doi:10.1016/j.jmb.2003.10.071 *J. Mol. Biol.* (2004) **335**, 399–408

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

COMMUNICATION

Recognition of 5′-YpG-3′ Sequences by Coupled Stacking/Hydrogen Bonding Interactions with Amino Acid Residues

Jason S. Lamoureux, Jason T. Maynes and J. N. Mark Glover*

Department of Biochemistry
University of Alberta
Edmonton, Alta., Canada T6G 2H7

The combined biochemical and structural study of hundreds of protein–DNA complexes has indicated that sequence-specific interactions are mediated by two mechanisms termed direct and indirect readout. Direct readout involves direct interactions between the protein and base-specific atoms exposed in the major and minor grooves of DNA. For indirect readout, the protein recognizes DNA by sensing conformational variations in the structure dependent on nucleotide sequence, typically through interactions with the phosphodiester backbone. Based on our recent structure of Ndt80 bound to DNA in conjunction with a search of the existing PDB database, we propose a new method of sequence-specific recognition that utilizes both direct and indirect readout. In this mode, a single amino acid side-chain recognizes two consecutive base-pairs. The 3′-base is recognized by canonical direct readout, while the 5′-base is recognized through a variation of indirect readout, whereby the conformational flexibility of the particular dinucleotide step, namely a 5′-pyrimidine–purine-3′ step, facilitates its recognition by the amino acid *via* cation–π interactions. In most cases, this mode of DNA recognition helps explain the sequence specificity of the protein for its target DNA.

*Corresponding author

Keywords: Ndt80; DNA–protein interactions; YpG steps

Proteins bind and recognize specific target DNA sequences through complex networks of electrostatic and hydrophobic interactions. In an early, theoretical study, Seeman *et al*. suggested that sequence-specific interactions might be attained through interactions between protein side-chains and the faces of the DNA base-pairs exposed in the major groove.[1] The determination of the first, high-resolution structures of protein–DNA complexes in the late 1980s validated many of these ideas, but also indicated that proteins might also sense sequence-dependent variations in the fine structure of the DNA double helix, or in its flexibility, in a type of recognition termed "indirect readout".[2–4] Over the last 15 years, the structures of hundreds of DNA-binding proteins in complex with their cognate DNAs have been determined by X-ray crystallography and NMR, and, while these studies have provided a rationale for why specific proteins recognize their particular target sequence, it has been extremely difficult to extract fundamental principles from this database that could be used to predict the DNA-binding preferences of a given protein *a priori* (for recent reviews of the database of protein–DNA complexes, see Refs. 5,6).

Arginine residue recognition of 5′-YpG-3′ by Ndt80

We recently determined the structure of a complex of Ndt80, a central regulator of sporulation in *Saccharomyces cerevisiae*, bound to a DNA containing its consensus site, the middle sporulation element, or MSE.[7] The structure unexpectedly revealed that Ndt80 is a member of the Ig-fold family of transcription factors, and binds DNA in a manner similar to other members of the family,

Supplementary data associated with this article can be found at doi: 10.1016/j.jmb.2003.10.071

Abbreviation used: MSE, middle sporulation element.

E-mail address of the corresponding author: mark.glover@ualberta.ca

such as p53, NF-κB, AML-Runt, and STAT transcription factors. The Ndt80 structure, determined at 1.4 Å resolution, offered an excellent opportunity to understand in detail the way in which this protein selectively binds the MSE. The consensus MSE[8,9] is 5′-gNCR CAAAW-3′ (where N refers to any nucleotide; R, a purine residue; Y, a pyrimidine residue; W either an adenine base or a thymine base, and lower case letters refer to semi-conserved positions, and the nucleotide positions are labeled 1–9). The protein recognizes the 3′ poly(A)–poly(T) portion of the MSE through minor groove interactions, while the 5′ CG-rich end of the MSE is recognized by three arginine residue side-chains that make bidentate hydrogen bonds to the major groove faces of the C·G base-pairs.

While these interactions are quite similar to the ways in which other transcription factors recognize DNA, the structure revealed an unexpected mode of recognition of the conserved pyrimidine residues immediately 5′ to the guanine residues at positions 3 and 5. Ndt80 recognizes the inherent flexibility in these 5′-YpG-3′ dinucleotide steps through hydrogen bonding interactions of arginine residue side-chains to the major groove face of the guanine residue, coupled to the loss of stacking of the 5′-pyrimidine residue with the guanine residue, and concomitant stacking of the pyrimidine residue on the co-planar guanidinium group of the arginine residue (Figure 1). 5′-YpR-3′ steps are significantly more flexible than other dinucleotide steps, probably due to the low degree of base-pair overlap within the 5′-YpR-3′ step.[10] As a result, it is less energetically costly to deform these dinucleotide steps, and 5′-YpR-3′ steps are often sites of DNA bending.[10,11] Interestingly, the 5′-YpR-3′ steps presented here are not sites of significant bending. The deformation of the 5′-YpR-3′ steps in Ndt80 is accompanied by a shift in the backbone conformation of the 5′-pyrimidine residue and/or its complementary purine residue from the common BI conformation, which is characterized by ε and ζ torsion angles in the (t/g^-) range, where $\varepsilon - \zeta \approx -90°$, to the less common BII conformation, where $\varepsilon, \zeta = (g^-/t)$ and $\varepsilon - \zeta \approx +90°$. BII conformations

