaces. TFmodeller builds interface-optimized homology models of protein-complexes. DISIS, DP-Bind, DNABindR, BindN, BindN-RF, DBS-PRED and MetaDBSite are machine learning approaches that make interface predictions taking a protein sequence as sole input.

A



B

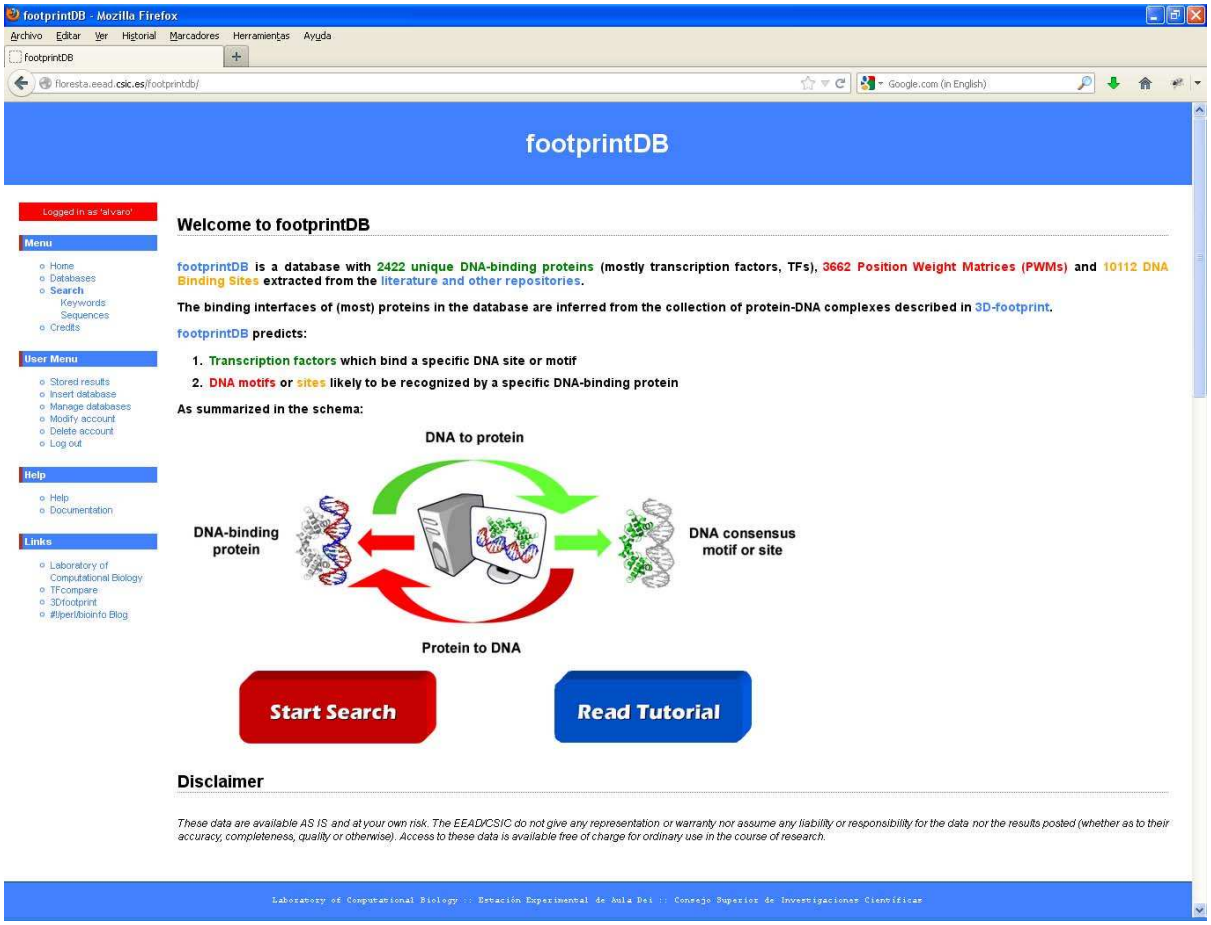


.

