RESEARCH ARTICLE

Multiple protein-DNA interfaces unravelled by evolutionary information, physico-chemical and geometrical properties

Flavia Corsi[1,2], **Richard Lavery**[3], **Elodie Laine**[1]*, **Alessandra Carbone**[1,4]*

1 Sorbonne Université, CNRS, IBPS, UMR 7238, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France, **2** Sorbonne Université, CNRS, Institut des Sciences, du Calcul et des Données (ISCD), 75005 Paris, France, **3** Lyon University, CNRS, IBCP, UMR 5086, Molecular Microbiology and Structural Biochemistry, 69367 Lyon, France, **4** Institut Universitaire de France, 75005 Paris, France

* elodie.laine@upmc.fr (EL); alessandra.carbone@lip6.fr (AC)

Citation: Corsi F, Lavery R, Laine E, Carbone A (2020) Multiple protein-DNA interfaces unravelled by evolutionary information, physico-chemical and geometrical properties. PLoS Comput Biol 16(2): e1007624. https://doi.org/10.1371/journal.pcbi.1007624

Editor: Bert L. de Groot, Max Planck Institute for Biophysical Chemistry, GERMANY

Received: September 13, 2019

Accepted: December 20, 2019

Published: February 3, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1007624

Data Availability Statement: The data underlying the results presented in the study are available from http://www.lcqb.upmc.fr/JET2DNA and http://www.lcqb.upmc.fr/PDNAbenchmarks.

Abstract

Interactions between proteins and nucleic acids are at the heart of many essential biological processes. Despite increasing structural information about how these interactions may take place, our understanding of the usage made of protein surfaces by nucleic acids is still very limited. This is in part due to the inherent complexity associated to protein surface deformability and evolution. In this work, we present a method that contributes to decipher such complexity by predicting protein-DNA interfaces and characterizing their properties. It relies on three biologically and physically meaningful descriptors, namely evolutionary conservation, physico-chemical properties and surface geometry. We carefully assessed its performance on several hundreds of protein structures and compared it to several machine-learning state-of-the-art methods. Our approach achieves a higher sensitivity compared to the other methods, with a similar precision. Importantly, we show that it is able to unravel 'hidden' binding sites by applying it to unbound protein structures and to proteins binding to DNA via multiple sites and in different conformations. It is also applicable to the detection of RNA-binding sites, without significant loss of performance. This confirms that DNA and RNA-binding sites share similar properties. Our method is implemented as a fully automated tool, JET^2_{DNA}, freely accessible at: http://www.lcqb.upmc.fr/JET2DNA. We also provide a new dataset of 187 protein-DNA complex structures, along with a subset of 82 associated unbound structures. The set represents the largest body of high-resolution crystallographic structures of protein-DNA complexes, use biological protein assemblies as DNA-binding units, and covers all major types of protein-DNA interactions. It is available at: http://www.lcqb.upmc.fr/PDNAbenchmarks.

Author summary

Protein-DNA interactions are essential to living organisms and their impairment is associated to many diseases. For these reasons, they have become increasingly important therapeutic targets. Experimental structure determination has revealed different binding

Funding: This work was funded by LabEx CALSIMLAB (public grant ANR-11-LABX-0037-01 constituting a part of the "Investissements d'Avenir" program - reference: ANR-11-IDEX-0004-02) (FC) and the Institut Universitaire de France (AC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

motifs and modes, associated to different functions. Yet, the available structural data gives us only a glimpse of the multiplicity and complexity of protein surface usage by DNA. In this work, we use a three-layer model to describe and predict DNA-binding sites at protein surfaces. Given a protein, we consider the way its residues are conserved through evolution, their physico-chemical properties and geometrical shapes to decrypt its surface. We are able to detect a large portion of interacting residues with good precision, even when they are 'hidden' by conformational changes. We highlight cases where one protein binds DNA via distinct regions to perform different functions. We are able to uncover the alternative binding sites and relate their properties with their specific roles. Our work can help guiding mutagenesis experiments and the development of new drugs specifically targeting one site while limiting possible side effects.

Introduction

Interactions between proteins and DNA play a fundamental role in essential biological processes [1] and their impairement is associated with human diseases [2, 3]. Thus, they represent increasingly important therapeutic targets. The experimental determination of protein-DNA complexes is a costly and time consuming process (<5000 protein-DNA complex structures available in the Protein Data Bank [4], release 2019). This has motivated the development of a large number of computational methods, most of them based on machine learning, to predict DNA-binding residues [5–24]. Both sequence-based properties (conservation, amino acid type, predicted secondary structure, solvent accessibility and disorder) and structural properties (electrostatic potential, dipole moment, surface shape and curvature, structural motifs, secondary structure, amino acid microenvironment, solvent accessibility and hydrogen-bonding potential) have been considered. Among those, amino acid composition is one of the most powerful feature. Indeed, positively charged and polar amino acids are largely over-represented in DNA-binding sites, in order to counterbalance the excess of negative charge coming from the DNA phosphate groups [25]. Methods using a large body of features can achieve very high accuracy on known binding sites but generally lack interpretability.

Our understanding of protein-DNA interactions is hampered by the fact that protein surfaces are complex dynamical objects which may interact with several DNA molecules at the same time or at different moments but via the same region [26–28], accommodate indifferently DNA and RNA [29] and undergo substantial conformational changes upon binding [30]. A protein may harbour multiple DNA-binding sites on its surface, each of which may have a different role with a different level of importance in the accomplishment of the protein's function [28, 31] and thus display particular properties. The relatively small amount of available structural data gives us only a glimpse of the many ways nucleic acid molecules use protein surfaces. This calls for the development of tools able to comprehensively identify and characterize protein-DNA interfaces and help decipher protein surface complexity.

Here, we present $\mathrm{JET}^2_{\mathrm{DNA}}$, a new method for predicting DNA-binding sites at protein surfaces based on a few biologically and physically meaningful parameters. Specifically, we use evolutionary conservation, physico-chemical properties and local/global geometry. These descriptors recently proved useful to identify and characterize protein-protein interfaces [32, 33]. $\mathrm{JET}^2_{\mathrm{DNA}}$ predicts surface patches following the *support-core-rim* model for experimental protein-protein interfaces, where interacting residues are classified based on their structural role in the interaction [34]. This three-layer model comes from a reexamination of the previously proposed *core-rim* one, which proved useful in dissecting protein-protein [35, 36] and protein-DNA interfaces [37] and in understanding their properties. To the best of our knowledge, this is the first time that the three-layer *support-core-rim* model is used for analyzing and predicting protein-DNA interfaces. JET^2_{DNA} implements three different scoring strategies, aimed at detecting different types of protein-DNA interfaces and different subregions of a given interface.

We assessed JET^2_{DNA} performance on two new representative benchmarks of high-resolution structures, comprised of 187 protein-DNA complexes (HR-PDNA187) and 82 associated bound-unbound pairs (HOLO-APO82), and on an independent test set of 82 nucleic acid binding proteins (TEST-NABP82), taken from [18]. We show that JET^2_{DNA} predictions are very accurate and robust to protein conformational and stoichiometry changes associated to DNA binding. Moreover, the predictive performances are equivalent for both DNA- and RNA-binding sites, showing that these sites share similar properties. In addition, rigorous comparison with three established machine learning based methods, namely DISPLAR [7], multiVORFFIP [15] and DRNApred [18] demonstrates the better predictive power of JET^2_{DNA} on all datasets. Specifically, JET^2_{DNA} is able to significantly detect more interacting residues while retaining similar precision and to unravel 'hidden' alternative binding sites. We show how one can learn about the origins and specificities of different types of protein-DNA interfaces through direct interpretation of JET^2_{DNA} predictions. With this work, we pave the way to the discovery of yet unknown binding sites, opening up new perspectives for drug design.

We provide the structures and the experimentally known DNA-binding sites for our two newly created datasets, HR-PDNA187 and HOLO-APO82 at: http://www.lcqb.upmc.fr/PDNAbenchmarks/. These datasets represent the largest body of non-redundant known high-resolution crystallographic protein-DNA complex structures. They were manually curated to ensure quality and biological relevance. They can be used as benchmark sets for evaluating DNA-binding site predictors and DNA-protein docking methods. The code of JET^2_{DNA} is available at: http://www.lcqb.upmc.fr/JET2DNA.

Materials and methods

Datasets

HR-PDNA187. The complete list of 1257 protein-double strand DNA complexes determined by X-ray crystallography with a resolution better than 2.5Å was downloaded from the Nucleic Acid Database [38] (February 2016 release). This list was filtered using PISCES sequence culling server [39] to define a set of 222 protein-DNA complexes non-redundant at 25% sequence identity, with an R-factor lower than 0.3 and with at least one protein chain longer than 40 amino acids. The complexes' 3D structures were downloaded from the PDB [4].

Subsequently, the dataset was manually curated to ensure its good quality. We removed entries where: (1) the asymmetric unit did not contain at least one biological unit or (2) the DNA molecule was single-stranded or contained less than 5 base pairs or (3) only Cα atoms were present. Moreover, only chains with more than one contact with the DNA were retained. When the biological unit contained multiple copies of the protein-DNA complex, only one copy was kept. The structure 4hc9 was excluded because it displays different DNA-binding sites in the asymmetric and biological units. The complex 4aik was substituted with the 100% homolog 4aij, where the DNA-binding site is twice bigger. Finally, we removed the only antibody present in the dataset (3vw3), since this class of proteins have very peculiar characteristics and should be treated separately. In total, we retained 187 complexes (Table A in S1 File).

HR-PDNA187 covers all major groups of DNA-protein interactions according to Luscombe *et al.* classification [1]: helix-turn-helix (HTH), zinc-coordinating, zipper type, other α-

helical, β-sheet, β-hairpin/ribbon, other. It spans a wide range of functional classes: it comprises 100 enzymes, 78 regulatory proteins, 7 structural proteins, 1 protein with other function and 1 unclassified protein (Table A in S1 File).

HOLO-APO82. We collected all available X-ray structures of the APO forms of the proteins from HR-PDNA187. We used the blastp program from the BLAST+ package [40] from NCBI with a threshold of 95% for seq. id., 10^{-3} for the E-value, a percentage of coverage $\geq$ 70% and a percentage of gaps $\leq$ 10% with respect to the query protein chain. Among the structures matching these criteria, only the ones having the same UniProt code [41] or belonging to the same organism as the query sequence were retained. If several structures passed all filters, only the closest one to the query or the one with the highest resolution was chosen. We found unbound forms for 81 complexes. For the complex 2isz, the unbound protein was found as a monomer and also as a dimer in two different PDB structures. Both structures were retained as they were both reported as present in equilibrium [42]. The resulting list comprises a total of 82 HOLO(bound)-APO(unbound) pairs. Within each pair, the APO form may be in the same oligomeric state as the HOLO form or may have fewer chains (Table A in S1 File).