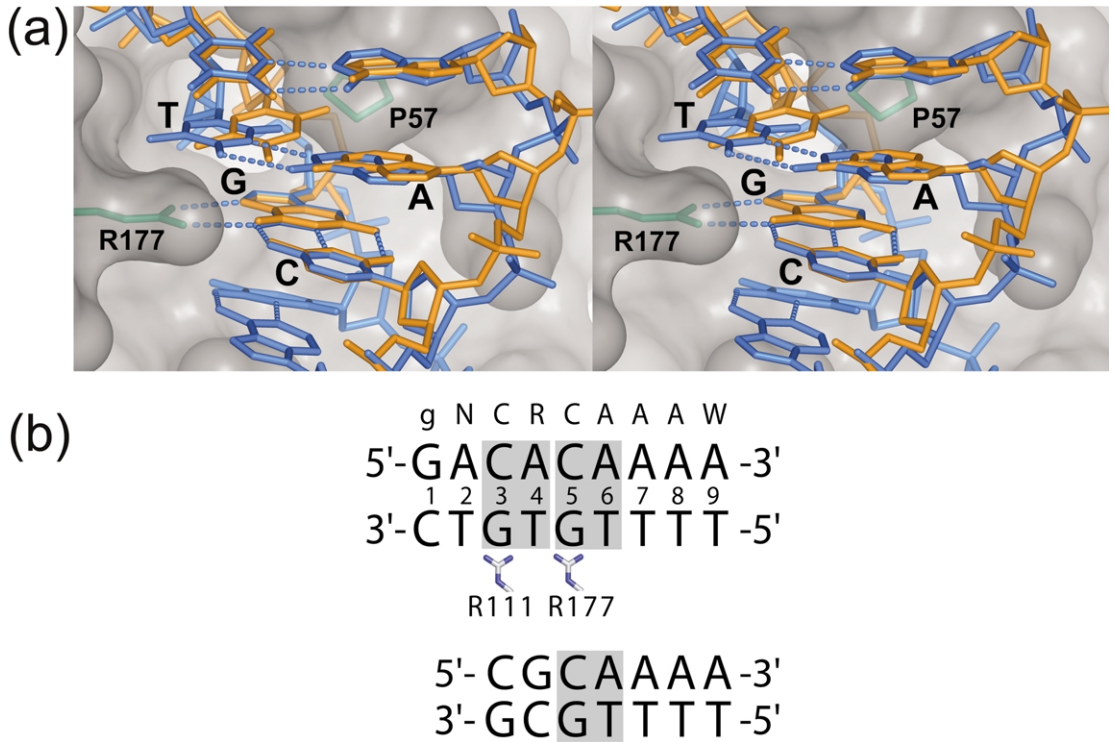

Figure 1. Alignment of Ndt80–DNA complex with reference DNA structure. (a) The Ndt80–DNA complex (blue) is aligned to the reference DNA structure (orange) using the 3′ G·C base-pair of the 5′-YpG-3′ step. The surface of Ndt80 is shown in transparent grey with a green stick representation for Arg177 and Pro57 involved in the recognition of the 5′-YpG-3′ step. The thymine base 5′ to the aligned guanine residue shifts about 1.5 Å into the major groove in the Ndt80 structure relative to the reference. This shift allows Arg177 to make cation–π interactions to the thymine base in addition to the typical bidentate hydrogen bonds to the guanine base. Pro57 also facilitates the shift by pushing the thymine base from the minor groove. Note that the base-pairs 5′ to the shifted thymine base align well with the reference DNA, indicating that this distortion is limited to the 5′-YpG-3′ step. A similar distortion of the 5′-YpG-3′ step is seen at the 5′-TpG-3′ step contacted by Arg111. (b) Consensus MSE sequence aligned to the actual sequence used in the crystal structure. The two unstacked 5′-TpG-3′ steps are on the bottom strand and are highlighted. The bottom duplex is the reference DNA structure with the 5′-TpG-3′ step highlighted. This reference DNA was chosen because its sequence closely matches the sequence of the MSE in the Ndt80–DNA structure, and its structure and helical parameters are similar to a standard 5′-YpG-3′ step constructed with 3DNA[51] using averaged dinucleotide step parameters obtained from the DNA structure database.[10]

have long been known to lead to base unstacking[12] and theoretical studies predict that these transitions will be most prevalent in flexible 5′-YpG-3′ dinucleotide steps.[13]

The importance of the 5′-pyrimidine residue within both of the 5′-YpG-3′ sequences for Ndt80 binding was demonstrated by the finding that mutation of either 5′-pyrimidine residue to a purine residue resulted in a three- to fivefold reduction in DNA-binding affinity. Moreover, mutation of the conserved thymine base at position 6 to a uracil residue, also caused a significant (twofold) reduction in binding affinity, consistent with the idea that the 5-methyl group of thymine base is critical for stacking interactions with the arginine residue.[7]

Quantum mechanical calculations have also provided support for the idea that arginine residue side-chains that are hydrogen bonded to guanine residues will interact favorably with bases immediately 5′ to the guanine nucleotide.[14,15] These studies, however, predicted that interactions between the arginine residue and a 5′-purine residue, not a pyrimidine base, would be more energetically favorable.

Database search for interactions between arginine residues and 5′-YpG-3′ steps

To ascertain whether recognition of 5′-YpG-3′ dinucleotide steps by arginine residues is utilized by other DNA-binding proteins, we searched a database of structures of proteins bound to DNA for 5′-YpG-3′ dinucleotide steps in which an arginine residue side-chain is simultaneously hydrogen-bonded to the guanine residue and in van der Waals contact with the 5′-pyrimidine residue which is shifted into the major groove. The criteria used to determine whether the pyrimidine residue was shifted into the major groove are summarized in Figure 2(a). The search of 553 protein–DNA complexes derived from the protein database† uncovered 13 distinct complexes which clearly display this kind of interaction (Table 1, see Methods). One of these proteins, the AML1/Runt domain, is structurally related to Ndt80. AML1/Runt, like Ndt80, is an Ig-fold transcription factor, and binds DNA using the same edge of the Ig-fold β-sandwich.[16,17] The other proteins uncovered in the search have a variety of unrelated structures. MATα2, Pbx, and Ubx are all homeobox proteins, while the *Escherichia coli* PurR repressor and CRE recombinase recognize DNA *via* helix-turn-helix motifs. E2F and DP proteins utilize a winged-helix motif to bind their target sequence, while ZIF268 recognizes DNA using three tandem Zn fingers which each recognize three consecutive base-pairs. C/EBPβ is an homodimeric bZIP transcription factor that contacts DNA using two highly positively charged α-helices. The *E. coli*

