Published online 18 September 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D91–D97***
doi:10.1093/nar/gkp781

3D-footprint: a database for the structural analysis of protein–DNA complexes

Bruno Contreras-Moreira[1,2,3,*]

[1]Estación Experimental de Aula Dei, Consejo Superior de Investigaciones Científicas, [2]Fundación ARAID, Paseo María Agustín 36, Zaragoza, Spain and [3]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico

Received August 14, 2009; Revised September 2, 2009; Accepted September 3, 2009

ABSTRACT

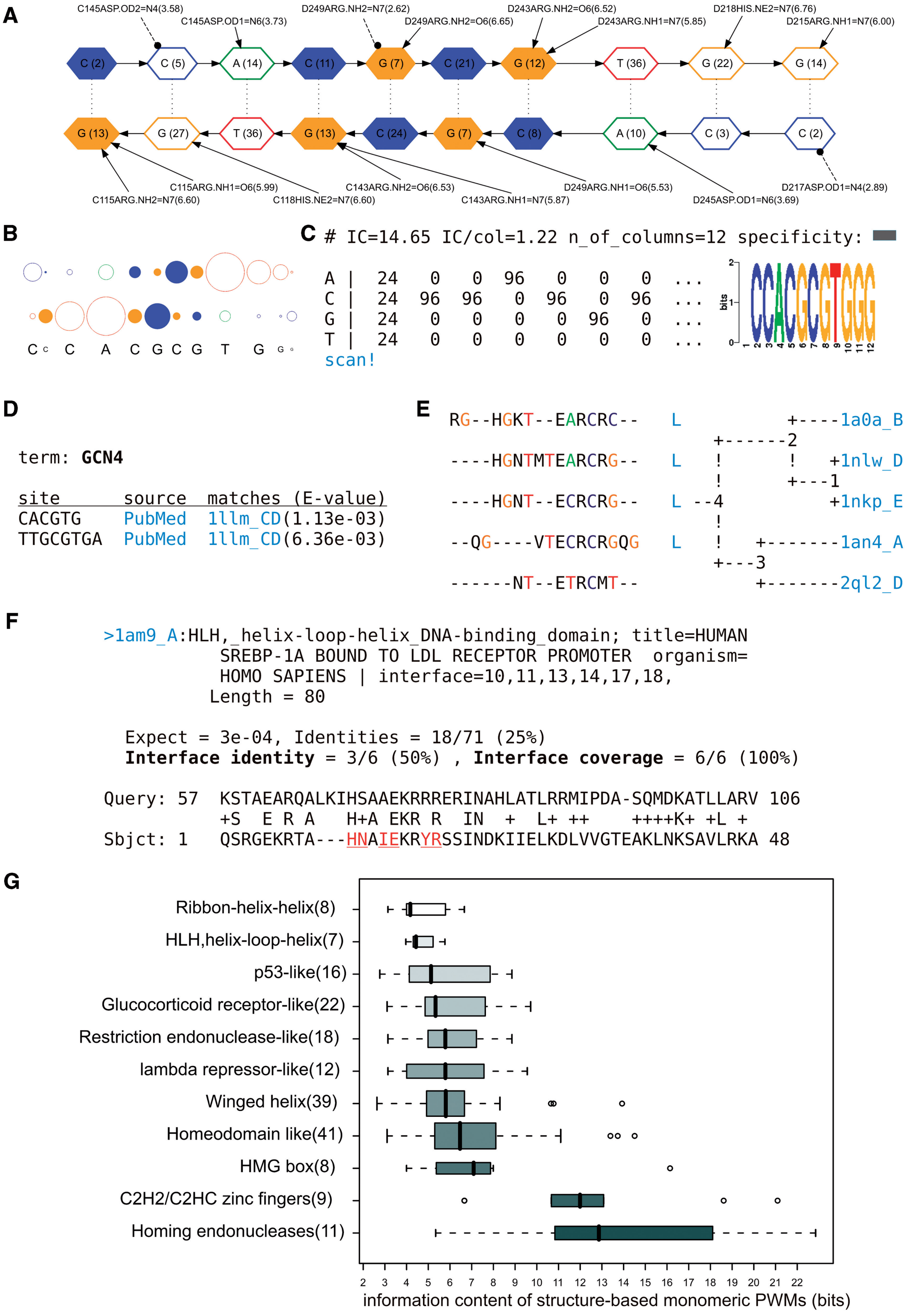
3D-footprint is a living database, updated and curated on a weekly basis, which provides estimates of binding specificity for all protein–DNA complexes available at the Protein Data Bank. The web interface allows the user to: (i) browse DNA-binding proteins by keyword; (ii) find proteins that recognize a similar DNA motif and (iii) BLAST similar DNA-binding proteins, highlighting interface residues in the resulting alignments. 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. Benchmark experiments, including comparisons with expert-curated databases RegulonDB and TRANSFAC, support the quality of structure-based estimates of specificity. The relevant content of the database is available for download as flat files and it is also possible to use the 3D-footprint pipeline to analyze protein coordinates input by the user. 3D-footprint is available at http://floresta.eead.csic.es/3dfootprint with demo buttons and a comprehensive tutorial that illustrates the main uses of this resource.

