doi:10.1016/j.jmb.2004.01.011

J. Mol. Biol. (2004) **337**, 65–76

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Protein–DNA Hydrophobic Recognition in the Minor Groove is Facilitated by Sugar Switching

Michael Y. Tolstorukov[1], Robert L. Jernigan[2] and Victor B. Zhurkin[1]*

[1]*Laboratory of Experimental and Computational Biology National Cancer Institute National Institutes of Health Bg. 12B, Rm. B116, Bethesda MD 20892-5677, USA*

[2]*Laurence H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011 USA*

Corresponding author

Information readout in the DNA minor groove is accompanied by substantial DNA deformations, such as sugar switching between the two conformational domains, *B*-like C2′-*endo* and *A*-like C3′-*endo*. The effect of sugar puckering on the sequence-dependent protein–DNA interactions has not been studied systematically, however. Here, we analyzed the structural role of *A*-like nucleotides in 156 protein–DNA complexes solved by X-ray crystallography and NMR. To this end, a new algorithm was developed to distinguish interactions in the minor groove from those in the major groove, and to calculate the solvent-accessible surface areas in each groove separately. Based on this approach, we found a striking difference between the sets of amino acids interacting with *B*-like and *A*-like nucleotides in the minor groove. Polar amino acids mostly interact with *B*-nucleotides, while hydrophobic amino acids interact extensively with *A*-nucleotides (a hydrophobicity–structure correlation). This tendency is consistent with the larger exposure of hydrophobic surfaces in the case of *A*-like sugars. Overall, the *A*-like nucleotides aid in achieving protein-induced fit in two major ways. First, hydrophobic clusters formed by several consecutive *A*-like sugars interact cooperatively with the non-polar surfaces in proteins. Second, the sugar switching occurs in large kinks promoted by direct protein contact, predominantly at the pyrimidine–purine dimeric steps. The sequence preference for the *B*-to-*A* sugar repuckering, observed for pyrimidines, suggests that the described DNA deformations contribute to specificity of the protein–DNA recognition in the minor groove.

Keywords: protein–DNA recognition; DNA deformability; hydrophobic interactions; sugar pucker; DNA accessible surface

Introduction

In addition to hydrogen bonding,[1] sequence-dependent deformability of DNA[2] contributes to protein–DNA recognition, including minor groove interactions. There, the base specificity is obscured:[1] pyrimidines have identical atom groups and the N2 of guanine is the only difference between purines. Still, the minor groove interactions are rather specific and important for complex formation with many proteins, from enzymes to transcription factors to architectural proteins.[3] The characteristic examples are the eukaryotic regulatory proteins TBP, LEF-1, HMG1, hSRY, etc., which recognize a DNA sequence through the minor groove, possibly binding to nucleosomal DNA on the outer side of the loop.[4] The molecular origins of recognition through the DNA sequence-dependent deformability were analyzed in each of these cases.[5–10] Yet, understanding of the general principles of the DNA mechanics underlying such an indirect readout of the sequence has remained on the "intuitive" level.

The slow progress in this direction is related to enormous complexity of the protein–DNA interactions, mostly implemented in electrostatic and van der Waals contacts, rather than hydrogen bonds.[11] Bulk of those contacts is made to sequence non-specific DNA groups, namely to sugar-phosphate backbone, especially in the minor groove. In these cases, the sequence specificity is brought into play through DNA deformability. The duplex deformations, which are strongly sequence-dependent,[2] enable protein contacts to various non-specific

Abbreviation used: ASA, accessible surface area.

E-mail address of the corresponding author: zhurkin@nih.gov

0022-2836/$ - see front matter Published by Elsevier Ltd.

DNA groups that would be impossible if DNA remained in "standard" *B* configuration. In other words, two schemes are being used to facilitate the recognition: in addition to "digital" H-bond recognition, the "analogue" one helps in fitting to structural constraints in the complexes.[12] So, there is no simple code for protein–DNA recognition.[13,14]

To solve this problem, thorough analyses of the available structural data on the protein–DNA complexes have been carried out.[11,14–21] This approach has been extensively exploited for the analysis of the relationship between the hydration sites and the sites of protein–DNA contacts,[15,16] for overall characterization of interacting surfaces of DNA and proteins,[17,18] for analysis of the amino acid–base specificity in the context of local geometry[11,14,19,20] and DNA structural features in protein–DNA complexes.[18,21]

To date, the main emphasis has been on the deformations in the geometry of base-pairs and dimeric steps, whereas the backbone variability has mostly been ignored, since it does not bear "direct chemical" information of the DNA sequence. On the other hand, to analyze deformability of the duplex and its role in recognition, one has to consider the pivotal elements (hinges) of its structure. Among the principal hinges are the sugar ring puckering and the torsion angle about the glycosyl C1′–N linkage:[22] they are strongly correlated in *B*-DNA crystal structures and differ for purines and pyrimidines.[23] Although the sugars *per se* are sequence non-specific, their closeness to the bases implies that whenever the sugar conformation is changed, the accessibility of the corresponding base is also influenced, especially in the minor groove (cf. Figure 1(a) and (b)). Further, changes in sugar pucker are often associated with larger alterations to DNA structure, such as the *B*-to-*A* transition that widens and flattens the minor groove, thereby providing a more readily accessible surface for interactions with proteins.[16] And, since the energy cost of this transition is sequence-dependent in solution,[24,25] it is plausible that the *B*-to-*A* switch can increase selectivity of the protein–DNA binding.

All these arguments, taken together, have encouraged us to investigate the role of the sugar puckering in the protein–DNA interactions. One of our hypotheses was that structural differences between the *B* and *A*-like conformations might be utilized by proteins for enhancing the indirect readout of DNA sequence. Indeed, the "partial *B*-to-*A*-like" transformation has been repetitively observed in complexes.[10,21,26–28] Generally, it controls the widths of the grooves, giving more space for interactions with proteins in the minor groove, and more specifically, proteins can use the increased exposures of the *A*-like sugars in the minor groove for extensive hydrophobic interactions (Figure 1(a) and (c)). At the same time, the partial *B*-to-*A* transformation can enhance and tighten interactions of the protein α-helices in the major groove by its narrowing.[2,26] Also, widening the minor groove and narrowing the major groove helps proteins introduce substantial bends into DNA.[29] At the mesoscopic level of one or two helical turns, the *B*-to-*A* transition causes shortening of the DNA spacer between two recognition sites interacting with the protein heads, as in the CRP–DNA complex.[30]

A closer look at DNA conformations reveals a more complicated picture, however, because in most of protein–DNA complexes there is no clear-cut *B*-to-*A* transition. Instead of forming 10–15 bp long *A*-DNA fragments as observed in solution,[31] the *A*-like nucleotides tend to aggregate in short clusters, often on one strand, while the complementary strand remains mostly *B*-like (the patterns are available on the web†). Thus, the observed DNA deformations cannot be reduced to the canonical *B*-to-*A* transition as often defined,[21,32] but rather reflect more complex ways of utilization of *A*-nucleotide clusters at the protein–DNA interface. Still, it remains unclear how proteins discriminate between *A* and *B*-like structures, and whether *A*-nucleotides in the complexes are protein-induced deformations, or whether they existed in the DNA before complexation (it is known that a fraction of *A*-like nucleotides is present in *B*-DNA in solution.[33–35]) In other words, can *A*-like nucleotides be signals for proteins and themselves initiate sequence recognition?

Here, we have analyzed the structural role of *A*-nucleotides in protein–DNA complexes. We have developed a novel approach to distinguish interactions in the minor groove from those in the major groove, and to calculate the solvent-accessible surface areas (ASA) in each groove separately. Based on this approach, we found that polar and hydrophobic amino acids demonstrate different preferences for *A* and *B*-like nucleotides, and observed particular schemes for the *B*/*A* selectivity by DNA-binding proteins. Our results imply that sequence preference for the *A*-like sugar puckers in free DNA is operative for the recognition in the minor groove.