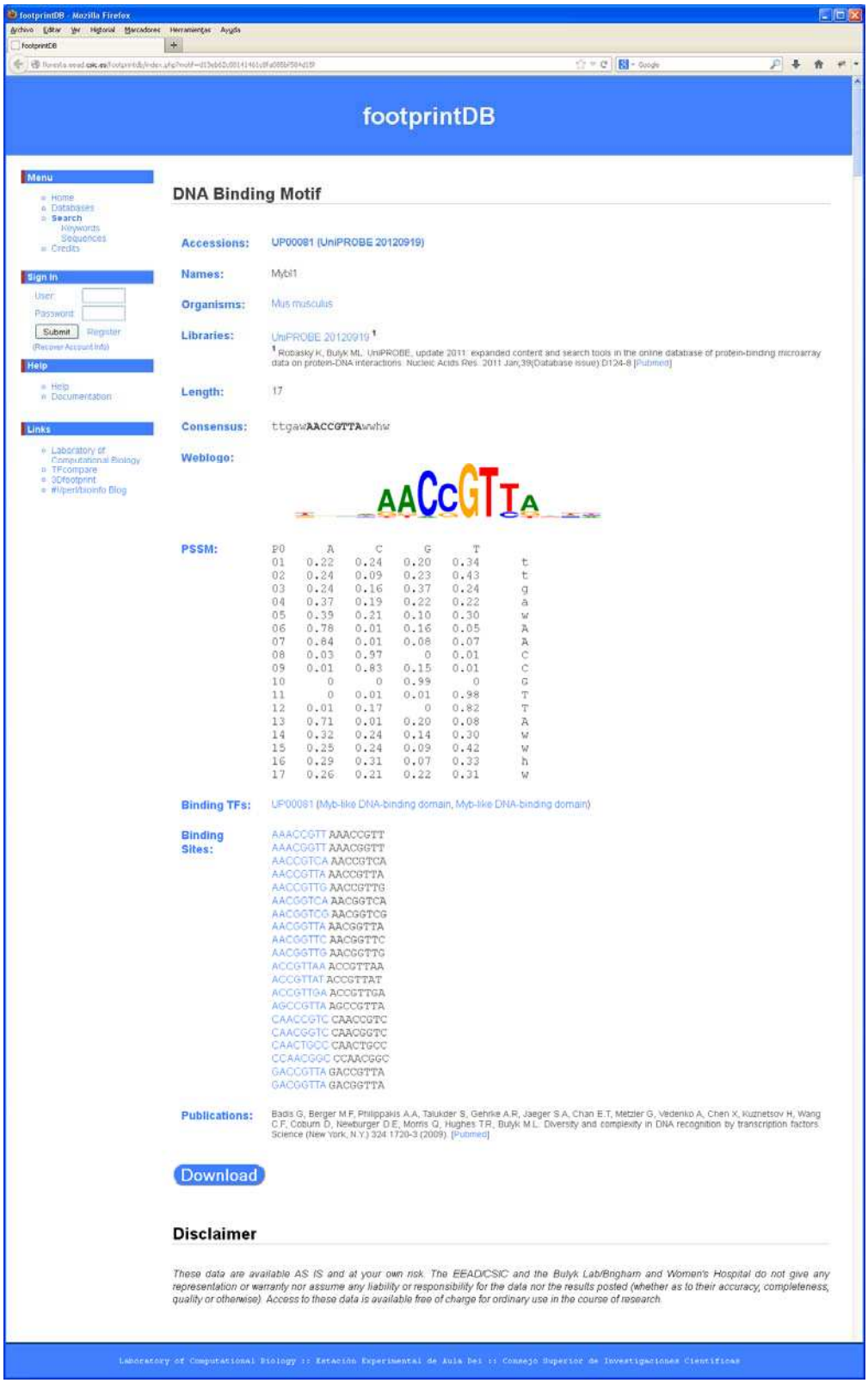
Supplementary figure S5. Benchmark of different HMMER versions and parameters when annotating Pfam domains within a non redundant set of representative transcription factor sequences. (A) Summary of overall results. (B) Number of retrieved domains for different transcription factor families with different settings.