† http://www.rcsb.org/pdb/

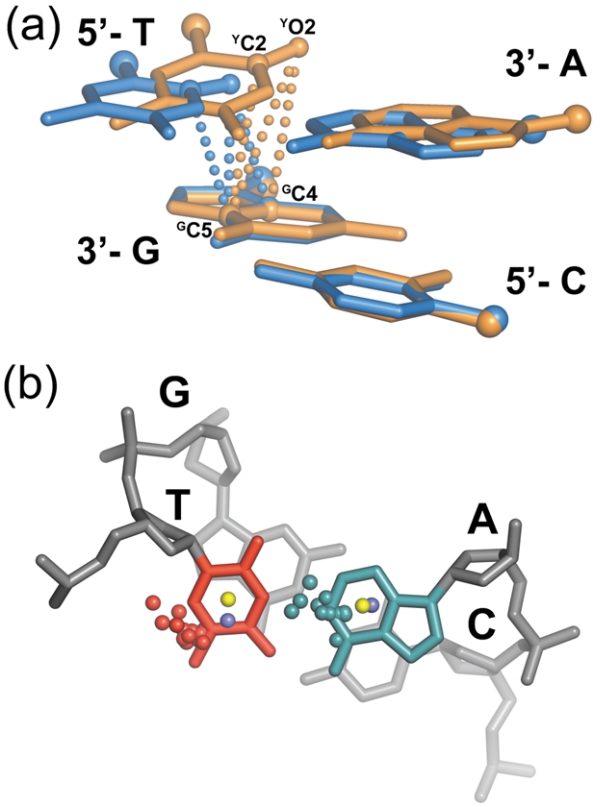

Figure 2. Searching the structural database for amino acid-assisted 5′-YpG-3′ unstacking. (a) Identification of unstacked pyrimidine–purine residue steps. In this panel, the 5′-YpG-3′ dinucleotide step is shown as a stick representation of the bases with the C1′ atom of the deoxyribose sugar depicted as a large sphere. The 5′-YpG-3′ step corresponding to base-pairs 5 and 6 of the Ndt80 complex is shown in blue and the reference base-pair step is shown in orange. Three interatomic distances (shown as dotted lines) were calculated: ^{Y}O2–^{G}C4, ^{Y}C2–^{G}C5, and ^{Y}O2–^{G}C5, where the superscript text indicates the identity of the base. In standard *B*-DNA, the ^{Y}O2–^{G}C4 and ^{Y}C2–^{G}C5 distances are approximately equal while the ^{Y}O2–^{G}C5 distance is greater. In the Ndt80–DNA complex, the pyrimidine residue shifts into the major groove and the ^{Y}O2–^{G}C4 and ^{Y}C2–^{G}C5 distances become larger while ^{Y}O2–^{G}C5 is shortened. In our analyses of the protein–DNA structure database, we consider a dinucleotide step to be unstacked if the ^{Y}O2–^{G}C5 distance is shorter than both the ^{Y}O2–^{G}C4 and the ^{Y}C2–^{G}C5 distances. (b) The summary of base displacements. The 5′-pyrimidine residue (red) and its complementary purine residue (green) of the reference DNA are shown viewed down the helical axis. The centroid of the pyrimidine residue ring and the centroid of the six member ring of the purine residue are shown as yellow spheres for the reference DNA. The corresponding centroids for the average 5′-YpG-3′ step, as calculated by 3DNA, are displayed as blue spheres, while the centroids of the shifted pyrimidine residue rings are shown as red spheres with the complementary purine residue centroids as green spheres. The movement of the pyrimidine residue is mostly a shift into the major groove while the purine residue compensates by a displacement along its long axis. Typically the movement of the pyrimidine residue is greater than that of the complementary purine residue.

Table 1. Summary of the database search and DNA parameters for the 5′-YpG-3′ steps

PDB ID	Fold type	DNA complex	Resolution	Residues: Arg residue[a]	Residues: His residue	DNA consensus	Y to ref Y distance[b]	Shift	Slide	Rise	Tilt	Roll	Twist	Energy ($k_BT/2$)
A														
1AKH	Homeo-domain	MAT a1**alpha2**	2.50	R54 (B185)		**TG**T	2.07	−0.06	−0.79	3.51	−1.23	9.82	41.1	7.1
1AKH		MAT **a1**alpha2	2.50	R55 (A124)		**TG**ATGT	0.70	0.32	0.11	3.02	−0.70	5.70	35.9	2.4
1B72		**Pbx1**-Hox1	2.35	R55 (B290)		R**TG**ATT	2.02	−0.83	0.42	3.34	−3.75	9.01	33.9	4.5
1B8I		Ubx-**Exd**	2.40	R55 (B258)		R**TG**ATT	1.58	−0.66	0.36	3.45	−1.96	10.77	35.1	3.6
1QPZ	Helix-turn-helix motif	PurR	2.50	R26 (A26)		A**YG**CAAAC	2.25	−1.49	−0.67	3.56	1.15	−6.52	34.4	17.1
5CRX		Cre	2.70	R259 (A259)		TATA**CG**AAGTTAT	1.78	−1.22	0.30	3.04	−1.51	0.99	32.8	5.7
1H9D	Ig-fold Beta sandwich	**AML1**-CBFβ	2.60	R174 (C174)		YG**YG**GTY	2.11	−1.24	−0.18	3.25	−4.45	2.28	33.7	4.9
1MNN		NDT80	1.40	R111 (A111)		WTTTG**YG**Nc	2.37	−0.95	−0.30	3.73	−5.47	3.98	39.8	10.6
1MNN		NDT80	1.40	R177 (A177)		WTT**TG**YGNc	1.97	−1.22	0.51	3.15	−3.69	3.32	34.7	8.6
1CF7	Winged-helix motif	E2F4-**DP2**	2.60	R121 (B121)		TTT**CG**CGCG	2.87	−1.95	1.05	3.66	−2.61	−4.35	34.9	13.1
1CF7		**E2F4**-DP2	2.60	R56 (A56)		TTTCGCG**CG**	2.52	−1.65	0.27	3.43	−6.21	−10.09	35.1	17.2
1ECR	Interdomain Beta-strands	TUS	2.70	R232 (A232)		TAGTA**TG**TTGTAACTA	2.75	−1.11	−0.86	4.27	−0.49	9.84	35.9	18.7
1AAY	Zn finger	ZIF268	1.60	R18 (A118)		GCGTGGG**CG**	2.23	−1.14	0.31	3.37	−4.50	6.05	33.7	2.3
1AAY			1.60	R74 (A174)		G**CG**TGGGCG	2.42	−0.95	−0.21	3.49	−7.23	7.82	36.2	4.8
1GU4	bZIP	C/EBP beta	1.80	R289 (A289)		RTTR**CG**CAAY	2.32	−1.13	−0.67	3.27	1.58	7.89	29.0	9.6
							AVG	−1.018666667	−0.023333333	3.436	−2.738	3.767333333	35.07333333	8.7
							STDEV	0.569358871	0.55802799	0.309787946	2.605420723	6.36013417	2.798427279	6
B														
1AAY	Zn finger	ZIF268	1.60		H49 (A149)	GCG**TG**GGCG	2.17	−0.94	−0.11	3.31	−5.55	3.48	34.13	4.8
1PDN	Homeo-domain	PRD	2.50		H47 (C47)	**CG**TCACGSTTSR	2.07	−1.1	0.38	3.49	−8.25	5.08	38.77	4.4
							AVG	−1.02	0.135	3.4	−6.9	4.28	36.45	4.590785129
							STDEV	0.113137085	0.346482323	0.127279221	1.909188309	1.13137085	3.280975465	0.255668782
						Average CG/CG[c]		0.00	0.41	3.39	0.00	5.40	36.10	
						Average TG/CA[c]		−0.09	0.53	3.33	−0.50	4.70	37.30	
						ref TG/CA (1D98)		0.32	0.85	3.01	−3.41	9.73	34.19	

All distances and resolutions are in Angstroms.

