3298–3308 *Nucleic Acids Research, 2018, Vol. 46, No. 7* *Published online 9 March 2018*
doi: 10.1093/nar/gky170

Structural changes in DNA-binding proteins on complexation

Sayan Poddar[1,†], Devlina Chakravarty[2,†] and Pinak Chakrabarti[1,2,*]

[1]Department of Biochemistry, Bose Institute, P1/12 CIT Scheme VIIM, Kolkata 700054, India and [2]Bioinformatics Centre, Bose Institute, P1/12CIT Scheme VIIM, Kolkata 700054, India

Received September 26, 2017; Revised February 21, 2018; Editorial Decision February 22, 2018; Accepted March 02, 2018

ABSTRACT

Characterization and prediction of the DNA-biding regions in proteins are essential for our understanding of how proteins recognize/bind DNA. We analyze the unbound (U) and the bound (B) forms of proteins from the protein–DNA docking benchmark that contains 66 binary protein–DNA complexes along with their unbound counterparts. Proteins binding DNA undergo greater structural changes on complexation (in particular, those in the enzyme category) than those involved in protein–protein interactions (PPI). While interface atoms involved in PPI exhibit an increase in their solvent-accessible surface area (ASA) in the bound form in the majority of the cases compared to the unbound interface, protein–DNA interactions indicate increase and decrease in equal measure. In 25% structures, the U form has missing residues which are located in the interface in the B form. The missing atoms contribute more toward the buried surface area compared to other interface atoms. Lys, Gly and Arg are prominent in disordered segments that get ordered in the interface on complexation. In going from U to B, there may be an increase in coil and helical content at the expense of turns and strands. Consideration of flexibility cannot distinguish the interface residues from the surface residues in the U form.

INTRODUCTION

The interactions between DNA and proteins play a pivotal role in almost every cellular process, such as regulation of gene expression, DNA replication, rearrangement, repair, chromatin formation and organization, etc. (1). DNA-binding proteins have evolved to have a specific or general affinity for either single or double stranded DNA (2). The most intensively studied of these are the various transcription factors, each of which binds to one particular set of DNA sequence and activates or inhibits the transcription of genes (3). Generally, these proteins bind to DNA in the major groove due to the greater accessibility of the bases; however, there are also some proteins which bind in the minor groove (4). Protein–DNA interactions are mainly of two types, specific and non-specific (5,6). In case of non-specific interactions, as far as the binding interactions are concerned, the nucleotide sequence does not matter. This is important in a variety of contexts related to DNA packaging and nucleoprotein complex formation (7), and the interactions occur between functional groups on the protein and the sugar-phosphate backbone of DNA. On the other hand specific DNA–protein interactions depend not only on the specific sequence of bases but also on the orientation of the bases in the nucleotide (8). These DNA–protein interactions are strong and are mediated by various types of bonding, such as hydrogen bonding which can be direct or indirect, mediated by water molecules, ionic interactions such as salt bridges, protein side chains-DNA backbone interactions, as well as others, like van der Waals and hydrophobic interactions.

Protein–DNA interactions have been characterized by analyzing the interface formed between the protein and DNA of a large number of protein–DNA complexes (1,9–11). In addition to physicochemical features and pattern of hydrogen bonding, conservation of residues, their clustering etc. have been found to be important in distinguishing DNA-binding patch from the rest of the protein surface (12). Many of these features are used to predict protein–DNA binding affinities (13–15), as well as distinguishing between single- or double-stranded DNA-binding proteins (16). These studies, however, have one bottleneck, which make their applicability in the development of a general docking algorithm rather strenuous (17)—they use the static and bound form of the protein as found in protein–DNA complexes. However, it is well-known that protein structure may undergo considerable changes while forming a complex, whether between protein molecules or between protein and DNA or RNA. Indeed, the genome-wide anal-

*To whom correspondence should be addressed. Tel: +91 33 25693253; Fax: +91 33 23553888; Email: pinak@jcbose.ac.in
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Present address: Devlina Chakravarty, Center for Computational Biology, The University of Kansas, Lawrence, KS, USA.

yses have indicated that many of the transcription factors are intrinsically disordered (18–20).

Recently, we have compared the bound and unbound forms of proteins involved in protein–protein interactions and shown that there are distinct changes in structural features accompanying complex formation (21,22). We extend the study to proteins involved in DNA binding using protein–DNA docking benchmark (23). We analyze the conformational changes that take place in the residues that constitute the interface on binding DNA, employing simple parameters such as accessible surface area (ASA) and root mean square deviations (RMSD). Missing segments as seen for the unbound protein and secondary structure assumed on binding are also studied, as also the change in crystallographic temperature factors (*B* factors) of interface atoms on complex formation. Compared to protein–protein interactions, interface residues in DNA-binding proteins may exhibit greater conformational changes (disorder-to-order transitions, in particular) and the interface atoms a larger change in accessible surface area (either increase or decrease) on complexation. Attempt has been made to correlate the structural changes with the change in free energy of DNA-binding of residues on mutation to Ala.

MATERIALS AND METHODS

We used the protein–DNA docking benchmark (23) that contains 66 binary protein–DNA complexes and their unbound counterparts. In the dataset, there are 16 NMR entries with multiple models for each structure; for our analysis only the first model was considered. In 41 structures the monomeric chain was bound to DNA, and dimers in 25 cases. For each protein, we designated the unbound structure as U and the bound structure (isolated from its partner DNA) as B; the bound structure (together with DNA) in the complex is C. For both bound [B] and unbound [U] structures the accessible surface area (ASA) values were calculated separately using the NACCESS program (24), which employs the Lee and Richards algorithm (25).

We used EMBOSS software (26) to perform the local alignment (Smith–Waterman algorithm) and global alignment (Needleman and Wunsch) of the polypeptide chains constituting the U/B pairs. 52 of the 66 U/B pairs have sequence identity ≥98%. Based on the sequence alignment, the interface residues as seen in the complex were mapped to those in the unbound state using PROFIT (27) and Biopython (28). We have also identified the interface atoms in our analysis, which are the atoms losing more than 0.1 Å^2 of surface area upon complexation formation (B to C) (29). Conformational changes were measured using both the backbone and interface RMSDs, which were calculated on equivalent C$^\alpha$ positions, as well as all atoms comprising the interface, respectively, after superposition (using all the non-hydrogen backbone atoms) with PROFIT. Residues showing large RMSDs were also identified. There are also some proteins that have modified residues (Supplementary Table S1A) which are tagged HETATM in PDB (instead of the usual ATOM); selenomethionine residues belonging to this category were manually edited to Met and tagged as ATOM. Supplementary Table S1b documents the structures where residue name differs in U and B. As a result of order-to-disorder transition in the two PDB files, some coordinates may be missing in the U state, the information on these missing residues/atoms are provided in Supplementary Table S1C. The residues given in Supplementary Table S1B and those in Supplementary Table S1C for which all the atoms are missing were not included in the ASA calculations; however, when all the atoms are not missing those present, common to both U and B states were included.

