BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 2 2014, pages 258–265
doi:10.1093/bioinformatics/btt663

Databases and ontologies

Advance Access publication November 14, 2013

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, Department of Genetics and Plant Production, Estación Experimental de Aula Dei/CSIC, Av. Montañana 1005, Zaragoza (http://www.eead.csic.es/compbio) and [2]Fundación ARAID, Paseo María Agustín 36, Zaragoza, Spain

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Traditional and high-throughput techniques for determining transcription factor (TF) 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 and systematically annotates the binding interfaces of the corresponding TFs. The first release contains 2422 unique TF sequences, 10 112 DNA binding sites 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 and implementation: Web site implemented in PHP, Perl, MySQL and Apache. Freely available from http://floresta.eead.csic.es/footprintdb.

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

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 4, 2013; revised on October 22, 2013; accepted on November 8, 2013

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 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 articles 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 nuclear magnetic resonance 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 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,

*To whom correspondence should be addressed.

the binding interfaces of the corresponding TFs. Furthermore, we survey the redundancy of all included databases and compare them with TRANSFAC, a subscription-based commercial alternative. Besides allowing users to compare DNA sequences/motifs with records in the database, as most included repositories do, footprintDB can also interrogate complete proteomes 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 as follows: (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 that 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 nuclear magnetic resonance complexes, that have been independently validated (AlQuraishi and McAdams, 2011; Lin and Chen, 2013). The remaining databases and repositories currently integrated in footprintDB are as follows:

1. JASPAR CORE (2009 version, all species redundant set): a high-quality collection of TF DNA binding preferences, modeled as PSSMs (Portales-Casamar *et al.*, 2010).
2. UniPROBE (Universal PBM Resource for Oligonucleotide Binding Evaluation, Sep 2012 version): it contains *in vitro* DNA binding specificities of proteins measured with universal protein binding microarrays (Robasky and Bulyk, 2011).
3. ‘HumanTF’: a 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: a genome-wide map of potential TF binding sites in *Arabidopsis thaliana* (Bulow *et al.*, 2009).
5. RegulonDB (7.5 version): it 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 *B.subtilis* (Sierro *et al.*, 2008).
7. ‘DrosophilaTF’: motifs for 56 *Drosophila melanogaster* TFs 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 TFs, 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 Figure 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 article 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 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 Figure 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 TF 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 Figure 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 Figure 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, whereas 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 versus B and B versus 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 Figure 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 of aligned matrix columns. Motif similarity is defined as the sum of column Pearson Correlation Coefficient 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, whereas 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 to benchmark the predictive ability of footprintDB: (i) *A.thaliana* data from Athamap; (ii) *E.coli* data from RegulonDB and (iii) 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 $E \leq 1E\text{-}5$), 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 percent sequence identity >90.

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

3 RESULTS

3.1 Database contents

The first release of footprintDB contains 2422 unique TF sequences, 3662 PSSMs and 10 112 DBSs. As we added the contents of TRANSFAC version 2012.1 to the analysis, these numbers increased significantly to 4923, 5349 and 21 988, respectively (see Table 1). In the next section, redundancy analyses are performed to fairly evaluate the richness of each data source.

The most populated species among footprintDB sources, as compared with TRANSFAC and JASPAR CORE, are shown in Figure 1A, together with the corresponding number of TF sequences annotated in each data source (full statistics are reported as Supplementary Material). 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 shallow coverage. Moreover, the analysis unveils that some organisms are better covered by TRANSFAC (for instance *Gallus gallus*, *Rattus norvegicus* or *Xenopus laevis*), whereas others, such as bacteria, are only considered by open-access libraries such as DBTBS. In the Section 2.7, we assess to what extent well-explored species can help make predictions on the remaining.