INTRODUCTION

DNA footprinting is a well-proven experimental methodology used to probe the specificity of proteins that bind DNA. The method takes its name from the footprint that bound proteins leave on electrophoresis gels after DNA digestion (1). In the same way, the structures of protein–DNA complexes, contributed by researchers worldwide and stored in the Protein Data Bank (PDB) (2), can be seen as molecular footprints at the atomic scale. The increasing collection of such complexes has encouraged computational approaches that take atomic coordinates as input to analyze sequence recognition from different perspectives (3–8). Two of the most recent approaches, the cumulative contact method (6)—which assumes that the consensus DNA sequence is captured in the experimental structure—and DNAPROT (7)—that samples oligonucleotides considering both direct and indirect readout mechanisms—are taken here to construct 3D-footprint, a weekly updated database that contains structure-based footprints and specificity estimates for all complexes available at the PDB. 3D-footprint entries are annotated according to the SCOP (9) and Pfam (10) databases, and their binding interfaces are clustered in terms of structural similarity. In addition, each complex in the database is dissected in order to draw interface graphs and footprint diagrams, a novel graphical representation that summarizes contacts across the double-stranded DNA segment at the interface. Oligonucleotides extracted from related scientific papers are also reported in order to show the range of variant sites bound by each protein and other related proteins. Furthermore, entries in the database are linked to external resources that provide valuable related information: NDB (11), PDBSum (12), ProNuc (13), NPIDB (14), BIPA (15) and the Protein Mutant Database (16). The web interface of the database, also mirrored at Universidad Nacional Autónoma de México, offers the following:

(1) an up-to-date repository of complexes that classifies complexes as non-redundant, multimeric and redundant
(2) interface graphs that plot indirect readout bases and atomic interactions responsible for direct readout
(3) footprint diagrams with bases depicted as circles of diameter proportional to the number of contacts observed at the interface, with DNA strands plotted separately

[*] To whom correspondence should be addressed. Tel: +34 976716089; Email: bcontreras@eead.csic.es



```
# IC=14.65 IC/col=1.22 n_of_columns=12 specificity:

A |  24   0   0  96   0   0   0  ...
C |  24  96  96   0  96   0  96  ...
G |  24   0   0   0   0  96   0  ...
T |  24   0   0   0   0   0   0  ...
scan!
```

```
term: GCN4

site       source  matches (E-value)
CACGTG     PubMed  1llm_CD(1.13e-03)
TTGCGTGA   PubMed  1llm_CD(6.36e-03)
```

```
RG--HGKT--EARCRC--    L          +----1a0a_B
                            +------2
----HGNTMTEARCRG--    L     !      !   +1nlw_D
                            !      +---1
----HGNT--ECRCRG--    L   --4          +1nkp_E
                            !
--QG----VTECRCRGQG    L     !   +-------1an4_A
                            +---3
------NT--ETRCMT--              +-------2ql2_D
```

```
>1am9_A:HLH,_helix-loop-helix_DNA-binding_domain; title=HUMAN
           SREBP-1A BOUND TO LDL RECEPTOR PROMOTER  organism=
           HOMO SAPIENS | interface=10,11,13,14,17,18,
          Length = 80

  Expect = 3e-04, Identities = 18/71 (25%)
  Interface identity = 3/6 (50%) , Interface coverage = 6/6 (100%)

Query: 57  KSTAEARQALKIHSAAEKRRRERINAHLATLRRMIPDA-SQMDKATLLARV 106
           +S  E R A   H+A EKR R  IN  +  L+ ++    ++++K+ +L +
Sbjct: 1   QSRGEKRTA---HNAIEKRYRSSINDKIIELKDLVVGTEAKLNKSAVLRKA 48
```