TEST-NABP82. We used an additional recent dataset [18] of 82 proteins bound to nucleic acids, with a resolution better than 2.5Å. Among these, 49 proteins were solved bound to DNA (TEST-DBP49) and 33 to RNA (TEST-RBP33). Among the 49 DNA-binding proteins, 25 share more than 30% sequence identity with proteins from HR-PDNA187. To fairly assess $\mathrm{JET}^2_{\mathrm{DNA}}$ performance, they were removed from the dataset and the remaining 24 were used as non-redundant test set of DNA-binding proteins (TEST-DBP24). The list of the 24 complexes present in TEST-DBP24, annotated with the reasons they were excluded from HR-PDNA187 despite their good (<2.5Å) resolution, is reported in Table B in S1 File.

Definition of interface residues

For each bound form from HR-PDNA187, we calculated the residues relative accessible surface areas in presence ($rasa_{DNA}$) and absence ($rasa_{free}$) of DNA, using NACCESS 2.1.1 [43] with a probe size of 1.4Å. Interface residues were defined as those being more buried in the presence of DNA than in the free form ($\Delta rasa > 0$). They were classified in three structural components [34]: *support* residues are buried both in presence ($rasa_{DNA}$<25%) and absence of DNA ($rasa_{free}$<25%); *core* residues are exposed in absence of DNA ($rasa_{free}\geq$25%) and become buried upon binding ($rasa_{DNA}$<25%); *rim* residues are exposed in presence ($rasa_{DNA}\geq$25%) and in absence ($rasa_{free}\geq$25%) of DNA (Fig 1a).

The interfaces for TEST-NABP82 were directly taken from [18]. They were defined from the PDB entries comprised in the dataset and also PDB structures of identical or similar complexes (seq. id.≥80% and TM-score≥0.5) [44, 45]. This allows to account for interface variability coming from molecular flexibility. Nevertheless, we should stress that these interfaces were defined using a very stringent distance criterion of 3.5Å, and hence they are substantially smaller than those we defined from HR-PDNA187. We used them to fairly assess $\mathrm{JET}^2_{\mathrm{DNA}}$ performance on an independent dataset, not created by us, and to fairly compare $\mathrm{JET}^2_{\mathrm{DNA}}$ with DRNApred [18], which was evaluated on them.

Residue descriptors

Evolutionary conservation. Conservation levels are computed using the Joint Evolutionary Trees (JET) method [46]. This measure is inspired but different from the evolutionary trace introduced in [47, 48]. Briefly, the algorithm performs a PSI-BLAST search [49] to retrieve a set of sequences homologous to the query. These sequences are then sampled by a Gibbs-like approach and from each sample a phylogenetic tree is constructed [32, 46]. From

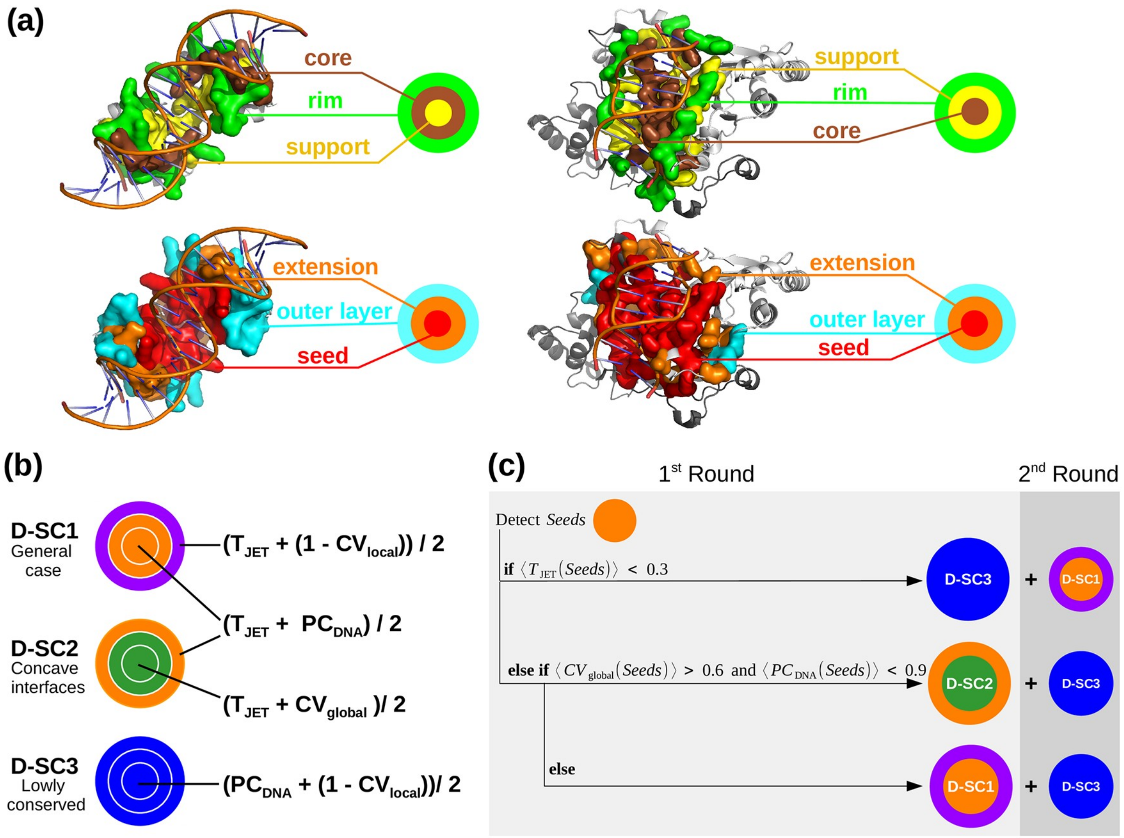
(a)
core
rim
support
support
rim
core
extension
outer layer
seed
extension
outer layer
seed
(b)
D-SC1
General case
D-SC2
Concave interfaces
D-SC3
Lowly conserved
$(T_{JET} + (1 - CV_{local})) / 2$
$(T_{JET} + PC_{DNA}) / 2$
$(T_{JET} + CV_{global}) / 2$
$(PC_{DNA} + (1 - CV_{local})) / 2$
(c)
1st Round
2nd Round
Detect Seeds
if $\langle T_{JET}(Seeds)\rangle < 0.3$
else if $\langle CV_{global}(Seeds)\rangle > 0.6$ and $\langle PC_{DNA}(Seeds)\rangle < 0.9$
else
D-SC3
+
D-SC1
D-SC2
+
D-SC3
D-SC1
+
D-SC3

Fig 1. Experimental interface definition, JET$^2_{DNA}$ scoring schemes and complete automated clustering procedure. (a) Top, sections of two experimental interfaces (on the left, PDB code: 1JE8; on the right, PDB code: 1D02). Bottom, the corresponding JET$^2_{DNA}$ prediction using D-SC2. The experimental and predicted interface residues are displayed in opaque surfaces: *support*, *core* and *rim* are in yellow, brown and green, respectively; cluster *seed*, *extension* and *outer layer* are in red, orange and cyan, respectively; (b) Schematic representation of the three scoring schemes provided in JET$^2_{DNA}$. T_{JET}: conservation level, PC_{DNA}: protein-DNA interface propensities, CV_{local} and CV_{global}: local and global circular variance computed with a radius of 12 Å and 100 Å, respectively. Different colors correspond to different combinations of properties used to predict interface residues in the three steps of the clustering procedure. (c) Schematic representation of the complete automated JET$^2_{DNA}$ clustering procedure.

https://doi.org/10.1371/journal.pcbi.1007624.g001

each tree, a *tree trace* is computed for each position in the query sequence: it corresponds to the tree level where the amino acid at this position appeared and remained conserved thereafter [32, 46]. *Tree traces* are averaged over all trees to get more statistically significant values, denoted as T_{JET}, which vary in the interval [0, 1].

Physico-chemical properties. Propensities specific to every amino acid to be located at a protein-DNA interface (PC_{DNA}) were taken from [50] (see Fig. A in S2 File, in red). The original values, ranging from 0 to 2.534, were scaled between 0 and 1.

Circular variance. The circular variance (CV) is a geometrical measure of the vectorial distribution of a set of neighboring points localized around the center of a three dimensional sphere of radius r_c [51, 52]. The CV value of an atom *i* is computed as:

$$CV(i) = 1 - \frac{1}{n_i}\left|\sum_{j\neq i, r_i \leq r_c} \frac{\vec{r_{ij}}}{\|\vec{r_{ij}}\|}\right| \tag{1}$$

where n_i is the number of all individual atoms distant by less than r_c Å from atom *i*. The CV value of a residue is then computed as the average of the atomic CVs, over all its atoms.

Given a residue, its CV value reflects the protein density around it. A low CV value (close to 0) indicates that a residue is located in a protruding region of the protein surface, while a high value (close to 1) indicates it is buried within the protein. By consequence, the complement measure (1- CV) shows low values for buried residues and high values for protruding ones. As described in the following, since the algorithm clusters together highly scored residues, the CV_{global} measure is then used to detect globally buried residues, while $(1 - CV_{local})$ is used to detect locally protruding ones.

Varying the radius cutoff r_c allows describing the local (CV_{local}, $r_c = 12$Å) and the global (CV_{global}, $r_c = 100$Å) geometry of the surface. The values of r_c for CV_{local} and CV_{global} are the same ones we already used in [32] to predict protein-protein interfaces. In [32], we showed that thresholds in the range 10-14Å for CV_{local} would not change the results. On the other hand, a CV_{global} value of 100Å ensures the inclusion of essentially all protein atoms in the calculation. Setting the threshold to a value larger than the standard protein size does not affect the efficiency nor the outcome of the calculation.

JET^2_{DNA} workflow

The JET^2_{DNA} method requires as input a protein query sequence for which three-dimensional (3D) structural coordinates are available in the PDB. It computes T_{JET}, PC_{DNA}, CV_{global} and CV_{local} values and it combines them to assign a score to each residue. Different combinations are implemented in three different scoring schemes (D-SC in Fig 1b) designed to detect different types of protein-DNA interfaces (see Results). The computed scores are used to rank, select and cluster protein residues. The clustering algorithm is adapted from the protein-protein interface prediction method JET/JET^2 [32, 46]. First, highly scored residues proximal in 3D space are clustered together to form *seeds*. Then, the *seeds* are extended by progressively adding highly scored neighboring residues. At this stage, if two predicted patches are in contact, they will be merged. Finally, the prediction is completed with an *outer layer* of residues. The *seed*, *extension* and *outer layer* (Fig 1a, in red, orange and cyan, respectively) approximate the *support*, *core* and *rim* (Fig 1a, in yellow, brown and green, respectively) detected in experimental interfaces. For a more detailed description of the three-steps clustering procedure see Table C in S1 File. JET^2_{DNA} implements an automated procedure to choose the most appropriate scoring scheme for the studied system (Fig 1c; see also *Automated clustering procedure*). Alternatively, the user can manually choose a scoring scheme. The user has the possibility to complement the predictions with another round of the clustering procedure using a complementary scoring scheme (Fig 1c; see also *Complete clustering procedure*). To get more robust predictions, several iterations of JET^2_{DNA} can be run ($iJET^2_{DNA}$; see also *Iterative mode*). A schematic representation of the JET^2_{DNA} pipeline is reported in Fig. B in S2 File. In the following, we explain in details the new criteria and new procedures implemented in JET^2_{DNA}, compared to JET^2 [32].

