Bioinformatic analysis of the protein/DNA interface.
Bohdan Schneider, Jirí Cerny, Daniel Svozil, Petr Cech, Jean-Christophe Gelly, Alexandre de Brevern

To cite this version:

HAL Id: inserm-00926088
https://inserm.hal.science/inserm-00926088
Submitted on 9 Jan 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Bioinformatic analysis of the protein/DNA interface

Bohdan Schneider, Jiří Černý, Daniel Svozil, Petr Čech, Jean-Christophe Gelly and Alexandre G. de Brevern

INTRODUCTION

Interactions between proteins and DNA are essential for molecular processes of replication, transcription, gene regulation or chromosome packaging. Despite an extensive effort to understand the principles governing protein/DNA recognition, no simple and general rules have been found. The paradigm of molecular biology, DNA self-recognition via Watson–Crick base pairing, has probably no analogy in protein/DNA recognition. According to Matthews, there is no simple ‘code of recognition’ between amino acids and nucleotides (1), and the reason might be that the interaction between these two structurally complicated molecules has too many degrees of freedom (2).

Proteins recognize specific DNA sequences by two strategies commonly referred to as ‘direct’ and ‘indirect’ readout (3). However useful, this classification is artificial, and all protein/DNA high-affinity interactions depend on the conformational flexibility of the binding partners. Intrinsic conformational flexibility is more frequent in protein regions binding to DNA than in regions that do not bind to DNA (4). DNA is also known to conformationally adapt to its binding partner, e.g. by varying double helical groove widths, the helical twist, other base-pair parameters and the backbone conformations (3). The knowledge accumulated about modulations of DNA structure and electrostatics has complicated the idea of straightforward sequence-dependent readout by hydrogen-bonding patterns (5) and ultimately led to understanding that proteins recognize sequence-dependent flexibility or deformability rather than the sequence by direct readout (6). Such a complex nature of protein/DNA interactions requires elaborate functional and structural analysis of complexes (7) that has led to identification of specific rules of recognition for various families of protein/DNA complexes. An algorithm revealing likely sequences of potential transcription factors has been published soon after their first structures had been solved (8). Later, with many more experimental structures available, protein structural, physicochemical characteristics and thermodynamic properties have been examined to determine the rules of residue conservation in DNA-binding.

Received April 30, 2013; Revised and Accepted November 14, 2013

ABSTRACT

To investigate the principles driving recognition between proteins and DNA, we analyzed more than thousand crystal structures of protein/DNA complexes. We classified protein and DNA conformations by structural alphabets, protein blocks [de Brevern, Etchebest and Hazout (2000) (Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Prots. Struct. Funct. Genet., 41:271–287)] and dinucleotide conformers [Svozil, Kalina, Omelka and Schneider (2008) (DNA conformations and their sequence preferences. Nucleic Acids Res., 36:3690–3706)], respectively. Assembling the mutually interacting protein blocks and dinucleotide conformers into ‘interaction matrices’ revealed their correlations and conformer preferences at the interface relative to their occurrence outside the interface. The analyzed data demonstrated important differences between complexes of various types of proteins such as transcription factors and nucleases, distinct interaction patterns for the DNA minor groove relative to the major groove and phosphate and importance of water-mediated contacts. Water molecules mediate proportionally the largest number of contacts in the minor groove and form the largest proportion of contacts in complexes of transcription factors. The generally known induction of A-DNA forms by complexation was more accurately attributed to A-like and intermediate A/B conformers rare in naked DNA molecules.
proteins (9,10); other studies analyzed the structural principles governing protein/DNA recognition (11) and classified protein motifs that bind to DNA (12). Rules determining recognition of DNA by some protein motifs, e.g. zinc fingers (13–15), or helix-turn-helix (16,17), have been discovered. These studies provide evidence that diverse structural descriptors have to be considered to describe origins of the binding specificity for different protein families.

Analysis of structural and physicochemical properties of the protein/DNA interface and of atom–atom interactions has demonstrated that amino acid and base compositions are correlated (18–20). The interface is formed mostly by positive and polar amino acids forming hydrogen bonds with bases and phosphates; the interface is more polar than basically lipophilic protein/protein interfaces (18,21); and contacts are often water-mediated. The importance of interactions between charged phosphate groups and charged or polar amino acid for the stability of complexes points to a key role of electrostatics in protein/DNA recognition, and modeling of electrostatic potentials has been used to predict DNA-binding sites (22–24). Another specific type of interaction, hydrogen bonding, has also attracted a considerable attention: networks of hydrogen bonds have been correlated to recognition of DNA by transcription factors (25) and direct amino acid–base contacts have been statistically analyzed (26). More specific types of interactions such as CH...O interactions (27) or pi/H-bond stacking motifs (28) have also been studied. Both proteins and DNA are heavily hydrated molecules, and an importance of water and of other solvent species for the binding has been recognized from early days of DNA structural research (29) and later recapitulated in several reviews (30–32).

The growing availability of structures of protein/DNA complexes has facilitated purely bioinformatics approaches to protein/DNA recognition. Many of these studies emphasize the active role of proteins in the recognition process, e.g. in graph representation of the interactions (33,34), or in structural classification of the interfaces from over a hundred protein/DNA structures (35). Structural alignment of interfacial protein and DNA residues has revealed surprising similarities between proteins of different folds (36). Similarly, surprising results have been obtained by using 11 structural descriptors that classify protein/DNA interfaces of 62 crystal complexes (37), concluding that DNA-binding proteins with the same binding motif (such as zinc-finger) may belong to different structural and functional classes. A recent work (4) has investigated local conformational changes at the interfaces of DNA-binding proteins classifying protein conformations by a protein structural alphabet but not distinguishing between different subfamilies of protein binding motifs and using subjective and coarse classification of DNA conformations.

