footprintDB: analysis of plant *cis* regulatory elements, transcription factors and binding interfaces

Bruno Contreras-Moreira^{1,2}, Alvaro Sebastian³

¹ Laboratory of Computational Biology, EEAD/CSIC, Zaragoza, Spain

² Fundación ARAID, Zaragoza, Spain

³ Evolutionary Biology Group, Adam Mickiewicz University, Poznan, Poland

Corresponding author: Bruno Contreras-Moreira bcontreras@eead.csic.es

Abstract FootprintDB is a database and search engine that compiles regulatory sequences from open access libraries of curated DNA *cis*-elements and motifs, and their associated transcription factors (TFs). It systematically annotates the binding interfaces of the TFs by exploiting protein-DNA complexes deposited in the Protein Data Bank. Each entry in footprintDB is thus a DNA motif linked to the protein sequence of the TF(s) known to recognize it, and in most cases, the set of predicted interface residues involved in specific recognition. This chapter explains step-by-step how to search for DNA motifs and protein sequences in footprintDB and how to focus the search to a particular organism. Two real-world examples are shown where this software was used to analyze transcriptional regulation in plants. Results are described with the aim of guiding users on their interpretation, and special attention is given to the choices users might face when performing similar analyzes.

Keywords: Bioinformatics, transcription factor, DNA binding, DNA motif, PSSM, protein domain, promoter, *cis*-element, database, open-access

Running head: Exploring plant transcriptional regulation with footprintDB

1. Introduction

Transcription is a central process in gene expression. It is modulated primarily by the binding of regulatory proteins called transcription factors (TFs) to short DNA sequences, called *cis*-regulatory elements. DNA recognition is a flexible mechanism, since most TFs can usually distinguish a collection of non-identical DNA binding sites (DBSs), which in turn define a DNA-binding motif (DBM). Position-specific scoring matrices (PSSMs) are a common way of representing DBMs, which in their simplest form tally the observed nucleotide frequencies at each position of the motif (1). DBMs are also frequently plotted as sequence logos, which graphically summarize the binding preferences of TFs (2), as shown on **Figure 1**. These are convenient models that hide some known complexities of TFs but are still useful. For instance, columns in a DBM might be correlated and thus not accurately modelled by a PSSM, which usually features independent columns. Moreover, the alignment of *cis*-elements in order to derive motifs has several pitfalls: 1) both DNA strands must be considered; 2) short DBSs are easily misaligned, particularly if structural constraints are not considered (3); and 3) mismatches are common due to TF binding plasticity. Nevertheless, because of their simplicity, PSSMs are usually the preferred representation of protein-DNA binding models.

Figure 1. Typical representations of DNA motifs. (A) Multiple alignment of DNA binding sites recognized by a TF, usually *cis*-elements located in different promoters. The motif core is in upper-case. (B) Position-specific scoring matrix in TRANSFAC-like notation. The last column is the consensus. (C) Sequence logo, with base heights proportional to conservation across sites.

Experimental methods to identify DBSs are technically challenging and have been traditionally limited to determining *cis*-regulatory sites for one TF at a time. Among such protocols are DNA footprinting, chromatin immunoprecipitation (ChIP) or electrophoretic mobility shift assays, which yield high quality data despite their low throughput (4-6). These approaches are being replaced by higher throughput protocols such as protein binding microarrays, HT-SELEX, ChIP-chip or ChIP-Seq techniques (7-10). These procedures often produce large volumes of raw sequence data, which must be pre-processed and filtered in order to derive DBMs employing a variety of recipes (11,12). Eventually, resulting PSSMs are collected and annotated in databases.

In addition, a number of algorithms have been developed in order to predict and annotate DBSs within genomic sequences. Some of them try to discover them *de novo* by detecting overrepresented DNA motifs (13-15). Others use previously known experimental DNA binding data to localize similar regions in genomes by sequence and PSSM alignments or machine learning techniques (16-19).

Figure 2. DNA-binding interface of PDB complex 9ANT (Antennapedia in complex with a *cis*-element) as annotated by 3D-footprint. Inter-atomic distances are calculated among atoms of amino acid side chains and nitrogen bases, and a matrix of interactions generated. Interface residues, in upper case, are extracted to show the DNA-binding interface core.

Other experimental approaches focus on characterizing the interface residues of TFs, those in charge of recognizing the nucleotide bases of DBSs (Figure 2). Besides site-directed

mutagenesis (20,21), 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), and can also be exploited to infer structure-based DBMs (22-27). An exhaustive analysis of these complexes shows that individual DNA-binding proteins typically bind a nucleotide segment 3 to 10 bases long (**Figure 3A**). However, TFs usually identify target sites in conjunction with other proteins. For this reason, biologically relevant motifs most often correspond to protein multimers or multi-domain TFs, which bind longer, contiguous regions in the DNA sequence as shown in **Figure 3B** for two plant TFs.

Figure 3. (A) Histogram of the length of *cis* elements recognized by non-redundant monomeric proteins in TF-DNA complexes deposited in the PDB and annotated in 3D-footprint as of July, 2015. (B) Sequence logos of experimentally determined binding sites of plant transcription factors LEAFY (top) and VRN1 (bottom) (available in footprintDB). Note that both PSSMs contain highly conserved sub-motifs with interleaved degenerate sequences.