Results and Discussion

Protein–DNA contacts in the minor and major grooves

A representative non-redundant set of 156 protein–DNA complexes was selected for the analysis. The numbers of amino acid–DNA contacts (contact profiles) were counted separately for two DNA atom groups: sugars and bases. Then, all nucleotides were placed into two classes according to their sugar puckers: the *B*-like S-domain (e.g. C2′-*endo*) and the *A*-like N-domain (e.g. C3′-*endo*) (see Methods for details).

† http://home.ccr.cancer.gov/lecb/tolstorukov/prot-dna/

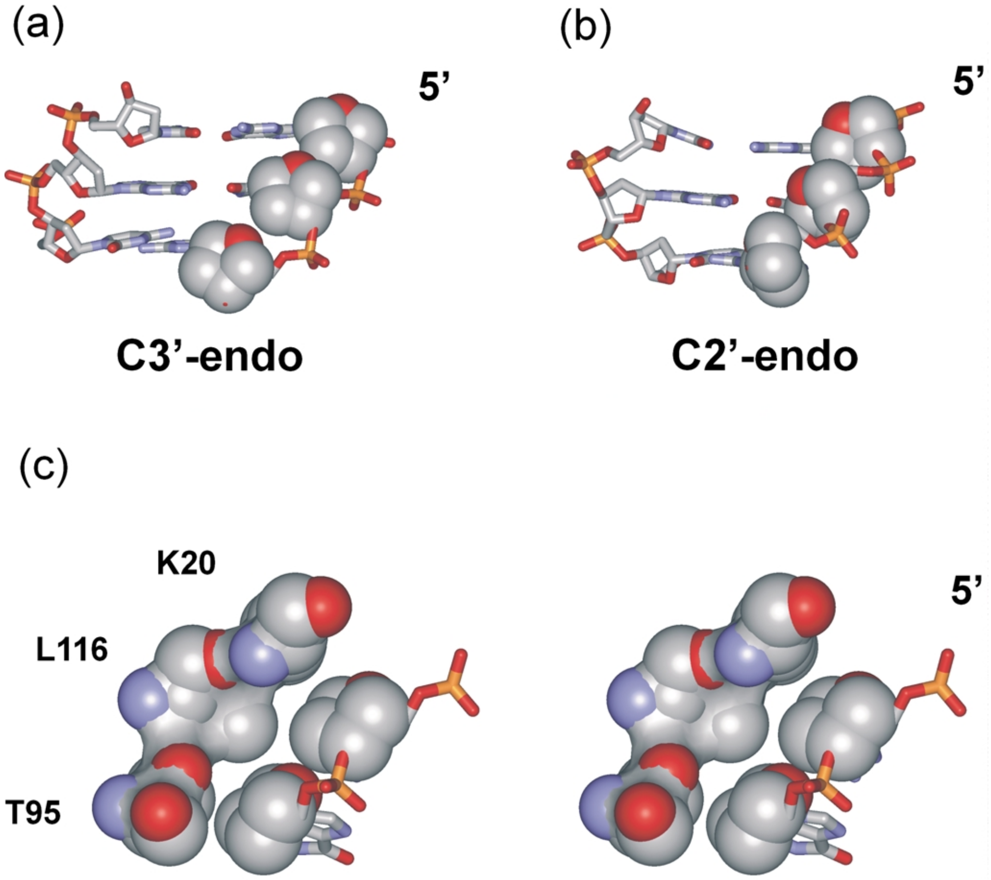

Figure 1. (a) and (b) Minor groove views of the DNA trimer CGC:GCG (the sugar rings in one strand are shown in the CPK representation). (a) *A*-form DNA;[61] (b) *B*-form.[62] The orientation of sugars is essentially different in *A*-DNA and *B*-DNA, with the greater accessibility of hydrophobic atoms in *A*-DNA being evident. (c) An example of the extensive hydrophobic interactions in protein–DNA complexes (stereoview). The complex of I-*Ppo* I with DNA is shown.[63] Notice that the sugar rings and the protein carbon atoms are in direct contact.

The minor groove contact profiles for nucleotides with *B* and *A*-like sugars are strikingly different (Figure 2(a)–(d)). The *B*-like nucleotides interact mostly with the polar amino acids, especially Arg and Lys: their occurrences (peaks) significantly dominate all other peaks in Figure 2(a) and (c). On the other hand, the *A*-like nucleotides interact with the hydrophobic amino acids more frequently than with the polar ones (Figure 2(b) and (d)). For example, the four hydrophobic amino acids (Val, Leu, Ala, and Phe) account for as much as 53% of all contacts with the bases of the *A*-like nucleotides in the minor groove, compared to only 12% for the *B*-like nucleotides (Figure 2(a) and (b), cutoff 4.0 Å).

Notice that predominance of the hydrophobic amino acids in the contact profiles for the *A*-like nucleotides becomes especially strong when the cutoff distance increases from 3.5 Å to 4.0–4.5 Å. In particular, the numbers of base contacts with Val are higher than those with polar amino acids at these distances (Figure 2(b)), which is contrary to previous analyses of protein–DNA interfaces[11,17–19] where interactions in both grooves were considered together. Predominance of the long-distance interactions between DNA bases and hydrophobic amino acids is understandable because they contain many aliphatic carbon atoms with relatively large van der Waals radii (1.85 Å). This tendency is similar to the increase in the equilibrium distances between the atoms of hydrophobic amino acids observed for the intra-protein interactions,[36] and is a manifestation of the lower specificity of hydrophobic interactions.

The fraction of *A*-like nucleotides is relatively small, being just 12% of the protein-interacting nucleotides (Table 1). However, proteins bind to *A*-nucleotides more intensively than to the *B*-nucleotides. On average, there are 2.5 amino acid contacts to an interacting *A*-nucleotide compared to 1.6 contacts to a *B*-nucleotide, within 4.0 Å (data are calculated from Table 1 and the contact profiles). This increase in the number of contacts is mostly due to protein–base interactions.

Consider the ratio between the numbers of protein contacts with the bases and those with the sugars. This ratio is 0.4 for the *B*-like and 0.6 for the *A*-like nucleotides (calculated from Figure 2(a)–(d) at cutoff 4.0 Å). That is, the sugar switching from the *B*-like C2′-*endo* conformation to the

