The EMBO Journal Vol.18 No.23 pp.6610–6618, 1999

The structure of a chromosomal high mobility group protein–DNA complex reveals sequence-neutral mechanisms important for non-sequence-specific DNA recognition

Frank V.Murphy IV[1,2], Robert M.Sweet[3] and Mair E.A.Churchill[1,4]

[1]Department of Pharmacology, The University of Colorado Health Sciences Center, 4200 E. Ninth Avenue, Denver, CO 80262, [2]Department of Biochemistry, University of Illinois at Urbana-Champaign, 600 South Matthews Avenue, Urbana, IL 61801 and [3]Department of Biology, Brookhaven National Laboratory, Upton, NY 11973, USA

[4]Corresponding author
e-mail: Mair.churchill@uchsc.edu

The high mobility group (HMG) chromosomal proteins, which are common to all eukaryotes, bind DNA in a non-sequence-specific fashion to promote chromatin function and gene regulation. They interact directly with nucleosomes and are believed to be modulators of chromatin structure. They are also important in V(D)J recombination and in activating a number of regulators of gene expression, including p53, Hox transcription factors and steroid hormone receptors, by increasing their affinity for DNA. The X-ray crystal structure, at 2.2 Å resolution, of the HMG domain of the *Drosophila melanogaster* protein, HMG-D, bound to DNA provides the first detailed view of a chromosomal HMG domain interacting with linear DNA and reveals the molecular basis of non-sequence-specific DNA recognition. Ser10 forms water-mediated hydrogen bonds to DNA bases, and Val32 with Thr33 partially intercalates the DNA. These two 'sequence-neutral' mechanisms of DNA binding substitute for base-specific hydrogen bonds made by equivalent residues of the sequence-specific HMG domain protein, lymphoid enhancer factor-1. The use of multiple intercalations and water-mediated DNA contacts may prove to be generally important mechanisms by which chromosomal proteins bind to DNA in the minor groove.

Keywords: chromatin/DNA binding/HMG domain/non-sequence-specific/X-ray crystallography

Introduction

Chromatin is the template on which all DNA-dependent processes occur in eukaryotes, but the structures and functions of many essential abundant proteins that form and modulate chromatin structure are still not understood (Kornberg and Lorch, 1992; Bustin and Reeves, 1996). With the exception of histone proteins, the high mobility group (HMG) proteins are among the most ubiquitous of the chromatin-associated proteins (Read *et al*., 1995; Bustin and Reeves, 1996). The HMG domain superfamily is composed of two families distinguished by their abundance, function and DNA specificity (Figure 1A) (Grosschedl *et al*., 1994; Baxevanis and Landsman, 1995; Bustin and Reeves, 1996). Both groups, the sequence-specific transcription factors typified by SRY (Goodfellow and Lovellbadge, 1993; Haqq *et al*., 1994) and lymphoid enhancer factor-1 (LEF-1) (Giese *et al*., 1991), and the non-sequence-specific chromosomal proteins typified by HMG1/2 (Bianchi *et al*., 1992), bind to DNA using the ~80 residue HMG domain.

Despite this common DNA-binding domain, the functions of the two families of HMG domain proteins are quite varied. The HMG domain transcription factors are less abundant, restricted to fewer cell types, usually contain a single HMG domain and function by binding site-specifically in promoter or enhancer regions of regulated genes (Grosschedl *et al*., 1994). In contrast, the archetype of the chromosomal HMG domain group, the HMG1/2 group of proteins, usually contain two or more tandem HMG domains and bind to DNA non-sequence-specifically to facilitate nucleosome function, DNA recombination and repair, as well as activation (Shykind *et al*., 1995) and repression (Ge and Roeder, 1994) of general transcription. They interact directly with nucleosomes (Nightingale *et al*., 1996) and are believed to be modulators of chromatin structure (Bustin and Reeves, 1996). They are also important in V(D)J recombination (van Gent *et al*., 1997) and in activating regulators of gene expression, including p53, Hox transcription factors and steroid hormonereceptors, by increasing their affinity for DNA (Zwilling *et al*., 1995; Zappavigna *et al*., 1996; Boonyaratanakornkit *et al*., 1998; Jayaraman *et al*., 1998; David-Cordonnier *et al*., 1999). The characteristics of the chromosomal HMG proteins that facilitate such diverse functions in bulk DNA as well as in particular sequence contexts include moderate DNA-binding affinity, minimal sequence specificity, DNA bending and protein–protein interactions.