In this work, we present a novel bioinformatics analysis of protein/DNA interactions. Both protein and DNA structures were classified using a well-established concept of structural alphabet (38–43). To characterize local conformations of proteins, we used the Protein Blocks (PBs) (44,45) that consist of 16 folding patterns of five consecutive amino acid residues; DNA local conformations were described at the dinucleotide (ntC) level (46). We then determined counts of mutually interacting PBs and ntCs, which form the protein/DNA interface, and compared their populations with numbers of non-interacting PBs and ntCs. The scope of over a thousand analyzed protein/DNA complexes and simultaneous objective classification of protein and DNA conformations offer a detailed insight into the protein/DNA interactions.

**MATERIALS AND METHODS**

**Selection of protein/DNA structures**

Protein/DNA complexes were retrieved from the Nucleic Acid Database (47) and the Protein Data Bank (PDB) (48). X-ray structures were selected containing protein and DNA longer than 6 nt, not RNA, and with crystallographic resolution better than 3.3 A. The resolution limit of 3.3 A was used to include as many functionally different complexes as possible. Short nucleotides were excluded for their low information content. The resulting 1475 structures are listed in Supplementary Table S1. Locally installed MolProbity suite (49,50) was used to add hydrogens, utilizing the option to flip oxygens and nitrogens in asparagine, glutamine and histidine residues.

**Elimination of sequence identities and similarities**

Sequence redundancy among 1475 structures was treated at two levels of stringency leading to two different datasets—Que and Umb. A list of selected structures is given in Supplementary Table S1.

1. **Que**–data set containing 339 complexes with sequentially unique proteins. Close evolutionary relationships among the protein sequences were avoided by removing structures with 50% or larger protein sequence identity. From two redundant structures, the one with higher crystallographic resolution was retained. If the resolution between two structures differed by <0.2 A, structure with lower MolProbity score (49) was selected.

2. **Umb**–data set containing 1018 complexes with unique interfaces. This selection was based only on the identity of DNA sequences. Two complexes were considered unique when they differed at least by two (for strands shorter than 24 nt) or by three (for strands longer than 25 nt) nucleotides. The rationale for this less stringent selection based primarily on DNA sequences lies in the fact that we studied the structural features of the protein/DNA interfaces, not the protein or DNA behavior per se. A larger size of the Umb data set allowed an additional classification of structures by a protein functional class and by crystallographic resolution.

**Protein classification**

In addition to Que and Umb data sets, data sets containing proteins with more specific functions were analyzed. Structures were divided into broad categories consisting of enzymes (Enz), proteins regulating transcription (TrF)...
and structural proteins (Str). Structures containing enzymes were further classified as nuclease (Nuc) and polymerases (Pol). Other groups of structures such as DNA complexes with DNA repair proteins (Air), proteins operating on DNA topology (Top) and histone particles (His) were created, but they were not large enough to perform statistically reliable analysis. Functional classification of proteins was based primarily on the Pfam database (51); ~15% of structures with missing Pfam annotations were classified manually based on the information in their original articles.

Because many structural features depend on the crystallographic resolution, the complexes were analyzed in three resolution bins: high-resolution structures up to 1.9 Å (labeled R1), middle-resolution structures between 1.9 and 2.8 Å (labeled R2) and low-resolution structures between 2.8 and 3.3 Å (labeled R3). Abbreviations and counts of structures in various functional groups and resolution bins are summarized in Table 1.

**Table 1. Number of analyzed structures**

<table>
<thead>
<tr>
<th>Group of structures</th>
<th>Description</th>
<th>Code</th>
<th>R1: up to 1.90 Å</th>
<th>R2: 1.90–2.80 Å</th>
<th>R3: 2.80–3.30 Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>Unique interface</td>
<td>Umb</td>
<td>200</td>
<td>636</td>
<td>182</td>
</tr>
<tr>
<td>Subsets of structures</td>
<td>Enzymes</td>
<td>Enz</td>
<td>121</td>
<td>351</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Regulatory</td>
<td>TrF</td>
<td>71</td>
<td>255</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>Structural</td>
<td>Str</td>
<td>8</td>
<td>32</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Nuclease</td>
<td>Nuc</td>
<td>46</td>
<td>101</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Polymerase</td>
<td>Pol</td>
<td>32</td>
<td>133</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Repair</td>
<td>Air</td>
<td>28</td>
<td>82</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Topology</td>
<td>Top</td>
<td>3</td>
<td>34</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Histone</td>
<td>His</td>
<td>2</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Sequentially unique</td>
<td>Que</td>
<td>100</td>
<td>205</td>
<td>34</td>
</tr>
</tbody>
</table>

Shown are the numbers of structures in the considered groups as a function of crystallographic resolution. Umb, ‘Unique interfaces’ represent the largest analyzed group, all others are just subsets.

Modified nucleotides and amino acids

Modified amino acid residues were not excluded from the analysis because they are rare, chemically homogeneous (mostly phosphorylated serines) and most of them occur outside the contact area with DNA. The identity of the modified amino acids was assigned to the parent natural amino acid.