Modification of the expected size of the interface. At each step of the clustering procedure, JET^2_{DNA} uses two thresholds, namely $score^{layer}_{res}$ and $score^{layer}_{clus}$, that respectively determine which residues should be considered as candidates for the predicted patches and when the process of growing each patch should be stopped. The set up of these threshold is based on the expected relative size of a protein-DNA interface, $f^{DNA}_{intfrac}(x)$. $f^{DNA}_{intfrac}(x)$ was determined by plotting the percentage of interface residues versus the total number of surface residues for HR-PDNA187 (Fig. C in S2 File, circles). The function that best approximates our data is $f^{DNA}_{intfrac}(x) = (2.66/\sqrt{x}) + 0.03$ (Fig. C in S2 File, solid line), where x is the number of protein surface residues. Compared to protein-protein interfaces [32], DNA-binding sites cover a larger portion of the protein surface (Fig. C in S2 File, compare dotted and solid lines).

Dynamic set up of the residue and cluster thresholds. Contrary to JET^2 algorithm, where $score_{res}^{layer}$ and $score_{clus}^{layer}$ thresholds were fixed at the beginning of the clustering procedure and could not be changed, in JET^2_{DNA} we decided to initially fix them more stringently and relax them in a second stage, if the prediction is smaller than two thirds (<70%) of the expected size (Fig. B in S2 File, in red). This dynamic set up limits false positives in cases where surface regions outside but close to the interface display a detectable signal. It should be noted that the thresholds in JET^2_{DNA}, even relaxed, remain stricter than the ones in JET^2. Details about threshold values used in JET^2_{DNA} are given in Table D in S1 File.

Avoiding small-ligand binding pockets. Like protein-protein interactions, protein-DNA interactions are often mediated or regulated by small ligands. As a result, a significant number of protein-protein and protein-DNA interfaces are close to or overlapping small ligand-binding pockets. These pockets are generally very conserved (*e.g.* active sites of enzymes) [32] and may thus confound the prediction. In JET^2, we resolved the issue by designing a specific scoring scheme exploiting the fact that small ligand-binding pockets are more deeply buried than protein-protein interfaces (see SC2 in Fig. 2 from [32]). The specific detection of protein-DNA interfaces is more difficult, as these interfaces are more concave than protein-protein interfaces (compare (1-CV) boxplots in Fig 2e–2g). To tackle the problem, we implemented a procedure in JET^2_{DNA} that redefines *seeds* when they are too buried. Specifically, if a *seed* comprises a significant proportion (>20% for D-SC1, >30% for D-SC2) of highly buried residues, then the clustering procedure restarts avoiding such residues (Fig. B in S2 File, in blue). The procedure does not apply to D-SC3 as this scoring scheme specifically selects protruding/exposed residues. Residues were considered as highly buried if their CV_{local} was higher than ${CV_{local}}^{high}$ = 0.9. They represent less than 3% of the protein surface (Fig. D in S2 File).

Filtering out putative false positive clusters. In JET^2, small patches were filtered out based on the comparison between their size and the size distribution of randomly generated patches [46] after the *seed* or *extension* steps, depending on the scoring scheme used. In the case of protein-DNA interfaces, we observed that this procedure removed too many true positives. Hence, we modified it in JET^2_{DNA}. Namely, we first detect all possible *seeds* and extend them. Then, if the patches represents more than two thirds (>70%) of the expected size of the interface, we iteratively filter them starting from the smallest one (Fig. B in S2 File, in green). To eliminate aberrant predictions, clusters composed of 1 or 2 residues are still systematically filtered out.

Automated clustering procedure

The implemented algorithm is described in Fig 1c and Table E in S1 File. By default, JET^2_{DNA} first detects *seeds* using D-SC1 (T_{JET} + PC_{DNA}). If these display a very low conservation signal (average T_{JET}<0.3) then the strategy is to exploit the other two descriptors and look for locally protruding residues that satisfy expected physico-chemical properties (D-SC3). Otherwise, the algorithm analyses the physico-chemical and the geometrical properties of the detected *seeds*. If the *seeds* are in a very concave region (average CV_{global}>0.6) and do not display highly favorable physico-chemical properties (average PC_{DNA}<0.9), then the algorithm switches from D-SC1 to D-SC2, where CV_{global} is employed to accurately detect an “enveloping” (concave) interface. Otherwise, physico-chemical properties are considered the driving force for an accurate prediction of the interface.

Selecting residues with a CV_{global}>0.6 allows to well define globally concave protein regions, giving a good compromise between too small and too large ones (see Fig. E(a) and Fig. E(b) in S2 File, on top).

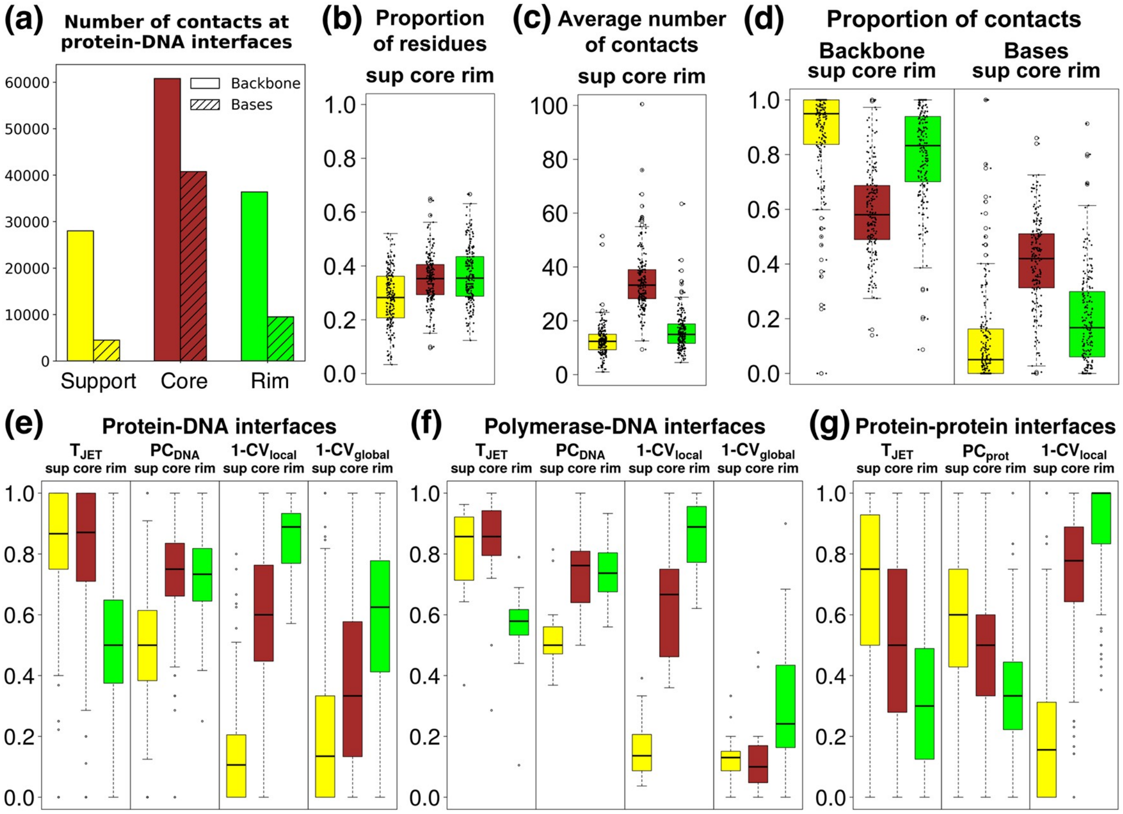

Fig 2. Signals detected in experimental interfaces. The calculations in (a)-(f) were performed on the HR-PDNA187 dataset. The plot in (g) was taken from Fig. 1a in [32]. The *support*, *core* and *rim* are in yellow, brown and green, respectively. Atomic contacts are defined by an atomic distance < 5Å. (a) Total number of atomic contacts between protein interface residues, divided in *support*, *core* and *rim*, and DNA backbone atoms (plain color) and DNA base atoms (diagonal hatching). (b) Distributions of the proportion of residues comprised in *support*, *core* and *rim*. (c) Distributions of the average number of atomic contacts per interface residue. Interface residues are divided in *support*, *core* and *rim*. One outlier point, reaching a value of 206, was removed in the boxplot of the distribution of *core* residues. (d) Distributions of the proportion of atomic contacts established by *support*, *core* and *rim* residues with DNA backbone atoms and DNA base atoms. (e-g) Distributions of the proportion of interface residues, divided in *support*, *core* and *rim*, having values above the median value computed over the entire protein. T_{JET}: conservation level, PC_{DNA}: protein-DNA interface propensities, PC_{prot}: protein-protein interface propensities, CV_{local} and CV_{global}: local and global circular variances computed with a radius of 12Å and 100Å, respectively. Distributions are computed on: (e) all protein-DNA interfaces from HR-PDNA187, (f) all polymerases interfaces from HR-PDNA187, (g) all the 176 protein-protein interfaces in PPDBv4 (see Fig. 1a in [32]).

https://doi.org/10.1371/journal.pcbi.1007624.g002

Complete clustering procedure

A complete procedure is available for both manual and automated clustering procedure and is described in Fig 1c and Table E in S1 File (see also Fig. B in S2 File, in yellow). It consists in complementing the main clusters predicted in the first round by secondary clusters detected by another scoring scheme. D-SC3 will be used in the second round if D-SC1 or D-SC2 were chosen as main scoring schemes. D-SC1 will be the complementary one when D-SC3 is used in the first round.

Iterative mode

Multiple JET^2_{DNA} runs may lead to slightly different predictions, due to the Gibbs sampling of the sequences. To get more robust predictions, JET^2_{DNA} can be run in an iterative mode of the program, which we call $iJET^2_{DNA}$. In this way, we can compute the number of times a given residue is detected in a cluster divided by the total number of runs. The result will be a number comprised between 0 and 1 and it reflects the probability of the given residue to be at an interface.