HMG-D is a *Drosophila melanogaster* homolog of HMG1, which contains a single HMG domain and is abundant in embryonic chromatin at a time when the embryo is transcriptionally silent but is undergoing rapid rounds of replication (Wagner *et al*., 1992; Ner and Travers, 1994). Like HMG1 (Pil and Lippard, 1992; Paull *et al*., 1993; Read *et al*., 1995; Teo *et al*., 1995), HMG-D binds to DNA with minimal sequence specificity (Churchill *et al*., 1995), preferring duplex DNA that is deformable, 'pre-bent' and/or underwound (Churchill *et al*., 1995, 1999; Wolfe *et al*., 1995; Payet and Travers, 1997). It binds cooperatively, and causes severe bending and unwinding of linear DNA, forming DNA circles as small as 55 bp (Churchill *et al*., 1999). The affinity range of the protein for linear DNA is micromolar to nanomolar depending on whether the C-terminal basic tail is intact (Payet and Travers, 1997; Churchill *et al*., 1999). The solution structure of the HMG-D HMG domain in the absence of DNA (Jones *et al*., 1994) revealed that it has

 © European Molecular Biology Organization

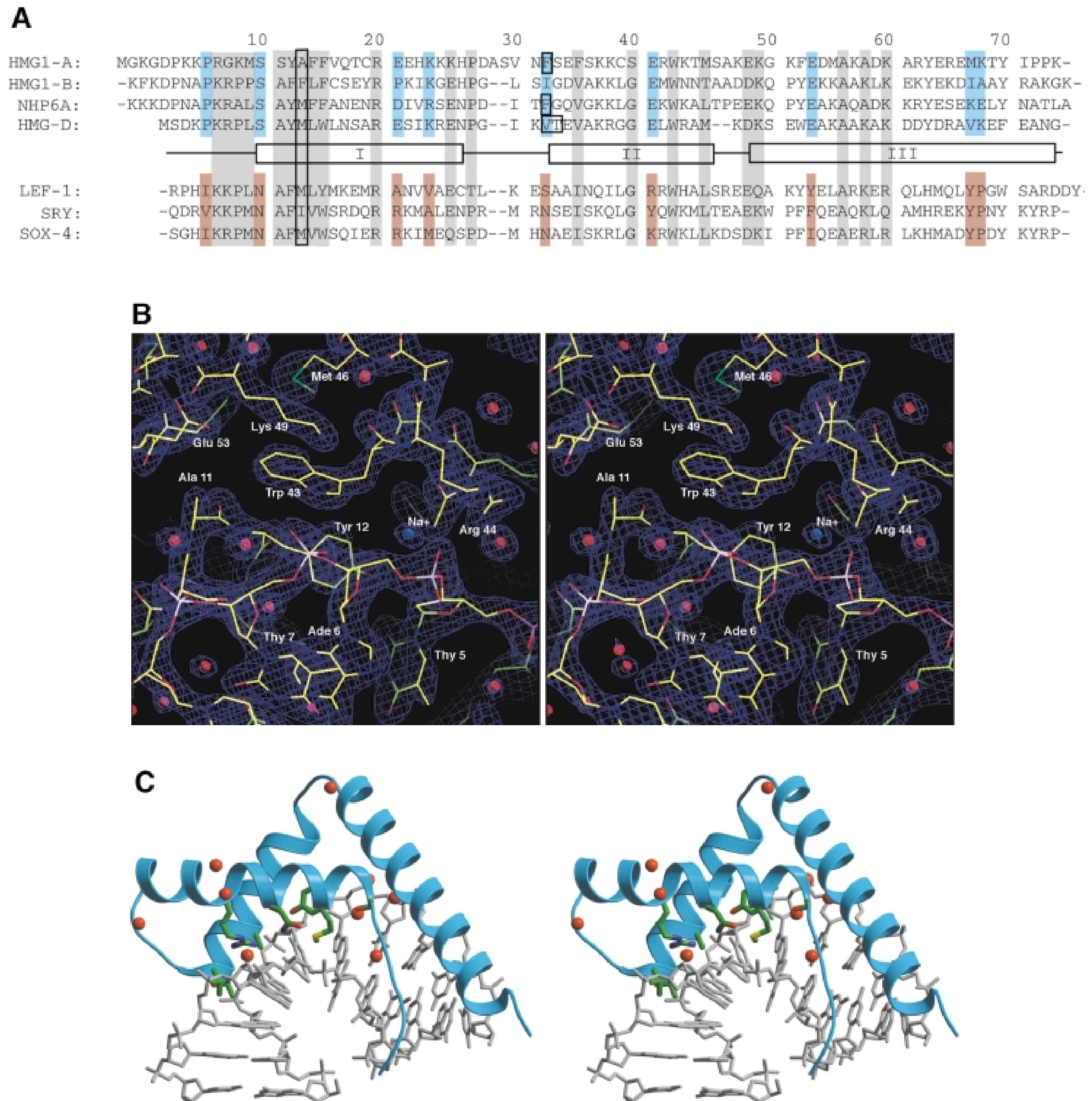

Fig. 1. Structure of the HMG-box of HMG-D bound to DNA. (**A**) Sequence comparison of sequence-specific and non-sequence-specific HMG domains. The sequences are aligned and numbered according to the HMG-D structure, with helices, I, II and II depicted by black boxes (Jones *et al*., 1994; Baxevanis and Landsman, 1995). Residues shown from structural and modeling studies to intercalate the DNA are outlined in black (Love *et al*., 1995; Werner *et al*., 1995a,b; Balaeff *et al*., 1998; Allain *et al*., 1999; Ohndorf *et al*., 1999). Residues that are conserved between the two HMG-box families are shaded in gray, whereas those residues that consistently differ between the two families of HMG domains are highlighted in cyan and brown (Balaeff *et al*., 1998; Churchill *et al*., 1999). (**B**) Stereo view of the refined ($2|F_o| - |F_c|$) electron density map contoured at a level of 1.9 σ. The protein and DNA are colored using standard CPK coloring, with water molecules and a sodium ion represented by red and blue spheres, respectively. (**C**) Ribbon diagram in stereo view of the complex. HMG-D is depicted in cyan, the DNA in gray, and structural water molecules found in the protein and at the DNA interface in red. Several side chains that interact with the DNA, Ser10, Tyr12, Met13, Asn17, Arg20, Val32, Thr33 and Ala36, are shown in green. The protein is well ordered from residue 4 to 72, and the DNA is well ordered throughout except for base cytosine 10, which adopts two conformations in the crystal (only one conformation is shown).