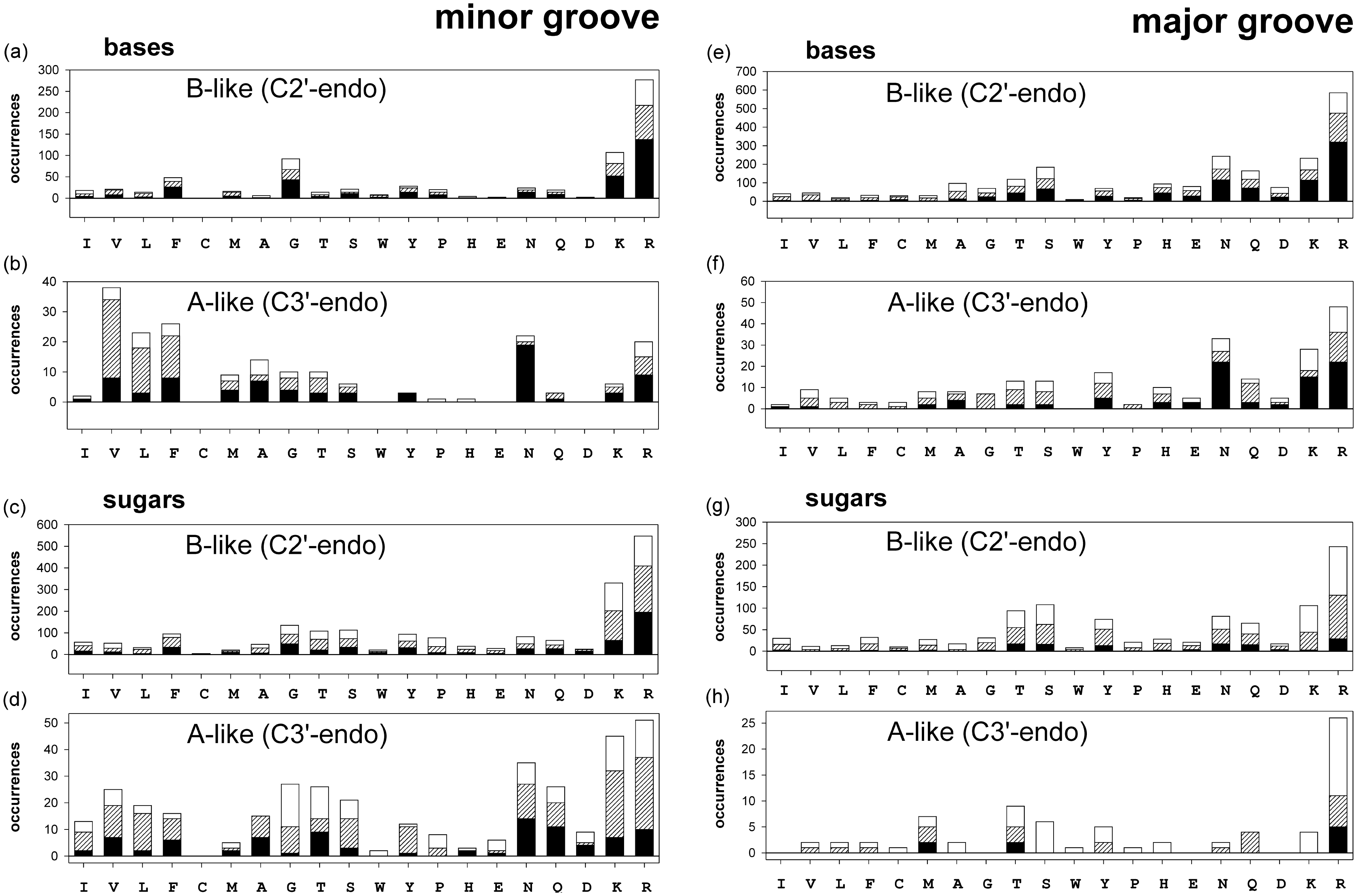
minor groove
major groove
(a)
bases
B-like (C2'-endo)
(b)
A-like (C3'-endo)
(c)
sugars
B-like (C2'-endo)
(d)
A-like (C3'-endo)
(e)
bases
B-like (C2'-endo)
(f)
A-like (C3'-endo)
(g)
sugars
B-like (C2'-endo)
(h)
A-like (C3'-endo)
occurrences
I V L F C M A G T S W Y P H E N Q D K R
0 50 100 150 200 250 300
0 10 20 30 40
0 100 200 300 400 500 600
0 10 20 30 40 50
0 100 200 300 400 500 600 700
0 10 20 30 40 50 60
0 50 100 150 200 250 300
0 5 10 15 20 25

Figure 2.

Table 1. Fractions of *A*-like nucleotides

	Minor groove interacting nucleotides in protein–DNA complexes P^+		Nucleotides without minor groove interactions, P^-		*B*-DNA crystals	
ATGC	12.4	(1106)	8.5	(1396)	7.4	(880)
R (A, G)	9.6	(603)	5.6	(773)	2.7	(440)
Y (T, C)	15.7	(503)	12.0	(623)	12.0	(440)

Fractions of the *A*-like nucleotides ($-30° < P_{sug} < 90°$) are given in percentage of the total numbers (given in parentheses). The data presented were summarized for all the four nucleotides (ATGC), purines (R), and pyrimidines (Y). A cutoff of 4.0 Å was used to determine minor groove interacting nucleotides.

A-like C3′-*endo* results in a 50% increase in the relative "accessibility" of the DNA bases for proteins approaching the duplex in the minor groove. This is an important result, since the protein–base contacts are essential for retrieving the sequence information encoded in the pattern of donor and acceptor groups along the base edges.

Contrary to observations in the minor groove, the contact profiles in the major groove practically do not depend on the sugar puckering: for both *B*-like and *A*-like conformations, the polar interactions are predominant (Figure 2(e)–(h)). Another distinction is that there are more contacts with bases than with sugars in the major groove. For example, the ratios of the numbers of contacts with bases and with sugars in the major groove are 2.8 for *B*-like and 5.4 for *A*-like nucleotides (the ratios calculated from Figure 2(e)–(h) at a cutoff of 4.0 Å; compare with the corresponding minor groove values of 0.4 and 0.6 given above). Thus, the sugars are poorly accessible for the interactions with proteins in the major groove, especially when switched into *A*-like puckers. Overall, our results for the major groove are consistent with the earlier studies where protein–DNA contacts were analyzed for both grooves together.[11,17–19]

To test the validity of the above results, we restricted the set of protein–DNA complexes to the recently refined 44 crystal structures having the highest resolution, i.e. 2.0 Å or better. This selection guarantees a more certain assignment of the sugar conformations. Importantly, the contact profiles for the reduced set of complexes (data not shown) are nearly the same as for the initial dataset. This indicates that our findings are not biased due to the poor resolution in some complexes, but rather reflect the details of the protein–DNA recognition.

Summarizing, there is a marked difference between the contact profiles for the *B* and *A*-like nucleotides in the minor groove. The latter ones demonstrate the richer amino acid "repertoire" involved in protein–DNA recognition. Proteins form denser networks of interactions with *A*-like nucleotides, and a larger fraction of hydrophobic contacts. This is in contrast to the mostly hydrophilic protein–DNA interactions in the major groove, which are not sensitive to the sugar puckers. Notice that, although the *A* and *B*-like nucleotides differ in their sugar puckering, it is at the minor groove base edges where the strongest differences in the contact profiles are observed (Figure 2(a) and (b)). That is, the sugar switching leads to

Figure 2. Cumulative numbers of amino acid–nucleotide contacts in 156 protein–DNA complexes (contact profiles), given separately for the *B*- and *A*-like nucleotides. (a)–(d) Interactions in the minor groove. (a) and (b) Contacts with bases involve atoms C2, N3, C4(Ade/Gua), N2(Gua) or O2, C2(Thy/Cyt). (c) and (d) Contacts with sugars were calculated as described in Methods. (e)–(h) Interactions in the major groove. (e) and (f) Contacts with bases involve atoms C5, C6, N7, C8(Ade/Gua), N6(Ade)/O6(Gua) or O4(Thy)/N4(Cyt), C4, C5, C6(Thy/Cyt), C5M(Thy). (g) and (h) Contacts with sugars. The contacts were counted within three cutoff distances: 3.5 Å (black areas in the bars), 4.0 Å (hatched areas), and 4.5 Å (white areas). Amino acids are denoted by one-letter codes and arranged according to the hydrophobicity scale of Kyte & Doolittle.[64] Results are based on the analysis of the following protein–DNA complexes in the NDB (134 crystal structures of 2.6 Å resolution or better and 22 structures solved by NMR): pd0002, pd0006, pd0007, pd0008, pd0012, pd0016, pd0020, pd0024, pd0028, pd0029, pd0034, pd0042, pd0045, pd0047, pd0050, pd0051, pd0056, pd0062, pd0068, pd0070, pd0073, pd0075, pd0076, pd0085, pd0086, pd0089, pd0091, pd0099, pd0101, pd0107, pd0108, pd0110, pd0111, pd0114, pd0115, pd0116, pd0117, pd0118, pd0121, pd0122, pd0125, pd0141, pd0142, pd0151, pd0153, pd0154, pd0165, pd0167, pd0169, pd0173, pd0177, pd0180, pd0187, pd0188, pd0189, pd0191, pd0192, pd0194, pd0200, pd0202, pd0207, pd0208, pd0210, pd0211, pd0212, pd0219, pd0220, pd0221, pd0225, pd0227, pd0231, pd0234, pd0241, pd0251, pd0253, pd0259, pd0264, pd0272, pd0287, pd0289, pd0293, pd0298, pd0311, pd0314, pd0334, pd0335, pd0341, pd0350, pd0371, pd0386, pde005, pde009, pde0124, pde0128, pde0131, pde0135, pde0145, pde025, pdr001, pdr009, pdr010, pdr011, pdr012, pdr018, pdr022, pdr032, pdr034, pdr036, pdr047, pdr051, pdr056, pdrc03, pdt008, pdt013, pdt015, pdt028, pdt029, pdt030, pdt031, pdt033, pdt034, pdt035, pdt036, pdt038, pdt039, pdt040, pdt044, pdt045, pdt048, pdt049, pdt062, pdt064, pdtb41, pdv001, 1a66, 1b69, 1bbx, 1c7u, 1e7j, 1f4s, 1g4d, 1gcc, 1hry, 1ig4, 1iv6, 1j5n, 1kqq, 2lef, 1lfu, 1lo1, 1mse, 1nk2, 1tf3, 1yui, 2ezd, 2gat. See the following URL for complete literature citations: http://home.ccr.cancer.gov/lecb/tolstorukov/prot-dna/

noticeable changes in the relative preferences between the bases and amino acids.