Evaluation of performances

We used six standard measures of performance:

$$Sens = \frac{TP}{TP + FN}; \; PPV = \frac{TP}{TP + FP}; \; Spe = \frac{TN}{TN + FP}; \; Acc = \frac{TP + TN}{TP + FN + TN + FP};$$
$$F1 = \frac{2 \cdot Sens \cdot PPV}{Sens + PPV}; \quad MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP (true positives) are the number of residues correctly predicted as interacting, TN (true negatives) are the number of residues correctly predicted as non-interacting, FP (false positives) are the number of non-interacting residues incorrectly predicted as interacting and FN (false negatives) are the number of interacting residues incorrectly predicted as non-interacting. Thus, in our case, *Sensitivity* (*Sens*) measures the proportion of true interface residues that were correctly predicted as interacting with respect to the total number of interface residues; the *Positive Predictive Value* (*PPV*) measures the proportion of true interface residues that were correctly predicted as interacting with respect to the total number of predicted residues; the *Specificity* (*Spe*) measures the proportion of true non-interface residues that were correctly predicted as non-interacting with respect to the total number of non-interface residues; *Accuracy* (*Acc*) measures the proportion of correctly predicted residues (interacting and non-interacting) with respect to the total number of residues; *F1* score is a weighted average, specifically the harmonic mean, between *Sensitivity* and *PPV*, measuring the balance between these two; *Matthew's Correlation Coefficient* (*MCC*) is a correlation coefficient between observed and predicted residues. The *MCC* is the only measure varying in the range [-1, 1], where 0 represents a prediction no better than random. All the other measures return values in the range [0, 1]. We multiplied by 100 all the statistical values reported in tables and figures, representing them in percentages varying in the range [-100, 100] for the *MCC*, and [0, 100] for the other measures.

To assess the statistical significance of the differences in performance between each pair of methods, we relied on the paired *t*-test when the distributions were normal and on the Wilcoxon signed-rank test [53] otherwise. To verify if the data were normally distributed we used the Anderson-Darling test [54]. The difference was considered significant when the p-value was lower than 0.05.

Choice of the other methods

To compare JET^2_{DNA} performance, we considered a large set of popular DNA-binding site predictors [5–18, 20–24, 55, 56]. Among those, we discarded BindN [8], BindN+ [9], BindN-RF [12], MetaDBsite [55], PreDs [21], DBindR [13], PreDNA [10], DBD-Hunter [23], RBscore [20, 56] and PDNAsite [16] as their web servers are no longer available or do not work. We finally retained three methods with a relatively short runtime, namely DISPLAR [7], multiVORFFIP [15] and DRNApred [18]. DISPLAR was reported to show better performance than DP-Bind [11] and is more recent than DBS-Pred [6] and DBS-PSSM [5]. multiVORFFIP was chosen because it showed good performance on protein-protein interfaces prediction [32]. DRNApred is a very recent predictor that uses only sequence information and which aims at specifically detecting DNA-binding residues, discriminating them from RNA-binding residues. We ran the three tools with the default parameters. Since multiVORFFIP does not provide binary prediction results (binding versus non-binding residues), predicted residues were defined by those with a normalized score (probability) > 0.5.

Results

Support-Core-Rim vs *Core-Support-Rim* model for protein-DNA interactions

To identify characteristic features of protein-DNA interfaces, we analysed the non-redundant set of protein structures bound to DNA available in the PDB. This dataset, HR-PDNA187, comprises 187 high-resolution crystallographic structures, covers all major types of DNA-interactions and spans a wide range of protein functions (see Materials and methods). Interface residues were classified based on the *support-core-rim* model (Fig 1a, on top) proposed for protein-protein interfaces [34]. The *support* (in yellow) comprises residues buried both in presence and absence of DNA, the *core* (in brown) comprises residues exposed in absence of DNA and becoming buried upon binding, and the *rim* (in green) comprises residues exposed both in presence and absence of DNA (see Materials and methods). In protein-protein interfaces, the three components are spatially organized in concentric layers, with the *support* at the center, the *core* in an intermediate position and the *rim* on the external border (see Fig. 2 in [32]). In protein-DNA interfaces, in addition to this spatial organization (Fig 1a, on top left), we also observe an organization where the *support* and the *core* switch positions (Fig 1a, on top right). This variety reflects the different ways a protein may bind to DNA.

However, *support*, *core* and *rim* residues seem to play the same role during the DNA binding in both structural organizations described above. Overall, *core* residues establish the majority of the atomic contacts (defined by an atomic distance<5Å) (Fig 2a and 2c) with both DNA backbone and DNA base atoms (Fig 2a), despite a comparable number of *support*, *core* and *rim* residues in HR-PDNA187 (Fig 2b) (*support*: 2501, *core*: 2958, *rim*: 2891). More than 75% of protein-DNA base contacts involve *core* residues (Fig 2a and 2d), while *support* and *rim* residues prevalently establish contacts with DNA backbone atoms (Fig 2a and 2d).

Characteristics of protein-DNA interfaces

To characterize experimental protein-DNA interfaces, we relied on evolutionary conservation (T_{JET}), DNA-binding propensity (PC_{DNA}) and local and global burial degree (CV_{local} and CV_{global}) (Materials and methods for precise definitions of the descriptors). Specifically, for each complex we computed the percentage of residues located in the *support*, *core* and *rim*, displaying higher values than the median computed over the entire protein (Fig 2e and 2f). The resulting distributions were compared with those obtained for the 176 protein-protein complexes of the Protein-Protein Docking Benchmark version 4 (PPDBv4) [57] (Fig 2g), already reported in Fig. 1a in [32].

DNA binding sites, especially the *support* and the *core*, are significantly more conserved than the rest of the entire protein (Fig 2e, T_{JET}). This conservation signal is stronger than in protein-protein interfaces (compare Fig 2e and 2g). Residues with high DNA-binding propensities tend to be located in the *core* and *rim* (Fig 2e, PC_{DNA}). By contrast, residues displaying physico-chemical properties favourable to protein-protein binding are mainly found in the *support*, and to a lesser extent in the *core*, of protein-protein interfaces (Fig 2g, PC_{prot}). These different geometrical distributions reflect differences between the two PC scales (Fig. A in S2 File). Protein-DNA interfaces are enriched in positively charged and polar residues, which tend to prefer regions exposed to the solvent. By contrast, protein-protein interfaces are enriched in hydrophobic residues, which prefer to be located towards the interior of the interface. Regarding the surface geometry, the *core* and *rim* of protein-DNA interfaces are enriched in locally protruding residues (Fig 2e, (1-CV_{local})), as observed for protein-protein interfaces (Fig 2g). When looking at the status of the interface with respect to the global shape of the protein surface (1-CV_{global}), we found that the enzymes, and in particular the polymerases and some nucleases, display a very specific profile (compare Fig 2e and 2f). All three interface components, especially the *support* and the *core*, are located in concave protein regions (Fig 2f), indicating that these proteins bind to DNA by "enveloping" it.

Overall, this analysis showed that protein-DNA interfaces encode signals that can be described by a few residue-based features and that the way these signals are distributed can be described using the *support-core-rim* model. It also confirmed that protein-DNA interfaces are more conserved than protein-protein interfaces, as reported in [37, 58, 59]. They display specific physico-chemical and geometrical characteristics that vary depending on the type of protein-DNA interaction considered. Hence, the correct detection of these interfaces requires the development of adapted scoring strategies.

Three strategies to detect protein-DNA interfaces

Following the *support-core-rim* model (Fig 1a, on top), we use a predictive model comprising three components (Fig 1a, on bottom), namely the *seed* (in red), the *extension* (in orange) and the *outer layer* (in cyan). JET^2_{DNA} implements different algorithms and scores to detect each one of these components (see Materials and methods). To be able to predict a wide range of protein-DNA interfaces, we devised three scoring schemes (see D-SC1-3, Fig 1b). Each D-SC combines T_{JET}, PC_{DNA}, CV_{local} and CV_{global} in a different way. As the *support* and the *core* may exchange their positions, the same combination is used for *seed* detection and *extension* in all D-SC (Fig 1b, same color for the two first layers). Specifically,

- **D-SC1** first detects highly conserved residues with good physico-chemical properties (high values of (T_{JET} + PC_{DNA})) and completes the prediction with conserved locally protruding residues (high values of (T_{JET} + (1 − CV_{local}))). It aims at detecting generic DNA-binding sites.
- **D-SC2** clusters together conserved residues located in highly concave regions (high values of (T_{JET} + CV_{global})); then, the prediction is completed by adding an *outer layer* of conserved residues displaying good physico-chemical properties (high values of (T_{JET} + PC_{DNA})). It is designed to specifically detect interfaces characteristic of the "enveloping" binding mode displayed by polymerases.
- **D-SC3** leaves out conservation, focusing on locally protruding residues displaying good physico-chemical properties (high values of (PC_{DNA} + (1 − CV_{local}))). It is intended to deal with cases where no evolutionary information is available or the whole protein displays a homogeneous conservation signal. Moreover, it is useful to complete predictions of the previous two scoring schemes, when some residues are less or no conserved than the rest of the binding site, sometimes because they are DNA sequence specific for a subfamily of the DNA-binding protein (see Fig 3, bottom panel, and *Discovery of alternative DNA-binding sites*).