a global fold of three helices stabilized in an 'L-shaped' configuration by two hydrophobic cores identical to the HMG domains of other chromosomal proteins, including HMG1 and NHP6A (Read *et al*., 1993; Weir *et al*., 1993; Hardman *et al*., 1995; Allain *et al*., 1999).

The general mode of DNA binding of the HMG domain superfamily was determined from biochemical binding studies (Giese *et al*., 1991, 1992), but the details of protein–DNA interactions were revealed by solution structures of DNA complexes formed with the HMG domains of the sequence-specific transcription factors LEF-1 (Love *et al*., 1995) and SRY (King and Weiss, 1993; Werner *et al*., 1995b). The HMG domain binds to a flattened, underwound and bent DNA minor groove using a large surface on the concave face of the protein, bending the DNA helix axis away from the site of contact. Numerous electrostatic interactions and van der Waals contacts make up the interface. Partial intercalation of an aliphatic residue at position 13 (HMG-D numbering) occurs at a single site, although a smooth DNA bend occurs throughout

Table I. Structure determination and refinement

	Native	Br-λ1	Br-λ2	Br-λ3	Br-λ4
Phasing statistics					
Wavelength (Å)	1.54	0.90	0.9211	0.92157	1.05
Resolution (Å)	20–2.2	40–2.4	40–2.4	40–2.4	40–2.6
Observed reflections	94 446	136 303	143 741	143 475	102 921
Unique reflections	10 672	8685	8829	8807	8657
Completeness (%)	98.0 (97.9)	95.5 (96.4)	95.0 (95.6)	94.6 (95.4)	91.6 (65.3)
Redundancy	8.8	18.5	16.3	16.3	11.9
R_{sym}[a]	5.2 (19.8)	8.2 (24.4)	7.3 (22.3)	7.8 (25.0)	6.5 (17.6)
Phasing power					
Centric			1.96	0.25	0.19
Acentric (Iso/Ano)		0.0/1.17	1.98/1.06	0.98/1.33	1.18/0.45
R_{Cullis}			0.447	0.388	0.628
Mean figure of merit[b] (centric/acentric) is 0.15/0.26 (20–2.4 Å)					
Refinement statistics					
Resolution range (Å)		20–2.2			
Average B (Å^2)		34.2			
Crystallographic R-value[c] (%)		23.8			
R_{free}[c] (%)		28.8			
No. of reflections used		10 672			
Luzatti coordinate error (Å)		0.30			

	Protein	DNA	Solvent	Ion
No. of atoms	1701	423	115	1
R.m.s.d. bond lengths (Å)	0.009	0.007		
R.m.s.d. bond angles (°)	1.32	1.19		

[a]$R_{sym} = \Sigma_i|I_i - \langle I\rangle|/\Sigma_i I_i$, where $\langle I\rangle$ is the mean intensity for equivalent reflections, I_i.
[b]Calculated by SHARP before solvent flattening.
[c]R-value $= \Sigma||F_o| - |F_c||/\Sigma|F_o|$, with R_{free} calculated with an excluded test set of 560 reflections.

the central region of the binding site. Sequence-specific interactions between DNA bases and residues Asn10, Ser31, Asn36 and Tyr76 also occur in the LEF-1 complex.