To facilitate the analysis of DNA-protein interactions, researchers can take advantage of *in silico* tools for designing experiments and engineering DNA-recognition. In this chapter we describe footprintDB, a database that compiles experimental data of thousands of TFs and their DNA motifs. FootprintDB has two main applications: 1) the prediction of TFs able to bind novel DNA motifs and 2) the prediction of DNA motifs for uncharacterized TFs. The first search type is illustrated with two protocols and their application to a real research problem concerning transcription factors predicted to regulate a set of co-expressed *Arabidopsis thaliana* promoters. For the second kind of query a generic protocol is also presented and then applied to the study of a stress-related promoter sequence in rice.

2. Materials

The only resources required to replicate the analysis described in this chapter are an Internet connection and a web browser. These will suffice to learn how to use footprintDB and associated tools, which are now presented.

2.1. FootprintDB

FootprintDB is a meta-database that integrates several open access repositories of curated *cis*-elements, DNA motifs and TFs into a unique repository (<http://floresta.eead.csic.es/footprintdb>, **Figure 4**) (28). The May 2015 release includes the following databases: JASPAR, 3D-footprint, HumanTF, HOCOMOCO, FlyZincFinger, UniPROBE, ArabidopsisPBM, AthaMap, DBTBS, RegulonDB, DrosophilaTF and EEADannot. JASPAR is the reference source of curated TFs and DNA motifs derived from published collections of experimentally defined DBS for eukaryotes (29). 3D-footprint annotates *cis* elements captured in protein-DNA complexes deposited in the PDB (26). UniPROBE (Universal PBM Resource for Oligonucleotide Binding Evaluation) hosts data generated by universal protein binding microarrays (PBM) with proteins from a diverse collection of organisms (30). The remaining repositories provide experimentally-supported DBMs for specific organisms and taxa, such as HumanTF and HOMOCOMO for human (31,32), RegulonDB and DBTBS for bacteria (33,34), Athamap and ArabidopsisPBM for *A. thaliana* (35,36), or DrosophilaTF and FlyZincFinger for fruit fly (37,38). Finally, EEADannot is a manually curated set of plant data compiled in our laboratory.

Figure 4. FootprintDB web interface. (A) Main page. (B) Search form.

Available DNA-binding data for plant TFs is scarce, and for this reason Athamap and ArabidopsisPBM collections, as well as EEADannot, are valuable resources for plant promoter analysis and *cis*-element discovery. We note that commercial database TRANSFAC annotates also a repertoire of plant DNA motifs (39), but a subscription fee is required. Other valuable plant-specific resources such as PLACE and AGRIS were considered. However, while the former contains single DBSs without annotated binding TFs, most of the data in the latter are already annotated by other resources like JASPAR and Athamap (40,41).

FootprintDB handles redundant data deposited in several repositories by annotating unique entries with multiple references to the original sources. The underlying database can model complex scenarios in which a single TF binds to several DBMs or where the same *cis*-element is targeted by multiple TFs.

All in all, footprintDB currently contains 3,095 unique TFs, 4,646 PSSMs and 18,840 DBSs (July 2015). Each data entry can be searched by name, identifier, sequence or other descriptors and visualized in a data sheet format including footprintDB annotations (see examples in **Figure 5**). Of these, the current release contains 275 non-redundant DNA motifs from plant TFs, which

Figure 6. The two main queries supported by the footprintDB search engine.

Users can also upload collections of TFs and DNA motifs into footprintDB, keeping the data for private use (default) or making it publicly available. The footprintDB search engine can also be accessed programmatically using a SOAP web services interface. These and more possibilities are further explained in the software documentation.

Only few organisms have available experimental TF-DNA binding data in the literature. For this reason, footprintDB searches can be extended to third-party organism proteomes, allowing search results to be extrapolated to a particular organism of interest. For example, we can use a known *A.thaliana* TF that might bind conserved *cis*-elements in promoters of over-expressed genes and then find its homologues in the proteome of *Oryza sativa*, if we are interested in conducting an experiment with rice. According to our benchmarks, homologous TFs from different species are more likely to bind a similar DNA motif when their interfaces are similar (28) .

In order to evaluate the contents of footprintDB, the initial 2013 release was compared to subscription-based TRANSFAC 2012.1, and a high degree of redundancy was found among TFs and DBMs stored in both repositories, as detailed in **Note 2 (28)**.

TF protein sequences have their DNA interfaces annotated in footprintDB. That means that residues involved in the recognition of nucleotide bases are predicted based on homology to proteins that take part on three-dimensional protein-DNA complexes annotated in the 3D-footprint database (see **Note 3**, <http://floresta.eead.csic.es/3dfootprint>) (26). FootprintDB lists the PDB accessions for these complexes, making it possible to produce structure-based alignments of *cis* elements and to compare binding interfaces with the TFcompare web server (see **Note 4**) (3). DNA binding protein domains are also annotated by scanning sequences against PFAM domains (46). Check **Figure 5** for the content of a typical entry of the database.