On the other hand, chemically modified nucleotides occur more frequently and their modifications may be significant. Hence, we analyzed chemical structure of all modified nucleotide residues individually; of all 84 types of chemically modified nucleotides, 38 were judged chemically close to their parent residues and sterically not too different from the natural nucleotides, so they were included in the analyzed sample, and the other 46 were excluded. The list of all modified residues and PDB IDs of structures where they occur is given in the Supplementary Table S2.

Protein/DNA contacts

Nucleotide and amino acid residues in contact define the protein/DNA interface. We calculated direct and water-mediated protein/DNA contacts using in-house scripts using the Visual Molecular Dynamics (VMD) program (52). A nucleotide and amino acid residues were considered in a direct contact if any of their non-hydrogen atoms were closer than 3.40 Å. The direct contacts were classified as polar between polar atoms and as van der Waals between non-polar atoms. Water-mediated protein/DNA contacts were assigned to nucleotide and amino acid atoms that were connected by water oxygen no further than 3.40 Å. Direct and water-mediated contacts were assigned independently, i.e. an atom may be involved in both. All contacts were determined considering the crystallographic symmetry.

Classification of local conformations

**Protein blocks**

PBs are pentapeptide conformers defined by five pairs of the \( \Phi, \Psi \) peptide dihedral angles. The 16 local prototypes of the alphabets labeled from \( a \) to \( p \) were obtained by an unsupervised classification similar to Kohonen Maps and hidden Markov models of 342 non-homologous protein structures (44). This structural alphabet allows a reasonable approximation of local protein 3D structures with a root-mean-square deviation evaluated to be 0.42 Å, and is currently the most widely used structural alphabet (53). The PBs were assigned to all protein chains in the analyzed set of complexes according to the published procedure (54). A brief qualitative description of PB conformations and their occurrence at and outside the protein/DNA interface are given in Table 2.

**Assignment of DNA conformer classes (ntC)**

A DNA structural alphabet characterizing local conformations of ntC units was developed by Svozil et al. (46). In the present work, we critically consolidated a larger set of originally published conformers into a group of 18 letters. Three Z-DNA conformers were assigned but not further analyzed, and an additional ntC (referred to as ‘NN’) was designated to conformations that could not be assigned to any of the existing classes. NtCs were assigned to DNA steps using a modified version of a \( k \)-nearest neighbor algorithm (55). The ntC classes are...
briefly characterized in Table 3 and their backbone torsions are summarized in the Supplementary Table S3. After the assignment, three conformers with \( \chi \) angle in the syn region (\( \chi < 180^\circ \)), ntCs 119, 121 and 122, were pooled into one ntC labeled 155. Together with structurally diverse ntC class NN, we analyzed 14 DNA conformational classes.

**Table 3.** PBs (44) assigned to proteins in the 1018 analyzed protein/DNA complexes with unique interface (Umb) and their occurrence

<table>
<thead>
<tr>
<th>PB label</th>
<th>Brief characterization</th>
<th>At the interface</th>
<th>Outside the interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, c</td>
<td>N-terminus of β-strand</td>
<td>4465</td>
<td>74 544</td>
</tr>
<tr>
<td>d</td>
<td>Center of β-strand</td>
<td>5163</td>
<td>78 833</td>
</tr>
<tr>
<td>e, f</td>
<td>C-terminus of β-strand</td>
<td>3097</td>
<td>38 039</td>
</tr>
<tr>
<td>g, h, i, j</td>
<td>C-terminus of α-helix</td>
<td>2241</td>
<td>22 072</td>
</tr>
<tr>
<td>k, l</td>
<td>C-terminus of α-helix</td>
<td>5884</td>
<td>50 877</td>
</tr>
<tr>
<td>m</td>
<td>Center of α-helix</td>
<td>7978</td>
<td>174 348</td>
</tr>
</tbody>
</table>

*Number of PBs identified at and outside the protein/DNA interface in 1018 analyzed structures.*

**Table 3.** Nucleotide conformers (ntC) used for the conformational assignment (55) of 57 797 DNA steps in the 1018 analyzed protein/DNA complexes (the Umb data set)