Examples of predictions are shown on Fig 3. D-SC1 is particularly suited for double-headed interfaces stacking in the DNA grooves and some single-headed interfaces. These binding modes are often adopted by transcription factors, regulatory proteins and some glycosilases. The "enveloping" mode captured by D-SC2 is found in many polymerases and nucleases. D-SC3 deals with cases where physico-chemical and geometrical properties have a better discriminative power than conservation. By default, JET^2_{DNA} will automatically determine the

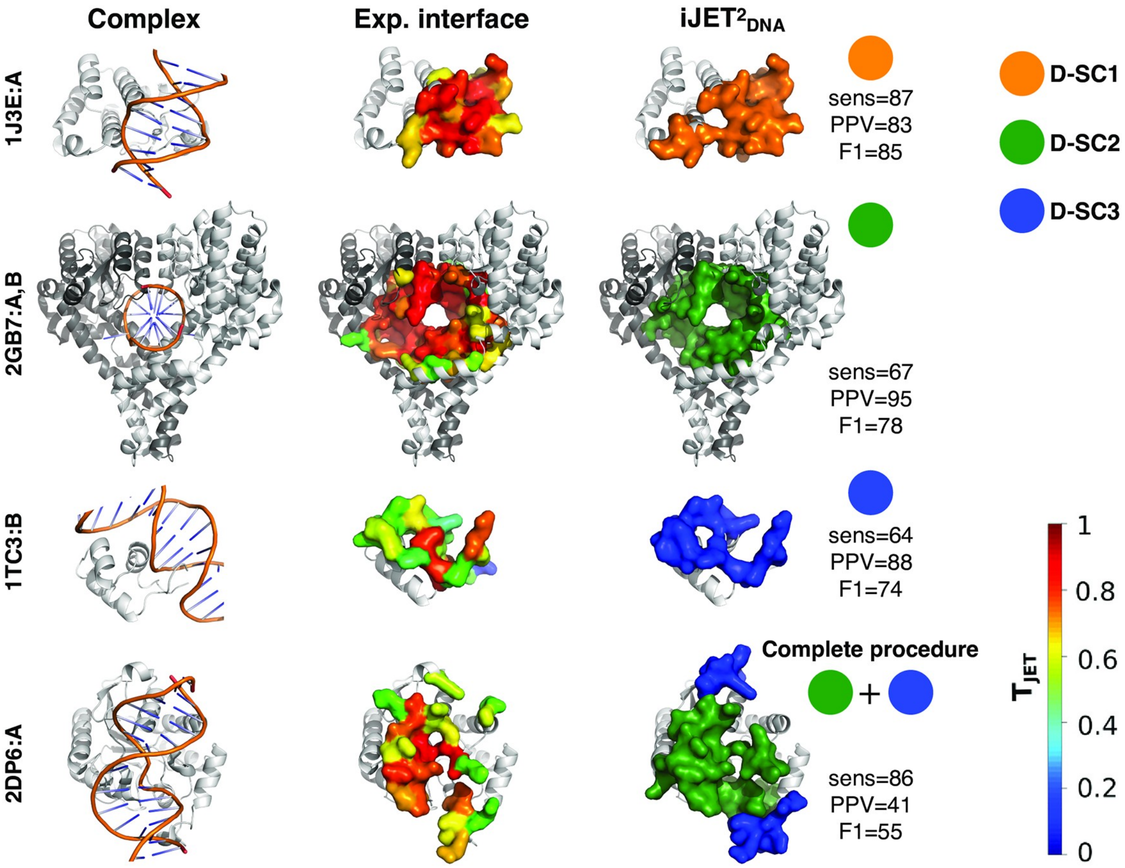

Fig 3. Examples of DNA-binding sites predicted by iJET$^2_{DNA}$. In the first column, the experimental complexes formed between the proteins of interest (greyscaled colored chains) and the DNA (orange) are represented as cartoons. In second and third columns, the experimental and predicted DNA-binding sites are displayed as opaque surfaces, respectively. The experimental interface residues are colored according to conservation levels (T_{JET} values) computed by iJET$^2_{DNA}$. iJET$^2_{DNA}$ predictions, obtained from a consensus of 2 runs out of 10, are colored according to the scoring scheme: D-SC1 in orange, D-SC2 in dark green and D-SC3 in blue. Statistical performance values are given in percentages.

https://doi.org/10.1371/journal.pcbi.1007624.g003

scoring scheme the most suited to the protein being analyzed (Fig 1c). In Fig. E in S2 File, we report five examples of predictions by JET$^2_{DNA}$ automated clustering procedure. In the first three cases, the protein concave regions coincide with the respective DNA-binding sites and the JET$^2_{DNA}$ automated clustering procedure is able to correctly choose, for the first and third cases, the D-SC2 scoring scheme designed to accurately predict these types of protein-DNA interfaces (Fig. E(a) in S2 File, on bottom). Concerning the second case of Fig. E(a) in S2 File, the JET$^2_{DNA}$ automated clustering procedure chose D-SC1, likely due to the physico-chemical properties of this region. However, the predicted region by D-SC1 mostly overlap with the D-SC2 one. In the other two cases, the concave regions are distant from the respective DNA-binding sites and the JET$^2_{DNA}$ automated clustering procedure is able to avoid them and predict the correct DNA-binding sites by choosing D-SC1 (Fig. E(b) in S2 File, on bottom), likely because the conservation and physico-chemical properties of the concave regions do not match with the ones expected for DNA-binding sites. Alternatively, the user can apply the scoring scheme of his/her choice.

Some DNA-binding sites may be comprised of several regions exhibiting different signals. To correctly detect those, JET$^2_{DNA}$ implements a *complete* procedure where patches predicted in a first round are complemented by patches predicted in a second round and using a complementary scoring scheme (Fig 1c). An example of prediction is given in Fig 3 (bottom panel). It may also happen that the DNA-binding site is overlapping or in close proximity to a small ligand binding pocket. JET$^{2}_{\mathrm{DNA}}$ implements a procedure to specifically avoid these pockets (see Materials and methods). The impact of such procedure can be appreciated on Fig. F in S2 File.

Overall assessment of JET$^{2}_{\mathrm{DNA}}$ performance

We report results obtained using the iterative mode of JET$^{2}_{\mathrm{DNA}}$ (iJET$^{2}_{\mathrm{DNA}}$, see Materials and methods), with a consensus of 2 and 8 iterations out of 10. For each protein, the best predicted patch or combination of patches among the three scoring schemes was retained for the evaluation. iJET$^{2}_{\mathrm{DNA}}$ reaches an average F1-score of 61% on HR-PDNA187 and of 58-59% on a subset of 82 proteins (HOLO-APO82) for which unbound forms are available (Table F in S1 File and Fig 4, on top). The performance is similar on bound and unbound forms, indicating that JET$^{2}_{\mathrm{DNA}}$ is robust to conformational changes associated with DNA-binding. For the vast majority of proteins (82%), the sensitivity attained on the unbound form is at least as high as 90% of the sensitivity on the bound form, and in some cases it is even higher than 100%. Hence, JET$^{2}_{\mathrm{DNA}}$ is able to detect interacting residues even when they are 'hidden' by conformational changes. Moreover, the extent of the conformational change is not correlated with the difference in performance (Fig. G-I in S2 File). The predictions are also robust to stoichiometry changes (Table G in S1 File). Varying the consensus threshold enables shifting the balance between sensitivity (Sens) and precision (or predictive positive value, PPV), such that the lower threshold (2/10) yields more extended predictions while the higher threshold (8/10) yields more precise ones. The *seeds* on their own cover about a third of the experimental

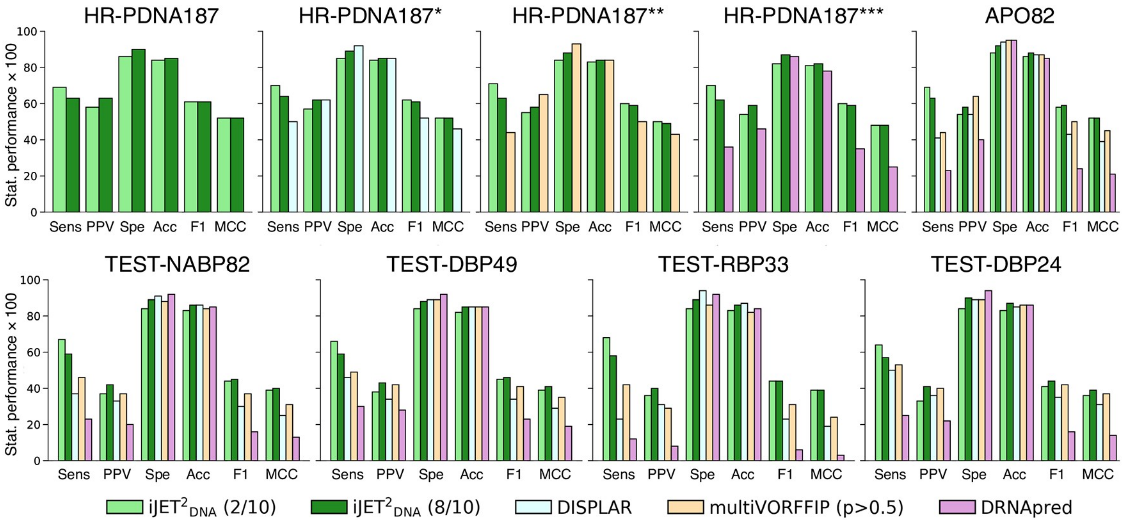

Fig 4. Comparison of iJET$^{2}_{\mathrm{DNA}}$, DISPLAR, multiVORFFIP and DRNApred performance. For iJET$^{2}_{\mathrm{DNA}}$, consensus predictions were obtained from 2 (in light green) and 8 (in dark green) runs out of 10. Statistical performance values are given in percentages. **Top panel**. HR-PDNA187 and APO82. To fairly assess DISPLAR, multiVORFFIP and DRNAPred performances, the proteins used for training these methods were removed from HR-PDNA187. We used a sequence identity cutoff of 95% and ended up with 106 (*), 87 (**) and 42 (***) proteins, respectively. **Bottom panel**. TEST-NABP82 and its subsets. TEST-DBP49 and TEST-DBP24 contain only DNA-binding sites, while TEST-RBP33 contains only RNA-binding sites. TEST-DBP24 and TEST-RBP33 contain only proteins sharing less than 30% sequence identity with proteins from HR-PDNA187 (see Materials and methods). DRNApred and multiVORFFIP were evaluated on their RNA-specific algorithms for the RNA-binding proteins, and on their DNA-specific ones for the DNA-binding proteins. iJET$^{2}_{\mathrm{DNA}}$ and DISPLAR, which do not have a specific algorithm to predict RNA-binding sites, were evaluated on the same algorithm for both DNA- and RNA-binding proteins.

https://doi.org/10.1371/journal.pcbi.1007624.g004

interfaces and contain very few positives (Table H in S1 File). On an independent test set comprising both DNA- and RNA-binding sites (TEST-NABP82), iJET$^2_{\mathrm{DNA}}$ achieves an average F1-score of 45 (Table I in S1 File and Fig 4, on bottom). The sensitivity is roughly the same as for HR-PDNA187 but the precision (PPV) is lower. This can be explained by the fact that the experimental reference interfaces are defined using a much more stringent distance cutoff (see Materials and methods). iJET$^2_{\mathrm{DNA}}$ detects equally well DNA- and RNA-binding sites (Fig 4, on bottom, compare TEST-DBP49/24 and TEST-RBP33).

Discovery of alternative DNA-binding sites

The complete and/or automated mode(s) of JET$^2_{\mathrm{DNA}}$ produce patches that do not match the experimental interfaces present in our datasets (Table J in S1 File). Due to the relatively small number of protein-DNA complexes available in the PDB, it is difficult to systematically assess the pertinence of these additional predictions. Nevertheless, we could identify several cases where the protein binds to DNA via two completely distinct binding sites. We report four such cases in the following (Fig 5 and see Table K in S1 File for F1 values).