	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	1	0	0	0	0	0	0	1	0	1	1	0	1	0	1	0	0	0	0	1
R	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
N	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
G	1	0	0	0	0	0	0	1	0	1	1	0	1	0	1	0	0	0	0	1
H	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0
I	1	0	0	0	0	0	0	0	0	1	1	0	1	0	1	0	0	0	0	1
L	1	0	0	0	0	0	0	1	0	1	1	0	1	0	1	0	0	0	0	1
K	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
M	1	0	0	0	0	0	0	1	0	1	1	0	1	0	1	0	0	0	0	1
F	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0
P	1	0	0	0	0	0	0	1	0	1	1	0	1	0	1	0	0	0	0	1
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
W	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0
Y	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0
V	1	0	0	0	0	0	0	1	0	1	1	0	1	0	1	0	0	0	0	1

Supplementary figure S6. Interface similarity scoring matrix based on an approximate physicochemical classification of amino acids, frequently used for column coloring by multiple alignment viewers.



Supplementary figure S7. Homepage of footprintDB. The main menu is on the left side.



Supplementary figure S8. Example of visualization of motif Mybl1.

SUPPLEMENTARY TABLES

Supplementary table S1. Number of TF sequences by main species and source in footprintDB. TRANSFAC data are included as a reference.

				footprintDB							
Organisms	**Total TFs**	**footprintDB**	**TRANSFAC**	**JASPAR CORE**	**3D-footprint**	**HumanTF**	**UniPROBE**	**RegulonDB**	**Athamap**	**DBTBS**	**DrosophilaTF**
Homo sapiens	1541	730	855	77	204	447	3	0	0	0	0
Mus musculus	1323	694	794	301	79	81	284	0	0	0	0
Saccharomyces cerevisiae	368	285	179	175	39	0	90	0	0	0	0
Rattus norvegicus	357	54	307	19	35	0	0	0	0	0	0
Drosophila melanogaster	242	184	100	121	34	0	0	0	0	0	57
Arabidopsis thaliana	151	71	102	5	11	0	0	0	58	0	0
Escherichia coli	135	134	1	0	52	0	0	82	0	0	0
Gallus gallus	132	9	125	6	3	0	0	0	0	0	0
Xenopus laevis	104	7	97	3	4	0	0	0	0	0	0
Bacillus subtilis	87	87	0	0	17	0	0	0	0	70	0
Caenorhabditis elegans	58	48	17	21	5	0	22	0	0	0	0
Bos taurus	36	2	34	2	0	0	0	0	0	0	0
Girella zebra	35	0	35	0	0	0	0	0	0	0	0
Danio rerio	35	0	35	0	0	0	0	0	0	0	0
Sus scrofa	20	0	20	0	0	0	0	0	0	0	0
Oryctolagus cuniculus	19	5	15	5	0	0	0	0	0	0	0
Mycobacterium tuberculosis	18	18	0	0	18	0	0	0	0	0	0
Mesocricetus auratus	18	1	17	0	0	0	0	0	0	0	0
Zea mays	14	7	14	6	0	0	0	0	3	0	0
Nicotiana tabacum	13	4	13	1	0	0	0	0	4	0	0
Oryza sativa	12	2	12	0	0	0	0	0	2	0	0

Supplementary table S2. Abundance of the most important Pfam DNA-binding domains from TFs annotated in footprintDB. TRANSFAC data are included as a reference.