<table>
<thead>
<tr>
<th>ntC</th>
<th>Symbol</th>
<th>Characterization</th>
<th>At the interface</th>
<th>Outside the interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>A</td>
<td>The most frequent A-DNA</td>
<td>1242</td>
<td>354</td>
</tr>
<tr>
<td>13</td>
<td>A</td>
<td>A-DNA, BI-like ( \chi )</td>
<td>727</td>
<td>202</td>
</tr>
<tr>
<td>19</td>
<td>A</td>
<td>A-DNA, ( \alpha+1/\gamma+1 ) crank</td>
<td>573</td>
<td>205</td>
</tr>
<tr>
<td>41</td>
<td>A2B</td>
<td>A-to-B, ( \delta &gt; C^3 )-, ( \delta+1 C^2 )-endo</td>
<td>2014</td>
<td>724</td>
</tr>
<tr>
<td>32</td>
<td>B12A</td>
<td>BI-to-A, ( \delta+1 O^4 )-endo</td>
<td>358</td>
<td>909</td>
</tr>
<tr>
<td>109</td>
<td>B12A</td>
<td>BI-to-A, ( \delta+1 C^3 )-endo</td>
<td>333</td>
<td>106</td>
</tr>
<tr>
<td>110</td>
<td>B12A</td>
<td>as 109 plus ( \alpha+1/\gamma+1 ) crank, high ( \beta+1 )</td>
<td>457</td>
<td>267</td>
</tr>
<tr>
<td>54</td>
<td>BI</td>
<td>The most frequent BI variant</td>
<td>9261</td>
<td>7529</td>
</tr>
<tr>
<td>50</td>
<td>BI</td>
<td>BI variant</td>
<td>3677</td>
<td>2073</td>
</tr>
<tr>
<td>86</td>
<td>BI</td>
<td>the most frequent BI variant</td>
<td>2805</td>
<td>2820</td>
</tr>
<tr>
<td>96</td>
<td>BI</td>
<td>BI variant</td>
<td>1620</td>
<td>1133</td>
</tr>
<tr>
<td>116</td>
<td>BI</td>
<td>BI, ( \alpha+1/\gamma+1 ) crank, ( \alpha/\gamma ) normal</td>
<td>2431</td>
<td>1935</td>
</tr>
<tr>
<td>155</td>
<td>B1syn</td>
<td>orig. 119: 5'-mismatches, BI, ( \chi ) syn, ( \alpha/\gamma ) crank</td>
<td>254</td>
<td>188</td>
</tr>
<tr>
<td>155</td>
<td>B1syn</td>
<td>orig. 121: 3'-mismatches, ( \delta O^4 )-endo, ( \chi+1 ) syn</td>
<td>1217</td>
<td>1113</td>
</tr>
<tr>
<td>155</td>
<td>B1syn</td>
<td>orig. 122: as 121 plus ( \alpha+1/\gamma+1 ) crank</td>
<td>3421</td>
<td>2854</td>
</tr>
</tbody>
</table>

*Number of ntCs identified at and outside the protein/DNA interface in the Umb data set.*
Figure 1. Occurrence of protein and DNA structural descriptors at and outside the protein/DNA interface for the group of structures Umb-R2 (636 complexes with crystallographic resolution between 1.90 and 2.80 Å). Histograms show distributions of amino acid residues (top), PBs (center) and ntCs (bottom) involved in direct polar contacts. Histograms on the left show the relative frequencies at the interface (in yellow) and outside the interface (in red). Histograms on the right show logodds of these frequencies, with underpopulation indicated by blue and overpopulation by red; hue indicates the significance of the effect. PBs are labeled by their one-letter codes (Table 2) and ntC by their numbers as defined in Table 3. Histograms for other groups of complexes are given in Supplementary Figure S1.
factors (TrF) and polymerases (Pol). The structures are divided into three groups based on their crystallographic resolution; the middle-resolution bin R2 comprising structures between 1.90 and 2.80 Å contains most structures (Table 1), so we primarily concentrate on the analysis of this bin.

Statistics of contacts for selected classes of structures

Table 4 shows selected statistics of direct polar contacts for selected groups of structures in the three resolution bins; a more detailed account of various statistical measures of the interactions can be found in Supplementary Table S4. In the high-resolution bin R1, only enzyme complexes are numerous enough to be analyzed as a separate subgroup. On the other hand, in the medium-resolution bin R2, we could also analyze transcription factors, nucleases and polymerases (TrF, Nuc and Pol) individually.

Table 4 shows that polar contacts are, on average, mediated by 1.3 atoms in amino acid residues, and by 1.7 atoms in nucleotides. For amino acids, these numbers are remarkably similar within all groups of structures, and slightly more variable for nucleotides. Water-mediated contacts are as common as direct polar contacts as demonstrated by numbers under the ‘HOH/polar’ column in Table 4, and their role is discussed in greater detail in ‘The role of water-mediated contacts’.

To test the robustness of the observed features of the large Umh group (group with sequentially unique interfaces), we compared them with the features of the Que group (sequentially unique proteins). Descriptors given in Table 4 show virtually identical values for Que-R2 and Umb-R2 data sets, and other descriptors analyzed in this work also demonstrate similar-to-identical characteristics of these two groups in all resolution bins (see also Supplementary Table S4).

Protein structure elements

Neither type of interactions (direct polar, water-mediated, van der Waals) nor resolution changes the general pattern of protein binding characteristics. As expected (18,19,26), most contacts to DNA are formed by arginine and lysine followed by other polar and/or charged amino acids (Figure 1, Supplementary Figure S1). Positively charged arginine is overpopulated at the negatively charged DNA surface regardless of the structural type or resolution, and lysine is overpopulated in most groups. Lipophilic amino acids, namely, leucine, valine, isoleucine, methionine and phenylalanine, have low occurrence at the polar interface and are statistically underrepresented. Strong underrepresentation of proline at the interface likely originates in its structural rather than lipophilic properties. In contrast to large differences in the presence of individual amino acids at and outside the interface, protein secondary structural elements do not show any preferences for the interface (not shown). In other words, no secondary structural element can be identified as a key building block for DNA recognition.

As Figure 1 shows, PBs have a larger discriminatory power in identifying structural elements recognizing DNA than secondary structure elements. PBs overpopulated at the interface are N-termini of α-helix and β-sheet (PBs k, l, h) and coil blocks (PBs h, j), and PBs underpopulated are central and especially C-terminal parts of α-helix (PBs p and n). We observed no real differences in the occurrence of these PBs between direct polar and water-mediated interactions.

Description of the protein local structure by PBs allowed observing differences between the general protein structure and structural features observed at the interface with DNA. Coil-related PB g, the second least frequent PB (56) associated with flexible regions, is even less present at the interface. Underrepresentation was also observed for some frequent sequences of PBs classified by de Brevern (57) as ‘Structural Words’, e.g. mnopac.