3. Methods

3.1. DNA motif search in footprintDB

The first run mode of footprintDB takes a DNA motif, *cis* element or DNA site as input. This kind of search is useful when we have data about DNA sequences (or a single sequence) recognized by an unknown DNA-binding protein and we want to predict TFs able to interact with them. In the next lines we explain how to feed a DNA motif or site in footprintDB and how to interpret the results using as example the *Antirrhinum majus* motif called ‘bZIP910’ (**Figure 5B**), originally annotated in the JASPAR database (29,47).

- 1) First, if you have a footprintDB account, log in to store your searches and reuse them.
- 2) Click on the ‘Start Search’ button or in the ‘Search Sequences’ link on the left menu at <http://floresta.eead.csic.es/footprintdb>.
- 3) Enter a name for the job and, optionally, an email address if you desire to receive the results by email.
- 4) Choose as input type ‘DNA sites or motifs’ and the number of results that you desire in ‘Limit number of results per query’ field.

- 5) Enter your DNA sites or motifs in the text area or upload them from a file. The accepted formats are FASTA and TRANSFAC (some simplifications of the original TRANSFAC format are also accepted). In the present example we will use the example data by choosing 'bZIP910 DNA Binding Motif in TRANSFAC-like format' and clicking the 'Demo' button. A matrix with 'bZIP910' TF binding preferences will be shown automatically in the text area.
- 6) To start the search, press the 'Search' button.

Additionally, options on the footprintDB search form such as 'Order results by' or 'Color results using twilight thresholds' can be changed (*see Note 5*). Search can also be limited by organism, source database or Pfam domains (*see Note 6*). Clicking on 'Search for homologues in a selected proteome' displays additional parameters which will be explained later in the **Section 3.2**.

Among the obtained results, shown in **Figure 7**, we notice that the first one is the query itself (the demo 'bZIP910' is a regular footprintDB entry), together with a list of transcription factors that are reported to bind very similar DNA motifs. Each row contains additional information about the motif alignment (E-value and similarity scores, *see Notes 5 and 7*), source organism and links to the original source, the motif datasheet in footprintDB, the annotated binding proteins, their binding interface residues and their binding protein domains. By clicking on 'Show interfaces' and 'Show domains' in the result table it can be seen that the results share a common DNA-binding interface (R[KL]x[SQK]NR[ev][SA]Axx[SCA]RxRK) within the Basic Leucine Zipper domain. Note that predicted binding residues are in uppercase in the consensus.

footprintDB results for Demo								
Query: bZIP910 (JASPAR CORE) mTGACGT								
footprintDB template	Source	Organisms	STAMP e-value	Motif similarity	footprintDB PWM / Consensus	Binding proteins	Interface sequences	Pfam domains
MA0096: bZIP910	JASPAR 2014	<i>Antirrhinum majus</i>	1.0e-12	7.00 / 7	ACGTCAk ACGTCAk	Show proteins	Show interfaces	Show domains
TGA2_2: TGA2	ArabidopsisPBM 20140210	<i>Arabidopsis thaliana</i>	3.0e-10	6.97 / 7	-ACGTCAk-- kACGTCAkCa	Show proteins	Show interfaces	Show domains
MF0002: bZIP CREB/IG-box-like subclass	JASPAR 2014	METAMODEL	8.9e-10	5.99 / 6	ACGTCAk ACGTCA-	METAMODEL		
CREM_f1: CREM	HOCOMOCO v9	<i>Homo sapiens</i> <i>Mus musculus</i>	1.8e-09	6.78 / 7	-ACGTCAk-- GACGTCAbys	Show proteins	Show interfaces	Show domains
XBP1_DBD_2: XBP1	HumanTF 1.0	<i>Homo sapiens</i>	9.6e-09	6.85 / 7	-----ACGTCAk- wrkGmCACGTCAkc	Show proteins	Show interfaces	Show domains
XBP1_DBD_1: XBP1	HumanTF 1.0	<i>Homo sapiens</i>	1.1e-08	6.76 / 7	----ACGTCAk- GaTGACGTCAtc	Show proteins	Show interfaces	Show domains
MA0588: TGA1	JASPAR 2014	<i>Arabidopsis thaliana</i>	1.4e-08	6.64 / 7	--ACGTCAk-- hyACGTCAbsm	Show proteins	Show interfaces	Show domains
TGA2: TGA2	ArabidopsisPBM 20140210	<i>Arabidopsis thaliana</i>	1.8e-08	6.52 / 7	---ACGTCAk rTGACGTCAy	Show proteins	Show interfaces	Show domains
MA0286: CST6	JASPAR 2014	<i>Saccharomyces cerevisiae</i>	2.6e-08	6.37 / 7	--ACGTCAk tkACGTCAy	Show proteins	Show interfaces	Show domains
TGA1: TGA1	Athamap 20091028	<i>Arabidopsis thaliana</i>	2.8e-08	6.65 / 7	--ACGTCAk--- t aaCGTCAbsaw	Show proteins	Show interfaces	Show domains

Figure 7. Example of results for the ‘bZIP910’ DNA motif search in footprintDB. Note that slightly different versions of *A. thaliana* motifs TGA 1 and 2 are reported. Results can vary in future versions of footprintDB.