1. **RNA polymerase from bacteriophage T7**. This protein first binds the DNA promoter via its recognition site [26, 60] (*site 1*, Fig 5a, left), then undergoes a large conformational change and starts the transcription at its catalytic active site [27] (*site 2*, Fig 5b, left), where a short strand of RNA is paired to the DNA strand. We considered the ensemble of nucleic acid binding residues as experimental interface. In both structures, D-SC1 (and D-SC2 with comparable values) correctly detected the catalytic active site (Fig 5a and 5b, right, in orange) while D-SC3 predicted the recognition site and some DNA-binding residues flanking the active site (Fig 5a and 5b, right, in blue). The fact that the two sites are detected by two different D-SC indicates that they are characterized by different properties, which correlate with their respective functions. Indeed, the active site is generic while the recognition site is specific of this protein family [26]. The very high conservation signal of the former masks the weaker signal of the latter (compare colors in Fig 5a and 5b, left), which is still detectable based on the other properties.

2. **N-terminal domain of the adeno-associated virus replication protein**. This protein contains three distinct DNA-binding sites [61]: a stem loop sequence specific binding site (*site 1*, Fig 5c, left), a tetranucleotide repeat recognition site (*site 2*, Fig 5d, left), and the Tyr153 active site (no PDB structure). D-SC1 and D-SC3 lead to accurate predictions of *site 1* and *site 2* (Fig 5c and 5d), which both display rather low conservation signal and are mostly protruding/exposed.

3. **Modification-dependent restriction endonuclease**. In the original structure from HR-PDNA187 (Fig 5e, left), the DNA is bound only to the winged-helix domain binding site [62] (*site 1*), while the catalytic domain (*site 2*) is partially disordered. In the alternative structure (Fig 5f, left), both DNA-binding sites are occupied. Although the relative domain orientations differ drastically between the two structures, JET$^2_{\mathrm{DNA}}$ accurately identified the two binding sites by combining D-SC1 with D-SC3 (Fig 5e and 5f, right, in orange and in blue). The fact that a combination of D-SC is required reflects the heterogeneity of the conservation signal within each site (Fig 5e and 5f, left, colored by conservation level). D-SC3 enables rescuing lowly conserved subregions that D-SC1 is not able to detect.

4. **Cyclic GMP-AMP synthase**. DNA binding to this protein is associated to a catalytic reaction producing a cyclic dinucleotide from ATP and GTP. The protein was solved bound to DNA both in monomeric (Fig 5g, left [63]) and dimeric (Fig 5h, left [28]) forms. Since the

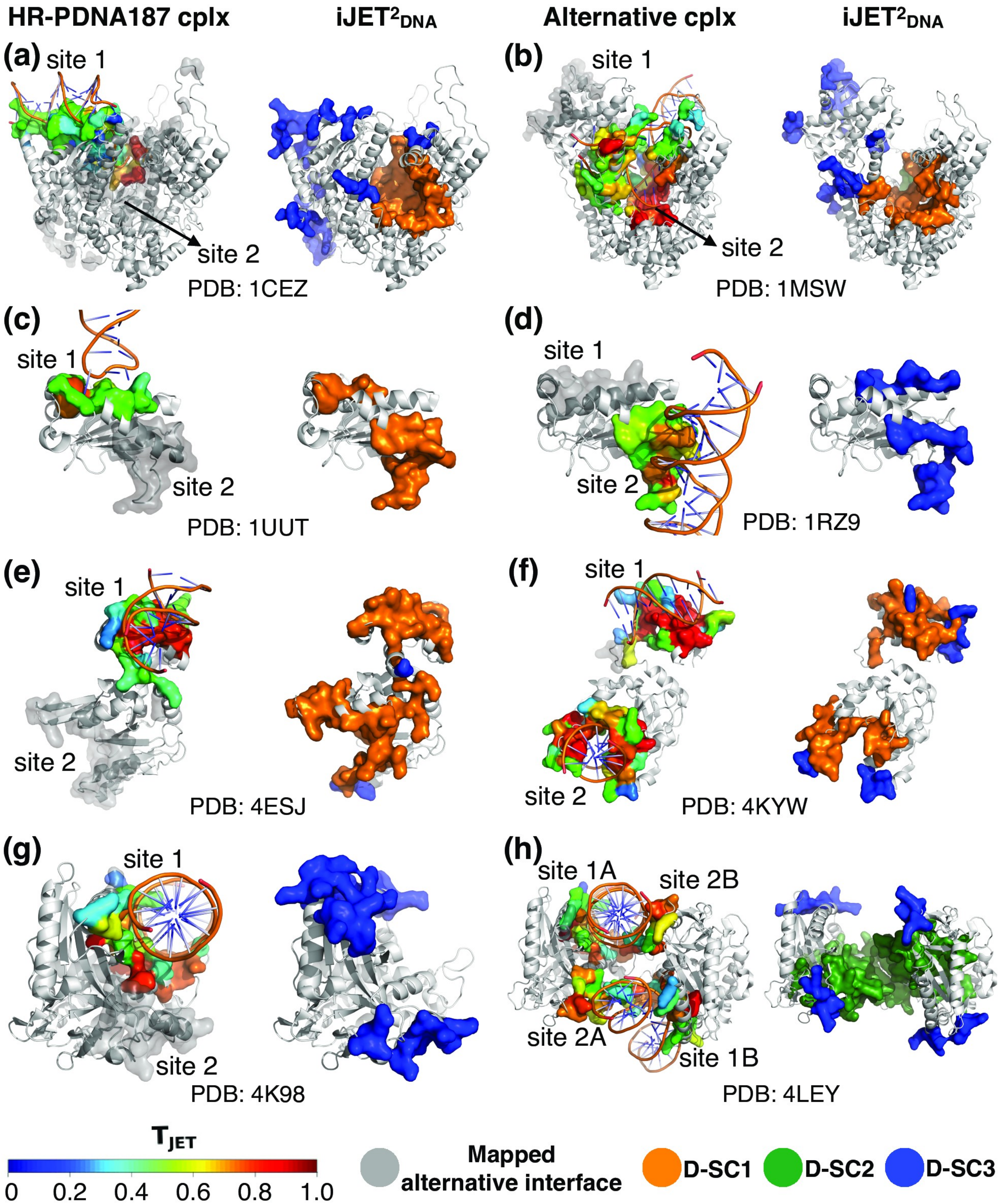

Fig 5. Prediction of multiple DNA-binding sites by iJET$^2_{DNA}$. For each protein (each row), the experimental complex present in HR-PDNA187 and another experimental complex displaying a distinct DNA-binding site are shown on the first and third columns, respectively. For each structure, the 'main' DNA-binding site is colored according to T_{JET} values and the site coming from the other structure is mapped and displayed in transparent grey. The iJET$^2_{DNA}$ predictions computed on each experimental structure are displayed on the second and fourth columns, respectively, and colored according to the scoring scheme. (a-b) RNA polymerase from bacteriophage T7 (PDB: 1CEZ and 1MSW); (c-d) N-terminal domain of the adeno-associated virus replication protein (1UUT and 1RZ9); (e-f) R.DpnI modification-dependent restriction endonuclease (4ESJ and 4KYW); (g-h) Cyclic

GMP-AMP synthase (cGAS) (4K98 and 4LEY). iJET$^2_{\mathrm{DNA}}$ predicted patches were obtained from a consensus of 2, for (a-b) and (e-f), or 5, for (c-d) and (g-h), runs out of 10. See Fig. J-L in S2 File and Table K in S1 File for comparison with other tools.

https://doi.org/10.1371/journal.pcbi.1007624.g005

small ligand-binding active site is much more conserved than the DNA-binding sites, this case is particularly difficult. In the monomeric form, only D-SC3 is able to identify the two DNA-binding sites (Fig 5g, right, in blue). Upon dimerization, the interface geometry changes such that the two sites form a large concave region (Fig 5h, left), detected by D-SC2. Complementing the prediction with D-SC3 further improves the prediction (Fig 5h, right, in dark green and in blue).

These examples illustrate the different usages of JET$^2_{\mathrm{DNA}}$ multiple scoring schemes and how one can learn about each DNA-binding site properties from JET$^2_{\mathrm{DNA}}$ predictions. Specifically, different D-SC may target different binding sites (recognition or catalytic), or different regions of the same binding site (conserved *core* or protruding *rim*), or even yield overlapping predictions. In all cases, JET$^2_{\mathrm{DNA}}$ detection proved to be robust to extensive conformational changes and stoichiometry changes.

Comparison with other tools

We compared JET$^2_{\mathrm{DNA}}$ with three other state-of-the-art machine learning based predictors, namely DISPLAR [7], multiVORFFIP [15] and DRNApred [18] (see Materials and methods). iJET$^2_{\mathrm{DNA}}$ achieves ~20% higher sensitivity and similar accuracy compared to DISPLAR and multiVORFFIP on all the datasets (Table F and Table I in S1 File and Fig 4). The predictions are slightly less precise (by ~5-10%) than those of multiVORFFIP on HR-PDNA187** and APO82, but equally or more precise than DISPLAR and multiVORFFIP on the other datasets (Table F and Table I in S1 File and Fig 4). This results in ~5-10% higher F1 values and most of the times ~5-10% higher MCC values compared to both tools (Table F and Table I in S1 File and Fig 4). Compared to DRNApred, iJET$^2_{\mathrm{DNA}}$'s detection is at least twice more sensitive and at least ~10% more precise on all the datasets (Table F and Table I in S1 File and Fig 4). We should stress that DRNApred was designed to specifically discriminate DNA- from RNA-binding residues, which is not the purpose of JET$^2_{\mathrm{DNA}}$. In particular, JET$^2_{\mathrm{DNA}}$ performs much better than the three other tools on the RNA-binding proteins of TEST-RBP33 (Table I in S1 File and Fig 4). Moreover, when looking at the four above mentioned cases of multiple DNA-binding sites, JET$^2_{\mathrm{DNA}}$ proved more powerful to detect alternative binding sites compared to the three other tools (Fig. J-L in S2 File and Table K in S1 File).

Comparison with other benchmarks

We compared HR-PDNA187 and HOLO-APO82 with the most popular benchmarks, namely PDNA62 [6, 64], DBP374 [13], MetaDBsite316 [55], Displar264 [7], PDDB1.2 [65], DBP206 [66], PDNA224 [10] and DNABINDPROT54 [24]. Contrary to our datasets, most of them comprise only single chains, even if the functional biological unit of the protein in complex with DNA is annotated as a multimer. Moreover, when applying to them the PISCES criteria used to construct HR-PDNA187, the number of resulting complexes was systematically smaller than 187 (Table L in S1 File). Finally, very few of them provide the APO forms of the proteins.