While calculating ASA(B) and ASA(U), we have used only the interface atoms common to both as we had done for the previous analysis of protein–protein structures (21,22). The analysis of atoms mainly depends on matching labels that may often be assigned in an arbitrary fashion. For example, the corresponding atom in U might be labelled OD1 or OD2 for an atom OD1 of an interface residue Asp in a bound structure B. As a result direct comparison of the ASA values of two atoms having similar label is not justified. Residues which have ambiguous atom label pairs are Asp (OD1/OD2), Arg (NH1/NH2), Asn (OD1/ND2), Glu (OE1/OE2), Gln (OE1/NE2), His (CE1/NE2, CD2/ND1), Phe (CD1/CD2, CE1/CE2), Leu (CD1/CD2), Val (CG1/CG2), Tyr (CD1/CD2, CE1/CE2). To circumvent the problem, atoms with both the labels were taken for calculation of ASA. This would mean that if OD1 is present in the interface, both OD1 and OD2 are taken to be a part of the interface. This increases the number of interface atoms by 7%. Based on the calculated ASA value for U and B, ΔASA and δASA values have been calculated as follows.

$$\Delta\mathrm{ASA} = [\mathrm{ASA(B)} - \mathrm{ASA(U)}],$$

where ASA(B) is the solvent accessible surface area of the interface atoms in the complex state, and ASA(U) is the accessible surface area of the equivalent mapped atoms in the isolated state.

$$\delta\mathrm{A} = \Delta\mathrm{ASA}/\mathrm{ASA(B)},$$

difference in ASA relative to the total value in the complex state. It may be mentioned that ΔASA and δA are the average values for all the interface atoms in a given structure. In some places we have used ΔASA for a given residue, or calculated δA using all the atoms of the interface residue, or for the residues located in protein surface—these have been mentioned explicitly.

Additionally, δA was also calculated for surface and interface residues (considering all the residue atoms). The buried surface area, BSA = [ASA(B) – ASA(C)], is calculated using all the interface atoms (22).

Secondary structure was calculated using DSSP (30) software. Changes in secondary structural composition were enumerated in terms of changes in helix, strand, turn and coil (ΔH, ΔS, ΔT and ΔC). To quantify the change in the percentage composition of the secondary structures for the interface residues between the bound ($n^{\mathrm{B}}{}_{\mathrm{i}}$) and the unbound ($n^{\mathrm{U}}{}_{\mathrm{i}}$) states the Euclidean distance was calculated as

$$D = \sqrt{\left(\Sigma_i^m (n_i^B - n_i^U)^2/(m-1)\right)},$$

where $m = 4$ (different forms of secondary structure, i.e. helix, strand, turn and coil).

B factors were analyzed to discern the flexibility of the interface and surface regions. The normalized values were used, defined as follows:

$$b_{fr'} = [b_{fr} - \mu(bf)]/\sigma(bf),$$

where b_{fr} is the average B factor of C, C^{α}, O, N and C^{β} of the residue r (C^{β} cannot be considered when the residue is Gly), $\mu(bf)$ and $\sigma(bf)$ are the mean and the standard deviation of B factors for that chain, respectively. After scaling, the b_{fr}' values were used to derive the averages over the interface, surface and core and rim regions of the interface (29). The Euclidean metrics, Δb, for the B factors of residues in different states/structural regions were calculated in a similar way, $\Delta b = \sqrt{(\Sigma_i^n (bf^{(1)}{}_i - bf^{(2)}{}_i)^2/(n\text{-}1))}$, where n represents the number of amino acid types, and $bf^{(1)}{}_i$, $bf^{(2)}{}_i$ are the scaled B factors of residue type i in states 1 and 2, respectively. The states compared were interface, non-interface, bound and unbound.

RESULTS

Changes in accessible surface area (ASA) and root mean square deviations (RMSD) in going from U to B states

The distribution of δA (Figure 1A) shows a rather bimodal distribution, with almost equal number of proteins having positive (34 cases) and negative (32) δA values, with an average of –0.15 ± 11%. (Instead of using the first model, if all the models were used for NMR structures, the average value would be –0.24 ± 11%, not a significant change, as was observed earlier on using NMR models in the analysis of protein–protein interactions (22)). Two examples of large δA values (both positive and negative) are shown in Figure 2 (31–33). As a control we have plotted the distribution of δA values for surface residues, which has a small average value (–0.003 ± 8.6%) as seen above, but the distribution now is quite normal (Supplementary Figure S1). δA, when calculated based on the whole interface residue (Figure 1B), indicates a trend towards having a negative value (–4.4 ± 10%).

We checked if depending on their type the residues may favour +ve or –ve δA values in Figure 1A. Supplementary Table S2 provides the number of +ve and –ve cases for all the 20 amino acids, which indicates that although overall there is not much distinction (P value = 0.89), there seems to be a slight excess of +ve values (P value = 0.28) for five hydrophobic residues (excluding Leu), whereas the negatively-charged Asp has distinctly more number of –ve cases. If we consider all the atoms of the interface residues (corresponding to Figure 1B), the trend becomes more prominent with all the charged residues showing an excess of –ve δA values.

Conformational changes were also measured based on RMSD values. The scattered plot of interface and backbone RMSDs (Figure 3A) shows that the former is mostly greater than the latter (points above the diagonal line, except for five structures). The average backbone RMSD (only C^{α}) is 2.5 ± 1.9 Å and average interface RMSD (using all interface atoms) is 3.7 ± 2.3 Å. There are 12 (out of 66) structures (18.2%) whose backbone RMSD is greater than 4 Å and 20 structures (30%) with interface RMSD > 4 Å. In comparison, for protein–protein complexes, the average RMSD is 1.4 ± 1.6 Å and 2.2 ± 1.5 Å for backbone and interface, respectively (Figure 3B). Histograms for both the backbone and interface RMSDs can be compared between protein–protein and protein–DNA datasets (Supplementary Figure S2), which shows that the changes in both backbone as well the interface are more for the protein–DNA interaction.

To understand what might cause the occurrence of large ΔASA values (positive, as well as negative) we calculated the residue-wise RMSD values. A scattered plot for ΔASA versus RMSD for a few such residues is shown in Figure 4 (34–42), which shows that residues with high RMSD values tend to have high ΔASA values also. The residues undergoing such extreme changes have been shown in Supplementary Figure S3. The structures are mostly of enzymes and some of the residues are discussed below.

In HhaI methyltransferase oligonucleotide complex (34,35), there are mismatched bases in the substrate that are flipped out of the DNA helix and pushed in the active-site pocket of the enzyme. This results in the entry of residues, such as Ile86 into the helix, which shows a large change in RMSD and a positive ΔASA (Supplementary Figure S3A).

Type II restriction endonucleases, such as BamHI (36,37) recognize short (four to eight base pairs) palindromic DNA sequences and cleave both the strands and contain at least three residues, mostly acidic that bind divalent cations, which are essential for activity. Interestingly, changes associated with DNA binding causes a large RMSD and positive ΔASA for a hydrophobic residue, Met198 (Supplementary Figure S3B). Similarly, in a monomeric endocuclease, BcnI (31), which introduces double-strand breaks by sequentially nicking individual DNA strands, exhibit large positive ΔASA values for Ile51 and Arg30 (Figure 2A)

Restriction endonuclease, HinP1I cleaves the palindromic tetranucleotide sequence G↓CGC. It is a 2-fold related dimer with two active sites and two DNA duplexes bound on the outer surfaces of the dimer facing away from each other (38,39). Phe91 intercalates the duplex from the major groove causing the DNA to be kinked by ∼60°. Upon binding to cognate DNA the largest change in HinP1I occurs in the N-terminal 17 residues, which become part of a long helix that binds to the minor groove. Phe15 lying in this region, as well as Phe91 mentioned earlier both have high RMSD values as well as large positive ΔASA values (Supplementary Figure S3C).