3.2. Proteome-specific DNA search in footprintDB

As illustrated in the previous section, a regular DNA search can give us valuable information about a novel DNA *cis*-element or motif by comparing it to similar motifs annotated in footprintDB. However, often we are interested in finding a list of transcription factors that most likely bind that DNA motif in a specific organism. For example, we might know an abiotic stress *cis*-element in the *Antirrhinum majus* genome (bZIP910) and we want to test rice TFs that potentially bind this sequence. For this purpose, we can extend the footprintDB search by selecting a specific target proteome:

- 1) Repeat the steps 1-5 of DNA motif search explained on **Section 3.1**.

- 6) Click on the link 'Search for homologues in a selected proteome' to expand proteome search options and select from the list of available proteomes 'Oryza sativa - MSU6.1' or upload a FASTA file with the desired proteome.
- 7) Click on the 'Search' button to start the search.

Results look like a regular DNA motif search, but footprintDB entries with homologous proteins within the rice proteome are shown on the top of the list, and entries without rice homologues at the bottom with the legend 'NO HITS'. Click on the link 'Show Oryza sativa - MSU6.1 homologues' displayed below high-ranking results on the leftmost column to expand a list of significant BLASTP hits (homologues) found in the proteome. Each one contains information about the binding interface and alignment scores with the footprintDB TF sequence. In the bZIP910 example, rice homologues show conserved interface residues and common DNA-binding domains as expected. Following these results, experiments can now be designed to test whether these TFs actually bind a bZIP910-like motif in our conditions of interest, thus reducing the search space from around fifty thousand rice transcripts to a few dozens of footprintDB predictions.

3.3. Protein sequence search in footprintDB

For most putative TFs in sequence databases there is little or no information regarding their DNA binding preferences. In these cases we can look for homologous TFs annotated in footprintDB and transfer their curated DNA binding data. To illustrate this type of search in footprintDB we will take the amino acid sequence of bZIP910 TF from *A. majus* (**Figure 5A**), extending the examples of the previous two sections.

- 1) Repeat the steps 1-3 listed on **Section 3.1**.
- 4) Choose as input type 'Proteins' and set the wished number of results in the 'Limit number of results per query' box.

- 5) Enter the protein sequences in the text area or upload them from a file in FASTA format. In the present example we will use the demo data by choosing “bZIP910 Protein Sequence in FASTA format” and clicking the ‘Demo’ button. The protein sequence will be shown automatically in the text area.
- 6) Press the ‘Search’ button.

As previously, search can be limited by organism, source database or Pfam domains (*see Note 6*). Additionally results can be ordered by E-value (*see Note 7*) or interface similarity (*see Note 8*). As expected, the first reported TF is bZIP910 and all remaining TFs belong to the same ‘bZIP domain’ family, share similar interface residues and recognize G-box (CACGTG) and C-box (GACGTC) motifs related to the cognate bZIP910 consensus: a G-box/C-box hybrid (GACGTG) (**29,47**). As illustrated on **Section 3.2**, an organism specific proteome can be chosen or uploaded to retrieve homologous proteins only from desired species.

3.4. *In silico* prediction of transcription factors for co-expressed gene promoters in *A. thaliana*

Let us review a recent study where 32 co-expressed drought-responsive *A.thaliana* gene clusters were analyzed (**48**). One of the aims of this work was the identification of transcription factors involved in drought response regulation and their experimental validation. This is an application of the protocol in **Section 3.2**. To achieve this goal, gene expression profiles in a large *A.thaliana* microarray set were clustered and upstream sequences (1Kb from transcription start site) of genes in each cluster searched for significantly overrepresented short DNA sequences. The chosen tool for motif discovery was BEST (<http://www.people.fas.harvard.edu/~junliu/BEST>), a meta-predictor which combines different motif-finding programs (**49**). Out of 179 motifs discovered, 15 putative regulatory sequences were selected to scan a large library of *A. thaliana* TFs -called REGIA (*see Note 9*) - in a yeast one-hybrid experiment. Here the first of these sequences (*shoots1hr_9*, CTCCACGTGC) is

further used to demonstrate the performance of footprintDB when looking in the *A. thaliana* proteome for TFs binding similar DNA targets:

- 1) Repeat the steps 1-4 of DNA motif search explained on **Section 3.1**. Set “Limit number of results per query” to 100. Paste the DNA sequence (*see Note 10*).
- 5) Select from the list of available proteomes ‘Arabidopsis thaliana - regia’ and set the BLASTP E-value threshold to 1E-10 (*see Note 7*) in order to reproduce the search strategy of the work of Dubos *et al.* Note that TRANSFAC 2012.1 was also used as a database on that project, but unfortunately that repository cannot be offered to footprintDB users due to license restrictions.
- 6) Click on the ‘Search’ button to launch the job.

If the specific proteome search option is turned off, the first plant TF is ranked 10th (MYC4 from JASPAR database, but it can vary in future versions). However, when the REGIA proteome is selected MYC4 is ranked first on the list of results, together with other TFs that in the preliminary search appeared lower in the rankings (*see Note 11*). Homologous TFs from REGIA can be visualized by expanding the link ‘Show Arabidopsis thaliana - regia homologues’, with binding interfaces highlighted. In this example, the top MYC4 homologues share the interface motif [ns]HV[ev]AE[rk][qr]RRekl(X)₁₂[vi][st][kr]Mdk.