[a] The first value is literature numbering and in parenthesis are the chain and residue number of the pdb file.

[b] The distance between the centroid of the shifted pyrimidine residue and the centroid of the reference DNA pyrimidine residue.

[c] As derived from values calculated by Olson *et al*.[10]

replication terminator protein Tus recognizes DNA using interdomain β-strands that contact a deformed DNA major groove.

Conformational analysis of unstacked 5′-YpG-3′ steps

To compare the conformations of the DNA in these complexes, we aligned each structure on a reference, unbound DNA structure,[18] the sequence of which is nearly identical with the MSE. The 5′-YpG-3′ step from this DNA that we have used for the reference structure is very similar to an averaged 5′-YpR-3′ step as derived from the DNA structure database.[10] With the structures aligned in this way, the 5′-pyrimidine bases all are displaced into the major groove between 0.7 Å and 2.9 Å. To maintain pairing with the shifted pyrimidine residue, the complementary purine residue slides between 0.5 Å and 2.9 Å along its long axis (Figure 2(b)). In general, the inherent stacking symmetry of the 5′-YpG-3′ step is broken such that base–base stacking is reduced in the strand that is contacted by the arginine residue, while it is maintained in the complementary strand. While these base displacements are in general quite large, all torsion angles remain in their most favored positions for *B*-DNA. None of these structures adopt a true BII conformation ($\varepsilon - \zeta > 50°$) like that seen in Ndt80. Nevertheless, a large proportion have $\varepsilon - \zeta$ values between 0° and 40° for the 5′-pyrimidine residue. This is a large deviation from the mean BI value of $-80(\pm 40)°$ and may indicate a shift toward a BII conformation. Because of the difficulty in accurately modeling the phosphate group backbone at the moderate resolutions of most of these structures, it is possible that some may indeed adopt a BII conformation. In general, the positions of the phosphate groups of both the guanine residue and the 5′-pyrimidine residue are also shifted towards the major groove. In each of the complexes, the shifted phosphate groups are contacted by the protein through either salt bridges or hydrogen bonding interactions. This suggests that the 5′-YpG-3′ deformation enhances stacking with the arginine, and facilitates backbone interactions that may be critical for specificity through indirect readout. Conversely it is also possible that these backbone interactions may facilitate the 5′-YpG-3′ deformation.

We have also analyzed the DNA conformation in terms of the six independent parameters that fully describe the conformation of two successive base-pairs within a double-stranded DNA structure[19] (Table 1). Intriguingly, most of the 5′-YpG-3′ steps show significant negative shift and negative tilt, but no consistent trend away from standard values is observed for any of the other parameters. The negative shift corresponds to a movement of the 5′-pyrimidine residue into the major groove. The negative tilt corresponds to a change in the angle between the adjacent base-pairs of the 5′-YpG-3′ step such that the bases in contact with the arginine residue on one strand have a smaller rise than their complementary bases on the opposite strand. This negative tilt is allowed because of the low degree of stacking between bases of the 5′-YpG-3′ step in contact with the arginine residue.

We also estimated the energy cost associated with each of these base-pair steps as derived from their helical parameters. The costs vary from 2.3 to 18.7 in terms of $k_BT/2$, and correspond to *Z* scores of 1.5–4.3 (Table 1). These values indicate that the conformations of each 5′-YpG-3′ step differ significantly from the average 5′-YpG-3′ structures, and provide additional support for the idea that the conformations of these base-pair steps have been deformed through interactions with the amino acid side-chain.

Arginine residue—5′-YpG-3′ interactions and sequence-specific recognition

We next analyzed the available biochemical data to determine whether these proteins selectively bind to DNA targets that have pyrimidine residues rather than purine residues immediately 5′ to the guanine base. In all cases, the available data strongly suggest that the 5′-pyrimidine residue is preferred and, in most cases, an analysis of the protein–DNA interface indicates that contacts between the arginine residue and the 5′-YpG-3′ plays a key role in this recognition.

For the AML/Runt protein, the consensus DNA-binding site has been defined as 5′-YG**YG**GTY-3′ (contacted bases in bold) through *in vitro* selection experiments,[20–22] where the 5′-YpG-3′ highlighted in bold is contacted by Arg174. In this case, no other contacts are made to the 5′-pyrimidine residue (or its complementary purine residue), although contacts are made to the phosphodiester backbone. We also note that the 5′-YpG-3′ step immediately 5′ to this step does not display significant unstacking, yet the 5′-pyrimidine residue is conserved.

For the homeodomain protein, MATα2, a 5′-**TG**T-3′ sequence forms the core of the recognition site[23,24] and Arg54 contacts the central guanine residue and stacks with the 5′-thymine base.[25,26] The O4 of the 5′-thymine base hydrogen bonds with a water molecule that in turn is hydrogen bonded by Ser50. The Ser50 interaction is not conserved in all the MATα2 structures; for example, in the structure of MATα2 determined in the absence of MATa1 (**1APL**),[27] this water molecule is missing and the T·A base-pair in question makes no direct or indirect contact with the protein, other than through the DNA backbone. It, therefore, seems very likely that the strong preference for thymine bases at the 5′ position is due to stacking interactions with Arg54. The homeodomain-binding partner of MATα2, MATa1, also shows this form of recognition between Arg55 and a 5′-TpG-3′ step within its consensus binding site, 5′-**TG**ATGA-3′. In this case, the 5′-thymine base is

not in contact with MATa1 other than Arg55, although the degree of displacement of the 5′-thymine base is the smallest of the structures examined here. Another family of heterodimeric homeodomain transcription factors, Hox-Pbx (human)[28] and Ubx-Exd (Drosophila),[29] show this kind of DNA interaction between the Pbx/Exd subunit the consensus DNA-binding site 5′-A**TG**ATT-3′.[30] Here, conserved Arg55 contacts the 5′-TpG-3′ in bold *via* the major groove, while the A·T base-pairs at the 5′-end of the site are also contacted by Arg5 *via* the minor groove. The minor groove contact by Arg5 is expected to exclude G·C base-pairs at this position, but cannot select T·A over A·T base-pairs. This selection is more likely provided by the Arg55 contact. This kind of interaction is reminiscent of the recognition of the 5′-TpG-3′ step at positions 5 and 6 of the Ndt80-binding site, where the thymine base appears to be pulled into the major groove through stacking interactions with Arg177, and at the same time, pushed *via* the minor groove by Pro57.