In the absence of a chromosomal HMG domain–DNA structure, predicted models of chromosomal HMG domain–DNA complexes aided identification of specificity determinants. Compared with sequence-specific complexes, the predicted structure of the HMG-D–DNA complex has similar interactions at the protein–DNA interface, with additional side chain partial intercalations (Balaeff *et al*., 1998). One of these putative intercalating residues (Val32) differs between the specific and non-sequence-specific HMG domain families (Figure 1A), and thus revealed a potential mechanism of sequence-neutral DNA recognition. A recent modeling study of NHP6A bound to DNA (Allain *et al*., 1999) also suggested that intercalation of the same residue may result in sequence-neutral DNA binding. The structure of the architecture-specific complex (HMG1 box A bound to cisplatin-modified duplex DNA) confirms this prediction, as residue 47 (equivalent to HMG-D 32) intercalates the severe kink created by the cisplatin lesion (Ohndorf *et al*., 1999). However, the architecture-specific complex differs from the known and predicted HMG–DNA structures in terms of DNA structure and bending, and interactions at the protein–DNA interface.

In order to understand the nature of non-sequence-specific DNA recognition by the chromosomal HMG proteins as well as the structural basis for the specificity difference between the two types of HMG domains, we co-crystallized the HMG domain of HMG-D with a linear duplex DNA fragment (Murphy *et al*., 1999). The structure was solved and refined to 2.2 Å resolution (Figure 1B; Table I), providing a novel detailed view of non-sequence-specific DNA recognition by this important family of proteins. Comparison with the architecture-specific complex reveals the differences in DNA recognition between the pre-kinked DNA and normal linear DNA. Comparison of this structure with the solution structure of the LEF-1–DNA complex reveals the differences in DNA interactions that contribute to differences in DNA specificity.

Results

Overall structure of the complex

The HMG domain of HMG-D formed the best quality co-crystals with a blunt-ended linear palindromic decamer of sequence GCGATATCGC. The estimated affinity of the protein for this short DNA fragment is in the micromolar range based on competitive electrophoretic mobility shift assays performed with DNA fragments of similar length and sequence (Churchill *et al*., 1995). The structure of the complex was determined using multiple wavelength anomalous dispersion (MAD) methods (Table I). The final refined structure is reported to 2.2 Å resolution (PDB accession No. 1qrv) with excellent stereochemistry (Table I). The quality of the data is illustrated by the electron density map of the protein–DNA interface, which shows well-ordered DNA bases, DNA backbone, amino acid side chains and water molecules (Figure 1B).

HMG-D binds in the minor groove on the outside of a smoothly bent AT-rich DNA sequence, consistent with previous DNA binding and footprinting studies (Churchill *et al*., 1995, 1999; Wolfe *et al*., 1995), modeling studies of the HMG-D–DNA complex (Balaeff *et al*., 1998)

Table II. DNA structure parameters

Base pair	Roll	Twist	Major groove		Minor groove		Sugar pucker
	(°)[a]	(°)[a]	Width (Å)[b]	Depth (Å)[b]	Width (Å)[b]	Depth (Å)[b]	
G1/C20	3.71	28.89					C2′-endo/C2′-endo
C2/G19	6.74	34.65			8.43	3.19	C2′-endo/C2′-endo
G3/C18	32.00	9.88	10.06	6.13	10.13	1.64	C2′-endo/C2′-endo
A4/T17	21.12	11.55	10.81	2.56	12.19	0.16	O1′-endo/C1′-exo
T5/A16	48.47	22.11	7.81	7.24	11.80	0.27	C4′-exo/C3′-exo
A6/T15	9.19	28.93	8.41	7.24	8.98	2.99	C2′-endo/C1′-exo
T7/A14	7.06	34.37	11.28	5.51	8.33	3.36	C2′-endo/C2′-endo
C8/G13	11.04	35.48			7.64	4.30	C2′-endo/C2′-endo
G9/C12	7.83	24.23					C2′-endo/C2′-endo
C10/G11							C3′-exo/C2′-endo

[a]The reported value is the global inter-base pair parameter for the base step starting with the assigned pair.
[b]The reported value is the average for the base pair, with the number of levels measured determined automatically by Curves 5.3 (Lavery and Sklenar, 1989).

and the structures of sequence-specific HMG-box–DNA complexes (Love *et al*., 1995; Werner *et al*., 1995b). Although HMG-D has a slight preference for AT-rich sequences containing TG dinucleotides, it binds with only slightly lower affinity to DNA sequences that lack TG, but are still AT-rich, presumably because the deformability of the sequence is more important. The characteristic L-shaped fold, comprised of three helices held together by two hydrophobic cores, is conserved upon binding to DNA (Figure 1C). All of the helices are relatively straight, except for helix I (see below). This is in contrast to the NMR structures and modeling studies of chromosomal HMG domains, in which helix III is usually bent near position 59 (Weir *et al*., 1993; Jones *et al*., 1994; Balaeff *et al*., 1998; Allain *et al*., 1999). Helix III is straight in both HMG-D molecules in the crystal asymmetric unit, despite different crystal packing environments, which suggests that the bend observed in the solution structures and modeling studies may be due to the effects of molecular dynamics calculations or solution dynamics.