TAL (transcription activator-like) effectors (40) are major virulence factors secreted by bacteria that cause diseases in plants. They recognize host DNA sequence through a central domain of tandem repeats, each comprising of 33–35 conserved amino acids that targets a specific base pair by using two hypervariable residues [known as repeat variable diresidues (RVD)] at positions 12 and 13. The structure of each repeat consists of two helices connected by a short RVD-containing loop, which contacts the DNA major groove. The 12th residue stabilizes the RVD loop, wherea the 13th makes a base-specific contact. The frequently occurring RVDs, His/Asp (HD), Asn/Gly (NG) and Asn/Ile (NI) recognize three distinct bases. In the structure with HD as the RVD (Supplementary Figure S3D), the Asp residues exhibit partner accommodation effect resulting in large negative ΔASA, as well as high RMSD values.

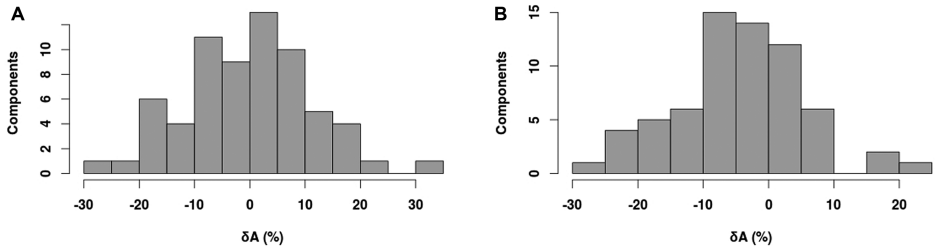

Figure 1. Distribution of δA values using interface (**A**) atoms, and (**B**) residues. The distribution of δA values using only interface atoms shows a bimodal distribution where both partner attraction and partner accommodation is taking place (average = –0.15 ± 11%), while the distribution for interface residues shows that partner accommodation effect prevails (average = –4.4 ± 10%).

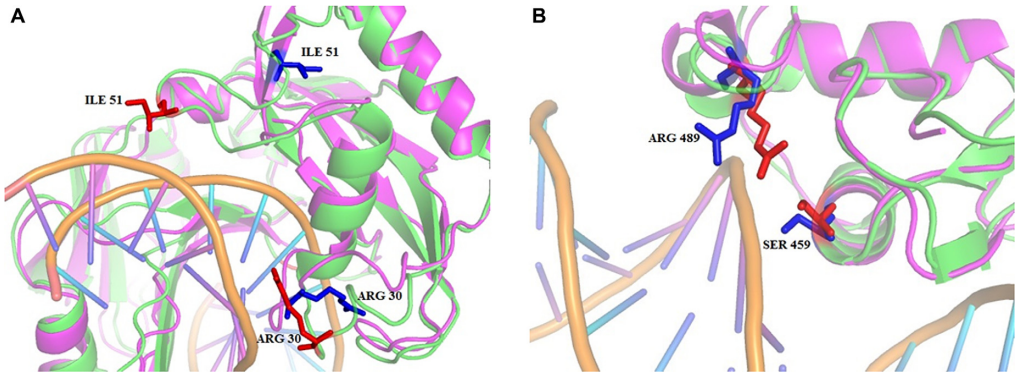

Figure 2. Examples of local movements leading large δA values, +ve (that makes interface residues more accessible in the bound state) in (**A**) and –ve in (**B**). The unbound protein is in pink, and the bound in green; DNA strands are in orange and the bases are shown as sticks. (A) Partner attraction effect—parts of two loops of Restriction Endonuclease (31) shift position on binding DNA. Interface residues, in stick representation, Ile51 and Arg30 are shown (in red) for the bound state (2odi), and (in blue) for unbound state of the enzyme (2odh). ASA increases from 58.2 to 260 Å^2 for these two residues. (B) Partner accommodation effect—glucocorticoid receptor rearranges on binding DNA. Interface residues Arg489 and Ser459 are shown for the bound state (1r4o) (32), and for unbound state, apo enzyme (1gdc) (33). ASA decreases from 155 to 76 Å^2 for these two residues.

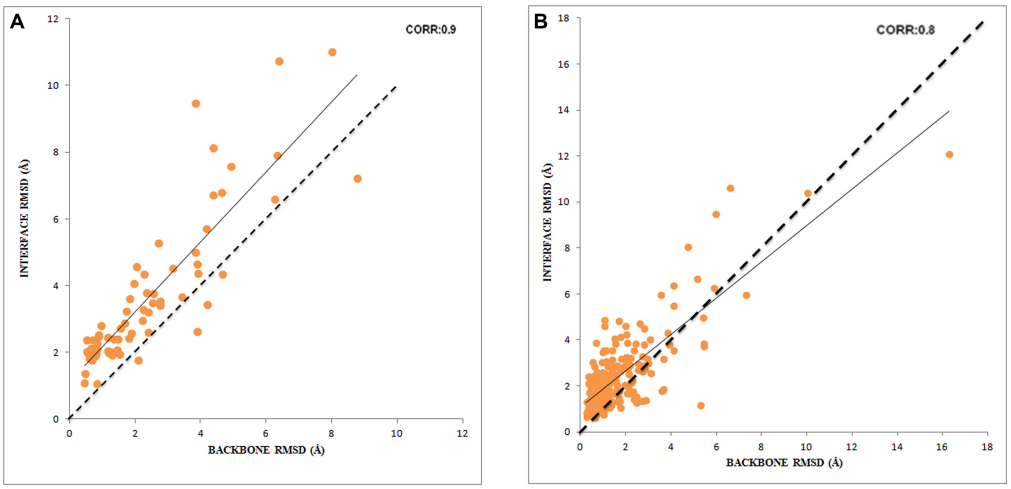

Figure 3. The scatter plot of RMSD between unbound and bound states in (**A**) protein–DNA and (**B**) protein–protein complexes. The line of regression is drawn along with the Spearman's correlation coefficient; the diagonal line is shown as dashed.

The *Staphylococcus aureus* multidrug binding protein QacR represses transcription of the *qacA* multidrug transporter gene and is induced by structurally diverse cationic lipophilic drugs. When bound to DNA (41,42), Tyr40 and Tyr41 of QacR form hydrogen bond with phosphates and van der Waals contacts with sugar moieties and the result is a decrease in ΔASA values (Supplementary Figure S3E).