					footprintDB							
Domain	**Full Description**	**Total TFs**	**footprintDB**	**TRANSFAC**	**JASPAR CORE**	**3Dfootprint**	**HumanTF**	**UniPROBE**	**RegulonDB**	**Athamap**	**DBTBS**	**DrosophilaTF**
PF00046	Homeobox domain	1152	655	635	264	64	185	171	0	4	0	14
PF00096, PF13894, PF13465	Zinc finger C2H2 type	434	219	255	86	32	57	40	0	1	0	13
PF00105	Zinc-finger C4 type	406	148	277	24	78	40	5	0	0	0	4
PF00010	Helix-loop-helix DNA-binding domain	345	142	225	40	28	46	30	0	2	0	1
PF00104	Ligand-binding domain of nuclear hormone receptor	303	44	277	23	6	13	0	0	0	0	4
PF00170, PF07716	bZIP transcription factor (basic region leucine zipper)	309	111	224	27	43	28	8	0	8	0	1
PF05920	Homeobox KN domain	215	132	108	46	18	34	40	0	0	0	1
PF00505	HMG (high mobility group) box	180	93	113	34	18	22	24	0	0	0	2
PF00178	Ets-domain	178	85	99	15	18	33	22	0	0	0	1
PF09011	Domain of unknown function (DUF1898)	178	93	111	34	18	22	24	0	0	0	2
PF00250	Fork head domain	136	62	82	21	13	21	8	0	0	0	2
PF00157	Pou domain - N-terminal to homeobox domain	117	49	75	12	13	14	10	0	0	0	0
PF00172	Fungal Zn(2)-Cys(6) binuclear cluster domain	116	88	50	41	20	0	30	0	0	0	0
PF03165	MH1 domain	90	14	77	2	8	3	1	0	0	0	0
PF00320	GATA zinc finger	77	42	47	14	11	4	9	0	5	0	2
PF00554	Rel homology domain (RHD)	77	34	44	5	23	4	0	0	0	0	3
PF00319	SRF-type transcription factor (DNA-binding and dimerisation domain)	75	35	49	8	16	5	3	0	5	0	0
PF00292	'Paired box' domain	69	23	51	6	4	10	3	0	0	0	1
PF00870	P53 DNA-binding domain	68	2	67	1	0	1	0	0	0	0	0
PF00847	AP2 domain	67	19	56	1	2	0	3	0	14	0	0
PF02198	Sterile alpha motif (SAM)/Pointed domain	64	22	45	7	0	15	0	0	0	0	0
PF00249	Myb-like DNA-binding domain	59	32	39	12	0	2	7	0	15	0	0
PF07710	P53 tetramerisation motif	59	1	59	1	0	0	0	0	0	0	0
PF00412	LIM domain	56	21	43	18	0	3	0	0	0	0	1
PF00605	Interferon regulatory factor transcription factor	56	29	31	7	10	7	5	0	0	0	0

Supplementary table S3. DNA motif internal redundancy for each footprintDB data source and TRANSFAC.

		Total Motifs	Non Redundant Motifs (E-value E-10)	% Redundant Motifs (E-value E-10)	Non Redundant Motifs (E-value E-5)	% Redundant Motifs (E-value E-5)
	footprintDB	3662	2755	25%	2246	39%
	TRANSFAC	2163	1700	21%	1369	37%
footprintDB	**JASPARCORE**	1312	1075	18%	810	38%
	3Dfootprint	802	748	7%	586	27%
	HumanTF	818	493	40%	468	43%
	UniPROBE	415	312	25%	276	33%
	RegulonDB	82	81	1%	66	20%
	Athamap	84	77	8%	64	24%
	DBTBS	88	87	1%	83	6%
	DrosophilaTF	61	61	0%	56	8%

Supplementary table S4. Internal redundancy of TF sequences for each footprintDB data source and TRANSFAC.

		Total TFs	Non Redundant Tfs (CDHIT90)	% Redundant Tfs (CDHIT90)	Non Redundant Tfs (CDHIT50)	% Redundant Tfs (CDHIT50)
	footprintDB	2422	1337	45%	1041	57%
	TRANSFAC	2919	1660	43%	945	68%
footprintDB	**JASPARCORE**	715	692	3%	621	13%
	3Dfootprint	605	236	61%	165	73%
	HumanTF	528	399	24%	247	53%
	UniPROBE	401	387	3%	247	38%
	RegulonDB	82	82	0%	80	2%
	Athamap	74	73	1%	60	19%
	DBTBS	70	70	0%	69	1%
	DrosophilaTF	57	56	2%	54	5%

Supplementary table S5. External redundancy of DNA motifs for each pair of footprintDB data sources. The main diagonal shows the total number of motifs of each source.