DNA structure elements

The dominant DNA form, BI-DNA, is represented here by ntCs 54 and 50. It is the most common form at the protein/DNA interface in all groups of structures. What distinguishes interacting DNA from unbound DNA is a larger relative occurrence of the A-forms in protein/DNA complexes (25,58–60). We observed an increased occurrence of the ‘canonical’ A-form (ntC 8), but owing to our finer classification of DNA conformers, also of deformed A-like and especially of mixed A/B conformers. The population of ntC 13 is notably increased. The occurrences ntCs 41 and 19 are also increased. NiC 41 with the A-like backbone but B-like values of the glycosidic torsion angle χ preserves perpendicular orientation of the base pairs relative to the helical axis; ntC 19 is an A-form with α and γ torsions switched from the 300°/60° canonical values to the 150°/180° combination (‘crankshaft’ motion). Although the most common BII-form (ntC 86)

Table 4. Protein/DNA contacts

<table>
<thead>
<tr>
<th>Structures</th>
<th>Residues in polar contacts</th>
<th>Atom-to-atom polar contacts per residue</th>
<th>HOH/polar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>Number</td>
<td>aa nt</td>
<td>aa nt</td>
</tr>
<tr>
<td>Umb-R1</td>
<td>200</td>
<td>3764 2445</td>
<td>1.33 1.81</td>
</tr>
<tr>
<td>Enz-R1</td>
<td>121</td>
<td>2399 1491</td>
<td>1.29 1.81</td>
</tr>
<tr>
<td>TrF-R1</td>
<td>71</td>
<td>1238 866</td>
<td>1.38 1.80</td>
</tr>
<tr>
<td>Pol-R1</td>
<td>32</td>
<td>562 378</td>
<td>1.24 1.42</td>
</tr>
<tr>
<td>Nuc-R1</td>
<td>46</td>
<td>1396 768</td>
<td>1.33 2.09</td>
</tr>
<tr>
<td>Que-R2</td>
<td>205</td>
<td>2707 2803</td>
<td>1.32 1.69</td>
</tr>
<tr>
<td>Umb-R2</td>
<td>636</td>
<td>14869 10039</td>
<td>1.35 1.71</td>
</tr>
<tr>
<td>Enz-R2</td>
<td>351</td>
<td>8342 5312</td>
<td>1.33 1.73</td>
</tr>
<tr>
<td>TrF-R2</td>
<td>255</td>
<td>5994 4056</td>
<td>1.35 1.68</td>
</tr>
<tr>
<td>Str-R2</td>
<td>32</td>
<td>975 699</td>
<td>1.45 1.65</td>
</tr>
<tr>
<td>Nuc-R2</td>
<td>101</td>
<td>2746 1726</td>
<td>1.34 1.91</td>
</tr>
<tr>
<td>Pol-R2</td>
<td>133</td>
<td>2843 1902</td>
<td>1.35 1.53</td>
</tr>
<tr>
<td>Umb-R3</td>
<td>182</td>
<td>4156 2997</td>
<td>1.32 1.63</td>
</tr>
</tbody>
</table>

*aStatistics for selected groups of structures, for abbreviations see Table 1.

1The columns list the total number of amino acids (aa) and nucleotides (nt) in direct polar contacts in selected groups of structures.

2The columns show how many protein (‘aa’) or DNA (‘nt’) atoms forming direct polar contacts interact per residue.

3HOH/polar show the number of water-mediated contacts divided by the number of direct polar contacts for protein (‘aa’) or DNA (‘nt’) atoms.
is disfavored at the interface, other BII conformers rare in
naked DNA (ntCs 109 and 110) are well represented in
protein/DNA complexes.

Unclassified nucleotides (ntC NN) representing extreme
structural variations are not significantly enriched at the
interface. Apparently, the interaction of proteins with
DNA does not induce any novel DNA local conformers,
but it stabilizes A (ntC 13) and A/B forms (ntCs 41, 32,
109, 110) that appear more often at the interface than in
uncomplexed DNA. Some of these conformers (namely
ntC 32) exhibit values of torsion ψ, which defines sugar
pucker, between 90° and 100° indicating high C3'-endo or
even O4'-endo pucker. Large number of these conformers
at the interface (especially in high-resolution structures)
refutes doubts about the existence of the O4'-endo sugar
pucker in DNA and demonstrates a smooth deformation
of the deoxyribose ring from the C3'-endo to C2'-endo
pucker via the O4'-endo observed in high-resolution
small nucleoside and nucleotide structures (61,62). In
this context, virtual absence of the O4'-endo pucker in
RNA structures (63) may be more a consequence of the
force fields used to refine RNA structures than reflection
of the actual distribution of sugar puckers.

Binding statistics in the group of low-resolution structures
Distributions of direct polar and van der Waals contacts
for structures at the lowest resolution bin R3 (2.80–3.30 Å)
show the same general features as distributions of struc-
tures at the higher resolution bins (Table 4 and
Supplementary Table S4). What discriminates low-reso-
lection structures is a larger number of unclassified ntC
NN that may be attributed to refinement difficulties with
poorly resolved electron density maps and incorrectly
fitted nucleotide conformations. Unexpected is a high fre-
quency of ntC 116, rare BI-form with alpha/gamma
crankshaft compensation. The low number of observed
water molecules in low-resolution structures does not
allow analysis of water-mediated contacts.