We see that the *A*-like nucleotides are the landmarks of the protein–DNA interactions in the minor groove. Now the questions arise: how do proteins discriminate between the two types of nucleotides, *A* and *B*-like? Is it possible, for example, that different atom exposures of *A* and *B*-like structures give rise to the differences in the affinities of amino acids to *A* and *B*-like nucleotides?

Sugar switching and accessibility in the minor groove

To quantify the difference between the exposures of the *A* and *B*-like nucleotides, we calculated their ASA in the minor groove (see Methods). Direct analysis of the DNA structural rearrangements upon protein binding (on a case-by-case basis) is statistically limited, since few oligonucleotides have ever been crystallized both with and without protein bound. Nevertheless, important information can be retrieved from an overall comparison of the two ensembles of structures: the protein-bound nucleotides and the free ones. Indeed, averaging over a large ensemble allows general trends to manifest themselves, while obscuring the details of specific interactions in individual complexes. Therefore, two sets of average ASA values were calculated: for the nucleotides interacting with proteins in the minor groove, P^+, and for those without such interactions, P^- (Figure 3). The ASA of the bases and sugars were calculated separately.

As follows from the right-hand panels in Figure 3, accessibilities of the *B*-like structures are not changed significantly by protein binding in the minor groove. On the other hand, in the comparisons of the *A*-like nucleotides, we find that the overall protein-induced increase in ASA is 11.2 Å^2, and this difference is statistically significant, both for the bases and for the sugars ($p < 0.01$, *t*-test; left-hand panels in Figure 3).

As a result of these apparent protein-induced deformations, the *A*-nucleotides (P^+) provide substantially larger surfaces for interaction with proteins than the *B*-nucleotides (compare filled columns in Figure 3). The increase in the average ASA value is 4.3 Å^2 for the bases, and 7.9 Å^2 for the sugars. Hence, the net effect in the minor groove is hydrophobic, because, although the surface of bases accessible in the minor groove is mostly polar, the highly exposed C3′-*endo*-like sugars are predominantly hydrophobic (see Figure 1 and Methods for details). This conclusion was independently verified when the ASA values were calculated separately for the polar and hydrophobic fractions of ASA (data not shown). Notice that the above comparison between the accessibilities of the bases and sugars is valid for "arbitrary" sequences, as the ASA values were normalized so as to correspond to an equal content of each nucleotide, 25%.

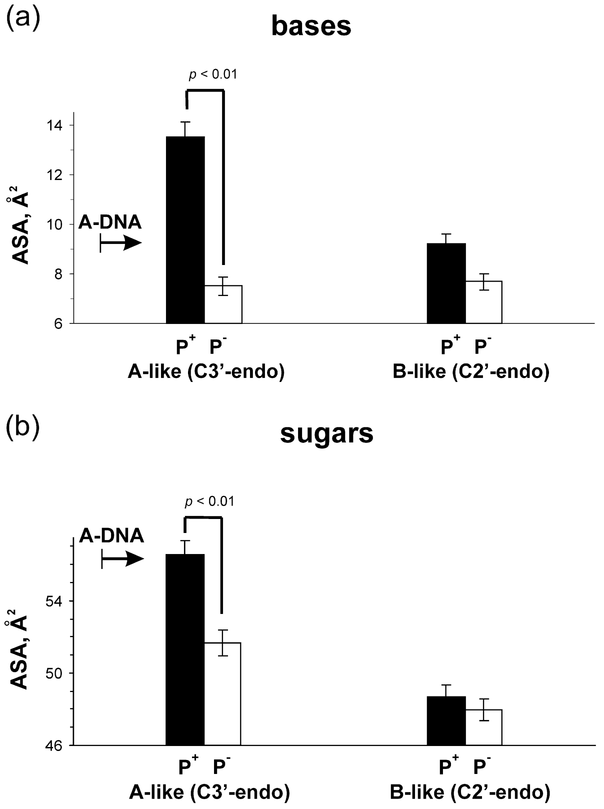

Figure 3. Average accessible surface areas, ASA, per nucleotide (columns) and the corresponding standard errors (bars) in the DNA minor groove for the *A* and *B*-like nucleotides. (a) ASA of bases and (b) ASA of sugars in the minor groove. Filled columns correspond to the ASA of nucleotides directly interacting with proteins in the minor groove, P^+; open columns to those without such interactions, P^- (interaction cutoff 4.0 Å). The corresponding ASA values for the free *A*-DNA in crystals are shown by arrows on the left. Average ASA per nucleotide and standard errors were normalized to have equal content of each nucleotide, 25%. The probe radius was 1.4 Å. Statistically significant differences ($p < 0.01$, *t*-test) are indicated.

The higher ASA values for the *A*-like P^+ nucleotides are consistent with the protein–DNA contact profiles (Figure 2). Indeed, as the protein contacts with the *A*-like nucleotides are more numerous and dense than with the *B*-nucleotides, there is a need for the larger DNA surface areas to accommodate the interacting amino acids, especially in the case of hydrophobic interactions with sugars. This comparison illustrates a clear-cut correlation between the preferential binding of the hydrophobic amino acids to the *A*-nucleotides (Figure 2), and the increased exposure of the non-polar sugars in the *A*-like structures (Figures 1 and 3).

The observed increase in the DNA accessibility cannot be explained entirely by the *B*-to-*A*-like transition, however. For example, comparing the *A* and *B*-like protein-free nucleotides, P^-, we see that the ASA values for the two conformations are much closer to each other than described above. The base accessibilities are practically identical, and there is only a modest difference of 3.7 Å^2 in the ASA of sugars (open columns in Figure 3). Thus, the sugar repuckering *per se* does not account for the increase in the ASA value detected for the protein-bound *A*-nucleotides.

Furthermore, in the bound *A*-like (P$^+$) nucleotides, the bases are more accessible (on average) than in other structures, including the canonical *A*-form (arrow in Figure 3(a)). Taken together, these observations indicate that in the complexes with proteins, DNA undergoes some striking deformations going beyond the conventional *A*-form.

Sugar repuckering associated with DNA kinks

One of the characteristic features of *A*-DNA distinguishing it from *B*-DNA is the base-pair rolling into the major groove (positive Roll), which increases accessibility to the bases in the minor groove. Indeed, we find that the steps with unusually large Roll (kinks) can be substantially responsible for the observed effect (i.e. extreme accessibility of bases in the *A*-like P$^+$ nucleotides). When the steps with Roll > 20° were eliminated from consideration, the bases' ASA of the P$^+$ nucleotides dropped from 13.5 Å^2 to 10.2 Å^2, which is close to the value for free *A*-DNA, 9.1 Å^2 (arrow in Figure 3(a)).

One particular way of DNA kinking into the major groove was repeatedly observed in the complexes. It is characterized by a specific pattern of the sugar ring conformations at the kinked step. A typical example is the complex of purine repressor (PurR) with DNA,[37] where two leucine side-chains intercalate between the bases in the minor groove and produce ~50° kink at the CG step (Figure 4(a)). Here, the 5′-cytosine sugars have *A*-like puckers, whereas the 3′-guanine sugars regain a *B*-like conformation. This 5′-*AB*-3′ sugar pattern is double-stranded; that is, the 5′-sugars are switched into the C3′-*endo*-like conformation in both strands.