The DNA in the HMG-D–DNA complex is severely distorted from a B-DNA conformation. The binding site extends from C2 to C8, but toward the middle of the binding site there are dramatic deviations from a canonical B-DNA structure, illustrated by the helical parameters in Table II. The sugar puckers and helix axis path are most similar to B-DNA, but the groove width parameters more closely resemble those of A-DNA. Overall, the DNA is severely bent and underwound, with a bend angle of 111.1° and average twist angle of 27°. The shallow and widened minor groove is 12.4 Å across at the widest point, creating a tight hydrophobic interface with HMG-D and burying a surface area of 905 Å^2. This bend angle is larger than that estimated for other chromosomal HMG domain proteins, which could be partly due to crystal packing interactions at the ends of the DNA. However, a large bending angle is consistent with solution experiments with HMG-D. For example, the smallest DNA circle that has been formed in the presence of HMG-D is 55 bp, considerably smaller (one DNA turn) than those reported for HMG1 (Paull *et al*., 1993) and NHP6A (Yen *et al*., 1998). This bend angle is similar in magnitude to that observed in the LEF-1–DNA complex (see below) (Love *et al*., 1995).

At 2.2 Å resolution, solvent molecules are well resolved and reveal a previously unknown role for structural water molecules in the HMG-box proteins; several well-ordered (*B*-values <32 Å^2) water molecules bridge key structural features of the protein (Figure 1C). Water molecules bridge backbone atoms of two pairs of residues in helix I, Asn17 O to Glu21 N and Ser18 O to Ser22 N, where a bend has been observed in all known HMG domain structures, with the exception of the recent HMG1–cisplatin-modified DNA complex structure (Ohndorf *et al*., 1999). Another water molecule bridges the N- and C-termini of the domain through residues Leu9 O to Ala56 O, which may help to stabilize the secondary hydrophobic core of the protein formed by residues 2–9 and 50–74. Two other water molecules appear to stabilize the loops joining helices I to II and II to III, through residues Lys24 O to Gly29 N and Asp48 O to Glu51 N.

Protein–DNA interactions

The structure of the HMG-D–DNA complex shows that the protein changes very little upon binding to linear DNA. The root-mean-squared deviation (r.m.s.d.) between the main chain atoms (residues 5–28, 30–46 and 49–69) of the bound HMG-D and the free protein structures, HMG-D determined by NMR (Jones *et al*., 1994), HMG1 box A (Hardman *et al*., 1995) and HMG1 box B (Weir *et al*., 1993) are 1.2, 1.7 and 1.5 Å, respectively (Figure 2A).

The protein–DNA interface is dominated by van der Waals interactions, with additional hydrogen bonds and electrostatic interactions with the DNA backbone as well as water-mediated contacts (Figure 2B). Six residues, Arg7, Asn17, Arg20, Ala36, Arg44 and Lys60, make direct hydrogen bonds or ion pairs with the DNA backbone. Water-mediated hydrogen bonds are observed to the phosphate backbone for Pro8, Tyr12, Arg20, Val32 and Lys49, and to the bases for Ser10 and Arg20. Direct hydrogen bonds to the bases are observed for residues Arg7 and Tyr12. The hydrophobic interactions at the interface with the base pairs are dominated by partially intercalating residues. Met13 intercalates at base step 5–6 as expected (Figure 2C). Owing to their proximity to the center of the base pair, and the associated base pair roll, Val32 and Thr33 partially intercalate the DNA at base step 3–4, and Ala36 partially intercalates at base step 4–5. This close interaction surface was not necessarily expected for non-