Figure 1. (A–E) Typical content of 3D-footprint entries, illustrated with dimeric complex 1llm_CD, a Zif23-GCN4 chimera (32), and with non-redundant monomeric complex 1a0a_B, positive regulatory protein PHO4 (33). (A) An interface graph dissecting atomic contacts and nucleotides at the interface responsible for specific DNA discrimination, where solid bases indicate indirect readout mechanisms. (B) Footprint logo diagram of 1llm_CD containing four central base-pairs subject to indirect readout. (C) Sequence logo and structure-based PWM obtained by averaging contact and readout PWMs for complex 1llm_CD. The calculated information content places this complex in the dark gray region of the boxplot in panel G (see below). Note that the underneath link exports this PWM to a RSAT form where the user can scan genomes of DNA fragments for occurrences

(4) structure-based position weight matrices (PWMs) that can be used to scan genomic sequences via RSAT (17)
(5) interface dendrograms that summarize the similarities between related DNA-binding proteins
(6) a keyword search tool that allows browsing the repertoire of protein–DNA complexes
(7) a protein sequence search engine that returns BLAST-like alignments which highlight interface residues
(8) a motif search facility which ranks 3D-footprint proteins that recognize similar DNA sequences to those input by the user
(9) an interactive footprint pipeline for the analysis of complexes provided by the user
(10) a miner form that retrieves DNA motifs related to a search term from PubMed abstracts
(11) a repository of flat files with all relevant data in the database available for download

The database is automatically updated by identifying new structures reported in the PDB; however, new complexes are manually curated before being added to the collection in order to spot and fix unusual problems with the coordinates—such as broken or reversed DNA strands—and to assign adequate protein names—and synonyms—that will drive the search for related motifs in scientific papers. PubMed abstracts frequently contain binding sites and consensus sequences that enrich those inferred from the molecular coordinates.

Benchmark experiments have been carried out in order to validate 3D-footprint specificity measurements, including a comparison with two expert-curated databases, RegulonDB (18) and TRANSFAC (19). The tests confirm that structure-based estimations of binding specificity correlate well with those derived from consensus alignments produced by biocurators. In addition, it is found that DNA-binding superfamilies display different specificities, in agreement with measurements obtained from TRANFAC motifs.

With respect to other databases, 3D-footprint occupies a unique position as to my knowledge it is currently the only up-to-date database providing structurally derived DNA motifs, in the form of PWMs, and sequence alignments that highlight interface residues responsible for DNA recognition. Moreover, the web interface allows custom analyses of protein–DNA complexes provided by the user.

DATA, METHODS AND IMPLEMENTATION

3D-footprint consists of several modules, which are built with custom-made programs, mostly written in Perl—making use of the CPAN module DB_File—and C++, plus third-party software. Figures B1 and B6 of the 'Benchmark' section at http://floresta.eead.csic.es/3dfootprint/benchmark.html explain how these modules integrate with each other; they are now described in more detail.

A non-redundant set of protein–DNA complexes

The PDB database is mirrored once a week, including the 95% non-redundant list of protein chains. For each cluster in the list, the best resolution chain with most protein–DNA contacts is taken, and any other chains are regarded as redundant. Protein chains docking the same DNA molecule are taken as components of multimeric complexes, which are also reported as they usually capture the most biologically relevant molecular docks. Further redundancy filters are applied in order to derive hydrogen bond and hydrophobic atomic preference matrices that are available in the download area of the web site, as published earlier (7). Note that no distinction is made between transcription factors and any other DNA-binding proteins. As of 19 July 2009, the database contains 500 non-redundant, 677 multimeric and 1045 redundant complexes.