Interaction matrices: correlations between interacting
PBs and ntCs
The counts of mutually interacting PBs and ntCs are pre-
sented in a form of 'interaction matrices' that show how
many protein and nucleotide conformers of certain type
interact and reflect therefore the local geometry of the
interface. Figure 2 shows interaction matrices for direct
polar contacts in the medium-resolution group of struc-
tures Umb-R2, and its subgroups TrF-R2 and Nuc-R2.
Interaction matrices for direct polar (Figure 2), water-
mediated (Supplementary Figure S2a) and van der Waals
(not shown) contacts are similar. Moreover, most
observations made for the medium resolution structures
are also valid for the high-resolution data set Umb-R1
(Supplementary Figure S2b).

The most frequent interactions occur between the main
architectural units of proteins and DNA, DNA BI form
ntC 54 and protein χ-helical PB m and β-strand PB d,
which form 15 and 12% of all contacts, respectively.
However, according to the logodds analysis, neither m54
nor d54 combination prefers or avoids the interface.

Combinations of conformers that characterize the inter-
face (occur at the interface with higher than expected fre-
quency and are therefore 'statistically overrepresented')
are A and mixed A/B DNA forms (mainly ntCs 8, 13,
19) associated with β-sheet (PBs b, d) and coil (PBs h, i,
j). Strongly overrepresented are also interactions between
less populated B-to-A ntCs 109 and 110 and PBs e
(C-terminus of β-strand), h (coil) and k (N-terminus of
χ-helix). In contrast, conformers that avoid the interface
are BII forms (ntCs 86, 96) and the C-terminal segments
of the χ-helix (PBs n, o and especially p). The most negatively
correlated associations are BII forms with the coil PB g
and the N-terminal χ-sheet PB a. The described pattern is
similar for medium- as well as high-resolution structures
and for direct polar and water-mediated contacts.

Figure 3a depicts examples of the most frequent PB/ntC
interaction partners. The dominant BI form (ntC 54) partic-
tipates frequently in contacts with χ-helical (m54, k54) as
well as β-sheet (d54, f54) PBs. The BI ntC 86 is common
at the interface (even when statistically underrepresented)
and its contacts with the main χ-helical PB m are frequent
(motif m86 in Figure 3a). A comparison of the three
binding motifs between the χ-helical PB m and three
B-DNA conformers, 54, 86 and 116 (less-populated BI
conformer), shows variability of the mutual orientation
between the B-DNA major groove and χ-helix. Arginine
contacting the major groove guanine O6 is, in most cases,
in its extended rotamer, but it can also accommodate more
compact rotameric forms as in motifs m86 and k54.

While motifs drawn in Figure 3a are common in all
types of complexes, Figure 3b depicts motifs typical for
complexes of transcription factors TrF-R2 (m41 and d13),
and for nucleases Nuc-R2 (f41, d19, k50 and i8).
Complexes of transcription factors have interaction
matrices similar to the matrices of the whole data set
Umb-R2 with dominating BI-DNA and χ-helical con-
formers. In contrast, complexes of nucleases (Nuc-R2) use
a wider spectrum of conformers at the interface, dominance
of BI ntC 54 is visibly weaker and more contacts are
actually formed by β-strand PB d than by otherwise
more populated χ-helical PB m; many contacts are also
formed by β-strand PB f. Preference for the A-like forms
measured by logodds is much stronger than in Umb or
TrF data sets, especially in combinations with β-strand
f, coil h and N-terminal χ-helical PBs k and l. The
population of undefined nucleotides NN is surprisingly
high. The BII forms are infrequent and statistically
disfavored. Conformational diversity of DNA/nuclease
interactions is underscored by their larger chemical
variability when fewer contacts are formed by arginine;
we show interacting lysine side chains (k50, i8) and also
a serine motif f41.

Contacts to the DNA minor groove, major groove and
phosphate
Protein interactions to DNA constituents, the minor
groove (mig), the major groove (MAG), the phosphate
(PH) and deoxyribose, are distributed unevenly. The phos-
phate atoms OP1 and OP2 form a large part of all polar
contacts to protein atoms, more than a half. On the other
Figure 2. Interaction matrices for direct polar contacts of the three groups of structures with crystallographic resolution between 1.90 and 2.80 Å (bin R2). Top: 636 protein/DNA complexes, Umb-R2. Center: 255 complexes of transcription factors, TrF-R2. Bottom: 101 complexes of nucleases, Nuc-R2. The matrices on the left show how many peptide blocks, PBs, interact with nucleotide conformers, nCs, the highest populations are highlighted in red. The matrices on the right show how much are the interactions statistically different from their expected frequencies estimated by
side, deoxyribose atoms O4', O5' and O3' together form ~5% of the contacts and are not important for protein binding. The proportion for direct polar contacts is mig:MAG:PH = 1:2:9 in the Umb-R1 data set, and comparable 1:3:15 in Umb-R2 (data for other datasets are given in Supplementary Table S5). Water-mediated contacts are distributed more evenly, and the corresponding ratios for water-mediated contacts are 1:2:6 and 1:2:7, respectively. Lower relative number of water-mediated contacts at phosphates shows that water molecules are better localized in the grooves than around more accessible phosphates.