Robustness of the results to parameter changes

JET$^2_{\mathrm{DNA}}$ parameters were set empirically based on our previous experience with JET and JET^2, our intuition and our analysis of the properties encoded in the experimental interfaces from

HR-PDNA187. However, we should stress that we did not use machine learning to infer them. Hence, HR-PDNA187 is not a training set *per se*. Moreover, JET^2_{DNA} parameters are few, physically explained and biologically intuitive. To validate our choices, we conducted a thorough analysis of the impact of varying parameters on the predictions (Fig. M-W in S2 File). In total, we ran about 220 000 additional JET^2_{DNA} calculations to assess the influence of seven parameters. These include the thresholds (% of residues and CV_{local}) used in the detection of small ligand pockets (Fig. Q-W in S2 File), the residue and cluster thresholds used to detect and grow patches (Fig. M-P in S2 File) and the confidence threshold used to filter out small patches (Fig. N-P in S2 File). This analysis showed that our choice of parameters is relevant, as the performances obtained with our default values are close to the best one can expect within the JET^2_{DNA} framework. Moreover, we found that our default values are consistent with the intervals/regions they fall in and that there are no abrupt changes in performance within these regions. Hence, our predictions are stable to small parameter changes.

Discussion

We have collected and carefully curated 187 high resolution protein-DNA complexes representative of all known types of protein-DNA interactions. This new dataset, supplemented by the 82 available protein unbound conformations, could serve as a reference benchmark for the community. Based on the analysis of the evolutionary and structural properties of the DNA-binding sites from this dataset, we have proposed an original method to predict them. Importantly, we do not estimate interacting probabilities for individual residues, but we predict ensembles of residues proximal in 3D space, in other words "patches". Predictions based on patches are justified by the fact that residues being conserved or displaying specific physico-chemical properties at protein-DNA interfaces tend to cluster together [58, 67, 68]. Moreover, we propose a "discretized model" of interfaces describing the role of each predicted residue in the interface. This model is inspired by the *support-core-rim* model defined for protein-protein interfaces. We have shown that it is also useful for protein-DNA interfaces, where the *support* and the *core* may switch their positions. We have defined three archetypal protein-DNA interfaces, namely conserved generic, conserved enveloping and not conserved, and have devised three scoring schemes to detect them.

We have thoroughly assessed our method's performance on bound and unbound protein forms and on an independent dataset. JET^2_{DNA} compares favorably with machine learning based state-of-the-art methods. Specifically, the predictions are more sensitive and remarkably robust to extensive conformational changes and to stochiometry changes. Moreover, we have highlighted several cases where the same complex was solved in different conditions or the same protein was able to bind DNA via different locations. JET^2_{DNA} was able to detect the 'alternative' binding sites despite major conformational changes between the different structures bound to DNA. A single crystallographic structure may reveal only one site or may even comprise a truncated or misplaced DNA, resulting in a "partial" associated binding site. In this context, JET^2_{DNA} can be used as a mean to get a more realistic description of known binding sites, when those are only partially covered experimentally, and to discover yet unknown binding sites.

Beyond predicting DNA-binding sites, JET^2_{DNA} provides a unique way to understand the origins and properties of these sites and interpret those in light of their functions. Transcription factors typically display single- or double-headed binding modes [69], with one or two highly conserved binding sites, which are well detected by D-SC1. Enzymes usually have larger interfaces to accommodate an exposed recognition site, detected by D-SC3, and a highly conserved active site, detected by D-SC1, or highly segmented protein-DNA interfaces, where the protein interacts with the DNA through multidomain units in addition to their active site [25, 69, 70]. Moreover, JET^2_{DNA} is useful to unravel the heterogeneity of signals comprised within given binding sites and to partition them in subregions displaying coherent properties. The predictions can help designing or repurposing small molecules to target protein-DNA interfaces in an intelligent way, *e.g.* specifically targeting the non-conserved subregions to avoid side effects.

There is growing experimental evidence that protein surfaces are used in many different ways by many partners. For example, for protein-protein interactions, it is now clear that several proteins can use the same region at the surface of a partner, possibly in different conformations, that binding sites can be overlapping and that different binding site properties relate to different interaction "types" or "functions" [33]. For protein-DNA interactions, the available structural data is much smaller. Still, the examples we showed let us envision a much larger complexity in the usage of protein surfaces by DNA than expected. Our work is inscribed in an effort to decipher such complexity.

Finally, we found that JET^2_{DNA} can be successfully applied to the prediction of RNA-binding sites. This is in line with the growing body of evidence showing that proteins that bind DNA are also likely to bind RNA. Indeed, although DNA-binding proteins used to be considered as functionally different from RNA-binding proteins and studied independently, this view has become outdated [29]. There are many examples of proteins binding both nucleic acids either via the same region at different times or simultaneously via distinct regions. Some crystallographic structures of complexes between proteins and hybrid D/RNA molecules are also available in the PDB. Hence, discriminating DNA- from RNA-binding residues is very challenging. The DRNApred method [18] represents a recent effort to address this issue. We found that it predicts very few residues, compared to JET^2_{DNA}. A future development for JET^2_{DNA} could be to include some RNA-specific features toward a better prediction of RNA-binding sites. It could be advantageous to include disorder information, an electrostatic potential in addition or replacing the residue propensities and/or restricting the calculation of the evolutionary conservation to the subfamily associated to the query protein. This could help in identifying family- or even protein-specific interfaces.

Supporting information

S1 File. Supplementary tables.
(PDF)

S2 File. Supplementary figures.
(PDF)

Acknowledgments

We thank Chloé Dequeker for sharing her scripts, N. Fernandez-Fuentes for indicating good practice for using the multiVORFFIP method and DISPLAR's authors for helping us in running the tool.

Author Contributions

Conceptualization: Richard Lavery, Elodie Laine, Alessandra Carbone.

Data curation: Flavia Corsi.

Funding acquisition: Alessandra Carbone.

Methodology: Elodie Laine, Alessandra Carbone.

Software: Flavia Corsi, Elodie Laine.

Supervision: Elodie Laine, Alessandra Carbone.

Validation: Flavia Corsi.

Writing – original draft: Flavia Corsi, Elodie Laine, Alessandra Carbone.

References

1. Luscombe NM, Austin SE, Berman HM, Thornton JM. An overview of the structures of protein-DNA complexes. Genome Biology. 2000; 1(1):reviews001–1. https://doi.org/10.1186/gb-2000-1-1-reviews001 PMID: 11104519
2. Bullock AN, Fersht AR. Rescuing the function of mutant p53. Nature Reviews Cancer. 2001; 1(1):68. https://doi.org/10.1038/35094077 PMID: 11900253
3. Chen-Plotkin AS, Lee VMY, Trojanowski JQ. TAR DNA-binding protein 43 in neurodegenerative disease. Nature Reviews Neurology. 2010; 6(4):211. https://doi.org/10.1038/nrneurol.2010.18 PMID: 20234357
4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Research. 2000; 28(1):235–242. https://doi.org/10.1093/nar/28.1.235 PMID: 10592235
5. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics. 2005; 6(1):33. https://doi.org/10.1186/1471-2105-6-33 PMID: 15720719
6. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics. 2004; 20(4):477–486. https://doi.org/10.1093/bioinformatics/btg432 PMID: 14990443
7. Tjong H, Zhou HX. DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. Nucleic Acids Research. 2007; 35(5):1465–1477. https://doi.org/10.1093/nar/gkm008 PMID: 17284455
8. Wang L, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Research. 2006; 34(suppl_2):W243–W248. https://doi.org/10.1093/nar/gkl298 PMID: 16845003
9. Wang L, Huang C, Yang MQ, Yang JY. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Systems Biology. 2010; 4(1):S3. https://doi.org/10.1186/1752-0509-4-S1-S3 PMID: 20522253
10. Li T, Li QZ, Liu S, Fan GL, Zuo YC, Peng Y. PreDNA: accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information. Bioinformatics. 2013; 29(6):678–685. https://doi.org/10.1093/bioinformatics/btt029 PMID: 23335013
11. Hwang S, Gou Z, Kuznetsov IB. DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics. 2007; 23(5):634–636. https://doi.org/10.1093/bioinformatics/btl672 PMID: 17237068
12. Wang L, Yang MQ, Yang JY. Prediction of DNA-binding residues from protein sequence information using random forests. BMC Genomics. 2009; 10(1):S1. https://doi.org/10.1186/1471-2164-10-S1-S1 PMID: 19594868
13. Wu J, Liu H, Duan X, Ding Y, Wu H, Bai Y, et al. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. Bioinformatics. 2008; 25(1):30–35. https://doi.org/10.1093/bioinformatics/btn583 PMID: 19008251
14. Nimrod G, Schushan M, Szilágyi A, Leslie C, Ben-Tal N. iDBPs: a web server for the identification of DNA binding proteins. Bioinformatics. 2010; 26(5):692–693. https://doi.org/10.1093/bioinformatics/btq019 PMID: 20089514
15. Segura J, Jones PF, Fernandez-Fuentes N. A holistic in silico approach to predict functional sites in protein structures. Bioinformatics. 2012; 28(14):1845–1850. https://doi.org/10.1093/bioinformatics/bts269 PMID: 22563069
16. Zhou J, Xu R, He Y, Lu Q, Wang H, Kong B. PDNAsite: identification of DNA-binding site from protein sequence by incorporating spatial and sequence context. Scientific reports. 2016; 6:27653. https://doi.org/10.1038/srep27653 PMID: 27282833
17. Zhu X, Ericksen SS, Mitchell JC. DBSI: DNA-binding site identifier. Nucleic acids research. 2013; 41(16):e160–e160. https://doi.org/10.1093/nar/gkt617 PMID: 23873960

18. Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA-and RNA-binding residues. Nucleic acids research. 2017; 45(10):e84–e84. https://doi.org/10.1093/nar/gkx059 PMID: 28132027

19. Miao Z, Westhof E. Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. Nucleic acids research. 2015; 43(11):5340–5351. https://doi.org/10.1093/nar/gkv446 PMID: 25940624

20. Miao Z, Westhof E. RBscore&NBench: a high-level web server for nucleic acid binding residues prediction with a large-scale benchmarking database. Nucleic acids research. 2016; 44(W1):W562–W567. https://doi.org/10.1093/nar/gkw251 PMID: 27084939

21. Tsuchiya Y, Kinoshita K, Nakamura H. PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces. Bioinformatics. 2004; 21(8):1721–1723. https://doi.org/10.1093/bioinformatics/bti232 PMID: 15613393

22. Chen YC, Wright JD, Lim C. DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry. Nucleic Acids Research. 2012; 40(W1): W249–W256. https://doi.org/10.1093/nar/gks481 PMID: 22661576

23. Gao M, Skolnick J. DBD-Hunter: a knowledge-based method for the prediction of DNA–protein interactions. Nucleic Acids Research. 2008; 36(12):3978–3992. https://doi.org/10.1093/nar/gkn332 PMID: 18515839