Beyond this example, an important result of the Dubos *et al.* study (**48**) was the systematic comparison of computer predictions and the Y1H results for all 15 *cis*-elements under study. First, it turned out that *in silico* TF predictions failed to correctly identify binding proteins whenever footprintDB contained no proteins with significantly similar DNA motifs (STAMP E-value > 1E-3, *see Note 7*). In other words, footprintDB can successfully predict TFs for input DNA motifs only if significantly similar motifs are already annotated in the databases. Second, in five cases where this occurred, footprintDB included the experimentally determined TFs among the predictions. However, it also incorporated a number of TFs which are false positives.

In summary, as long as the interface is similar, footprintDB will retrieve TFs from the same family, even if they are not expressed or they do not bind the *cis* element under the studied experimental conditions. **Figure 8** illustrates the agreement between *in silico* and yeast one-hybrid TF predictions for *shoots1hr_9* as in the original article by Dubos et al., 2014 (including TRANSFAC search results).

Figure 8. Comparison of footprintDB *in silico* predictions of binding TFs and yeast one-hybrid experimental results within a library of cloned *A.thaliana* TFs for the *cis* element ‘*shoots1hr_9*’. (A) Example TF predictions by both methods, with common ones in green. (B) Chart comparing all unique and common TF predictions obtained with both methods.

3.5. Prediction of OsEREBP1 and OsEREBP2 regulatory sites within the OsRMC promoter

In this last example, taken from Serra et al. (50), we show how footprintDB can be used to identify target *cis*-elements of a TF of interest. This is an application of the protocol in **Section 3.3**. In that work some experiments were performed to unveil the regulation of rice gene OsRMC under high salinity conditions. Thus, a salt-induced rice cDNA expression library was constructed and subsequently screened using the yeast one-hybrid system and the OsRMC promoter as bait. As a result, OsEREBP1 and OsEREBP2, two putative TFs of the AP2/ERF family, were identified to bind a 648bp region of the OsRMC promoter. In a first approach, the OsRMC promoter sequence was scanned with DNA motifs in the PlantPAN database, which included curated data from PLACE, TRANSFAC 7, AGRIS, and JASPAR 2008 databases (51,41,40,52,53). Several potential DBSs were identified in the promoter window, but none of

them contained the GCC box typical of ERF specificity. This section explains how footprintDB was used to design subsequent experiments which confirmed the precise locations of GCC-like *cis*-elements in this promoter:

- 1) Visit UniProt (<http://www.uniprot.org>) and download the amino acid sequences of rice proteins OsEREBP1 and OsEREBP2, which correspond to accessions Q6K7E6 and Q5N965 respectively. OsEREBP2 has only been partially annotated and for that reason appears as ‘AP2 domain transcription factor-like’. Scroll down the respective UniProt pages, locate the ‘Sequence’ section and download the sequences in FASTA format.
- 2) Repeat the steps 1-4 of **Section 3.3**.
- 6) Paste the downloaded sequences in FASTA format, name the job and select “Athamap” and “3D-footprint” as databases (*see Note 6*).
- 7) Press the ‘Search’ button.

The best candidate motif retrieved from Athamap and 3D-footprint are ERF4 (plus other ERF-like TFs) and 1gcc_A, respectively. Both are related motifs that contain the GCC box and their cognate TFs have all their interface residues conserved. OsEREBP1, OsEREBP2, ERF4 and 1gcc_A have the following common consensus interface: [iv]R[qk]RpWg[kr]xaaEiRdp(x)₄₅RvWlgt. The DNA motifs of ERF4 and 1gcc_A TFs were subsequently used to locate the most likely *cis*-elements within the OsRMC promoter using tools from the RSAT::Plants web server.

- 1) Click on ERF4 and 1gcc_A links in the ‘footprintDB PWM’ column from the previous results and download their PSSMs in TRANSFAC-like format.
- 2) Download the OsRMC gene promoter region from the original Serra et al. paper (**50**) (*see Note 12*).
- 3) Go to RSAT::Plants server (<http://plants.rsat.eu>), and select from the left menu: ‘Pattern-matching – matrix-scan (quick)’.

- 4) Paste the OsRMC promoter sequence in FASTA format and the two motif matrices selecting TRANSFAC format from the menu.
- 5) Select 'Background model estimation method - Organism-specific' and 'Oryza sativa IRGSP' with sequence type 'upstream-noorf'. Leave other parameters with default values.
- 6) Click on 'GO' button to start the calculations.

Among the obtained results, the *cis*-sequence TGCCTGCTC, found by both input motifs with -481,-473 coordinates, showed binding activity to OsEREBP1 and OsEREBP2 proteins in electrophoretic mobility shift assays (EMSA) (50).

3.6 Conclusions and perspectives

In this chapter we have presented some examples and real case studies in plants of the possibilities of footprintDB, a unified and open-access online database designed for the analysis of transcription factors and their genomic DNA targets. The main value of footprintDB is probably the integration of a variety of libraries of curated DNA motifs and their associated TFs, which have been increased and updated since the original publication, and in addition, the systematic annotation of interfaces residues of the corresponding TFs. FootprintDB has an open source philosophy, encouraging scientists to contribute with their DNA-binding data to the expansion of the database.