The *B*-to-*A*-like switching of two cytosine sugars at the CG step is equivalent to pulling their C2′ atoms into the minor groove (to the left in Figure

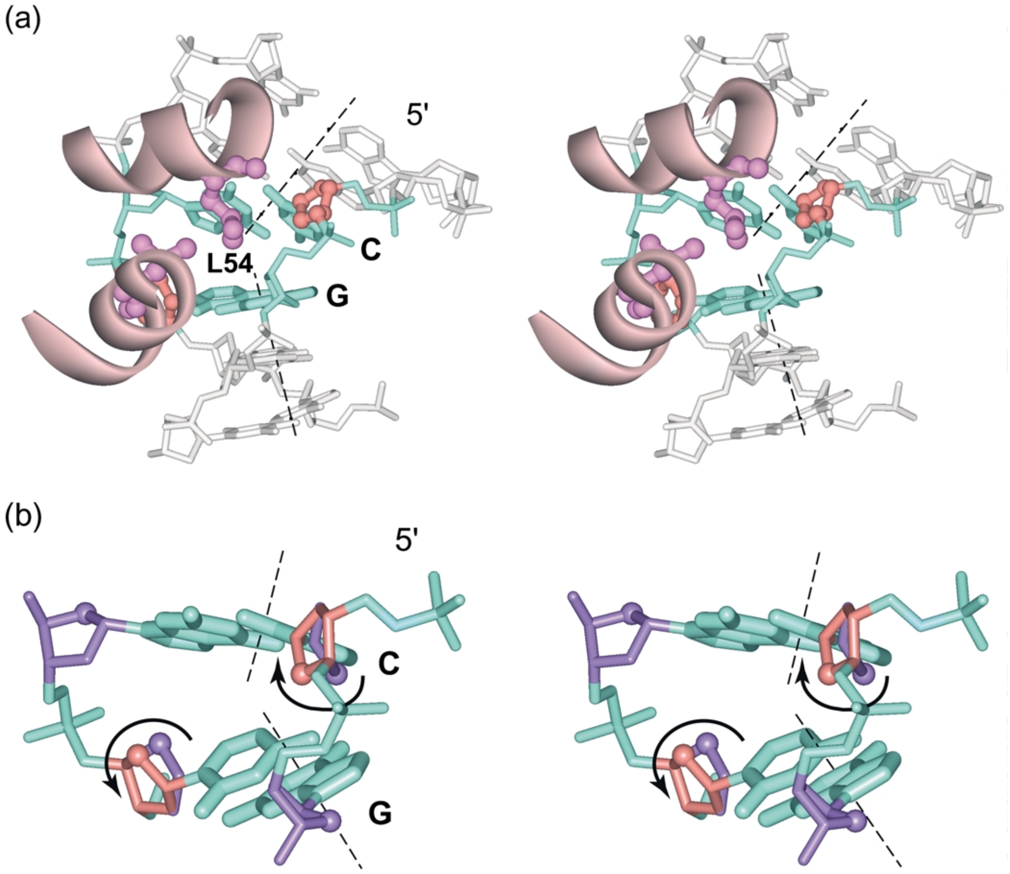

Figure 4. DNA kinking with sugar switches at the pyrimidine–purine dimeric step, as exemplified by the PurR–DNA complex.[37] (a) The side-chains of Leu54 (magenta) penetrate into the DNA minor groove and favorably interact with the *A*-like sugars (adobe). (b) Stereochemical mechanism of the kink with the 5′-*AB*-3′ sugar pattern. The *A*-like 5′-sugars are shown in adobe, the *B*-like 3′-sugars are lilac. To emphasize the sugar displacements, additional *B*-like sugars are aligned with the *A*-sugars by their atoms C3′, C4′, C5′, and O4′. The curved arrows indicate the clockwise and the counter-clockwise rotations of the top and bottom 5′-terminal strands, respectively (see the text). The DNA kinking is represented with broken lines going through the base-pair centers (DNA helical axes).

4(b)). Such a pulling would cause a clockwise rotation of the top 5′-terminal strand and a counter-clockwise rotation of the bottom 5′-terminal strand. At the local level, this will produce large positive Roll and Buckle angles, and at the global level, overall DNA kinking into the major groove. The described mechanism is similar to that presented in the pioneering work of Dickerson & Drew[23] to explain the connection between the sugar puckers/glycosyl angles and rolling of base-pair planes in the *B*-DNA dodecamer. Ours, however, incorporates additionally base-pair buckling at the kinked step in the direction permitting the intercalation of amino acids. Besides, our mechanism accounts for the global bending of the DNA axis,[38] contrasted to the compensatory base-pair rolling previously proposed.[23]

This 5′-*AB*-3′ pattern is found at the severely kinked steps (Roll > 40°) in DNA bound to many other proteins: TBP,[5] HMG1 domain A,[8] Sso7d,[39] Sac7d,[40] *Eco* RV,[41] CAP/CRP.[42] In most cases, DNA kinking is coupled to the intercalation of hydrophobic amino acids, such as Met and Val in the Sso7d[39] and Sac7d[40] complexes, or Phe in the case of TBP.[5,6] The 5′-*AB*-3′ sugar switching and, hence, the kinked DNA configuration are additionally stabilized by direct contacts between the hydrophobic amino acids and sugars. For example, in the PurR–DNA complex, side-chains of two leucine residues intercalating between the bases favorably interact with the cytosine sugars in the *A*-like puckers (Figure 4(a)). Similarly, in the complexes with TBP,[5] Sso7d[39] and Sac7d[40] the Phe rings and the Ala methyl groups interact with the DNA sugars, stabilizing them in the *A*-domain. On the other hand, DNA kinks can occur without amino acid intercalation and interaction with the sugar rings: in the CAP/CRP–DNA[42] and *Eco* RV–DNA[41] complexes the kinks are stabilized solely by interactions in the major groove. Nevertheless, the 5′-*AB*-3′ sugar switch apparently works in those cases as well, facilitating the DNA kink. The detailed analysis of this phenomenon will be given elsewhere.

Overall, the 5′-*AB*-3′ sugar pattern occurs only in 2% of all dimeric steps in our database (~2000 steps). This fraction dramatically increases for the kinked steps: up to 30% for steps with Roll > 20° and to 64% for steps with Roll > 40°. Thus, the larger the Roll angle (bending into the major groove), the higher the probability of observing the 5′-*AB*-3′ pattern at the kinked step. Moreover, there are no occurrences of steps having a 5′-*AB*-3′ sugar pattern and a negative Roll. The 5′-*AB*-3′ sugar pattern has also been observed in molecular dynamics simulations of DNA kinking.[43] This further confirms our assertion that the 5′-*AB*-3′ sugar pattern is favorable for DNA bending into the major groove.

In seven out of 13 steps with Roll > 20°, the 5′-*AB*-3′ kinking occurs at the pyrimidine–purine (YR) dimer, or in 54% of all the cases (67% for the steps with Roll > 40°), compared to the 25% expected for random (four YR steps out of 16 possible ones). It is easy to explain the observed sequence preference using the data from Table 1. Indeed, 16% of the pyrimidines interacting with proteins in the minor groove are *A*-like, compared to 10% of purines (P^+ nucleotides). In the *B*-DNA crystals this preference is even stronger: the fraction of *A*-like pyrimidines, 12%, is higher than the fraction of *A*-like purines by more than fourfold. In turn, cytosine is more frequently found with the *A*-like sugar puckers than thymine. This sequence dependence cannot be explained solely by crystal packing, since it agrees both with energy calculations for free DNA[44,45] and experimental NMR observations in solution.[34,35]

The higher propensity of pyrimidines to adopt the *A*-like conformation makes a YR step the optimal choice for 5′-*AB*-3′ sugar pattern, because those are pyrimidines that are switching to *A*-like puckers in this case. The YR steps have long been known to reveal anisotropic flexibility; they are most easily bent into the major groove.[2,18,46–49] One of the well-known reasons for this anisotropy is the purine–purine clash arising in the minor groove of a YR step if the two base-pairs are parallel with one another.[50] The sequence preference for the 5′-*AB*-3′ sugar pattern provides yet another stereochemical origin for the anisotropic flexibility of the YR steps.