24. Ozbek P, Soner S, Erman B, Haliloglu T. DNABINDPROT: fluctuation-based predictor of DNA-binding residues within a network of interacting residues. Nucleic Acids Research. 2010; 38(suppl_2):W417–W423. https://doi.org/10.1093/nar/gkq396 PMID: 20478828

25. Wodak SJ, Janin J. Structural basis of macromolecular recognition. Advances in Protein Chemistry. 2002; 61:9–73. https://doi.org/10.1016/s0065-3233(02)61001-0 PMID: 12461820

26. Cheetham GM, Jeruzalmi D, Steitz TA. Structural basis for initiation of transcription from an RNA polymerase–promoter complex. Nature. 1999; 399(6731):80. https://doi.org/10.1038/19999 PMID: 10331394

27. Yin YW, Steitz TA. Structural basis for the transition from initiation to elongation transcription in T7 RNA polymerase. Science. 2002; 298(5597):1387–1395. https://doi.org/10.1126/science.1077464 PMID: 12242451

28. Li X, Shu C, Yi G, Chaton CT, Shelton CL, Diao J, et al. Cyclic GMP-AMP synthase is activated by double-stranded DNA-induced oligomerization. Immunity. 2013; 39(6):1019–1031. https://doi.org/10.1016/j.immuni.2013.10.019 PMID: 24332030

29. Hudson WH, Ortlund EA. The structure, function and evolution of proteins that bind DNA and RNA. Nat Rev Mol Cell Biol. 2014; 15(11):749–760. https://doi.org/10.1038/nrm3884 PMID: 25269475

30. Andrabi M, Mizuguchi K, Ahmad S. Conformational changes in DNA-binding proteins: Relationships with precomplex features and contributions to specificity and stability. Proteins: Structure, Function, and Bioinformatics. 2014; 82(5):841–857. https://doi.org/10.1002/prot.24462

31. Mierzejewska K, Siwek W, Czapinska H, Kaus-Drobek M, Radlinska M, Skowronek K, et al. Structural basis of the methylation specificity of R. DpnI. Nucleic Acids Research. 2014; 42(13):8745–8754. https://doi.org/10.1093/nar/gku546 PMID: 24966351

32. Laine E, Carbone A. Local geometry and evolutionary conservation of protein surfaces reveal the multiple recognition patches in protein-protein interactions. PLoS Computational Biology. 2015; 11(12): e1004580. https://doi.org/10.1371/journal.pcbi.1004580 PMID: 26690684

33. Dequeker C, Laine E, Carbone A. Decrypting protein surfaces by combining evolution, geometry and molecular docking. Proteins: Structure, Function, and Bioinformatics. 2019; 87(11):952–965. https://doi.org/10.1002/prot.25757

34. Levy ED. A simple definition of structural regions in proteins and its use in analyzing interface evolution. Journal of Molecular Biology. 2010; 403(4):660–670. https://doi.org/10.1016/j.jmb.2010.09.028 PMID: 20868694

35. Chakrabarti P, Janin J. Dissecting protein–protein recognition sites. Proteins: Structure, Function, and Bioinformatics. 2002; 47(3):334–343. https://doi.org/10.1002/prot.10085

36. Bahadur RP, Chakrabarti P, Rodier F, Janin J. Dissecting subunit interfaces in homodimeric proteins. Proteins: Structure, Function, and Bioinformatics. 2003; 53(3):708–719. https://doi.org/10.1002/prot.10461

37. Biswas S, Guharoy M, Chakrabarti P. Dissection, residue conservation, and structural classification of protein-DNA interfaces. Proteins: Structure, Function, and Bioinformatics. 2009; 74(3):643–654. https://doi.org/10.1002/prot.22180

38. Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, et al. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophysical Journal. 1992; 63(3):751–759. https://doi.org/10.1016/S0006-3495(92)81649-1 PMID: 1384741

39. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics. 2003; 19 (12):1589–1591. https://doi.org/10.1093/bioinformatics/btg224 PMID: 12912846

40. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10(1):421. https://doi.org/10.1186/1471-2105-10-421 PMID: 20003500

41. Consortium U. UniProt: the universal protein knowledgebase. Nucleic Acids Research. 2016; 45(D1): D158–D169.

42. Chou CJ, Wisedchaisri G, Monfeli RR, Oram DM, Holmes RK, Hol WG, et al. Functional studies of the Mycobacterium tuberculosis iron-dependent regulator. Journal of Biological Chemistry. 2004; 279 (51):53554–53561. https://doi.org/10.1074/jbc.M407385200 PMID: 15456786

43. Hubbard S, Thornton J. NACCESS, 2.1. 1. Dept of Biochemistry and Molecular Biology: University College London. 1993;.

44. Yan J, Friedrich S, Kurgan L. A comprehensive comparative review of sequence-based predictors of DNA-and RNA-binding residues. Briefings in bioinformatics. 2015; 17(1):88–105. https://doi.org/10.1093/bib/bbv023 PMID: 25935161

45. Chen K, Mizianty MJ, Gao J, Kurgan L. A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. Structure. 2011; 19(5):613–621. https://doi.org/10.1016/j.str.2011.02.015 PMID: 21565696

46. Engelen S, Trojan LA, Sacquin-Mora S, Lavery R, Carbone A. Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. PLoS Computational Biology. 2009; 5(1):e1000267. https://doi.org/10.1371/journal.pcbi.1000267 PMID: 19165315

47. Mihalek I, Reš I, Lichtarge O. A family of evolution–entropy hybrid methods for ranking protein residues by importance. Journal of Molecular Biology. 2004; 336(5):1265–1282. https://doi.org/10.1016/j.jmb.2003.12.078 PMID: 15037084

48. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. Journal of Molecular Biology. 1996; 257(2):342–358. https://doi.org/10.1006/jmbi.1996.0167 PMID: 8609628

49. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997; 25 (17):3389–3402. https://doi.org/10.1093/nar/25.17.3389 PMID: 9254694

50. Park B, Kim H, Han K. DBBP: database of binding pairs in protein-nucleic acid interactions. BMC Bioinformatics. 2014; 15(15):S5. https://doi.org/10.1186/1471-2105-15-S15-S5 PMID: 25474259

51. Ceres N, Pasi M, Lavery R. A protein solvation model based on residue burial. Journal of Chemical Theory and Computation. 2012; 8(6):2141–2144. https://doi.org/10.1021/ct3001552 PMID: 26593844

52. Mezei M. A new method for mapping macromolecular topography. Journal of Molecular Graphics and Modelling. 2003; 21(5):463–472. https://doi.org/10.1016/s1093-3263(02)00203-6 PMID: 12543141

53. Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1945; 1(6):80–83. https://doi.org/10.2307/3001968

54. Anderson TW, Darling DA, et al. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. The annals of mathematical statistics. 1952; 23(2):193–212. https://doi.org/10.1214/aoms/1177729437

55. Si J, Zhang Z, Lin B, Schroeder M, Huang B. MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. BMC Systems Biology. 2011; 5(1):S7. https://doi.org/10.1186/1752-0509-5-S1-S7 PMID: 21689482

56. Miao Z, Westhof E. A large-scale assessment of nucleic acids binding site prediction programs. PLoS computational biology. 2015; 11(12):e1004639. https://doi.org/10.1371/journal.pcbi.1004639 PMID: 26681179

57. Hwang H, Vreven T, Janin J, Weng Z. Protein–protein docking benchmark version 4.0. Proteins: Structure, Function, and Bioinformatics. 2010; 78(15):3111–3114. https://doi.org/10.1002/prot.22830

58. Ahmad S, Keskin O, Sarai A, Nussinov R. Protein–DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. Nucleic Acids Research. 2008; 36 (18):5922–5932. https://doi.org/10.1093/nar/gkn573 PMID: 18801847

59. Luscombe NM, Thornton JM. Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. Journal of Molecular Biology. 2002; 320(5):991–1009. https://doi.org/10.1016/s0022-2836(02)00571-5 PMID: 12126620

60. Durniak KJ, Bailey S, Steitz TA. The structure of a transcribing T7 RNA polymerase in transition from initiation to elongation. Science. 2008; 322(5901):553–557. https://doi.org/10.1126/science.1163433 PMID: 18948533

61. Hickman AB, Ronning DR, Perez ZN, Kotin RM, Dyda F. The nuclease domain of adeno-associated virus rep coordinates replication initiation using two distinct DNA recognition interfaces. Molecular Cell. 2004; 13(3):403–414. https://doi.org/10.1016/s1097-2765(04)00023-1 PMID: 14967147

62. Siwek W, Czapinska H, Bochtler M, Bujnicki JM, Skowronek K. Crystal structure and mechanism of action of the N6-methyladenine-dependent type IIM restriction endonuclease R. DpnI. Nucleic Acids Research. 2012; 40(15):7563–7572. https://doi.org/10.1093/nar/gks428 PMID: 22610857

63. Gao P, Ascano M, Wu Y, Barchet W, Gaffney BL, Zillinger T, et al. Cyclic G (2', 5') pA (3', 5') p is the metazoan second messenger produced by DNA-activated cyclic GMP-AMP synthase. Cell. 2013; 153 (5):1094–1107. https://doi.org/10.1016/j.cell.2013.04.046 PMID: 23647843

64. Selvaraj S, Kono H, Sarai A. Specificity of protein–DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/non-cognate binding. Journal of Molecular Biology. 2002; 322 (5):907–915. https://doi.org/10.1016/s0022-2836(02)00846-x PMID: 12367517

65. Van Dijk M, Bonvin AM. A protein–DNA docking benchmark. Nucleic Acids Research. 2008; 36(14): e88–e88. https://doi.org/10.1093/nar/gkn386 PMID: 18583363

66. Xiong Y, Liu J, Wei DQ. An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins: Structure, Function, and Bioinformatics. 2011; 79(2):509–517. https://doi.org/10.1002/prot.22898

67. Dey S, Pal A, Guharoy M, Sonavane S, Chakrabarti P. Characterization and prediction of the binding site in DNA-binding proteins: improvement of accuracy by combining residue composition, evolutionary conservation and structural parameters. Nucleic Acids Research. 2012; 40(15):7150–7161. https://doi.org/10.1093/nar/gks405 PMID: 22641851

68. Jones S, Shanahan HP, Berman HM, Thornton JM. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Research. 2003; 31(24):7189–7198. https://doi.org/10.1093/nar/gkg922 PMID: 14654694

69. Jones S, van Heyningen P, Berman HM, Thornton JM. Protein-DNA interactions: A structural analysis1. Journal of Molecular Biology. 1999; 287(5):877–896. https://doi.org/10.1006/jmbi.1999.2659 PMID: 10222198

70. Nadassy K, Wodak SJ, Janin J. Structural features of protein- nucleic acid recognition sites. Biochemistry. 1999; 38(7):1999–2017. https://doi.org/10.1021/bi982362d PMID: 10026283