The footprintDB search engine allows querying the database for unknown TFs that are likely to bind input DNA motifs, and also the opposite (**Figure 6**). *In silico* predictions can save time and money when designing laboratory experiments to probe TF DNA binding specificities as in the study cases shown. Search results can be valuable as reported or taken for further analysis in external tools such as RSAT::Plants, TFcompare, etc.

While footprintDB stores data for a variety of organisms, here we have demonstrated typical use cases on plants, reviewing real-world problems that we have encountered with our collaborators. We hope the examples described here can aid other users with related research problems.

4. Notes

1. Due to the drastic increase of available genomes and to improve maintenance and update tasks, RSAT has recently divided in taxon-specific servers, one of them plant specific (42).

2. TRANSFAC (BIOBASE) is a subscription database with curated annotations of transcription factors, experimentally-proven binding sites and the corresponding PSSMs (39). Additionally, TRANSFAC contains annotation of miRNA and their target sites, together with functional annotations of TFs, predicted promoter binding sites and additional software tools for DBS prediction and discovery. A comparison between TRANSFAC version 2012.1 and footprintDB initial version showed a high degree of data redundancy between both databases, which shared around 71% of motifs (STAMP E-value $\leq 1E-10$) and 56% of TFs (% sequence identity ≥ 90). Additionally, some internal redundancy was detected in both databases, accounting for 20-25% of DBMs and 43-45% of TFs. These values indicate that a large proportion of the underlying experimental studies focus on a small number of regulation-related protein families, and that probably there are still many families and *cis* elements to be discovered.

3. 3D-footprint is a database which provides estimates of binding specificity for all protein–DNA complexes available at the Protein Data Bank (27). Each complex in the database is dissected to draw interface graphs and footprint logos, and two complementary algorithms are employed to characterize binding specificity. Moreover, oligonucleotide sequences extracted from literature abstracts are reported in order to show the range of variant sites bound by each protein and other related proteins. 3D-footprint is updated and curated on a weekly basis.

4. If three dimensional structures of two DNA-protein complexes of the same family are available, it is possible to compare their binding interfaces by a structure-based alignment. This can be done by feeding TFcompare server (<http://floresta.eead.csic.es/tfcompare>) with the PDB identifiers from two protein-DNA complexes (3). It first extracts individual DNA-binding protein domains to calculate their optimal fit and then returns their structural alignments. The

superposition of protein chains is used to generate the structure-based alignment of the bound DNA sequences. As a result, nucleotides that are recognized by equivalent interface residues in both complexes are aligned together. The resulting DNA alignment does not rely in the nucleotide sequences and differs in many cases from a pure sequence alignment. This kind of alignment makes sense only for binding domains, not for the whole protein, and for this reason the original structures are trimmed according to their PFAM-defined domain boundaries. All DNA-contacting domains from the first structure are aligned to those in the second and produced alignments are scored in terms of the number of identical superposed nucleotides and the sum of N9 (nitrogen 9 in purines) and N1 (nitrogen 1 in pyrimidines) atom pairs within 3.5 Å.

5. Results are sorted by default by STAMP E-value, but can also be sorted on computed DNA similarity. By default rows in the results table are colored after the alignment quality thresholds reported previously (3). Motifs returned to the user are shown with a green background when the obtained score is above the twilight zone; otherwise the background is in red. While the twilight cut-offs were calculated on a large set of alignments, it is still possible that a correct alignment turns out in red.

6. ‘Multiple Organisms’ can be selected by pressing the Ctrl key, as well as ‘Original Databases’ and ‘Pfam domains’. The option of restricting the search by “Organisms” should be used with caution because some TFs and DNA motifs are not associated to a specific species.

7. The Expect value (E-value) is an estimate of the number of false positive results we can expect by chance when searching a database. For example, an E-value of 1 can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance. The lower the E-value, or the closer it is to zero, the more probable that a match is a true positive and not a random result. Short sequences have a higher probability of occurring in the database purely by chance, that is why short DNA searches usually yield larger E-values than protein ones.

8. Interface similarity can be useful to score candidate protein sequences within a proteome of interest. This variable takes values from 0 to 100% and captures the sequence similarity of residues possibly involved in specific DNA recognition. In our benchmarks with human, *A.thaliana* and *E.coli* TFs (28), the interface similarity of correct predictions was significantly higher ($\approx 70\%$) than the similarity of all aligned TFs ($\approx 50\%$).

9. REGIA (REgulatory Gene Initiative in Arabidopsis) was an EU-funded project involving 29 European laboratories with the objective of determining the function of virtually all transcription factors from the model plant *Arabidopsis thaliana* (54). The REGIA consortium provided a normalized full size TF library (more than 800 full length ORFs cloned) available to the scientific community for screening for additional interactions, particularly with non-TF proteins.

10. In the experiment by Dubos et al (48) two types of input were tested: single *cis*-elements and PSSMs compiled from a group of aligned *cis*-elements. In this context single elements yielded fewer false positives, although our benchmarks suggest that PSSMs produce more accurate alignments than individual sequences (3). We therefore recommend trying both inputs if possible.