The most frequent DNA binding domains found in the analyzed databases are also summarized in Figure 1B (see also Supplementary Material). It can be seen that Homeobox and

Table 1. Number of unique TF sequences, DNA motifs and DBSs in footprintDB sources

Source	TFs	Motifs	Sites
TRANSFAC	2919	2163	11 949
footprintDB	2422	3662	10 112
JASPAR CORE	715	1312	2388
3D-footprint	605	802	722
HumanTF	528	818	0
UniPROBE	401	415	2963
RegulonDB	82	82	1862
Athamap	74	84	84
DBTBS	70	88	1234
DrosophilaTF	57	61	863
Total unique	4923	5349	21 988

Note: TRANSFAC data are included as a reference. Second row corresponds to the union of all individual data sources that contribute to footprintDB.

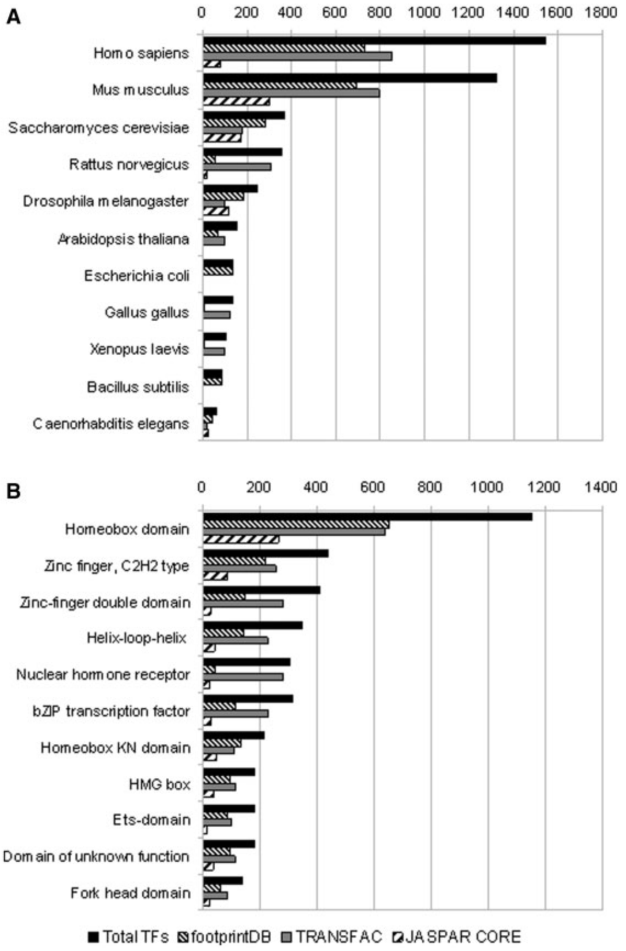

Fig. 1. Species and domain composition of footprintDB entries compared with TRANSFAC and JASPAR CORE. (**A**) Most represented species among database entries. (**B**) Most frequent DNA binding domains

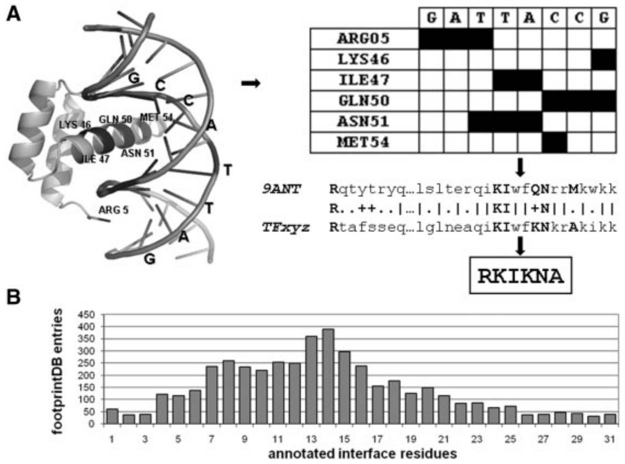

Fig. 2. Annotation of interface residues. (**A**) 3D-footprint interface of PDB entry 9ANT, which corresponds to Homebox protein Antennapedia in complex with a *cis* element. First, interatomic distances are calculated among atoms of amino acid side chains and nitrogen bases. Second, a matrix of interacting residues and their target bases is generated. Third, interface residues are marked as uppercase 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

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 TFs. This annotation relies on the 3D-footprint database, which routinely dissects the interfaces of protein–DNA complexes deposited at the PDB, as explained in Figure 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 7–16 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 Material). 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.