The *E. coli* purine repressor, PurR, contains a conserved 5′-YpG-3′ sequence at positions 3 and 4 of its consensus binding site.[31,32] This 5′-YpG-3′ is only contacted by a single arginine residue, Arg26, and is highly unstacked in a number of independently determined crystal structures.[33–35] Moreover, it seems likely that this kind of recognition is conserved within the large LacI family of transcriptional repressors, as many of these proteins recognize a conserved 5′-YpG-3′ step at positions 3 and 4 which they contact with conserved arginine residue side-chains.[31]

The *E. coli* CRE recombinase also contacts its DNA targets using a helix-turn-helix motif.[36,37] The recombinase recognizes a large, palindromic DNA target sequence, yet, paradoxically, makes only one major groove contact. The single contact is between Arg259 and a conserved 5′-CpG-3′ step in the consensus sequence. Glu262 is in the vicinity of the 5′-cytosine base, but its carboxylate group is too far from the cytosine base exocyclic amine group (>4 Å) to exert a significant sequence selectivity.

The human E2F and related DP protein are cell-cycle transcription factors that cooperatively recognize their target DNA (5′-TTT**CG**CG**CG**-3′)[38–40] with a winged helix motif. In a heterodimeric structure of E2F-DP there are two 5′-YpG-3′ steps that show arginine residue mediated unstacking.[41] One involves Arg56 of E2F and the second involves Arg121 of DP, both of which recognize 5′-CpG-3′ steps. The step recognized by Arg121 of DP is simultaneously contacted from the minor groove by Arg17 of E2F. The major difference in this case is the arginine residue in the minor groove comes from a different subunit, E2F. The ability of these two subunits to cooperate in the recognition of the 5′-CpG-3′ step is dependent on the flexibility of the 5′-YpR-3′ step to accommodate both contacts simultaneously.

ZIF268 contains three tandem Zn fingers that each recognize a three base-pair DNA target within a nine base-pair site.[42,43] The N-terminal finger contacts the 3′-most three base-pairs of the binding site (5′-G**CG**-3′). Arg18 immediately N-terminal to the recognition helix (at the " − 1" position) contacts the 5′-CpG-3′ step *via* hydrogen bonding and stacking. However, while the 5′-cytosine base is clearly recognized in a sequence specific manner, at least part of this recognition is from major groove van der Waals interactions between the cytosine base and Glu21. It seems plausible that in this case Arg18 and Glu21 cooperate to recognize the 5′-CpG-3′ step, as the shift of the cytosine base into the major groove (facilitated by stacking with Arg18) is required to achieve van der Waals contacts with Glu21. In addition, the third Zn finger, which recognizes the 5′-most binding site (5′-G**CG**-3′), contacts the highlighted guanine residue through Arg74. Here, the 5′ C·G base-pair does not make any other direct sequence specific contacts and the cytosine base of this pair is shifted into the major groove. Therefore, specificity for the 5′-cytosine base is likely achieved through arginine residue stacking.

The C/EBPβ homodimeric transcription factor preferentially binds a near palindromic DNA consensus sequence, 5′-RTTR**CG**CAAY-3′.[44] The centre of symmetry of the DNA site is a 5′-CpG-3′ step which is recognized in a surprisingly asymmetric manner by the protein. Arg289 from one of the monomers recognizes this base-pair step through hydrogen bonding interactions with the guanine residue and van der Waals interactions with the 5′-cytosine base, which induces unstacking between the contacted bases. In contrast, the 5′-CpG-3′ on the complementary strand remains stacked and, as a result, Arg289 from the other monomer from the complex does not contact the complementary strand but instead adopts a different conformation to contact an adjacent 5′-TpG-3′ step.

The *E. coli* replication termination protein Tus recognizes DNA through a novel β-strand motif that interacts with a distorted DNA major groove.[45] A highly conserved 5′-TpG-3′ step[46] is recognized by Arg232 from the β-strand motif. No other amino acid side-chains contact this base-pair step from either the major or minor groove. 5′-TpG-3′ unstacking induced by Arg232 may also be involved in generating the DNA distortion that is recognized by Tus.

Evidence for arginine residue-induced distortion 5′-YpG-3′ steps

The X-ray crystal structure of a mutant form of the MATα2 homeodomain bound to DNA provides strong evidence that arginine residue recognition of a 5′-TpG-3′ step directly leads to the unstacking of the dinucleotide step. In the mutated protein, the arginine residue (Arg54) which normally contacts the 5′-TpG-3′ step in the wild-type protein, is mutated to alanine, as are two other

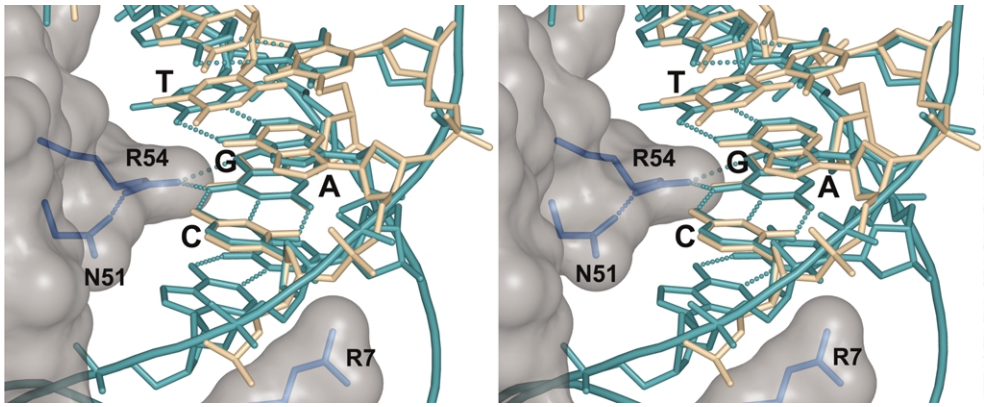

Figure 3. Evidence for arginine residue-induced unstacking 5′-YpG-3′ steps. A triple alanine residue mutant of MATα2 was aligned to a wild-type MATα2 structure using the G·C base-pair of the 5′-YpG-3′ step as in Figure 1. The surface and protein side-chains are from the wild-type structure and the wild-type DNA is shown in green. The DNA from the alanine residue mutant is shown in gold. The 5′-thymine base moves just over 1 Å towards the minor groove when Arg54 is mutated to an alanine.

major groove recognition residues, Ser50 and Asn51.[47] The DNA sequence of this structure is identical with the MATa1/α2/DNA structure described above. The wild-type and mutant protein structures are very similar as indicated by an RMSD value of 0.48 Å over 58 aligned C^{α} atoms. However, when the G·C base-pair of the mutant structure is aligned with the native structure, the extent of the unstacking induced by the arginine residue becomes apparent. The 5′-thymine base of the wild-type MATα2 is shifted more than 1 Å into the major groove when compared with the mutant structure (Figure 3). Since neither Ser50 nor Asn51 make direct contact to the shifted base-pair in the wild-type structure, it is most likely that the displacement of the 5′-thymine base into the major groove in the wild-type structure is a direct consequence of its interactions with Arg54.