Implications for DNA sequence recognition in the minor groove

The results presented above allow us to address the broader question of how interactions with *A*-like nucleotides contribute to sequence specificity of protein–DNA recognition. To this end, we compared the distribution of the *A*-like nucleotides along the DNA chains in the free state and in complexes. In free *B*-DNA in crystal, the fraction of *A*-nucleotides is 7.4% (Table 1), and they are mostly isolated, while in complexes the corresponding fraction increases up to 12.4%, and these nucleotides occur in clusters†. In other words, proteins interact predominantly with short *A*-nucleotide clusters, rather than with isolated single *A*-nucleotides. Clearly, interaction of two continuous hydrophobic surfaces is favorable because substantial areas of both DNA and protein are buried, thereby diminishing the effect of “solvent aversion” of the sugar rings and hydrophobic amino acids. Such a cooperativity of the hydrophobic interactions often results in changing the minor groove geometry at several base-pairs scale and, thus, facilitates the global fitting of the protein and DNA, as illustrated by the TBP–DNA complex.[5,6]

Another example of the *A*-like nucleotide clustering is the switch of four consecutive nucleotides in one of the strands in the LEF-1–DNA complex.[7] The four sugars form a hydrophobic surface interacting with the hydrophobic protein “patch” (Figure 5). Note that all the four *A*-like nucleotides in Figure 5 are pyrimidines, that is, the *B*-to-*A*

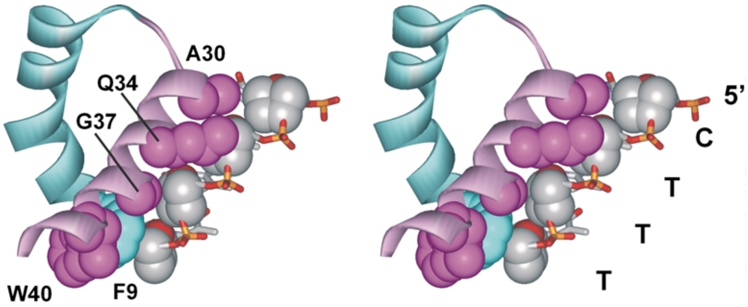

Figure 5. Formation of the two complementary hydrophobic surfaces in the LEF-1–DNA complex.[7] Four consecutive nucleotides are switched to the *A*-like conformation. (DNA atoms are colored according to their atom types; sugars shown in CPK representation. For clarity, only the carbon atoms of selected amino acids are shown in CPK representation.) Such a configuration of the minor groove facilitates accommodation of the two α-helices (cyan, helix-1 and magenta, helix-2). Phe9 and Trp40 are components of the hydrophobic core stabilizing the protein fold. Notice that Ala30, Gln34, Gly37 and Phe9 directly contact the *A*-like sugars.

sugar switching is preferable in this case (Table 1). This mechanism of recognition is utilized by other proteins as well. In the hSRY–DNA complex,[9,10] extensive interactions between the hydrophobic amino acids and the DNA minor groove are accompanied by sugar switching at two adjacent thymine bases.

In this context, it is feasible that *A*-like pyrimidines, comprising ~80% of all the *A*-like nucleotides in the crystalline *B*-DNA (Table 1), serve as nuclei for the clusters of *A*-like nucleotides observed in the complexes. This hypothesis is substantiated by the following considerations. A fraction of *A*-like sugars is always present in the free DNA in solution.[33–35] The spontaneous switch of the isolated deoxyribose to the *A*-like pucker exposes the sugar carbon atoms in the minor groove and widens the groove.[51] During complexation with protein, water molecules would be expelled from the DNA surface, partially dehydrating the duplex. Such dehydration would further promote the *B*-to-*A* sugar switching, thus inducing the protein–DNA fit by altering the interaction surface in the minor groove. (It is well known that the water activity is the key factor directing the equilibrium between the *B* and *A*-conformations in fibers and in solution:[52–54] dehydration of the duplex promotes the *B*-to-*A* transition.) Therefore, sugar repuckering in isolated pyrimidines in free DNA, and subsequent clustering of the *A*-like nucleotides in the complexes would greatly facilitate the DNA sequence recognition by proteins.

Concluding Remarks

The relationship between sugar switching and the global DNA conformation is well known.[22] Nevertheless, the possibility that the sugar *B*-to-*A* repuckering is critical for the minor groove recognition has not previously been tested, partly because the conventional approaches do not distinguish interactions in the minor groove from those in the major groove (especially for the sugar-phosphate backbone). To this end, we have developed a novel algorithm to analyze the interactions and to calculate the solvent ASA in each groove separately.

Although the fraction of *A*-like sugars is relatively modest, 10–15% of all nucleotides, the effect of the sugar switching is essential; it reveals a clear indirect readout mechanism for the minor groove, where hydrogen bonds *per se* are apparently insufficient for rigorous selection of the DNA sequence.[1] Available high-resolution structures of the protein–DNA complexes suggest that this mechanism is widely used in eukaryotes, where the minor groove is particularly exposed due to packaging of DNA in nucleosomes.[4]

Our results provide new insights into the role of *A*-like nucleotides in sequence-specific protein–DNA recognition. Indeed, we have shown that: (i) protein contacts with *A*-like nucleotides are predominantly with hydrophobic amino acids (Figure 2); (ii) accordingly, *A*-like nucleotides provide larger hydrophobic surfaces for interactions than *B*-like nucleotides (Figure 3); (iii) *A*-like nucleotides are operative in the protein-induced kinks (Figure 4); and (iv) clusters of *A*-like nucleotides cooperatively interact with non-polar surfaces in proteins, increasing favorability of hydrophobic contacts between proteins and DNA (Figure 5).

The findings presented here do imply that this class of minor groove–protein interactions is important. There may be some similarity between clustering of *A*-like nucleotides during protein–DNA complexation and formation of the hydrophobic core in proteins. In the latter case, hydrophobic collapse promotes protein folding.[55] By analogy, we suggest that in the case of protein–DNA recognition, specific clustering of *A*-like nucleotides reinforced by hydrophobic interactions in the minor groove, may aid in achieving an induced fit between protein and DNA.

Methods

The set of structures used here includes the protein–DNA complexes solved by X-ray and NMR analyses with coordinates available in the NDB Biomolecular

Resource.[56] All complexes in our set contain double-stranded DNA with more than three nucleotides in each strand. The dataset obtained under such restrictions included about 450 entries and was redundant due to the presence of complexes containing highly similar protein sequences, or different DNA sequences bound to the same protein, etc.

To select a non-redundant set, only one structure was taken from any subset of the several complexes containing the same protein or its mutants. The preference was given first to a complex with wild-type protein, and then to the one having the highest resolution. The final set contains 134 protein–DNA co-crystals with resolution 2.6 Å or better and 22 structures solved by NMR (see Figure 2 legend and URL†). Sets of structures of free *A*-DNA (64 entries) and *B*-DNA (54 entries) include those with resolution 2.5 Å or better available at the NDB (see URL†). In all these structures, the terminal pairs of nucleotides were excluded to avoid end effects. The DNA structural parameters, such as the Roll angle, were calculated with CompDNA.[57]

The sugar pucker was considered to be in the S-domain (*B*-like sugars) if its pseudorotation angle P_{sug} was between 90° and 210°. The N-domain (*A*-like sugars) was taken to span the interval between −30° and 90°. The nucleotides with stereochemically unfavorable P_{sug} values outside the S or N-domains were not included (~1% of the total).

Another frequent source of the DNA backbone flexibility is the switch of the torsional angle γ[O5′–C5′–C4′–C3′] from *gauche*(+) to *trans* or *gauche*(−) conformations. It influences accessibility of sugar in the minor groove, therefore, for the sake of clarity, only *gauche*(+) structures with $\gamma = 60(\pm 30)°$ are included in the present study. This reduces the number of nucleotides in the dataset by ~15%.

To characterize the protein–DNA interfaces, amino acid–nucleotide interactions were counted (contact profiles). The contact profiles were calculated separately for each of the two DNA atom groups: (i) the atoms of bases and (ii) the sugars (atoms C1′–C5′ and O4′). Contacts with the phosphate oxygen atoms were not counted because their accessibilities for protein amino acids are nearly independent of the DNA conformation. Thus, the subclass of the nucleotides interacting in the minor groove, P^+, includes those nucleotides that contact proteins with their bases or sugars.