It is found that most of the residues with high RMSD and ΔASA values (positive or negative) belong to the enzyme category, such as restriction endonucleases. Overall, in the 66 complexes, two major classes are enzymes (27) and the helix-turn-helix proteins (20), the rest being divided among zinc-coordinating (3), other α-helix (6), β-sheet (4) and β hairpin/ribbon (6) structures (1). The max-

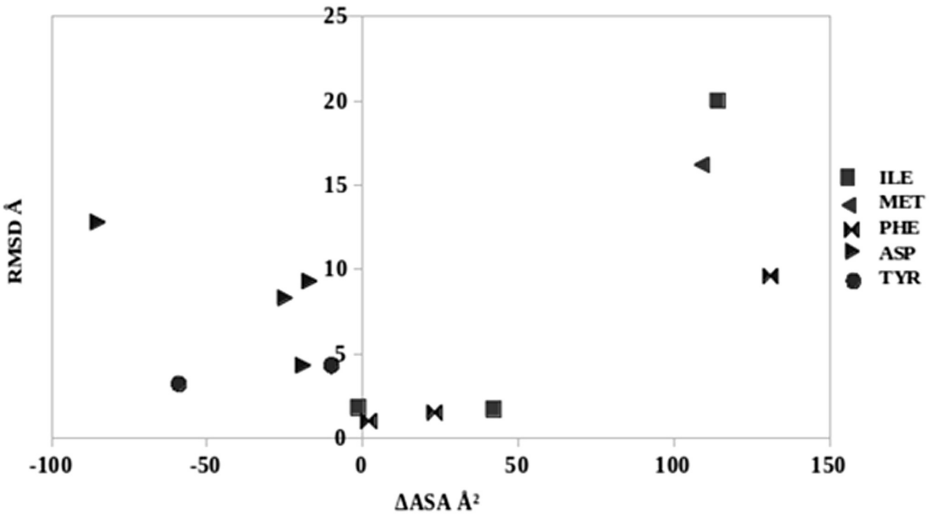

Figure 4. Scatter plot of ΔASA versus RMSD for a few selected residues (having large ΔASA); symbols for five residue types are indicated. The PDB codes (U/B) and their residues are: 2hmy (34)/7mht (35) (ILE 86, ILE 249, ILE 258), 1bam (36) /3bam (37) (MET 198), 1ynm (38) /2fl3 (39) (PHE 15, PHE 91, PHE 137), 3v6p/3v6t (40) (ASP 335, ASP 369, ASP 573, ASP 641), 1jus (41)/1jt0 (42) (TYR 40, TYR 41). Molecular diagrams indicating changes in these structures are shown in Supplementary Figure S3.

imum RMSD values for interface residues are observed for the enzyme category (4.3 ± 2.9 Å), followed by zinc-coordinating proteins (4.2 ± 2.6 Å); the helix-turn-helix proteins exhibit a value of 3.1 ± 1.6 Å. In a related study conformational changes associated with DNA binding were categorized into six classes, and the members exhibiting the largest changes were found to be either endonuleases or polymerases (43).

Analysis of the missing residues in the unbound form

The segments missing in the unbound proteins, which upon binding DNA are structured were also analysed. A missing segment is defined as the one with three (or more) missing residues (lying in the interface or elsewhere). An example is presented in Figure 5, where a helical portion (58–63) of the C-terminal region of the structure and a loop in the N-terminal part (residues 0–6, Supplementary Table S3) are ordered in the complex (44,45). Mostly, the interface and non-interface residues occur interspersed in the same stretch (Figure 6A) (46,47), but in Glucocorticoid receptor, these occur in two separate stretches (Figure 6B) (32,33). 23 structures have missing segments (Supplementary Table S3) in the unbound form of the protein. In 17 of these structures the segments contain interface residues. On an average the missing stretches constitutes 6% residues of these 23 structures. Missing atoms constitute 7% of the total interface atoms in the whole dataset and 18% in 19 structures (with one or more residues, missing entirely, Supplementary Table S1C). On average the contribution to BSA from missing residues (9.9 ± 3.4 $Å^2$ per atom) is greater than that from non-missing residues (8.8 ± 0.9 $Å^2$ per atom, P value = 0.07). 54% of the polypeptide chains with missing stretches have segments from the chain termini (60% of which constitute the interface also), similar to what was observed in protein–protein interactions (22).

We had observed in our previous analysis of protein–protein interactions (22), the number (197) of missing residues in the interface are more compared to those (131) in the non-interface region, whereas for protein–DNA structures, the opposite trend is seen (126 in interface vs. 258 in non-interface regions, considering those structures which have at least one missing residue in the interface) (Table 1). The interface residues which are found in greater number in these disordered stretches are Lys, Gly and Arg. Interestingly, Ala scores high if non-interface regions are only considered. As in protein–protein interactions (22), the secondary structures attained by the missing residues in the complex are mostly irregular, followed by helix and turn, strand being the least observed.

Changes in secondary structure

The change in percentage composition of the secondary structural elements during U to B transition was calculated. 58 (88%) structures showed some changes. The Euclidean distance (D) between the compositions of the four structural elements in the two states was also calculated and the average was found to be 6.2 (±5.5) for all, but the average increased to 8.0 (±5.1) for the structures (26) where regular secondary structural were formed at the expense of turns or coils. For understanding structural changes during complex formation we have used structural pairs with D > 6 (Supplementary Figure S4) where we can see that there is an increase in coil and helical content at the cost of turns and strands. In these 26 structures, we mostly find that extension of either existing helix or strand to be more frequent than formation of new helix or strand, and this was more for helices compared to strands (Figure 7A). Also it was observed that in case of helix mostly the extension takes place in the N-terminal, whereas for strand C-terminal extension is preferred (Figure 7B). Two examples of helix extension and helix formation are shown in Figure 8 (48–50).

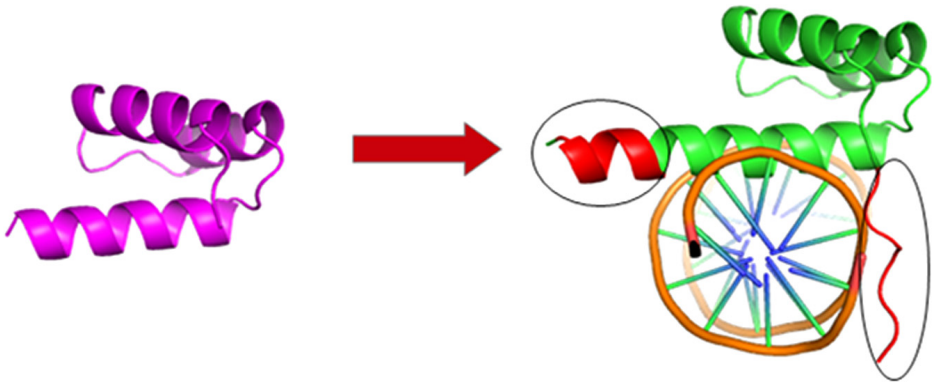

Figure 5. Two missing regions (encircled) in Aristaless Homeodomain (3a02) (left) (44) are ordered (in red) in the DNA bound complex (1fjl) (right) (45).

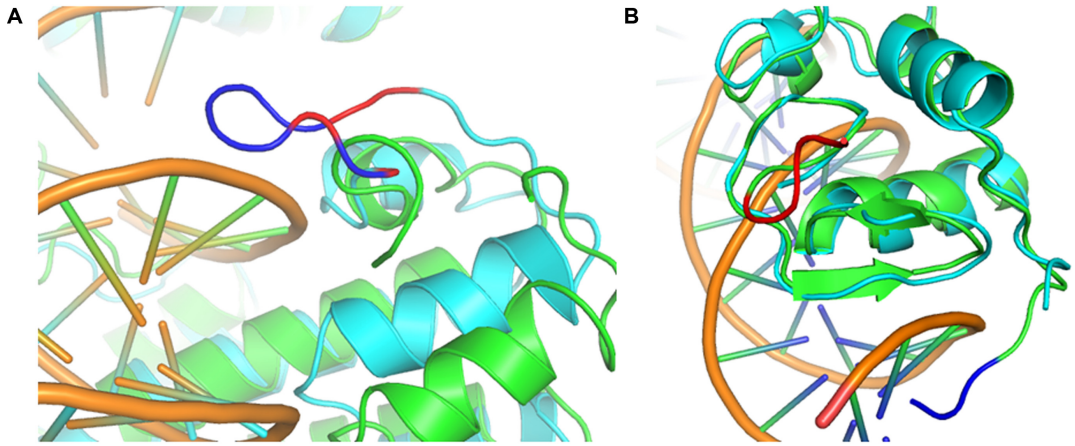

Figure 6. Two examples with missing regions in the unbound structure (green cartoon) that are seen in the B form (cyan); the protein region that gets ordered in the interface is indicated in blue, and the region that is not part of the interface is in red. (**A**) The interface residues are interspersed with the non-interface in the missing segment of DNA polymerase I (complexed with DNA, 4ktq (46) and unbound protein, 1ktq) (47). (**B**) The missing interface and non-interface residues form separate segments in glucocorticoid receptor (bound to DNA, 1r4o and the unbound protein, 1gdc) (32,33).