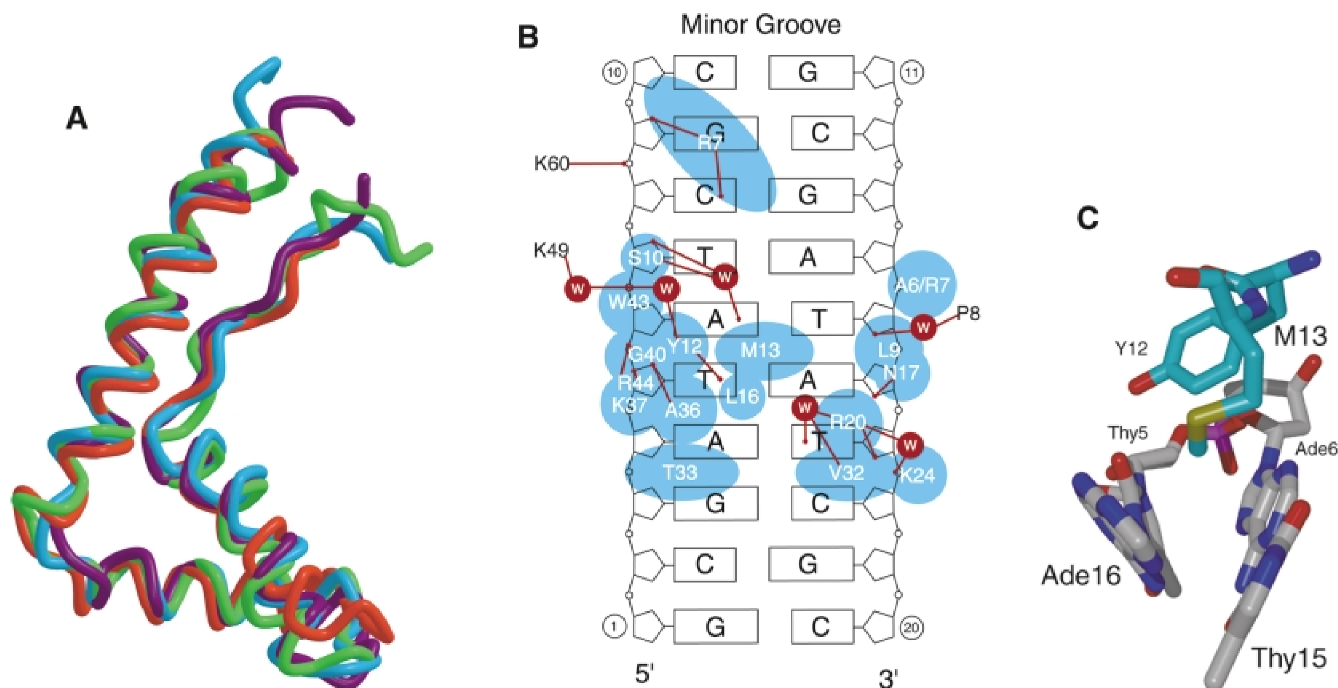

Fig. 2. HMG-D–DNA contacts. (**A**) The HMG-D protein from this structure (cyan) superimposed on the NMR structures of HMG-D free (Jones *et al*., 1994) (PDB accession No. 1hma; violet) and HMG1 box B (Weir *et al*., 1993) (PDB accession No. 1hmf; green), and the X-ray crystal structure of HMG1 box A from the structure of HMG1 box A bound to cisplatin-modified DNA (Ohndorf *et al*., 1999) (PDB accession No. 1ckt; red). (**B**) Ladder diagram of the DNA minor groove showing HMG-D residues with significant van der Waals DNA contacts mapped to the regions of DNA interaction. Proposed hydrogen bonds, electrostatic interactions and water molecules are illustrated as red lines and spheres, respectively. (**C**) Detailed view of the base step between base pairs 5 and 6 (roll angle of 36°) with the C_ε of the intercalating residue Met13 a distance of 3.8 Å from the closest base center.

sequence-specific DNA-binding proteins (reviewed in Revzin, 1990). The large buried surface area with the DNA bases and numerous side chain–non-phosphate DNA contacts prove that non-sequence-specific proteins interact with DNA using the same general mode of binding as homologous sequence-specific proteins.

The normal and structure-specific DNA-binding modes of the HMG domain are different

The structure of the architecture-specific complex of HMG1 box A bound to cisplatin-modified DNA (Ohndorf *et al*., 1999) is surprisingly different from that of HMG-D bound to linear DNA. The protein is the most similar part, with an r.m.s.d. of 2.1 Å for the backbone atoms of the HMG domain (residues 5–28, 30–46 and 49–69 of HMG-D). This r.m.s.d. is significantly greater than any of those between HMG-D and the free HMG domains described above. The largest deviation is in the C-terminal portion of helix I, the N-terminal portion of helix II and the connecting loop, caused primarily by a smaller bend in helix I of HMG1 box A (Figure 2A). In contrast, the DNA structures are markedly different, with an r.m.s.d. for the backbone atoms of 5.0 Å. These differences can be attributed to the nature of the nucleic acid substrates used in the co-crystal structures. The structure-specific complex contains a cisplatin lesion, which creates a dramatic distortion at one point, which is recognized by the HMG1 box A. The protein–DNA interface is dominated by the interaction of Phe32 with the site of cisplatin distortion of the DNA duplex. In contrast, the non-sequence-specific duplex DNA is smoothly bent over several base pairs by HMG-D, with three sites of partial intercalation where bending of the linear duplex DNA is more pronounced. The structure-specific complex has a buried surface area at the DNA–protein interface of 864 Å^2, which is similar to the 905Å^2 buried by HMG-D. However, the van der Waals contacts made by HMG-D indicate a more intimate complex, with more than twice the number of amino acids contacting bases. Further, the interactions of HMG-D with the DNA are located more deeply within the minor groove than those of HMG1. Therefore, it appears that the HMG domain has enough flexibility to adopt two slightly different structures, individually suited to binding to two DNA fragments with very different structures.