Clusters of interfaces

A cluster is defined as a set of DNA-binding proteins with significant structural similarity. Every monomer in the repository is structurally compared to the non-redundant set of complexes and all significant hits, with MAMMOTH (20) ln(E-value) < -7, are stacked in order to compile a multiple alignment or matrix of equivalent interface residues and bases. For alignment display purposes, only one base is linked to each interface residue, overlooking the cases in which a single residue contacts several bases. Expectation values among members of a cluster are converted to distances that support an unrooted dendrogram calculated with PHYLIP (21), as depicted in Figure 1E.

Interface graphs and footprint logo diagrams

A detailed view of the interface is provided for every complex in the form of a graph that shows the DNA molecule encoded in the PDB coordinates and a list of labeled atomic contacts that confer sequence discrimination. Contacts can be of three types: hydrogen bonds, water-mediated hydrogen bonds and hydrophobic interactions. Each contact has assigned a log-odd score, extracted from the preference matrices mentioned earlier and distance-corrected that evaluates its statistical significance. Often it is not possible to automatically classify some side-chain contacts, which

of this motif. (**D**) Examples of literature-extracted DNA sequences associated to the term 'GCN4' and their E-values, corresponding to non-redundant entry 1llm_D. (**E**) Dendrogram of similar interfaces for entry 1a0a_B, where the distance tree is based on the estimated structural similarity between binding domains and interface residues—those with 4.5Å heavy atom contacts with nitrogen bases—are aligned coloring their nucleotide partners. (**F**) Querying 3D-footprint with the protein sequence of *Zea mays* transcription factor PTZm00668.1 (34). Note that all six interface residues are covered in the alignment, but only three are conserved. (**G**) Scale of specificity observed for SCOP superfamilies in the database, computed over the parenthesized number of non-redundant of complexes, after excluding superfamilies with less than seven complexes. An up-to-date scale is available at http://floresta.eead.csic.es/3dfootprint/stats.html.

are still reported without a score. The graph itself is produced using the Graphviz library at http://www.graphviz.org and a modified version of HBPLUS (22) that handles hydrophobic contacts. The 3DNA package (23) is employed for labeling indirect readout bases, shown as solid hexagons—circles in footprint logos—as previously explained in detail (7). Interface bases also display the number of side-chain heavy atom contacts within 4.5Å, which is proportional to the diameter of nitrogen bases in footprint diagrams. Sample interface and footprint diagrams are displayed in Figure 1A and B.

PMWs (3D-footprints)

It is possible to convert the tally of side-chain contacts for every base in the DNA molecule into a column of a PMW, assuming that the coordinates file actually harbors the consensus sequence, as it is reasonable. This conversion follows the formulation of Morozov and Siggia (6) that states that bases with high contact numbers are more conserved in the consensus, as stated in previous reports (24). These are named contact PWMs, in contrast with readout PWMs, which are derived from the scores of the interface contacts that result when 4N nucleotide sequences are threaded into the backbone of the DNA molecule in the complex, of length N. These later PWMs consider both direct and indirect readout mechanisms and are calculated using the DNAPROT protocol (7). When reliable readout matrices are available—those derived from interfaces with at least five atomic interactions, see the 'Benchmark' section—a mean matrix is obtained by averaging both contact and readout PWMs. Otherwise, only contact PWMs are considered. In either case, 3D-footprints capture the binding specificity of proteins in the database and can be used to scan genomic sequences. However, not all proteins are equally specific. To guide the user, the information content (IC) of each footprint is reported as calculated by RSAT (17). IC has previously been shown to be a good approximation of binding specificity within bacterial regulatory networks (25). Sequence logos calculated with WEBLOGO (26) are more intuitive representations for PWMs, as they actually plot the IC of each motif and are calculated by taking the best B sequences as scored by the reference PWM. B is set by default to 50, but can in some cases be a smaller number *b*, if the (*b* + 1)-th site has a score that is worse than the worst single nucleotide mutation. A structure-based PWM and its corresponding sequence logo, overlaid, are shown in Figure 1C. Note that the IC of this example PWM corresponds to a highly specific binding class as shown in Figure 1G.

Motif search engine

A library of structure-based PWMs, or 3D-footprints, is maintained, encompassing both monomeric and multimeric protein–DNA complexes. This library can be searched by taking an input motif provided by the user, running STAMP (27) with appropriate background score distributions for local (JaspRand_PCC_SWU) or global (JaspRand_SSD_NW_go1000_ge1000) alignments. If the input data is originally an oligonucleotide or consensus sequence, it is first converted to a PWM based on the minimum number of sites required to convert all sequence letters to integer values.