To distinguish interactions in the minor groove from those in the major groove, special criteria were applied for the interactions with sugars. The protein atom under consideration had to be located on the minor groove side of the "surface" formed by four atoms from two adjacent nucleotides: O4′(*i*), P(*i*), P(*i* + 1), O4′(*i* + 1) (Figure 6). In general, these four atoms do not belong to the same plane, therefore, the following procedure was used. Each three atoms out of these four define a plane and a corresponding normal vector (black arrows in Figure 6; for example, the normal vector placed at P(*i*) is perpendicular to the plane defined by the atoms O4′(*i*), P(*i*), P(*i* + 1)). The sum of these four vectors gives the "overall normal vector" **n** (red arrow in Figure 6). Finally, the plane perpendicular to this vector and going through the geometric center of the four atoms was considered to be a "surface" separating the two grooves. If the projection of the protein atom (green sphere in Figure 6) on the vector **n** was positive, then the atom was ascribed to the minor groove; otherwise, it was assumed to be in the major groove.

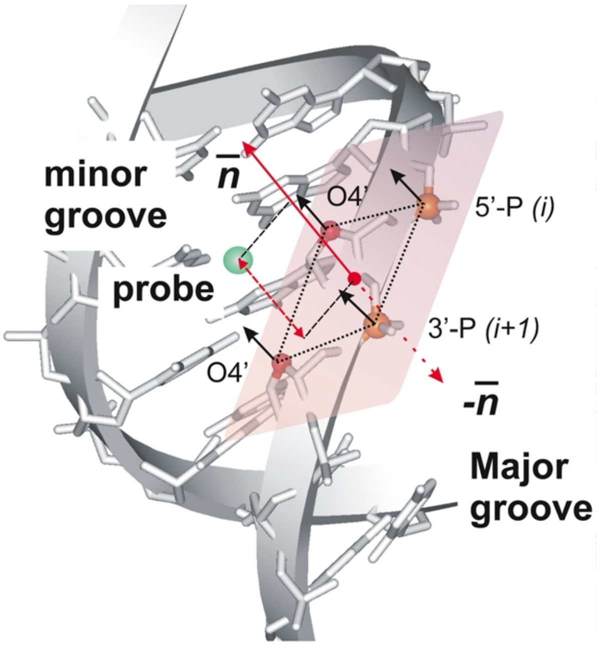

Figure 6. Scheme for distinguishing between minor groove and major groove protein–DNA contacts. The probe (green sphere) represents a protein atom interacting with DNA. The four normal vectors to the planes defined by each three out of four atoms O4′(*i*), P(*i*), P(*i* + 1), and O4′(*i* + 1) are shown in black. Their vector sum (red arrow) and the geometric center of the four atoms (red dot) determine the plane (shaded) used to distinguish between the minor and major groove locations of the probe. See Methods for details.

The solvent ASA were calculated according to the algorithm described by Higo & Go.[58] The main idea of the algorithm is filling the volume with small cubes and deciding for each cube whether it is located inside, outside or lies at the surface of the macromolecule. For the whole DNA duplex, this algorithm gives results similar to those produced by the frequently used approach of Lee & Richards[59] and Connolly.[60] We selected the "cube-based" algorithm because it enables one to separate easily the ASA values for each DNA atom or atom group into the major and minor groove fractions. This separation was made according to the same procedure as for the atom–atom contacts (Figure 6). The probe radius was 1.4 Å. The following atom radii were chosen for calculations: 1.85 Å for carbon atoms, 1.5 Å for nitrogen atoms, 1.4 Å for oxygen atoms, and 2.0 Å for phosphorus atoms (hydrogen atoms were omitted).

Acknowledgements

The authors are grateful to Sankar Adhya, Ad Bax, Marius Clore, Valery Ivanov, Wilma Olson, Ramaswamy Sarma and Don Rau for insightful discussions.

References

1. Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA*, **73**, 804–808.
2. Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M. & Zhurkin, V. B. (1998). DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
3. Murphy, F. V. & Churchill, M. E. (2000). Non-sequence-specific DNA recognition: a structural perspective. *Struct. Fold. Des.* **8**, R83–R89.
4. Beato, M. & Eisfeld, K. (1997). Transcription factor access to chromatin. *Nucl. Acids Res.* **25**, 3559–3563.
5. Kim, J. L., Nikolov, D. B. & Burley, S. K. (1993). Co-crystal structure of TBP recognizing the minor groove of TATA element. *Nature*, **365**, 520–527.
6. Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. (1993). Crystal structure of a yeast TBP/TATA–box complex. *Nature*, **365**, 512–520.
7. Love, J. J., Li, X., Case, D. A., Giese, K., Grosschedl, R. & Wright, P. E. (1995). Structural basis for DNA bending by the architectural transcription factor LEF-1. *Nature*, **376**, 791–795.
8. Ohndorf, U.-M., Rould, M. A., He, Q., Pabo, C. O. & Lippard, S. J. (1999). Basis for recognition of cisplatin-modified DNA by high-mobility-group proteins. *Nature*, **399**, 708–712.
9. Werner, M. H., Huth, J. R., Gronenborn, A. M. & Clore, G. M. (1995). Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY–DNA complex. *Cell*, **81**, 705–714.
10. Murphy, E. C., Zhurkin, V. B., Louis, J. M., Cornilescu, G. & Clore, G. M. (2001). Structural basis for SRY-dependent 46-X,Y sex reversal: modulation of DNA bending by a naturally occurring point mutation. *J. Mol. Biol.* **312**, 481–499.
11. Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (2001). Amino acid–base interactions: a three dimensional analysis of protein–DNA interactions at an atomic level. *Nucl. Acids Res.* **29**, 2860–2874.
12. Drew, H. R. & Travers, A. A. (1985). Structural junctions in DNA: the influence of flanking sequence on nuclease digestion specificities. *Nucl. Acids Res.* **13**, 4445–4467.
13. Matthews, B. W. (1988). Protein–DNA interaction. No code for recognition. *Nature*, **335**, 294–295.
14. Pabo, C. O. & Nekludova, L. (2000). Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.* **301**, 597–624.
15. Woda, J., Schneider, B., Patel, K., Mistry, K. & Berman, H. M. (1998). An analysis of the relationship between hydration and protein–DNA interactions. *Biophys. J.* **75**, 2170–2177.
16. Eisenstein, M. & Shakked, Z. (1995). Hydration patterns and intermolecular interactions in A-DNA crystal structures. Implications for DNA recognition. *J. Mol. Biol.* **248**, 662–678.
17. Nadassy, K., Wodak, S. J. & Janin, J. (1999). Structural features of protein–nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
18. Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. (1999). Protein–DNA interactions: a structural analysis. *J. Mol. Biol.* **287**, 877–896.
19. Mandel-Gutfreund, Y., Schueler, O. & Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J. Mol. Biol.* **253**, 370–382.
20. Mandel-Gutfreund, Y., Margalit, H., Jernigan, R. L. & Zhurkin, V. B. (1998). A role for CH···O interactions in protein–DNA recognition. *J. Mol. Biol.* **277**, 1129–1140.
21. Lu, X. J., Shakked, Z. & Olson, W. K. (2000). *A*-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.* **300**, 819–840.
22. Rich, A. (2003). The double helix: a tale of two puckers. *Nature Struct. Biol.* **10**, 247–249.
23. Dickerson, R. E. & Drew, H. R. (1981). Kinematic model for *B*-DNA. *Proc. Natl Acad. Sci. USA*, **78**, 7318–7322.
24. Minchenkova, L. E., Schyolkina, A. K., Chernov, B. K. & Ivanov, V. I. (1986). CC/GG contacts facilitate the *B* to *A* transition of DNA in solution. *J. Biomol. Struct. Dynam.* **4**, 463–476.
25. Tolstorukov, M. Y., Ivanov, V. I., Malenkov, G. G., Jernigan, R. L. & Zhurkin, V. B. (2001). Sequence-dependent *B* ↔ *A* transition in DNA evaluated with dimeric and trimeric scales. *Biophys. J.* **81**, 3409–3421.
26. Nekludova, L. & Pabo, C. O. (1994). Distinctive DNA conformation with enlarged major groove is found in Zn-finger–DNA and other protein–DNA complexes. *Proc. Natl Acad. Sci. USA*, **91**, 6948–6952.
27. Shakked, Z., Guzikevich-Guerstein, G., Frolow, F., Rabinovich, D., Joachimiak, A. & Sigler, P. B. (1994). Determinants of repressor/operator recognition from the structure of the trp operator binding site. *Nature*, **368**, 469–473.
28. Guzikevich-Guerstein, G. & Shakked, Z. (1996). A novel form of the DNA double helix imposed on the TATA-box by the TATA-binding protein. *Nature Struct. Biol.* **3**, 32–37.
29. Travers, A. A. (1995). Reading the minor groove. *Nature Struct. Biol.* **2**, 615–618.
30. Ivanov, V. I., Minchenkova, L. E., Chernov, B. K., McPhie, P., Ryu, S., Garges, S. *et al.* (1995). CRP–DNA complexes: inducing the *A*-like form in the binding sites with an extended central spacer. *J. Mol. Biol.* **245**, 228–240.
31. Ivanov, V. I., Minchenkova, L. E., Burckhardt, G., Birch-Hirschfeld, E., Fritzsche, H. & Zimmer, C. (1996). The detection of *B*-form/*A*-form junction in a deoxyribonucleotide duplex. *Biophys. J.* **71**, 3344–3349.
32. El Hassan, M. A. & Calladine, C. R. (1997). Conformational characteristics of DNA: empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. *Philos. Trans. Roy. Soc. London*, **355**, 43–100.
33. Zhou, N., Manogaran, S., Zon, G. & James, T. L. (1988). Deoxyribose ring conformation of [d(GGTA TACC)]$_2$: an analysis of vicinal proton–proton coupling constants from two-dimensional proton nuclear magnetic resonance. *Biochemistry*, **27**, 6013–6020.
34. Ojha, R. P., Dhingra, M. M., Sarma, M. H., Shibata, M., Farrar, M., Turner, C. J. & Sarma, R. H. (1999). DNA bending and sequence-dependent backbone conformation NMR and computer experiments. *Eur. J. Biochem.* **265**, 35–53.
35. Wu, Z., Delaglio, F., Tjandra, N., Zhurkin, V. B. & Bax, A. (2003). Overall structure and sugar dynamics of a DNA dodecamer from homo- and heteronuclear dipolar couplings and ^{31}P chemical shift anisotropy. *J. Biomol. NMR*, **26**, 297–315.
36. Bahar, I. & Jernigan, R. L. (1997). Inter-residue

potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.* **266**, 195–214.

37. Glasfeld, A., Koehler, A. N., Schumacher, M. A. & Brennan, R. G. (1999). The role of lysine 55 in determining the specificity of the purine repressor for its operators through minor groove interactions. *J. Mol. Biol.* **291**, 347–361.
38. Ulyanov, N. B. & Zhurkin, V. B. (1982). Flexibility of complementary dinucleoside phosphates–a Monte Carlo study. *Mol. Biol. (Engl. transl.)*, **16**, 857–867.
39. Gao, Y. G., Su, S. Y., Robinson, H., Padmanabhan, S., Lim, L., McCrary, B. S. *et al.* (1998). The crystal structure of the hyperthermophile chromosomal protein Sso7d bound to DNA. *Nature Struct. Biol.* **5**, 782–786.
40. Robinson, H., Gao, Y. G., McCrary, B. S., Edmondson, S. P., Shriver, J. W. & Wang, A. H. (1998). The hyperthermophile chromosomal protein Sac7d sharply kinks DNA. *Nature*, **392**, 202–205.
41. Martin, A. M., Sam, M. D., Reich, N. O. & Perona, J. J. (1999). Structural and energetic origins of indirect readout in site-specific DNA cleavage by a restriction endonuclease. *Nature Struct. Biol.* **6**, 269–277.
42. Schultz, S. C., Shields, G. C. & Steitz, T. A. (1991). Crystal structure of a CAP–DNA complex: the DNA is bent by 90 degrees. *Science*, **253**, 1001–1007.
43. Bosch, D., Campillo, M. & Pardo, L. (2003). Binding of proteins to the minor groove of DNA: what are the structural and energetic determinants for kinking a basepair step? *J. Comput. Chem.* **24**, 682–691.
44. Gorin, A. A., Ulyanov, N. B. & Zhurkin, V. B. (1990). S–N transition of the sugar ring in *B*-form DNA. *Mol. Biol. (Engl. transl.)*, **24**, 1036–1047.
45. Foloppe, N. & MacKerell, A. D. (1999). Intrinsic conformational properties of deoxyribonucleosides: implicated role for cytosine in the equilibrium among the *A*, *B*, and *Z* forms of DNA. *Biophys. J.* **76**, 3206–3218.
46. Suzuki, M. & Yagi, N. (1995). Stereochemical basis of DNA bending by transcription factors. *Nucl. Acids Res.* **23**, 2083–2091.
47. Dickerson, R. E. (1998). DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucl. Acids Res.* **26**, 1906–1926.
48. Sarai, A., Mazur, J., Nussinov, R. & Jernigan, R. L. (1989). Sequence dependence of DNA conformational flexibility. *Biochemistry*, **28**, 7842–7849.
49. Ulyanov, N. B. & Zhurkin, V. B. (1984). Sequence-dependent anisotropic flexibility of *B*-DNA. A conformational study. *J. Biomol. Struct. Dynam.* **2**, 361–385.
50. Calladine, C. R. (1982). Mechanics of sequence-dependent stacking of bases in *B*-DNA. *J. Mol. Biol.* **161**, 343–352.
51. Kamath, S., Sarma, M. H., Zhurkin, V. B., Turner, C. J. & Sarma, R. H. (2000). DNA bending and sugar switching. *J. Biomol. Struct. Dynam.* **11**, 317–325.
52. Franklin, R. E. & Gosling, R. G. (1953). The structure of sodium thymonucleate fibers. I. The influence of water content. *Acta Crystallog.* **6**, 673–677.
53. Ivanov, V. I., Minchenkova, L. E., Minyat, E. E., Frank-Kamenetskii, M. D. & Schyolkina, A. K. (1974). The *B* to *A* transition of DNA in solution. *J. Mol. Biol.* **87**, 817–833.
54. Malenkov, G., Minchenkova, L., Minyat, E., Schyolkina, A. & Ivanov, V. (1975). The nature of the *B–A* transition of DNA in solution. *FEBS Letters*, **51**, 38–42.
55. Chan, H. S. & Dill, K. A. (1990). Origins of structure in globular-proteins. *Proc. Natl Acad. Sci. USA*, **87**, 6388–6392.
56. Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T. *et al.* (1992). The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.* **63**, 751–759.
57. Gorin, A. A., Zhurkin, V. B. & Olson, W. K. (1995). *B*-DNA twisting correlates with base pair morphology. *J. Mol. Biol.* **247**, 34–48.
58. Higo, J. & Go, N. (1989). Algorithm for rapid calculation of excluded volume of large molecules. *J. Comput. Chem.* **10**, 376–379.
59. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
60. Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science*, **221**, 709–713.
61. Bingman, C., Li, X., Zon, G. & Sundaralingam, M. (1992). Crystal and molecular structure of d(GTGCG CAC): investigation of the effects of base sequence on the conformation of octamer duplexes. *Biochemistry*, **31**, 12803–12812.
62. Wood, A. A., Nunn, C. M., Trent, J. O. & Neidle, S. (1997). Sequence-dependent crossed helix packing in the crystal structure of a *B*-DNA decamer yields a detailed model for the Holliday junction. *J. Mol. Biol.* **269**, 827–841.
63. Galburt, E. A., Chevalier, B., Tang, W., Jurica, M. S., Flick, K. E., Monnat, R. J., Jr & Stoddard, B. L. (1999). A novel endonuclease mechanism directly visualized for I-*Ppo*I. *Nature Struct. Biol.* **6**, 1096–1099.
64. Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.

Edited by J. Thornton

(Received 11 July 2003; received in revised form 6 January 2004; accepted 6 January 2004)