Structural basis of HMG domain specificity

The comparison of the bound HMG-D protein with the LEF-1–DNA complex (Love *et al*., 1995) (Figure 3A) shows that the overall structure of the protein, with the exception of the C-terminus and loop regions, is quite similar in the two complexes, with r.m.s.d. values for the main chain atoms of 1.7 Å (residues 7–27, 30–46 and 51–69). The structure of the DNA is also similar, with r.m.s.d. values of 2.1 Å for the backbone atoms. The similarity of the two complexes is remarkable considering the differences in specificity. In both structures, the DNA bend is spread over several base steps, in contrast to the complex of HMG-1 bound to cisplatin-modified DNA, where the bend occurs at primarily one base step (Ohndorf *et al*., 1999).

The basis for differences in specificity between the

specific and non-specific HMG domains is evident from inspection of the few residues that differ between them (Figures 1A and 3A) (Baxevanis and Landsman, 1995; Balaeff *et al*., 1998; Churchill *et al*., 1999). We identify three specificity determinants, composed of residues at the protein–DNA interface with equivalent positions but different interactions. First is a single residue at position 10. This region of the HMG domain has been shown to affect DNA specificity (Read *et al*., 1994) and, to our knowledge, all sequence-specific HMG proteins have an asparagine at this position, whereas all non-sequence-specific HMG proteins have a serine (Baxevanis and Landsman, 1995). In HMG-D, Ser10 forms water-mediated hydrogen bonds to the thymine 7 ribose O4′ and adenine 6 N3 position of the DNA base (Figure 3B), unlike Asn10 in LEF-1, which directs a hydrogen bond network to three of the four bases of the CA base step, ensuring specificity for a CA dinucleotide (Figure 3C). The ability to accommodate other DNA sequences at this position is suggested because any base at this position could be bound by serine either directly for cytosine and thymine or through a water molecule for adenine and guanosine. It is also possible that in the context of other DNA sequences, there may be direct base N3 hydrogen bonds as suggested by the NHP6A–DNA model (Allain *et al*., 1999). However, both direct and water-mediated interactions with Ser10 would be sequence-neutral.

The second determinant includes Val32, which partially intercalates in the HMG-D–DNA structure. This residue is hydrophobic in all known non-sequence-specific proteins, but in the sequence-specific proteins it is hydrophilic (Balaeff *et al*., 1998). The equivalent residue in LEF-1, Ser32, is involved in a hydrogen bond network with Asn36, which together make direct hydrogen bonds to three of the four bases of an AA DNA base step providing further sequence specificity.

The third region is the C-terminal hydrophobic core. The DNA contacts made in this region by LEF-1 appear to include sequence-specific hydrogen bonds from Tyr76 to the guanosine and thymine bases of a CT base step. This interaction is not observed for HMG-D because this residue is not in the HMG domain. Further, the structure at the C-terminus differs in this region and is not near the DNA (Figure 1C). Therefore, different families of HMG domain proteins can bind to DNA in a similar fashion, and the degree of specificity appears to be controlled by changes in only a few direct protein–DNA interactions.