5′-TpG-3′ *versus* 5′-CpG-3′ recognition

In theory, either 5′-TpG-3′ or 5′-CpG-3′ steps could be recognized equally well by simultaneous π–cation and hydrogen bonding interactions with arginine residues. In most cases, however, the proteins are specific for either a 5′-thymine base or a 5′-cytosine base. This specificity is often achieved by other elements of the protein. For example, in ZIF268-, Cre-, E2F- and DP–DNA complexes, substitution of a 5′-thymine base for the consensus cytosine base would result in a steric clash between the 5-methyl group of the thymine base and a side-chain in the major groove. For ZIF268 and Cre, glutamic acid residues provide this additional specificity, while tyrosine residues provide this function in E2F and DP. For the homeodomain proteins Pbx and Ubx, specificity for thymine bases is achieved through the minor groove contacts that would exclude the 2-amino group of guanine residue. In Ndt80 and MATα2, the specificity for thymine bases is in part accomplished with the aliphatic portion of the arginine residue side-chain, which forms a hydrophobic half-pocket for the 5-methyl group of the thymine base. For Ndt80, the importance of the 5-methyl group has been directly demonstrated at the sixth position of the consensus DNA target sequence. Substitution of the conserved thymine base at this position with uracil, effectively replacing the 5-methyl group with a hydrogen atom, results in a twofold reduction in binding affinity.[7] However, there are some cases (Ndt80, AML1 and PurR, see Table 1) where the consensus sequence indicates that either pyrimidine residue can be accommodated. These examples show that recognition of 5′-YpG-3′ steps by arginine has the flexibility to allow either pyrimidine residue in the 5′ position.

Histidine residue—5′-YpG-3′ recognition

The planar, aromatic nature of histidine, together with its ability to hydrogen bond to nucleic acid bases, suggested that this side-chain might also recognize 5′-YpR-3′ steps by hydrogen bonding to the 3′-purine residue and stacking with the 5′-pyrimidine residue. To test this idea, we searched the protein–DNA sequence database for 5′-YpR-3′ steps in which the purine residue N7 is within hydrogen bonding distance to a histidine residue, and the 5′-pyrimidine residue is displaced into the major groove. Two examples of such an interaction were found (Table 1B). One is within the ZIF268–DNA complex, where His49 of the central finger contacts the central guanine residue of the three base-pair site recognized by this finger, 5′-TGG-3′. The 5′-thymine base is specifically recognized by the protein and this recognition was previously thought to involve stacking interactions with His49. We suggest that this stacking is made possible by the shift of the thymine base into the major groove. The second example is found in the structure of the Paired (PRD) homeodomain protein bound to its consensus DNA.[48] In this structure, His47 contacts the 5′-most 5′-CpG-3′

step of the DNA consensus (**CG**TCACGSTTSR, where S is a guanine base or cytosine base).[49] Neither the 5′-cytosine base, nor its complementary guanine base is contacted by the protein, other than by His47. This 5′-CpG-3′ step is not, however, conserved in the binding sites for Pax proteins that are highly similar to the PRD protein.

Generality of protein-induced unstacking of 5′-YpG-3′

Here, we have described a way in which arginine or histidine residues can recognize the inherent flexibility of 5′-YpG-3′ dinucleotide steps. Might other amino acid residues also induce similar distortions in 5′-YpR-3′ steps? Glutamine and asparagine residue side-chains can recognize adenine bases *via* a pair of hydrogen bonds to the major groove face of the DNA, in a manner that is structurally similar to arginine–guanine base recognition.[1] We searched the protein–DNA structure database for examples of glutamine or asparagine residue recognition of 5′-YpA-3′ steps that induced unstacking between the 5′-pyrimidine bases and the adenine bases, and enhanced contact between the glutamine/asparagine residue and the pyrimidine residue. No such examples were found. Thus, it may be that distortion of the 5′-YpG-3′ step requires the relatively strong cation–π interactions between an arginine residue or histidine residue side-chain and the shifted nucleic acid base.[14,15]

Unstacking (as defined in Figure 2(a)) occurs in approximately 34% of all 5′-YpG-3′ steps that are contacted by arginine residues. In contrast, only about 10% of 5′-YpG-3′ steps in the free DNA structure database display this kind of unstacking, indicating that arginine-induced 5′-YpG-3′ unstacking may be a common but not universal consequence of these interactions. Nevertheless, we have shown that this mode of protein–DNA recognition is utilized by many of the major classes of transcription factors and in most cases helps to explain hitherto inexplicable protein specificity for its consensus DNA. Our analysis has focused on relatively dramatic examples of unstacking, however, it is possible that smaller distortions could also allow recognition of 5′-YpG-3′ steps by these side-chains. If this is the case, then 5′-YpG-3′ recognition may be a much more general phenomena than reported here.