MOTIF REDUNDANCE (IDENTICAL)

footprintDB	footprintDB	TRANSFAC	JASPARCORE	3Dfootprint	HumanTF	UniPROBE	RegulonDB	Athamap	DBTBS	DrosophilaTF
footprintDB	**3662**	476	1312	802	818	415	82	84	88	61
TRANSFAC	476	**2163**	438	0	0	10	0	28	0	0
JASPARCORE	1312	438	**1312**	0	0	0	0	0	0	0
3Dfootprint	802	0	0	**802**	0	0	0	0	0	0
HumanTF	818	0	0	0	**818**	0	0	0	0	0
UniPROBE	415	10	0	0	0	**415**	0	0	0	0
RegulonDB	82	0	0	0	0	0	**82**	0	0	0
Athamap	84	28	0	0	0	0	0	**84**	0	0
DBTBS	88	0	0	0	0	0	0	0	**88**	0
DrosophilaTF	61	0	0	0	0	0	0	0	0	**61**

MOTIF REDUNDANCE (EVALUE<=E-10)

footprintDB	footprintDB	TRANSFAC	JASPARCORE	3Dfootprint	HumanTF	UniPROBE	RegulonDB	Athamap	DBTBS	DrosophilaTF
footprintDB	**3662**	1531	1295	446	818	412	81	84	76	57
TRANSFAC	2111	**2163**	963	78	577	401	8	49	5	30
JASPARCORE	2299	1332	**1312**	60	536	359	5	21	3	20
3Dfootprint	672	154	89	**802**	96	22	4	5	3	7
HumanTF	1453	651	386	45	**818**	174	4	11	1	14
UniPROBE	1086	628	386	7	265	**415**	3	5	2	6
RegulonDB	116	13	10	4	8	8	**82**	0	5	0
Athamap	130	108	22	6	14	3	0	**84**	1	0
DBTBS	94	6	4	6	1	3	3	1	**88**	0
DrosophilaTF	142	64	31	6	34	14	0	0	0	**61**

MOTIF REDUNDANCE (EVALUE<=E-5)

footprintDB	footprintDB	TRANSFAC	JASPARCORE	3Dfootprint	HumanTF	UniPROBE	RegulonDB	Athamap	DBTBS	DrosophilaTF
footprintDB	**3662**	2119	1311	686	818	412	82	84	84	60
TRANSFAC	3276	**2163**	1295	476	815	412	73	80	68	57
JASPARCORE	3229	2062	**1312**	449	807	408	70	70	61	53
3Dfootprint	2768	1482	947	**802**	652	300	47	52	45	39
HumanTF	2740	1631	1050	339	**818**	339	50	52	45	47
UniPROBE	2380	1529	950	184	676	**415**	52	37	32	37
RegulonDB	983	452	351	105	237	159	**82**	12	23	14
Athamap	731	588	271	83	178	75	13	**84**	17	10
DBTBS	718	300	222	115	163	75	35	13	**88**	11
DrosophilaTF	1230	714	452	107	376	201	16	8	10	**61**

Supplementary table S6. External redundancy of TF sequences for each pair of footprintDB data sources. The main diagonal shows the total number of TFs of each source.

TF REDUNDANCE (IDENTICAL)

		footprintDB	TRANSFAC	JASPARCORE	3Dfootprint	HumanTF	UniPROBE	RegulonDB	Athamap	DBTBS	DrosophilaTF
footprintDB	**footprintDB**	**2422**	418	715	605	528	401	82	74	70	57
	TRANSFAC	418	**2919**	368	1	3	37	0	38	0	16
	JASPARCORE	715	368	**715**	0	0	27	0	10	0	28
	3Dfootprint	605	1	0	**605**	0	1	0	0	1	0
	HumanTF	528	3	0	0	**528**	43	0	0	0	0
	UniPROBE	401	37	27	1	43	**401**	0	0	0	0
	RegulonDB	82	0	0	0	0	0	**82**	0	0	0
	Athamap	74	38	10	0	0	0	0	**74**	0	0
	DBTBS	70	0	0	1	0	0	0	0	**70**	0
	DrosophilaTF	57	16	28	0	0	0	0	0	0	**57**