Protein search engine

The collection of protein sequences of non-redundant monomeric complexes can be searched taking as input a protein sequence. The search engine is powered by PSI-BLAST (28) and returns similar DNA-binding proteins evaluated in terms of E-value and also in terms of interface identity, that scores the proportion of residues found to be in contact with DNA nitrogen bases that are conserved, as defined in a previous paper (29).

Consensus miner

3D-footprint includes a text-mining facility designed to extract oligonucleotide sequences, presumably consensus sequences, from the set of PubMed abstracts related to some input term, usually a protein name. This search engine builds on the Entrez eUtils toolbox at http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html and uses keywords to guide abstract retrieval. The set of keywords includes words found to be over-represented in papers reporting binding sites curated in RegulonDB 6.0 (18) and TRANSFAC 9.3 (19) and also terms identified by A.Santos-Zavaleta, curator of the RegulonDB team (see http://floresta.eead.csic.es/3dfootprint/miner.html). Non-redundant entries of the database include a table with similar DNA sequences reported in the literature, with *E*-values associated to STAMP local alignments, as shown in Figure 1D. The terms associated to non-redundant entries, which drive the PubMed search, are extracted from the corresponding title in the PDB file by running the GAPSCORE tagger (30) and are then manually curated when entries are added to the database for the first time. As of 19 July 2009, the database contains 184 PubMed reports linked to non-redundant entries.

WEB INTERFACE: EXAMPLES OF USE

This section presents a few examples of how the database can help with queries related with protein–DNA recognition. Typically, the result of a query will be a list of entries in the database, which contains several diagrams and reports as summarized in Figure 1. The possible queries and the interpretation of results are further explained in the online tutorial at http://floresta.eead.csic.es/3dfootprint/tutorial.html. The tutorial also includes sample Perl code to query the web services engine.

Text search

The user can type a single term, such as a name, a PDB identifier, a superfamily or a species that is related to a protein of interest and the server will return a list of links to relevant entries. Each entry contains a summary of the protein–DNA complex including the elements explained above, which will include a structure-based PWM if evidence for specific DNA binding is found.

Motif search

Another way of interrogating 3D-footprint is by asking: is there a protein that binds a DNA motif or sequence similar to this? For this purpose the user can simply input an oligonucleotide sequence, in which degenerate bases are allowed, or rather a PWM in CONSENSUS or TRANSFAC format. By default, searches run local motif alignments, but it is often useful trying global alignments, particularly when spaced short repeats, such as the CGGNNNNNNNNNNNCGG yeast Gal4 motif, are searched. The expectation value cutoff for motif similarity can be changed by the user.

Protein sequence search

The third basic query to 3D-footprint can be done by pasting a protein sequence. This will trigger a search for similar proteins in the database, that are likely to bind similar DNA motifs provided that they conserve the residues at the interface (6,29). A sample alignment is included in Figure 1F. The motifs of matched DNA-binding proteins are only displayed when the percent interface identity (%IID) is at least 50%. The server returns a list of complexes ranked by overall similarity, reported as a BLAST *E*-value. %IID and percentage of interface coverage values are also shown to assist in the interpretation of the alignments, which are printed in a BLAST-like format. If any similar complexes are found the server provides a link that exports the input protein sequence to the TFmodeller web server (31), in case the user wishes to build a comparative model of the input protein docked to a DNA molecule.

Analyzing a protein–DNA complex provided by the user (interactive footprint)

The pipeline of analysis is also available for interactive use. The required input is a file containing the coordinates of a protein–DNA complex in PDB format. For a correct function, it is necessary that protein and DNA atoms have different chain identifiers. This anonymous form can be used to analyze experimentally determined complexes or models generated by computational approaches. There is an option, set by default, to sample interface side-chain rotamers during the analysis, as this was found to be important during previous experiments (7). The returned results include interface diagrams and estimated binding preferences in PWM and sequence logo format.