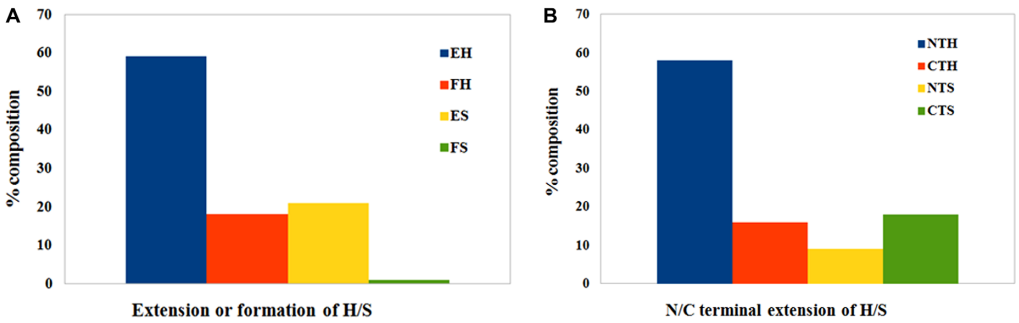

Figure 7. Percentage composition of (**A**) helix/strand extension and formation, (**B**) extension of helix/strand in N or C terminal.

Comparison of B factors

The relative vibrational motion in different parts of the protein structure is determined by *B* factor, also known as the temperature factor. The parts of the molecules which are highly flexible have high *B* factors.

B factors are generally used to assess the difference between the interface and rest of the protein surface. Usually interface residues, involved in protein–protein interactions, are less flexible and have lower *B* factors compared to those in the surface regions (51,22). Similar result was also obtained in protein–DNA interaction (52). While supporting this observation, another study that also compared the binding and the non-binding regions in the apo form did not find any clear trend in the *B* factors in the two regions (53). This necessitated a systematic analysis of *B* factors between the two forms, as well as between different regions in the structure in them. We have taken the scaled mean *B* factor of the C, C^{α}, O, N and C^{β} atoms as the representative for the whole residue, and the average values were calculated for

Table 1. Statistics on residues missing in the U form and their secondary structure in the B form

Residue	Number missing[a]	% relative to total number of		Secondary structure in the B form of residues missing in U (%)[b]			
		Interface residues of the same type	Missing residues[b]	H	S	T	C
Ala	6 (25)	5.6	4.8 (8.1)	33.3 (67.7)	0	16.7 (9.7)	50 (22.6)
Arg	14 (13)	5.1	11.1 (7.0)	35.7 (40.7)	0 (3.7)	28.6 (29.6)	35.7 (25.9)
Asn	4 (18)	2.4	3.2 (5.7)	0 (13.6)	50(13.6)	25(22.7)	20 (50)
Asp	3 (10)	2.9	2.4 (3.4)	33.3 (38.5)	0 (7.7)	33.3(23.1)	33.3 (30.8)
Cys	0 (2)	0.0	0 (0.5)	0 (50)	0	0	0 (50)
Gln	3 (17)	2.0	2.4 (5.2)	33.3 (30)	33.3(15)	0(10)	33.3 (45)
Glu	4 (16)	4.3	3.2 (5.2)	0 (40)	0(10)	75(45)	25 (5)
Gly	20 (21)	10.3	15.9 (10.7)	20(14.6)	0	40(51.2)	40 (34.1)
His	3 (5)	4.2	2.4 (2.1)	0 (37.5)	0	33.3(25)	66.7 (37.5)
Ile	2 (12)	2.3	1.6 (3.6)	0 (42.9)	0	50(28.6)	50 (28.6)
Leu	3 (35)	2.7	2.4 (9.9)	100 (52.6)	0(5.3)	0(13.2)	0 (28.9)
Lys	29 (20)	8.5	23 (12.8)	13.8 (24.5)	6.9(6.1)	31(34.7)	48.5 (34.7)
Met	0 (6)	0	0 (1.6)	0	0	0	0 (100)
Phe	2 (2)	2.9	1.6 (1.0)	0	0(25)	0	100 (75)
Pro	6 (11)	7.7	4.8 (4.4)	33.3 (29.4)	0	16.7(29.4)	50 (41.2)
Ser	13 (17)	6.3	10.31 (7.8)	23.1 (30)	0(3.3)	23.1(23.3)	53.9 (43.3)
Thr	8 (12)	3.6	6.35 (5.2)	25(30)	0 (10)	37.5 (20)	37.5 (40)
Trp	2 (2)	4.7	1.6 (1.0)	0	0	0 (25)	100 (75)
Tyr	1 (2)	0.9	0.8 (0.8)	0	0 (33.3)	100(66.7)	0
Val	3 (12)	3.1	2.4 (3.9)	66.7(73.3)	33.3(6.7)	0(13.3)	0 (6.7)
Total	**126 (258)**						

If the two types (interface and non-interface) of missing stretches are represented by o-o-o-x-o-o and x-x-x-x-x (where o indicates a residue in the interface, and x a non-interface residue), the table gives statistics using all the residues of type 'o'. Additionally, within the parentheses are values using all the residues (o + x) (footnote b below) or only the x residues (footnote a below). Six structures have missing residues (46 in number) only in the non-interface region.
[a]The numbers in parentheses correspond to the non-interface residues of the missing stretches in Supplementary Table S3.
[b]The numbers in parentheses are calculated considering all the residues of missing stretches in Supplementary Table S3.