TF REDUNDANCE (90%)

		footprintDB	TRANSFAC	JASPARCORE	3Dfootprint	HumanTF	UniPROBE	RegulonDB	Athamap	DBTBS	DrosophilaTF
footprintDB	**footprintDB**	**2422**	1632	714	604	528	401	82	74	70	57
	TRANSFAC	1545	**2919**	536	328	370	348	0	44	0	32
	JASPARCORE	1427	1062	**715**	226	232	328	0	11	0	28
	3Dfootprint	829	406	96	**605**	92	54	14	1	3	9
	HumanTF	1000	769	187	181	**528**	176	0	0	0	0
	UniPROBE	987	685	341	113	240	**401**	0	0	0	0
	RegulonDB	123	0	0	41	0	0	**82**	0	0	0
	Athamap	76	47	11	1	0	0	0	**74**	0	0
	DBTBS	74	0	0	5	0	0	0	0	**70**	0
	DrosophilaTF	76	34	28	20	0	0	0	0	0	**57**

TF REDUNDANCE (50%)

		footprintDB	TRANSFAC	JASPARCORE	3Dfootprint	HumanTF	UniPROBE	RegulonDB	Athamap	DBTBS	DrosophilaTF
footprintDB	**footprintDB**	**2422**	2351	715	605	528	401	82	74	70	57
	TRANSFAC	1747	**2919**	572	410	442	365	0	51	0	40
	JASPARCORE	1678	1705	**715**	358	354	362	0	14	0	31
	3Dfootprint	1011	832	152	**605**	234	145	15	6	4	22
	HumanTF	1184	1319	251	303	**528**	242	0	1	0	7
	UniPROBE	1195	1095	376	231	341	**401**	0	1	0	19
	RegulonDB	125	0	0	43	0	0	**82**	0	0	0
	Athamap	82	68	11	6	1	2	0	**74**	0	0
	DBTBS	79	0	0	10	0	0	0	0	**70**	0
	DrosophilaTF	91	51	29	68	15	46	0	0	0	**57**

A

```
>bZIP910
gATGACGTggca
gATGACGTggcc
gATGACGTggca
gATGACGTggcc
gATGACGTggca
gATGACGTggcc
gATGACGTggcc
gATGACGTggca
gATGACGTggcc
gATGACGTggca
gATGACGTgtcc
gATGACGTggca
gATGACGTggcc
gATGACGTggca
ggtggATGACGT
gcatgCTGACGT
gtatgCTGACGT
gcatgCTGACGT
gacggCTGACGT
gggtgCTGACGT
gggagCTGACGT
gggagCTGACGT
gcatgCTGACGT
gcatgCTGACGT
gggtgCTGACGT
gggtgCTGACGT
gtatgCTGACGT
gggtgCTGACGT
gggtgCTGACGT
gggcgCTGACGT
gGTGACGTggcc
gGTGACGTggca
gGTGACGTggca
gGTGACGTggca
gggtcGTGACGT
```

B

```
>bZIP910
ATGACGT
ATGACGT
ATGACGT
ATGACGT
ATGACGT
ATGACGT
ATGACGT
ATGACGT
ATGACGT
ATGACGT
ATGACGT
ATGACGT
ATGACGT
ATGACGT
ATGACGT
CTGACGT
CTGACGT
CTGACGT
CTGACGT
CTGACGT
CTGACGT
CTGACGT
CTGACGT
CTGACGT
CTGACGT
CTGACGT
CTGACGT
CTGACGT
CTGACGT
CTGACGT
CTGACGT
GTGACGT
GTGACGT
GTGACGT
GTGACGT
GTGACGT
```

Supplementary table S7. Experimentally obtained DNA-binding sites of transcription factor bZIP910 from snapdragon (*Antirrhinum majus*) in FASTA format. (A) Original sites from JASPAR CORE with core nucleotides in uppercase. (B) Aligned and trimmed sites to use as input in footprintDB search.