Mining DNA motifs in PubMed abstracts

A consensus miner is bundled in the website, where the user can paste a protein name to launch a PubMed search that will return a list of short DNA sequences related with the input term. As explained in the 'Data, methods and implementation' section, this search is driven by keywords known to be associated to binding sites in prokaryotes and eukaryotes. Results are tabulated and each oligonucleotide reported has pointers to the original literature sources.

Downloading data from 3D-footprint

Most data in 3D-footprint can be obtained at the download area, which is updated every week. The collection of structure-based PWMs is distributed in CONSENSUS and TRANSFAC format, for use with motif scanning software. The set of protein sequences for the non-redundant set of monomeric complexes is also available, including a list of interface residues in the FASTA header. The set of atomic interaction matrices, used to evaluate interface contacts, and the list of complexes used to derive them, are also available. The 'Statistics' section at http://floresta.eead.csic.es/3dfootprint/stats.html includes a report, concerning the specificity of 3D-footprints, which is also updated weekly.

BENCHMARK

The 'Benchmark' and 'Stats' sections of the website describe experiments that evaluate the quality of structure-based estimations of binding specificity; this section makes references to figures included there due to space restrictions.

Comparing contact and readout PWMs

3D-footprint uses two different algorithms in order to infer binding preferences from sets of atomic coordinates, but how do they compare to each other? Figure B4, panel A, of the 'Benchmark' section indicates that contact PWM and readout PWMs, inferred with both algorithms, are more similar as the number of atomic interactions at the interface increases, as measured with STAMP local alignments. Figure B4, panel B, provides further information, as it shows that the IC of readout PWMs follows the same trend. Taken together, these results suggest that the DNAPROT method is more sensitive to the quality of input data, as it requires a minimum number of atomic interactions in order to produce PWMs with high similarity to contact PWMs, which are assumed to describe cognate sites. As a consequence, a minimum five atomic interactions were required in order to further consider readout PWMs, which according to the regression lines is equivalent to a −log(*E*-value) of 4.3 and information content of 3.3 bits. In addition, it was necessary to evaluate the specificity estimations of 3D-footprint, by means of comparison to external datasets of known quality. To this end the set of 22 transcriptions factors in 3D-footprint for which PWMs were available from RegulonDB (seven proteins from *Escherichia coli*) and TRANSFAC (15 eukaryotic proteins) were taken to plot Figure B5, panel A, that unveils a strong correlation between both measurements. Figure B5, *panel* B, shows that the mean IC values for transcription factor superfamilies in 3D-footprint match those calculated from collections of TRANSFAC motifs, although it also hints that two superfamilies—homeodomains and zinc fingers are often over-estimated. These data support the scale of specificity presented in Figure 1G, which includes additional DNA-binding protein superfamilies, such as restriction and homing endonucleases. Furthermore, Figure S3 in the 'Stats' section provides one more independent evaluation of the quality of 3D-footprint, based on the set of entries for which a PubMed report is available. The histogram shows the *E*-values obtained after comparing 3D-footprints and literature motifs by means of local STAMP alignments, demonstrating that structure-based PWMs are indeed significantly similar to the motifs published in the literature for those proteins.

CONCLUSIONS

Existing databases (13–15) have already demonstrated that structures deposited at the PDB contain a wealth of information that can be exploited in order to study the mechanisms of DNA binding in biological systems. Moreover, benchmark experiments presented in this and earlier (25) work suggest that atomic coordinates can be effectively used to compute binding specificity and justify the main mission of 3D-footprint: to enrich the repertoire of known regulatory elements with those embedded in the atomic descriptions of protein–DNA complexes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The author would like to thank Vladimir Espinosa Angarica, Carlos Rodríguez Penagos, Alberto Santos Zavaleta, Osvaldo Graña, Jacques Van Helden, Heladia Salgado and Julio Collado-Vides for their help, support, suggestions and reviews of 3D-footprint. He would also like to acknowledge Alberto Fuentes, Alberto Relancio and Víctor del Moral for their technical assistance.