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 Section 2. 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 cutoff *E*-values used 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 percent 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 $E \leq$ 1E-10, which we considered as a cutoff for nearly identical motifs. Besides, these analyses confirm the minimum redundancy between bacterial

Table 2. External redundancy of DNA motifs across data sources integrated in footprintDB

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

[a] A redundant version of JASPAR was tested.

Note: The main diagonal shows the total number of PSSMs in each source (bold). Motifs aligned with STAMP E-values < 1E-10 were called redundant. Second 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.

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 Figure 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 and Supplementary Tables S5 and S6). They share 22% of motifs and 14% of TRANSFAC TFs (Fig. 3A and D). If nearly identical DNA motifs ($E \leq$ 1E-10) and TFs (with percent sequence identity $\geq$90) are considered, these percentages increase to 71% of motifs and 56% of TFs (Fig. 3B and E). Further relaxing these thresholds to $E \leq$ 1E-5 (short motifs or motifs that share a common pattern but not the whole motif) and sequence identity $\geq$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 and F). Such data overlap is also noticeable for JASPAR.

3.3 Web site

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 Web sites. From this table it is possible to browse TFs, DNA binding motifs (PSSMs) and 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) Key word 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 Figure 4. The search engine is designed primarily to receive two types of queries: (i) a DNA consensus motif, PSSM or site and (ii) a protein sequence of a putative DNA binding protein. Therefore, two kinds of output will be produced, respectively, (i) a list of DNA binding proteins predicted to bind a similar DNA motif and (ii) 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 homologs 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

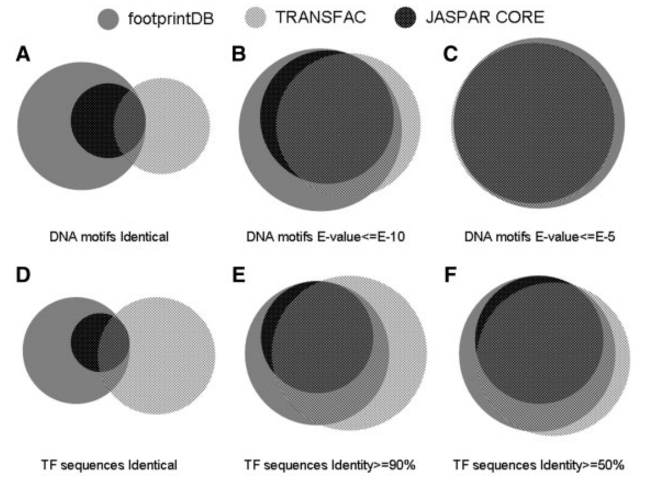

Fig. 3. Euler diagrams representing redundancy for DNA motifs (**A–C**) and TF sequences (**D–F**) annotated in footprintDB, TRANSFAC and JASPAR CORE databases

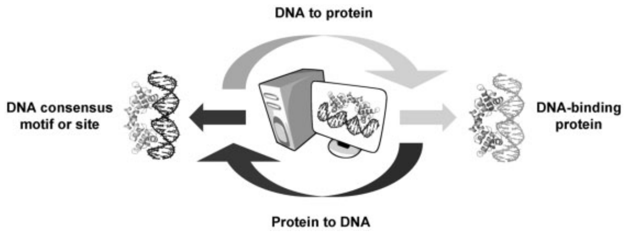

Fig. 4. Main search types supported by footprintDB. Light arrow (top): if the input data are a DNA sequence or motif, the search is powered by STAMP, and the output are proteins likely to bind sequence similar to the input. These proteins might be primary entries in footprintDB or rather endogenous TFs 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

DNA binding residues. This kind of search is useful 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.

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 *A.thaliana*, and want to identify the endogenous TFs that might recognize them, so that we can test them in the laboratory. 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), using 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 *A.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 laboratory 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*. Figure 5A and B 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 with overall values of 58 and 78%, respectively, as shown in Figure 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.