Methods

The 553 protein–DNA structures were obtained from the 15 September, 2003 release of the protein database†. The database contains all structures solved by X-ray crystallography with a resolution of 3.0 Å or better containing both DNA and protein. See Supplementary Material for listings of PDB files included in each of the databases used in this study. A Perl script (ArgStack.pl, see Supplementary Material) was used to find all 5′-YpG-3′ steps with an arginine residue side-chain within hydrogen bonding distance of the O6 and/or N7 of the guanine residue and simultaneously in close proximity to the 5′-pyrimidine residue (as determined by arginine residue CZ to pyrimidine residue C5 distance in the case of arginine residues). The distance cutoffs used for the hydrogen bonding distance and proximity to 5′-pyrimidine were 3.0 Å and 6.0 Å, respectively. Next, the degree of stacking of the 5′-pyrimidine on the 3′-purine residue was assessed using the criteria described in Figure 2(a). We considered a 5′-YpG-3′ step unstacked if both the ^{Y}C2–^{G}C5 and ^{Y}O2–^{G}C4 interatomic distances were longer than the ^{Y}O2–^{G}C5 distance. Typical values for these distances are 3.7 Å for ^{Y}C2–^{G}C5, 3.8 Å for ^{Y}O2–^{G}C4, and 3.9 Å for ^{Y}O2–^{G}C5. 5′-YpG-3′ steps with any of these distances greater than 5 Å were removed to eliminate non-*B* DNA structures from the analysis. In this way, 269 5′-YpG-3′ steps in contact with arginine residues were retrieved from the database, of which 81 (or 30.1%) are unstacked. We next repeated the search using a non-redundant database in which structures with a sequence identity of 90% or greater were removed. The non-redundant database, containing 209 structures, contained 142 5′-YpG-3′ steps in contact with arginine, of which 30 (or 21%) are unstacked. In a similar fashion we assessed the stacking for all 5′-YpG-3′ steps, regardless of proximity to amino acid side-chains, in both the redundant and non-redundant databases. 428/2124 (or 20%) of 5′-YpG-3′ steps in the redundant dataset and 176/907 (or 19%) of 5′-YpG-3′ steps in the non-redundant dataset were unstacked by our criteria. Finally, a DNA-only database was extracted from the RCSB database, which consisted of all structures containing only DNA, solved by X-ray crystallography to 3.0 Å resolution or better, and identified as *B*-DNA in the PDB header. 142/1444 (or 10%) of 5′-YpG-3′ steps of the DNA-only database are unstacked.

The structures retrieved from this automated procedure were then visually inspected to ascertain if the steps are unstacked independently of the criteria above. Each structure was aligned to the reference DNA by least-squares superposition (as implemented in O[50]) of the 3′-purine base and its Watson–Crick partner onto the guanine–cytosine base-pair of the reference step (residue B22 (G) and A3 (C) of the reference). The displacements of the centroids of the 5′-pyrimidine base and its base-paired partner relative to the reference structure were then determined. For the protein–DNA complexes for which there are multiple independent structures deposited in the RCSB database, we have only discussed those structures for which unstacking is consistently observed in a majority of the deposited structures.

DNA helical parameters and torsion angles of the extracted structures were calculated using

† http://www.rcsb.org/pdb/

3DNA.[51] The energy of displacement of the helical parameters from their mean values, expressed in terms of $k_BT/2$, was calculated as described.[10] The square root of this value gives the number of standard deviations from the minimum energy represented by the average parameters (i.e. *Z*-score).

Similar searches were also performed for 5′-YpR-3′ steps in contact with histidine, glutamine, lysine or asparagine residues *via* the major groove. Significant unstacking was only observed for 5′-YpG-3′ steps in contact with histidine, but not for any of the other side-chains.

Acknowledgements

This work was supported by grants from the Canadian Institutes of Health Research (CIHR), the Alberta Heritage Foundation for Medical Research (AHFMR) and the Canada Research Chairs program to J.N.M.G. J.S.L. was supported by studentships from CIHR and AHFMR.

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. Anderson, J. E., Ptashne, M. & Harrison, S. C. (1987). Structure of the repressor–operator complex of bacteriophage 434. *Nature*, **326**, 846–852.
3. Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q. *et al.* (1988). Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, **335**, 321–329.
4. Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M. & Harrison, S. C. (1988). Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science*, **242**, 899–907.
5. Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. (2000). An overview of the structures of protein–DNA complexes. *Genome Biol.* **1** REVIEWS001.
6. 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.
7. Lamoureux, J. S., Stuart, D., Tsang, R., Wu, C. & Glover, J. N. (2002). Structure of the sporulation-specific transcription factor Ndt80 bound to DNA. *EMBO J.* **21**, 5721–5732.
8. Hepworth, S. R., Ebisuzaki, L. K. & Segall, J. (1995). A 15-base-pair element activates the SPS4 gene midway through sporulation in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **15**, 3934–3944.
9. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O. & Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast [published erratum appears in Science 1998 Nov 20; 282(5393): 1421]. *Science*, **282**, 699–705.
10. 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.
11. Dickerson, R. E. & Chiu, T. K. (1997). Helix bending as a factor in protein/DNA recognition. *Biopolymers*, **44**, 361–403.
12. Prive, G. G., Heinemann, U., Chandrasegaran, S., Kan, L. S., Kopka, M. L. & Dickerson, R. E. (1987). Helix geometry, hydration, and G·A mismatch in a B-DNA decamer. *Science*, **238**, 498–504.
13. Bertrand, H., Ha-Duong, T., Fermandjian, S. & Hartmann, B. (1998). Flexibility of the B-DNA backbone: effects of local and neighbouring sequences on pyrimidine–purine steps. *Nucl. Acids Res.* **26**, 1261–1267.
14. Wintjens, R., Lievin, J., Rooman, M. & Buisine, E. (2000). Contribution of cation–pi interactions to the stability of protein–DNA complexes. *J. Mol. Biol.* **302**, 395–410.
15. Rooman, M., Lievin, J., Buisine, E. & Wintjens, R. (2002). Cation–pi/H-bond stair motifs at protein–DNA interfaces. *J. Mol. Biol.* **319**, 67–76.
16. Tahirov, T. H., Inoue-Bungo, T., Morii, H., Fujikawa, A., Sasaki, M., Kimura, K. *et al.* (2001). Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFbeta. *Cell*, **104**, 755–767.
17. Bravo, J., Li, Z., Speck, N. A. & Warren, A. J. (2001). The leukemia-associated AML1 (Runx1)–CBF beta complex functions as a DNA-induced molecular clamp. *Nature Struct. Biol.* **8**, 371–378.
18. Nelson, H. C., Finch, J. T., Luisi, B. F. & Klug, A. (1987). The structure of an oligo(dA)·oligo(dT) tract and its biological implications. *Nature*, **330**, 221–226.
19. Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. N., Kennard, O. *et al.* (1989). Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.* **208**, 787–791.
20. Kamachi, Y., Ogawa, E., Asano, M., Ishida, S., Murakami, Y., Satake, M. *et al.* (1990). Purification of a mouse nuclear factor that binds to both the A and B cores of the polyomavirus enhancer. *J. Virol.* **64**, 4808–4819.
21. Melnikova, I. N., Crute, B. E., Wang, S. & Speck, N. A. (1993). Sequence specificity of the core-binding factor. *J. Virol.* **67**, 2408–2411.
22. Speck, N. A. & Terryl, S. (1995). A new transcription factor family associated with human leukemias. *Crit. Rev. Eukaryot. Gene Expr.* **5**, 337–364.
23. Goutte, C. & Johnson, A. D. (1993). Yeast a1 and alpha 2 homeodomain proteins form a DNA-binding activity with properties distinct from those of either protein. *J. Mol. Biol.* **233**, 359–371.
24. Goutte, C. & Johnson, A. D. (1994). Recognition of a DNA operator by a dimer composed of two different homeodomain proteins. *EMBO J.* **13**, 1434–1442.
25. Li, T., Stark, M. R., Johnson, A. D. & Wolberger, C. (1995). Crystal structure of the MATa1/MAT alpha 2 homeodomain heterodimer bound to DNA. *Science*, **270**, 262–269.
26. Li, T., Jin, Y., Vershon, A. K. & Wolberger, C. (1998). Crystal structure of the MATa1/MATalpha2 homeodomain heterodimer in complex with DNA containing an A-tract. *Nucl. Acids Res.* **26**, 5707–5718.
27. Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D. & Pabo, C. O. (1991). Crystal structure of a MAT alpha 2 homeodomain–operator complex suggests a general model for homeodomain–DNA interactions. *Cell*, **67**, 517–528.