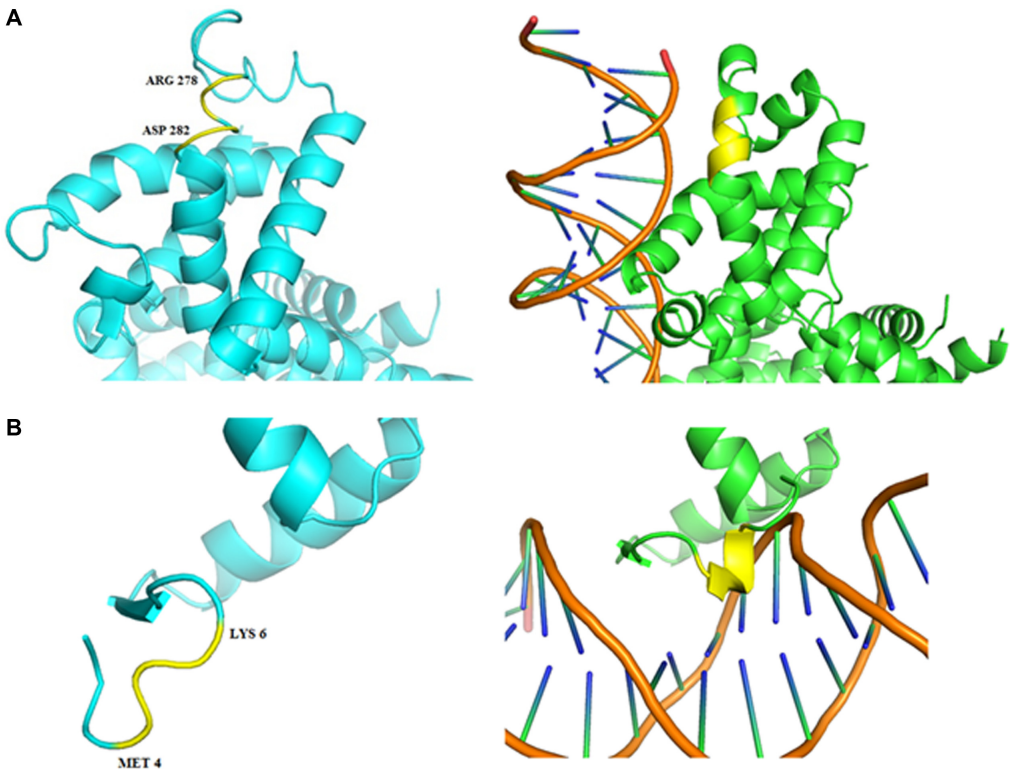

Figure 8. Examples showing change in secondary structural elements (left side, U; right side, B). Interface stretches have been marked in yellow. (**A**) Bacteriophage lambda cII protein (1zpq/1zs4) (48) in complex exhibits extension of its helix from Asp282 to Arg278 (which consisted of bends and coils in the U state) in N-terminal. (**B**) Wild type gene-regulating protein ARC (1arq) (49) in complex with the DNA (1bdt) (50) shows formation of helix from Met4 to Lys6.

each residue type for the interface and the surface regions in both U and B states.

On an average *B* factor in the bound form was found to be higher for the surface residues as opposed to interface residues (P value $< 2.2 \times 10^{-16}$, Supplementary Table S4). The normalized *B* factors for all the interface residues are found to be negative in the bound form against mostly positive values in the unbound form (P value $< 2.2 \times 10^{-16}$). So we can conclude that during unbound to bound transition the interface residues experience a decrease in their *B* factor. However for surface residues, an overall opposite trend was observed, i.e. while going from unbound to bound state these residues experience an increase in their *B* factor (P value = 0.0057). The same trend was reported in protein–protein interactions (22). In the unbound form *B* factor between the surface and interface regions are almost similar (P value = 0.2). Euclidean distances between the interface and surface residues (Figure 9A) were calculated for both bound and unbound structure, and it is found that the maximum change takes place between the surface and interface regions in the complex.

We have further divided the interface residues into core and rim regions (29), and the *B* factors for them were also compared between these two regions in the B and U form (Figure 9B, Supplementary Table S5). From the Euclidean distance we can see that the decrease in flexibility is more pronounced in the core region between the two forms while the rim residues show a smaller difference. This is also reflected in the P values (2.2×10^{-16} for the core and 6.117×10^{-7} for the rim). However, no significant difference was observed in the U form between the *B* factors of core and rim residues (P value = 0.4521).

Correlation between structural changes of residues on DNA binding and free energy of binding

To get an insight into binding affinity one has to understand thermodynamics data in terms of structural changes that occur on binding. Alanine-scanning data are unavailable for protein–DNA interfaces. To circumvent the issue, a data set of free energy changes upon point mutations in a general context was resorted to. Such a data set was compiled by Kumar *et al.* and presents a list of mutations in protein–DNA complexes for which experimental free-energy changes are available in ProNIT (54). From these data, single point mutations, where ΔG values for complex formation was available in the wild type as well as protein mutant, were extracted to relate them to structural differences between unbound and bound forms in a previous study (55). The authors computed the free-energy change ($\Delta\Delta G$) upon mutation as $\Delta\Delta G = \Delta G\_\text{mutant} - \Delta G\_\text{wild}$. A higher value of $\Delta\Delta G$ for a given mutation would indicate larger destabilization caused by the mutation. In that selected list, 11 of our 66 structures were found, but only 7 had ΔG_mutant and ΔG_wild values enabling us to calculate $\Delta\Delta G$ for the residues. Considering only the point mutations to Ala, and excluding two with multiple observations that matched rather poorly, we were left with eight entries (Supplementary Table S6). Supplementary Figure S5 shows a plot of absolute values of ΔASA (considering all atoms of the residue) versus $\Delta\Delta G$, the correlation coefficient being 0.79. Although the correlation is heavily biased by a single outlier, there seems to be an indication that a mutation in the interface residue undergoing change in ASA on complex formation is destabilizing as reflected by a positive $\Delta\Delta G$. It may be pertinent to add that for the data points $\Delta\Delta G$ is negatively correlated with BSA (–0.75) (Supplementary Table S6).

DISCUSSION

δA indicates the relative percentage change of the accessible surface area of the interface atoms in going from the unbound to bound state of the protein. The large values (Figure 1A) have been explained in terms of partner attraction (when δA is positive) and partner accommodation (δA is negative) effects (21,22); when the interface residues of the protein are drawn towards DNA, the phenomenon is referred to as partner attraction, whereas partner accommodation is when protein residues move away from the DNA to accommodate it. Although the distribution of δA is rather bimodoal, considering all the atoms of the interface residues in the calculation of δA (–4.4 ± 10%, Figure 1B) it can be seen that the partner accommodation effect prevails. Protein-DNA interactions are mediated by both charged and hydrophobic residues, and Asp, belonging to the former class, indicates a reduction in ASA (Supplementary Table S2), while the latter residues, in general, tend to display an increase in ASA. It can be seen in Figure 4 that ΔASA is negative for Asp, whereas the values are positive for Ile, Met and Phe. On the other hand protein–protein interactions (in particular the interfaces in homodimeric associations) are enriched in nonpolar contacts/residues (56) and U-to-B transition is accompanied by an increase in ASA (22). The –ve δA value of Asp could be due to the its moving away from the negatively charged DNA chain—the longer Glu side chain can accomplish this without much change in the solvent accessibility of its interface atoms (Supplementary Table S2).

It is also of interest to understand the effect of the change of accessible surface area of interface residues accompanying U to B transition on free energy of binding. Based on a very limited amount of data of ΔΔG of binding on mutation of residues to Ala (Supplementary Figure S5 and Table S6), there appears to be a trend of ΔΔG increasing with increase in the absolute value of ΔASA.

The results obtained from analysis of protein–DNA complexes can be compared with the analysis involving protein–protein interaction affinity dataset (21,22). The distribution of δA values for protein–protein complexes showed that the partner attraction effect (δA = 3.3 ± 4.9%) prevails and number of structures undergoing large RMSD changes (≥4 Å) both in case of backbone and interface are less than what is observed for protein–DNA interactions. For the protein–protein dataset, there are 23 and 15 structures with interface and backbone RMSD values higher than 4 Å, this constitutes 8.2% and 5.34% respectively as opposed to 30% and 18.2% for interface and backbone RMSD respectively in case of protein–DNA interaction which is due to greater number of protein components undergoing greater structural changes on binding DNA.