The second benchmark consisted in scanning *E.coli* sequences from RegulonDB against footprintDB+TRANSFAC, after excluding all *E.coli* records. In total, only 9 of 82 tested TFs (12%) were successfully retrieved in the TF search, and 8 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 ~69 compared with 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 of 100 tested TFs, and 90 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

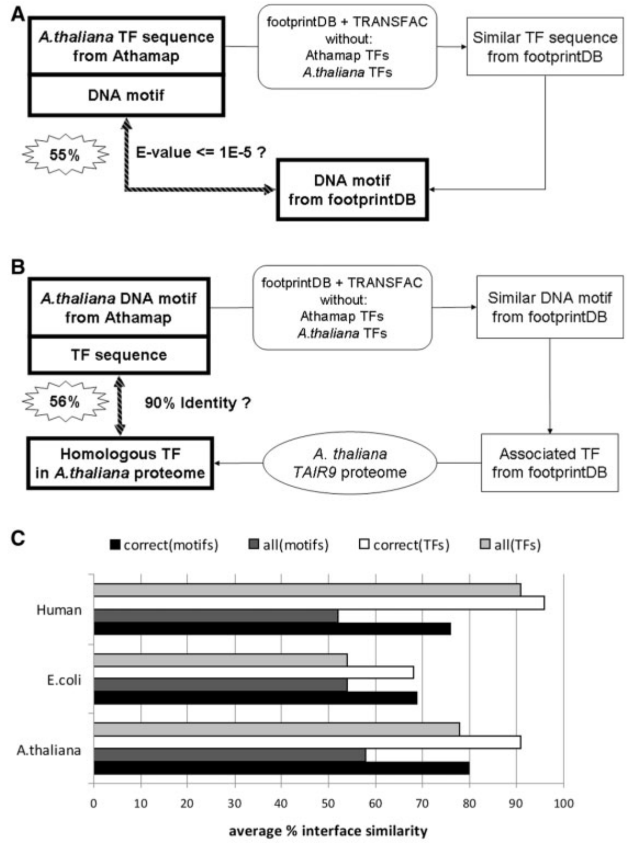

Fig. 5. Benchmark of footprintDB performance using *A.thaliana* data annotated in Athamap. (**A**) One TF sequence from Athamap is searched against footprintDB. (B) A DNA motif from Athamap is scanned against footprintDB. In both cases, red arrows represent the comparison of predictions to the cognate sequences, which are taken to be correct or false in terms of STAMP *E*-value (A) and percent sequence identity (**B**), as explained in Section 2. In the TF experiment (A), 31 of 56 Athamap sequences (55%) were successfully recovered. In the corresponding experiment with DNA motifs (B), the footprintDB pipeline recovered 27 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

or JASPAR. In both experiments, average interface similarities of recovered TFs were 86 and 96% compared with 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 with 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 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 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*, *A.majus*, *Hordeum vulgare*, *Solanum lycopersicum*, *Daucus carota* and *Oryza sativa*. In contrast, the *E.coli* test had limited success, as the only available reference organism was *B.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), whereas other articles 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 Figure 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. Inspection of the resulting clusters suggests that most redundant TFs at 90% of sequence identity are probably inparalogs and orthologs from phylogenetically close organisms. Inspection of relaxed clusters (50% sequence identity cutoff) unveils that they gain more divergent homologous proteins of the same family, which bind to regulatory elements using the same Pfam domains.

Although 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

The authors 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. They 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: 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.

Conflict of Interest: none declared

REFERENCES

AlQuraishi,M. and McAdams,H.H. (2011) Direct inference of protein-DNA interactions using compressed sensing methods. *Proc. Natl Acad. Sci. USA*, **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–D986.

Contreras-Moreira,B. (2010) 3D-footprint: a database for the structural analysis of protein-DNA complexes. *Nucleic Acids Res.*, **38**, D91–D97.

Contreras-Moreira,B. *et al.* (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. *et al.* (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.

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. *et al.* (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.*, **41**, W523–W530.

Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, W253–W258.

Matys,V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

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. *et al.* (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–D110.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

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–D161.

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–D128.

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–D213.

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. *et al.* (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. *et al.* (1981) Directed mutagenesis. *Ann. Rev. Genet.*, **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–D96.

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. *Nat. Rev. Genet.*, **10**, 252–263.