Interaction matrices of the minor groove contacts have distinct patterns, and also other statistics of contacts to mig differ from matrices constructed for MAG and PH (Supplementary Figure S2c versus S2d and S2e). The interaction matrices are formed by more β-sheet than α-helix contacts and also BI dominance is much lower than for contacts to MAG or PH. The second most populated nucleotide conformer is ntC NN that strongly correlates with β-sheet PB d; we do not have explanation for this observation. The differences observed between interaction matrices of TrF and Nuc for all contacts are more pronounced in mig; despite

Figure 3. Examples of the common protein/DNA interactions. Interacting motifs are labeled by the codes of the interacting PB and ntC (Tables 2 and 3, respectively) and by PDB id of structures in which they were identified. Interacting PBs are drawn as green cartoon with atom of the central amino acid in light green and the nucleotide step as a stick model using commonly used ‘chemical’ colors; the contacts (black sticks) are directed to the major groove edge of guanines in the right-handed double helical DNA. The 5'-end phosphates are on the left top of each motif. The N-ends of the PBs are labeled; the complementary DNA strand and amino acids adjacent to the depicted PB are in light gray. Molecular graphics was created by program Chimera.

(a) Motifs common to all types of structures approximately in order of their occurrence in the group of all 1018 structures. All contacts shown are between the guanine atom O6 and the arginine NH observed in crystal structures 3exj (65), 1nfk (66), 1bc8 (67), 1run (68), 1g2d (69) and 2i13 (70). (b) Motifs m41 and d13 are highly populated in transcription factors (TrF-R2) and underrepresented in nucleases (Nuc-R2), motifs f41, d19, k50 and l8 are highly populated in Nuc-R2 and less in TrF-R2. They appear in crystal structures 1au7 (71), 1mjq (72), 1sa3 (73), 2fkc (74), 2e52 (76). The motifs m41, d13 and d19 show interaction between the guanine O6 and arginine NH, k50 and l8 between the guanine O6 and lysine NZ, and f41 between the guanine N7 and serine OG.
the lower counts in the mig matrices, it seems clear that these interactions disfavor the BI-form, may induce unusual DNA conformers (ntC NN) and generally prefer β-sheet over α-helix.

Water-mediated contacts to the minor groove show fewer of these extreme features, and their interaction matrices resemble the interaction matrices of major groove and phosphates. A notable overall feature of the minor groove atoms is that they actually form more water-mediated than direct polar contacts, 1.5 times more in the medium-resolution structures (Umbr-R2), the corresponding ratios are 1.1 in MAG, and 0.7 in PH. High-resolution structures (Umbr-R1) show the same trend. Interaction of the narrow mig with proteins, therefore, requires either its substantial deformations or alleviation of the steric constraints by water-mediated contact.

Distribution of protein contacts to the grooves and phosphates is in some groups of structures different from the average values given above. Extreme behavior was observed for transcription factors (TrF) that have direct polar and water-mediated contacts distributed similarly between mig, MAG and PH, and for polymerases (Pol) with different distributions (ratios are listed in Supplementary Table S5). Because polymerases distribute fewer water contacts per residue than transcription factors (0.66 versus 0.90, Table 4), their interface is ‘more’ dehydrated than the interface of transcription factors, and this dehydration of polymerases is most pronounced for phosphate atoms.

The role of water-mediated contacts

The number of residues linked by direct polar contacts and by water bridges is comparable even for the medium-resolution structures (Umbr-R2) where 20 000 amino acids contact DNA directly and 16 000 via water. The last two columns of Table 4 (‘HOH/polar’) show that the number of water-mediated contacts divided by the number of direct polar contacts varies between various groups of structures. The highest proportion of water-mediated contacts was observed for complexes of transcription factors and nucleases, the lowest for polymerases (extremely low value for Str-R2 may be skewed by histone complexes). High proportion of water-mediated contacts in transcription factors in both relevant resolution bins, TrF-R1 and TrF-R2, is perhaps surprising, especially in the light of the fact that polymerases with arguably less stringent demand for specificity of interaction have their proportion of water contacts lower.

High proportion of water-mediated contacts in all complexes, and especially in complexes of transcription factors, suggests that these structured water molecules play an active role in the process of protein/DNA recognition and do not serve as mere fillers of cavities formed at imperfectly matching protein and DNA molecular surfaces as has been sometimes suggested (77). Similarity of the PB/ntC interaction matrices for direct polar and water-mediated contacts (Figure 2 and Supplementary Figure S2a) further demonstrates that interaction by direct polar contacts and via the interface waters has similar conformational constraints on both protein and DNA partners and indirectly points again to the active role of water to the recognition.

On complexation, heavily hydrated surfaces of protein and DNA molecules release a large number of water molecules and ions increasing entropy of the interaction and thus compensating for the entropy loss caused by the complex formation (32,78–80). Around the naked DNA double helices, water and cations lie in spatially localized hydration sites (81–83) that coincide largely with protein interaction sites (84). The waters trapped at the interface represent the remains of the first-shell waters and cations that have specific physical properties (79,85–87), and become an ‘integral part’ (29) of the protein/DNA interface (30). The packing of atoms at protein–DNA interfaces is as high as in the protein interior, and cavities at the interface are filled with water more frequently than the protein interior (88). Therefore, it is plausible to state that water contributes significantly to the protein/DNA recognition (84,89) and participates in protein/DNA interactions (90,91).