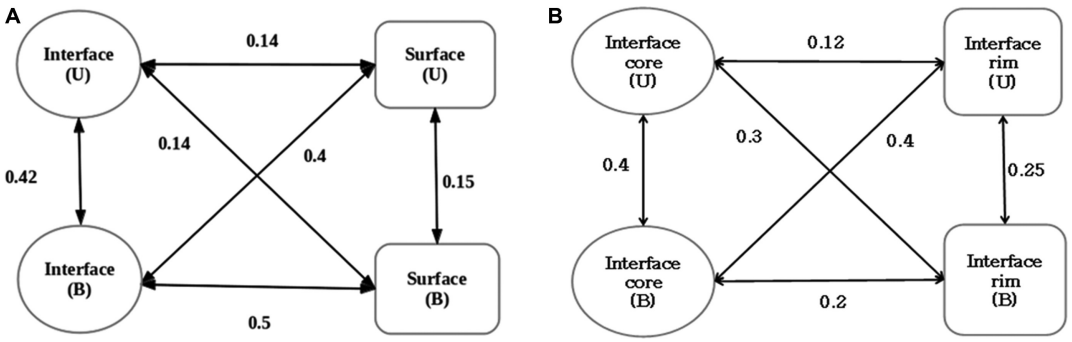

Figure 9. Euclidean distances involving B-factors between (**A**) interface and surface regions, and (**B**) core and rim region of the interface, in the U and B forms.

Figure 4 shows that some structures have residues with large |ΔASA| and RMSD values. These are not due to any crystallization artefacts, but have biological significance, many of these belonging to the enzyme category (restriction endonuclease and methyltransferase). It has been reported that highly specific and multi-specific DNA binding domains exhibit large conformational changes upon DNA-binding, and Asp is enriched in specific DNA-binding proteins (57). It is interesting to observe that residues having large changes in RMSD and ΔASA, aspartates in particular (Figure 4), all belong to the specific category.

A comparison of U and B forms of proteins allowed us to identify the residues that undergo disorder to order transition on binding DNA. For structures having missing residues in the U form the missing atoms constitute 18% of the interface atoms, somewhat greater than 12% observed in protein–protein interactions (22). The missing atoms contribute more (9.9 ± 3.4 Å^2) to the BSA in the bound state as compared to the non-missing atoms (8.8 ± 0.9 Å^2), similar to what was observed in protein–protein interactions (the corresponding values being (11.5 ± 6.8 and 9.4 ± 1.5 Å^2, respectively), indicating a greater degree of surface burial by missing atoms that can offset the entropic penalty associated with U-to-B transition (22). Intrinsically disordered regions (IDRs) and proteins have functional repertoire that complements ordered proteins, and there are attempts to predict and characterize such short regions (58,59). Our analysis have identified the disordered segments that are involved in binding DNA (Supplementary Table S3), and though the examples are rather limited in number, some residues such as Lys, Gly and Arg have been shown to interact with DNA. Such information could be incorporated into high-throughput methods for predicting DNA binding residues located in IDRs from protein sequence.

CONCLUSION

In this paper, we have compared the unbound and bound forms of DNA-binding proteins. While the interface atoms undergo an increase in ASA in going from U to B states in the majority of cases in protein–protein interactions, here the increase and decrease are found to equal extent. In general residues exhibit greater RMSDs and change in ASA values in protein–DNA interactions (PDIs) than protein–protein interactions (PPIs). However, during U-to-B transition both the types of interactions bring about similar changes in secondary structures and flexibility of residues located in the interface and the surface. 16% of the proteins in PPIs have missing (disordered) residues in the U form which get ordered in the B form (22); the number is higher 25% in PDIs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Prof. Shandar Ahmad for discussion on ProNIT database. This work is dedicated to the Centenary of Bose Institute.

FUNDING

Department of Science and Technology, India (Research grant to P.C.) [SR/S2/JCB-12/2006]; Department of Biotechnology (for funding the Centre). Funding for open access charge: Department of Science and Technology, India.
Conflict of interest statement. None declared.

REFERENCES

1. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein–DNA complexes. *Genome Biol.*, **1**, 1–37.
2. Berg,O.G. and von Hippel,P.H. (1988) Selection of DNA binding sites by regulatory proteins. *Trends Biochem. Sci.*, **13**, 207–211.
3. Huffman,J.L. and Brennan,R. G. (2002) Prokaryotic transcription regulators: more than just the helix-turn-helix motif. *Curr. Opin. Struct. Biol.*, **12**, 98–106.
4. Bewley,C.A., Gronenborn,A.M. and Clore,G.M. (1998) Minor groove-binding architectural proteins: structure, function, and DNA recognition 1. *Annu. Rev. Biophys. Biomol. Struct.*, **27**, 105–131.
5. Sarai,A. and Kono,H. (2005) Protein-DNA recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.*, **34**, 379–398.
6. Paillard,G. and Lavery,R. (2004) Analyzing protein–DNA recognition mechanisms. *Structure*, **12**, 113–122.
7. Iwahara,J., Schwieters,C.D. and Clore,G.M. (2004) Characterization of nonspecific protein–DNA interactions by 1H paramagnetic relaxation enhancement. *J. Am. Chem. Soc.*, **126**, 12800–12808.
8. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein–DNA recognition. *Annu. Rev. Biochem.*, **79**, 233.