11. Looking for homologous proteins in *A. thaliana* (or any other target species) helps filtering results, as it favors higher rankings for TFs with homologues in the chosen proteome, which are more likely to be true predictions.

12. OsRMC corresponds to gene OS04T0659300 of the International Rice Genome Sequencing Project (IRGSP, <http://rgp.dna.affrc.go.jp/IRGSP>). The promoter fragment used in the original Serra *et al.* paper can be retrieved with help from RSAT::Plants (<http://plants.rsat.eu>). On the left menu select 'Sequence tools – retrieve sequence', set 'Organism' to 'Oryza sativa IRGSP', choose 'Gene' selection and paste 'OS04T0659300-01'. Leave all other options with default values except 'From' and 'To', which should be set to -1321 and -674, respectively.

Acknowledgments

We would like to thank our colleagues C. Dubos, L Bülow, N. Saibo, T. Serra and J. van Helden for past and current collaborations. This work was funded by grant Euroinvestigación EUI2008-03612 under the framework of the Transnational (Germany, France, Spain) Cooperation within the PLANT-KBBE Initiative.

References

1. Stormo GD (2000) DNA binding sites: representation and discovery. *Bioinformatics* 16 (1):16-23
2. Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 18 (20):6097-6100
3. Sebastian A, Contreras-Moreira B (2013) The twilight zone of cis element alignments. *Nucleic Acids Res* 41 (3):1438-1449. doi:10.1093/nar/gks1301
4. Galas DJ, Schmitz A (1978) DNase footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* 5 (9):3157-3170
5. Garner MM, 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 (13):3047-3060
6. O'Neill LP, Turner BM (1996) Immunoprecipitation of chromatin. *Methods Enzymol* 274:189-197
7. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000) Genome-wide location and function of DNA binding proteins. *Science* 290 (5500):2306-2309. doi:10.1126/science.290.5500.2306
8. Berger MF, Bulyk ML (2006) Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol Biol* 338:245-260. doi:10.1385/1-59745-097-9:245
9. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316 (5830):1497-1502. doi:10.1126/science.1141319
10. Ogawa N, Biggin MD (2012) High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro. *Methods Mol Biol* 786:51-63. doi:10.1007/978-1-61779-292-2_3
11. Machanick P, Bailey TL (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 27 (12):1696-1697. doi:10.1093/bioinformatics/btr189
12. Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J (2011) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res*. doi:10.1093/nar/gkr1104
13. Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15 (7-8):563-577. doi:btc069 [pii]
14. Bailey TL, Williams N, Mischel C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* 34 (Web Server issue):W369-373. doi:34/suppl_2/W369 [pii]

10.1093/nar/gkl198

15. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* 32 (4):1372-1381. doi:10.1093/nar/gkh299

16. Chen QK, Hertz GZ, Stormo GD (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput Appl Biosci* 11 (5):563-566

17. Mahony S, Auron PE, Benos PV (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput Biol* 3 (3):e61. doi:06-PLCB-RA-0200R3 [pii]

10.1371/journal.pcbi.0030061

18. Turatsinze JV, Thomas-Chollier M, Defrance M, van Helden J (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat Protoc* 3 (10):1578-1588. doi:10.1038/nprot.2008.97

19. Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 14 (1):48-54

20. Shortle D, DiMaio D, Nathans D (1981) Directed mutagenesis. *Annual review of genetics* 15:265-294. doi:10.1146/annurev.ge.15.120181.001405

21. O'Neill M, Dryden DT, Murray NE (1998) Localization of a protein-DNA interface by random mutagenesis. *EMBO J* 17 (23):7118-7127. doi:10.1093/emboj/17.23.7118

22. Morozov AV, Havranek JJ, Baker D, Siggia ED (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res* 33 (18):5781-5798. doi:33/18/5781 [pii]

10.1093/nar/gki875

23. Alamanova D, Stegmaier P, Kel A (2010) Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. *BMC Bioinformatics* 11:225. doi:10.1186/1471-2105-11-225

24. Contreras-Moreira B, Collado-Vides J (2006) Comparative footprinting of DNA-binding proteins. *Bioinformatics* 22 (14):e74-80. doi:22/14/e74 [pii]

10.1093/bioinformatics/btl215

25. Angarica VE, Perez AG, Vasconcelos AT, Collado-Vides J, Contreras-Moreira B (2008) Prediction of TF target sites based on atomistic models of protein-DNA complexes. *BMC bioinformatics* 9:436. doi:1471-2105-9-436 [pii]
10.1186/1471-2105-9-436
26. Contreras-Moreira B (2010) 3D-footprint: a database for the structural analysis of protein-DNA complexes. *Nucleic Acids Res* 38 (Database issue):D91-97. doi:gkp781 [pii]
10.1093/nar/gkp781
27. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28 (1):235-242
28. Sebastian A, Contreras-Moreira B (2014) footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics* 30 (2):258-265. doi:10.1093/bioinformatics/btt663
29. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B, Sandelin A, Wasserman WW (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* 42 (Database issue):D142-147. doi:10.1093/nar/gkt997
30. Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res* 43 (Database issue):D117-122. doi:10.1093/nar/gku1045
31. Jolma A, Yan J, Whittington T, Toivonen J, Nitta Kazuhiro R, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas Juan M, Vincentelli R, Luscombe Nicholas M, Hughes Timothy R, Lemaire P, Ukkonen E, Kivioja T, Taipale J (2013) DNA-Binding Specificities of Human Transcription Factors. *Cell* 152 (1):327-339
32. Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, Makeev VJ (2013) HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res* 41 (Database issue):D195-202. doi:10.1093/nar/gks1089
33. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernandez S, Alquicira-Hernandez K, Lopez-Fuentes A, Porron-Sotelo L, Huerta AM, Bonavides-Martinez C, Balderas-Martinez YI, Pannier L, Olvera M, Labastida A, Jimenez-Jacinto V, Vega-Alvarado L, Del