28. Piper, D. E., Batchelor, A. H., Chang, C. P., Cleary, M. L. & Wolberger, C. (1999). Structure of a HoxB1-Pbx1 heterodimer bound to DNA: role of the hexapeptide and a fourth homeodomain helix in complex formation. *Cell*, **96**, 587–597.
29. Passner, J. M., Ryoo, H. D., Shen, L., Mann, R. S. & Aggarwal, A. K. (1999). Structure of a DNA-bound Ultrabithorax–Extradenticle homeodomain complex. *Nature*, **397**, 714–719.
30. Chang, C. P., Brocchieri, L., Shen, W. F., Largman, C. & Cleary, M. L. (1996). Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the Hox locus. *Mol. Cell. Biol.* **16**, 1734–1745.
31. Rolfes, R. J. & Zalkin, H. (1988). *Escherichia coli* gene purR encoding a repressor protein for purine nucleotide synthesis. Cloning, nucleotide sequence, and interaction with the purF operator. *J. Biol. Chem.* **263**, 19653–19661.
32. Rolfes, R. J. & Zalkin, H. (1990). Autoregulation of *Escherichia coli* purR requires two control sites downstream of the promoter. *J. Bacteriol.* **172**, 5758–5766.
33. Schumacher, M. A., Choi, K. Y., Zalkin, H. & Brennan, R. G. (1994). Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. *Science*, **266**, 763–770.
34. Schumacher, M. A., Glasfeld, A., Zalkin, H. & Brennan, R. G. (1997). The X-ray structure of the PurR–guanine–purF operator complex reveals the contributions of complementary electrostatic surfaces and a water-mediated hydrogen bond to corepressor specificity and binding affinity. *J. Biol. Chem.* **272**, 22648–22653.
35. 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.
36. Guo, F., Gopaul, D. N. & van Duyne, G. D. (1997). Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. *Nature*, **389**, 40–46.
37. Guo, F., Gopaul, D. N. & Van Duyne, G. D. (1999). Asymmetric DNA bending in the Cre-loxP site-specific recombination synapse. *Proc. Natl Acad. Sci. USA*, **96**, 7143–7148.
38. Lees, J. A., Saito, M., Vidal, M., Valentine, M., Look, T., Harlow, E. *et al.* (1993). The retinoblastoma protein binds to a family of E2F transcription factors. *Mol. Cell. Biol.* **13**, 7813–7825.
39. Buck, V., Allen, K. E., Sorensen, T., Bybee, A., Hijmans, E. M., Voorhoeve, P. M. *et al.* (1995). Molecular and functional characterisation of E2F-5, a new member of the E2F family. *Oncogene*, **11**, 31–38.
40. Zhang, Y. & Chellappan, S. P. (1995). Cloning and characterization of human DP2, a novel dimerization partner of E2F. *Oncogene*, **10**, 2085–2093.
41. Zheng, N., Fraenkel, E., Pabo, C. O. & Pavletich, N. P. (1999). Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. *Genes Dev.* **13**, 666–674.
42. Pavletich, N. P. & Pabo, C. O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268–DNA complex at 2.1 Å. *Science*, **252**, 809–817.
43. Elrod-Erickson, M., Rould, M. A., Nekludova, L. & Pabo, C. O. (1996). Zif268 protein–DNA complex refined at 1.6 Å: a model system for understanding zinc finger–DNA interactions. *Structure*, **4**, 1171–1180.
44. Osada, S., Yamamoto, H., Nishihara, T. & Imagawa, M. (1996). DNA binding specificity of the CCAAT/enhancer-binding protein transcription factor family. *J. Biol. Chem.* **271**, 3891–3896.
45. Kamada, K., Horiuchi, T., Ohsumi, K., Shimamoto, N. & Morikawa, K. (1996). Structure of a replication–terminator protein complexed with DNA. *Nature*, **383**, 598–603.
46. Hill, T. M., Pelletier, A. J., Tecklenburg, M. L. & Kuempel, P. L. (1988). Identification of the DNA sequence from the *E. coli* terminus region that halts replication forks. *Cell*, **55**, 459–466.
47. Ke, A., Mathias, J. R., Vershon, A. K. & Wolberger, C. (2002). Structural and thermodynamic characterization of the DNA binding properties of a triple alanine mutant of MATalpha2. *Structure (Camb)*, **10**, 961–971.
48. Xu, W., Rould, M. A., Jun, S., Desplan, C. & Pabo, C. O. (1995). Crystal structure of a paired domain–DNA complex at 2.5 Å resolution reveals structural basis for Pax developmental mutations. *Cell*, **80**, 639–650.
49. Jun, S. & Desplan, C. (1996). Cooperative interactions between paired domain and homeodomain. *Development*, **122**, 2639–2650.
50. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallog. sect. A*, **47**, 110–119.
51. Lu, X. J., Shakked, Z. & Olson, W. K. (2000). A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.* **300**, 819–840.

Edited by J. Thornton

(Received 10 June 2003; received in revised form 13 October 2003; accepted 30 October 2003)

Supplementary Material for this paper comprising listings of PDB files included in each of the databases used in this study and a sample Perl script used to analyze the PDB files are available on Science Direct