9. Nadassy,K., Wodak,S.J. and Janin,J. (1999) Structural features of protein-nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
10. Biswas,S., Guharoy,M. and Chakrabarti,P. (2009) Dissection, residue conservation, and structural classification of protein-DNA interfaces. *Proteins*, **74**, 643–654.
11. Sagendorf,J.M., Berman,H.M. and Rohs,R. (2017) DNAproDB: an interactive tool for structural analysis of DNA–protein complexes. *Nucleic Acids Res.*, **45**, W89–W97.
12. Dey,S., Pal,A., Guharoy,M., Sonavane,S. and Chakrabarti,P. (2012) 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 Res.*, **40**, 7150–7161.
13. 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.
14. Liu,R. and Hu,J. (2013) DNABind: a hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning- and template-based approaches. *Proteins*, **81**, 1885–1899.
15. Nagarajan,R., Ahmad,S. and Michael Gromiha,M. (2013) Novel approach for selecting the best predictor for identifying the binding sites in DNA binding proteins. *Nucleic Acids Res.*, **41**, 7606–7614.
16. Wang,W., Liu,J. and Sun,L. (2016) Surface shapes and surrounding environment analysis of single-and double-stranded DNA-binding proteins in protein-DNA interface. *Proteins*, **84**, 979–989.
17. Ding,X.M., Pan,X.Y., Xu,C. and Shen,H.B. (2010) Computational prediction of DNA–protein interactions: a review. *Curr. Computer-aided Drug Des.*, **6**, 197–206.
18. Liu,J., Perumal,N.B., Oldfield,C.J., Su,E.W., Uversky,V.N. and Dunker,A.K. (2006) Intrinsic disorder in transcription factors. *Biochemistry*, **45**, 6873–6888.
19. Theillet,F.X., Binolfi,A., Frembgen-Kesner,T., Hingorani,K., Sarkar,M., Kyne,C., Li,C., Crowley,P.B., Gierasch,L., Pielak,G.J. *et al.* (2014) Physicochemical properties of cells and their effects on intrinsically disordered proteins (IDPs). *Chem. Rev.*, **114**, 6661–6714.
20. Tompa,P. (2012) Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci.* **37**, 509–517
21. Chakravarty,D., Guharoy,M., Robert,C.H., Chakrabarti,P. and Janin,J. (2013) Reassessing buried surface areas in protein–protein complexes. *Protein Sci.*, **22**, 1453–1457.
22. Chakravarty,D., Janin,J., Robert,C.H. and Chakrabarti,P. (2015). Changes in protein structure at the interface accompanying complex formation. *IUCrJ*, **2**, 643–652.
23. van Dijk,M. and Bonvin,A.M. (2008) A protein–DNA docking benchmark. *Nucleic Acids Res.*, **36**, e88–e88.
24. Hubbard,S.J. (1992) *NACCESS: program for calculating accessibilities*. Department of Biochemistry and Molecular Biology University College of London.
25. Lee,B. and Richards,F.M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.
26. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.
27. McLachlan,A.D. (1982). Rapid comparison of protein structures. *Acta Crystallogr. A*, **38**, 871–873.
28. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F. and Wilczynski,B. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
29. Chakrabarti,P. and Janin,J. (2002) Dissecting protein–protein recognition sites. *Proteins*, **47**, 334–343.
30. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
31. Sokolowska,M., Kaus-Drobek,M., Czapinska,H., Tamulaitis,G., Szczepanowski,R.H., Urbanke,C., Siksnys,V. and Bochtler,M. (2007) Monomeric restriction endonuclease BcnI in the apo form and in an asymmetric complex with target DNA. *J. Mol. Biol.*, **369**, 722–734.
32. Luisi,B.F., Xu,W., Otwinowski,Z., Freedman,L.P., Yamamoto,K.R. and Sigler,P.B. (1991) Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature*, **352**, 497.
33. Baumann,H., Paulsen,K., Kovacs,H., Berglund,H., Wright,A.P., Gustafsson,J A. and Haerd,T. (1993) Refined solution structure of the glucocorticoid receptor DNA-binding domain. *Biochemistry*, **32**, 13463–13471.
34. O'Gara,M., Zhang,X., Roberts,R.J. and Cheng,X. (1999) Structure of a binary complex of HhaI methyltransferase with S-adnosyl-l-methionine formed in the presence of a short non-specific DNA oligonucleotide. *J. Mol. Biol.*, **287**, 201–209.
35. O'Gara,M., Horton,J.R., Roberts,R.J. and Cheng,X. (1998) Structures of HhaI methyltransferase complexed with substrates containing mismatches at the target base. *Nat. Struct. Mol. Biol.*, **5**, 872–877.
36. Newman,M., Strzelecka,T., Dorner,L.F., Schildkraut,I. and Aggarwal,A.K. (1994) Structure of restriction endonuclease *Bam*HI phased at 1.95 Å resolution by MAD analysis. *Structure*, **2**, 439–452.
37. Viadiu,H. and Aggarwal,A.K. (1998) The role of metals in catalysis by the restriction endonuclease Bam HI. *Nat. Struct. Mol. Biol.*, **5**, 910–916.
38. Yang,Z., Horton,J R., Maunus,R., Wilson,G.G., Roberts,R.J. and Cheng,X. (2005) Structure of HinP1I endonuclease reveals a striking similarity to the monomeric restriction enzyme MspI. *Nucleic Acids Res.*, **33**, 1892–1901.
39. Horton,J.R., Zhang,X., Maunus,R., Yang,Z., Wilson,G.G., Roberts,R.J. and Cheng,X. (2006) DNA nicking by HinP1I endonuclease: bending, base flipping and minor groove expansion. *Nucleic Acids Res.*, **34**, 939–948.
40. Deng,D., Yan,C., Pan,X., Mahfouz,M., Wang,J., Zhu,J.K., Shi,Y. and Yan,N. (2012) Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science*, **335**, 720–723.
41. Schumacher,M.A., Miller,M.C., Grkovic,S., Brown,M.H., Skurray,R.A. and Brennan,R.G. (2001). Structural mechanisms of QacR induction and multidrug recognition. *Science*, **294**, 2158–2163.
42. Schumacher,M.A., Miller,M.C., Grkovic,S., Brown,M.H., Skurray,R.A. and Brennan,R.G. (2002) Structural basis for cooperative DNA binding by two dimers of the multidrug-binding protein QacR. *EMBO J.*, **21**, 1210–1218.
43. Andrabi,M., Mizuguchi,K. and Ahmad,S. (2014) Conformational changes in DNA-binding proteins: Relationships with precomplex features and contributions to specificity and stability. *Proteins*, **82**, 841–857.
44. Miyazono,K.I., Zhi,Y., Takamura,Y., Nagata,K., Saigo,K., Kojima,T. and Tanokura,M. (2010) Cooperative DNA-binding and sequence-recognition mechanism of aristaless and clawless. *EMBO J.*, **29**, 1613–1623.
45. Wilson,D.S., Guenther,B., Desplan,C. and Kuriyan,J. (1995) High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. *Cell*, **82**, 709–719.
46. Li,Y., Korolev,S. and Waksman,G. (1998) Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of Thermus aquaticus DNA polymerase I: structural basis for nucleotide incorporation. *EMBO J.*, **17**, 7514–7525.
47. Korolev,S., Nayal,M., Barnes,W.M., Di Cera,E. and Waksman,G. (1995) Crystal structure of the large fragment of Thermus aquaticus DNA polymerase I at 2.5-A resolution: structural basis for thermostability. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 9264–9268.
48. Jain,D., Kim,Y., Maxwell,K.L., Beasley,S., Zhang,R., Gussin,G.N., Edwards,A.M. and Darst,S.A. (2005) Crystal structure of bacteriophage λcII and its DNA complex. *Mol. Cell*, **19**, 259–269.
49. Bonvin,A.M., Vis,H., Breg,J.N., Burgering,M.J., Boelens,R. and Kaptein,R. (1994) Nuclear magnetic resonance solution structure of the Arc repressor using relaxation matrix calculations. *J. Mol. Biol.*, **236**, 328–341.
50. Schildbach,J.F., Karzai,A.W., Raumann,B.E. and Sauer,R.T. (1999) Origins of DNA-binding specificity: role of protein contacts with the DNA backbone. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 811–817.
51. Jones,S. and Thornton,J.M. (1995) protein–protein interactions: a review of protein dimer structures. *Progress Biophys. Mol. Biol.*, **63**, 3161–5965.
52. Schneider,B., Gelly,J.C., de Brevern,A.G. and Černý,J. (2014) Local dynamics of proteins and DNA evaluated from crystallographic B factors. *Acta Crystallogr. D*, **70**, 2413–2419.
53. Xiong,Y., Liu,J. and Wei,D.Q. (2011) An accurate feature-based method for identifying DNA-binding residues on protein surfaces. *Proteins*, **79**, 509–517.
54. 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–D206.
55. Ahmad,S., Keskin,O., Sarai,A. and Nussinov,R. (2008) Protein–DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Res.*, **36**, 5922–5932.
56. Janin,J., Bahadur,R.P. and Chakrabarti,P. (2008) protein–protein interaction and quaternary structure. *Q. Rev. Biophys.*, **41**, 133–180.
57. Corona,R.I. and Guo,J.T. (2016) Statistical analysis of structural determinants for protein–DNA-binding specificity. *Proteins*, **84**, 1147–1161.
58. Peng,Z. and Kurgan,L. (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nuleic Acids Res.*, **43**, e121.
59. Khan,W., Duffy,F., Pollastri,G., Shields,D.C. and Mooney,C. (2013) Predicting binding within disordered protein regions to structurally characterized peptide-binding domains. *PLoS ONE*, **8**, e72838.