- Moral-Chavez V, Hernandez-Alvarez A, Morett E, Collado-Vides J (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res* 41 (Database issue):D203-213. doi:10.1093/nar/gks1201
34. Siervo N, Makita Y, de Hoon M, Nakai K (2008) DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res* 36 (Database issue):D93-96. doi:10.1093/nar/gkm910
35. Bulow L, Engelmann S, Schindler M, Hehl R (2009) AthaMap, integrating transcriptional and post-transcriptional data. *Nucleic Acids Res* 37 (Database issue):D983-986. doi:gkn709 [pii]
10.1093/nar/gkn709
36. Franco-Zorrilla JM, Lopez-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc Natl Acad Sci U S A* 111 (6):2367-2372. doi:10.1073/pnas.1316278111
37. Down TA, Bergman CM, Su J, Hubbard TJ (2007) Large-scale discovery of promoter motifs in *Drosophila melanogaster*. *PLoS Comput Biol* 3 (1):e7. doi:06-PLCB-RA-0344R2 [pii]
10.1371/journal.pcbi.0030007
38. Enuameh MS, Asriyan Y, Richards A, Christensen RG, Hall VL, Kazemian M, Zhu C, Pham H, Cheng Q, Blatti C, Brasefield JA, Basciotta MD, Ou J, McNulty JC, Zhu LJ, Celniker SE, Sinha S, Stormo GD, Brodsky MH, Wolfe SA (2013) Global analysis of *Drosophila* Cys(2)-His(2) zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants. *Genome Res* 23 (6):928-940. doi:10.1101/gr.151472.112
39. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E (2006) TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34 (Database issue):D108-110. doi:10.1093/nar/gkj143
40. Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res* 27 (1):297-300. doi:gkc003 [pii]
41. Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E (2003) AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics* 4:25. doi:10.1186/1471-2105-4-25

42. Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon J, Delerce J, Spinelli L, Jaeger S, Blanchet C, Vincens P, Caron C, Staines D, Contreras-Moreira B, Artufel M, Charbonnier L, Hernandez C, Thieffry D, Thomas-Chollier M, van Helden J (2015) RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res* 43:W50-56
43. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37 (Web Server issue):W202-208. doi:10.1093/nar/gkp335
44. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215 (3):403-410. doi:10.1006/jmbi.1990.9999
S0022283680799990 [pii]
45. Mahony S, Benos PV (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* 35 (Web Server issue):W253-258
46. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2014) Pfam: the protein families database. *Nucleic Acids Res* 42 (Database issue):D222-230. doi:10.1093/nar/gkt1223
47. Martinez-Garcia JF, Moyano E, Alcocer MJ, Martin C (1998) Two bZIP proteins from *Antirrhinum* flowers preferentially bind a hybrid C-box/G-box motif and help to define a new sub-family of bZIP transcription factors. *Plant J* 13 (4):489-505
48. Dubos C, Kelemen Z, Sebastian A, Bulow L, Huep G, Xu W, Grain D, Salsac F, Brousse C, Lepiniec L, Weisshaar B, Contreras-Moreira B, Hehl R (2014) Integrating bioinformatic resources to predict transcription factors interacting with cis-sequences conserved in co-regulated genes. *BMC Genomics* 15 (1):317. doi:10.1186/1471-2164-15-317
49. Che D, Jensen S, Cai L, Liu JS (2005) BEST: binding-site estimation suite of tools. *Bioinformatics* 21 (12):2909-2911
50. Serra TS, Figueiredo DD, Cordeiro AM, Almeida DM, Lourenco T, Abreu IA, Sebastian A, Fernandes L, Contreras-Moreira B, Oliveira MM, Saibo NJ (2013) OsRMC, a negative regulator of salt stress response in rice, is regulated by two AP2/ERF transcription factors. *Plant Mol Biol* 82 (4-5):439-455. doi:10.1007/s11103-013-0073-9

51. Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res* 36 (Database issue):D102-106. doi:gkm955 [pii]
10.1093/nar/gkm955
52. Wingender E, Karas H, Knuppel R (1997) TRANSFAC database as a bridge between sequence data libraries and biological function. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*:477-485
53. Chang WC, Lee TY, Huang HD, Huang HY, Pan RL (2008) PlantPAN: Plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups. *BMC Genomics* 9:561. doi:10.1186/1471-2164-9-561
54. Paz-Ares J, Regia C (2002) REGIA, an EU project on functional genomics of transcription factors from *Arabidopsis Thaliana*. *Comparative and functional genomics* 3 (2):102-108. doi:10.1002/cfg.146