Stabilization of the A-forms at the interface

High relative occurrence of A- and A/B DNA forms at the protein/DNA interface observed in the interaction matrices can be interpreted as remodeling of the B-form to the A-form. Almost continuous plastic deformation from B-to-A state through several minor conformational states (46) is accompanied by bending of the duplex that modifies the widths of the major and minor grooves and changes the exposition of the base pairs, deoxyribose and mainly phosphate atoms (59). The narrowing of the major groove of the protein-induced A and A/B conformers could provide one mechanism for forming specific contacts to a protein-binding motif preserving the essential stacking interactions of the base pairs (18). In some complexes, binding requires a high degree of DNA distortion (92,93), and a shift in the distribution of conformers from naked to complexed DNA suggests that conformational deformability and flexibility of DNA are essential for the recognition (94–96). The tendency to induce A-like conformers at the interface is accompanied by a shift from the C2'-endo sugar pucker typical for B-forms toward the C3'-endo pucker family, the effect described as the ‘sugar switching’ that facilitates hydrophobic recognition in the minor groove (97,98).

The driving force of the A-to-B transformation in naked DNA, partial dehydration of the DNA surface, is well known (99) (100) so that partial dehydration of DNA on complexation with proteins works in accord with the aforementioned steric reasons, and may contribute to the relative preference of the A- over the B-forms at the interface. The fact that the A-like structures are similarly overrepresented at the interface for direct polar and water-mediated contacts does not directly confirm or exclude such possibility, and in our opinion, the A and A/B conformers are induced in the protein/DNA complexes likely by a combination of two factors, the partial dehydration required by the complexation and the ability of DNA to adjust its conformation to protein (58,59) and in a broader sense, to reflect the environment (101,102).
CONCLUSIONS

We analyzed structural features of the protein/DNA interface and compared them with the features of non-interacting parts of proteins and DNA. Structures of proteins and DNA were classified by structural alphabets. Protein local conformers were classified into 16 pentapeptide PBs (44,53), and DNA into 14 ntCs (46,55). These structural alphabets describe biopolymer conformations at greater detail than elements of protein secondary structure and than crude and sometimes subjective DNA structural types such as A, BI and BII. Direct polar and water-mediated protein–DNA contacts were analyzed in >1000 protein/DNA crystal structures in three bins of crystallographic resolution. The counts of mutually interacting PBs and ntCs were assembled into ‘interaction matrices’ that serve as comprehensive description of structural features of the interface. The matrices demonstrate that minor DNA conformers are often significantly enriched at the interface so that the ability of DNA to adopt non-canonical conformers rare in naked DNA is clearly essential for the recognition by proteins. Rare DNA forms introduce significant deformations to the DNA regular structure and the occurrence of these rare forms was characterized here enabling better understanding of the role of non-B-DNA structures for genetic instability and evolution (103).

The well-known tendency of DNA to adopt A-like forms on protein binding (58,59) should be understood as a relative preference because the BI forms are the most frequent even at the interface (Figures 1 and 2). Our detailed structural classification of DNA conformers allowed a specific characterization of A-like forms enriched at the interface. We showed that the interaction with proteins induces more gradual deformations of the B form into B-A, A-B and exotic A conformations rather than solely into the canonical A-DNA. Importantly, unclassified conformers (ntC NN) representing rare or incorrectly refined conformers are not overpopulated at the interface so that interactions with proteins do not induce conformations unseen in naked DNA but only stabilize the less stable forms. The relative stabilization of the A-like forms at the interface is likely facilitated by synergy of the steric accommodation to the interacting protein and dehydration occurring during the interaction that also stabilizes the A-form.

The interaction matrices of direct polar and water-mediated contacts are remarkably similar, and water-mediated contacts are nearly as numerous as direct polar ones. Water molecules trapped at the interface are important for the binding by alleviating steric incompatibility between protein and DNA so that the interacting peptide and nucleotide fragments can remain in their energetically low-lying conformations. An important role of water molecules for the recognition is further underscored by their high occurrence at the interfaces with transcription factors (Table 4, column HOH/polar).

Both features characterizing protein/DNA binding, i.e. reduction of the mutual steric incompatibility by water bridges and induction of the B-to-A transition, are best visible in interaction matrices constructed for contacts to the narrow minor groove. They are conspicuously different from the matrices constructed for contacts to the major groove and phosphate atoms. Remarkably, water-mediated interactions form more than a half of all the contacts in the minor groove, while the proportion of ordered waters around the major groove and especially phosphate atoms is lower.

Interaction matrices counting contacts between protein and DNA residues classified into structural alphabets represent robust and comprehensive description of the interface and contribute to the understanding of principles underlying protein/DNA recognition.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work is dedicated to Prof. Helen M. Berman (RCSB PDB, Rutgers University), a great tutor and true aficionado of structural biology.

FUNDING

The Czech-France collaboration Barrande [MEB021032]; BIOCEV CZ.1.05/1.1.00/02.0109 from the ERDF, [P305/80 12/1801] from the Czech Science Foundation, and institutional [AV0Z50520701]; supported by [MSM 6046137302 to D.S. and P.Cˇ ]; supported by the Ministry of Research (France), University Paris Diderot, Sorbonne Paris Cité (France), National Institute of Blood Transfusion (INTS, France), National Institute of Health and Medical Research (INSERM, France) and ‘Investissements d’avenir’, Laboratories of Excellence GR-Ex (France) (to J.C.G. and A.G.dB.). Funding for open access charge: Czech Science Foundation and Academy of Sciences of the Czech Republic.

Conflict of interest statement. None declared.

REFERENCES