FUNDING

Consejo Superior de Investigaciones Científicas [grant number 200720I038]. Funding to pay the Open Access publication charges for this article was provided by [grant A06] from Gobierno de Aragón to the research group of J.M. Lasa in 2009.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
3. Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, **35**, 114–131.
4. Paillard,G. and Lavery,R. (2004) Analyzing protein-DNA recognition mechanisms. *Structure (Camb)*, **12**, 113–122.
5. Morozov,A.V., Havranek,J.J., Baker,D. and Siggia,E.D. (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, **33**, 5781–5798.
6. Morozov,A.V. and Siggia,E.D. (2007) Connecting protein structure with predictions of regulatory sites. *Proc. Natl Acad. Sci. USA*, **104**, 7068–7073.
7. Angarica,V.E., Perez,A.G., Vasconcelos,A.T., Collado-Vides,J. and Contreras-Moreira,B. (2008) Prediction of TF target sites based on atomistic models of protein-DNA complexes. *BMC Bioinformatics*, **9**, 436.
8. Mackerell,A.D. Jr and Nilsson,L. (2008) Molecular dynamics simulations of nucleic acid-protein complexes. *Curr. Opin. Struct. Biol.*, **18**, 194–199.
9. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
10. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–288.
11. Berman,H.M., Olson,W.K., Beveridge,D.L., Westbrook,J., Gelbin,A., Demeny,T., Hsieh,S.H., Srinivasan,A.R. and Schneider,B. (1992) The nucleic acid database. *A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys. J.*, **63**, 751–759.
12. Laskowski,R.A. (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.*, **29**, 221–222.
13. Kumar,M.D., Bava,K.A., Gromiha,M.M., Prabakaran,P., Kitajima,K., Uedaira,H. and Sarai,A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–206.
14. Spirin,S., Titov,M., Karyagina,A. and Alexeevski,A. (2007) NPIDB: a database of nucleic acids-protein interactions. *Bioinformatics*, **23**, 3247–3248.
15. Lee,S. and Blundell,T.L. (2009) BIPA: a database for protein-nucleic acid interaction in 3D structures. *Bioinformatics*, **25**, 1559–1560.
16. Kawabata,T., Ota,M. and Nishikawa,K. (1999) The Protein Mutant Database. *Nucleic Acids Res.*, **27**, 355–357.
17. Thomas-Chollier,M., Sand,O., Turatsinze,J.V., Janky,R., Defrance,M., Vervisch,E., Brohee,S. and van Helden,J. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.*, **36**, W119–127.
18. Gama-Castro,S., Jimenez-Jacinto,V., Peralta-Gil,M., Santos-Zavaleta,A., Penaloza-Spinola,M.I., Contreras-Moreira,B., Segura-Salazar,J., Muniz-Rascado,L., Martinez-Flores,I., Salgado,H. *et al.* (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **36**, D120–124.
19. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
20. Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
21. Felsenstein,J. (2005) Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
22. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
23. Lu,X.J. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.
24. Mirny,L.A. and Gelfand,M.S. (2002) Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Res.*, **30**, 1704–1711.
25. Lozada-Chavez,I., Angarica,V.E., Collado-Vides,J. and Contreras-Moreira,B. (2008) The role of DNA-binding specificity in the evolution of bacterial regulatory networks. *J. Mol. Biol.*, **379**, 627–643.

26. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
27. Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, W253–W258.
28. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
29. Contreras-Moreira,B. and Collado-Vides,J. (2006) Comparative footprinting of DNA-binding proteins. *Bioinformatics*, **22**, e74–e80.
30. Chang,J.T., Schutze,H. and Altman,R.B. (2004) GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics*, **20**, 216–225.
31. Contreras-Moreira,B., Branger,P.A. and Collado-Vides,J. (2007) TFmodeller: comparative modelling of protein-DNA complexes. *Bioinformatics*, **23**, 1694–1696.
32. Wolfe,S.A., Grant,R.A. and Pabo,C.O. (2003) Structure of a designed dimeric zinc finger protein bound to DNA. *Biochemistry*, **42**, 13401–13409.
33. Shimizu,T., Toumoto,A., Ihara,K., Shimizu,M., Kyogoku,Y., Ogawa,N., Oshima,Y. and Hakoshima,T. (1997) Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. *EMBO J.*, **16**, 4689–4697.
34. Guo,A.Y., Chen,X., Gao,G., Zhang,H., Zhu,Q.H., Liu,X.C., Zhong,Y.F., Gu,X., He,K. and Luo,J. (2008) PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Res.*, **36**, D966–D969.