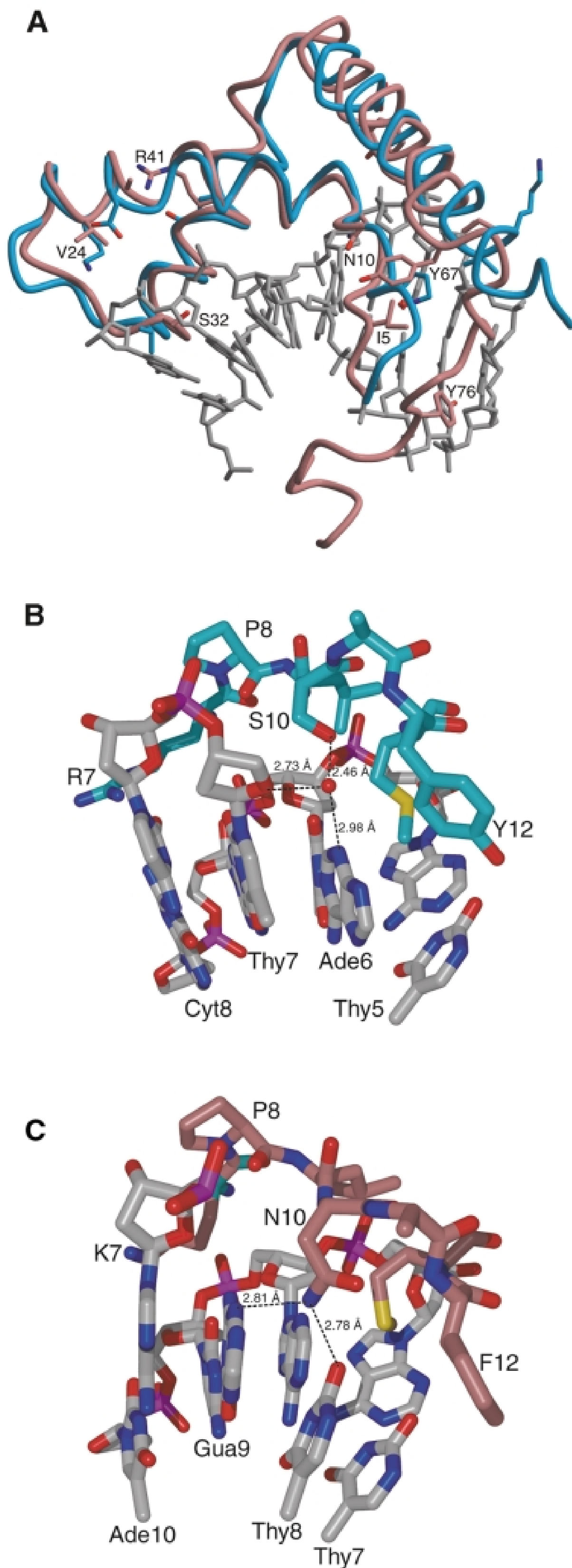

Fig. 3. Structural features involved in non-sequence-specific DNA recognition. (**A**) View of HMG-D protein from this structure (cyan) superimposed on the structure of the LEF-1–DNA complex (Love *et al*., 1995) (PDB accession No. 2lef; coral) in the same orientation as Figure 1C. Side chains, selected on the basis of their potential involvement in DNA specificity, are shown. Detailed view of the interaction of residue 10 from both HMG-D (**B**) and LEF-1 (**C**). HMG-D protein is in cyan, LEF-1 protein is in coral, DNA is in gray, and black dashed lines depict proposed hydrogen bonds with distances between donors and acceptors shown. The Ser10 hydroxyl oxygen of HMG-D makes water-mediated interactions with adenine 6 N3 and thymine 7 O4′. The LEF-1 Asn10 makes direct hydrogen bonds to guanine 9 N3 and thymine 8 O2 (in this LEF-1 model).

Discussion

This high-resolution structure of HMG-D bound to linear DNA reveals novel interactions that enable HMG proteins to bind to many different DNA sequences, a critical feature of their function. Several protein families use homologous

DNA-binding domains to recognize DNA with different specificities. These include both chromosomal protein–sequence-specific protein pairs, histone H5–HNF-3γ (Clark *et al*., 1993; Ramakrishnan *et al*., 1993), bacterial HU–integration host factor (IHF) (Tanaka *et al*., 1984; Rice *et al*., 1996) and the HMG domain superfamily. To date, only the structures of the free proteins of each group, the DNA-bound forms of the sequence-specific proteins and a structure-specific complex of HMG1 box A and cisplatin-modified DNA have been determined. Therefore, this structure of HMG-D bound to DNA provides a unique opportunity to examine closely which interactions differ and may be responsible for the differences in specificity observed for homologous proteins.

Comparison of the HMG-D–DNA structure with the structure of a homologous transcription factor (sequence-specific), LEF-1, reveals surprising similarity of the overall complexes (r.m.s.d. ~1.9 Å). This similarity made it possible to distinguish very clearly the precise side chain interactions that are responsible for the differences in DNA specificity observed between the proteins, insights that are unprecedented in the field of protein–nucleic acid recognition. The two sequence-neutral mechanisms that emerge from this study are multiple partial intercalations of varying degrees and the use of water-mediated base contacts at a sequence-neutral base position (Seeman *et al*., 1976). These interactions contribute to non-sequence-specific DNA binding by substituting for sequence-specific base contacts observed in the sequence-specific protein–DNA complex.

Partial intercalation of aliphatic and aromatic residues in the minor groove has been observed in several protein families, e.g. HMG proteins, TATA-box-binding protein (Kim *et al*., 1993a,b), Sac7d (Robinson *et al*., 1998) and IHF (Rice *et al*., 1996). In each case, intercalation appears at a DNA bend toward the major groove, and contributes significantly to a relatively large buried surface area and greater stability of the protein–DNA complex. Partial intercalation is important for DNA binding and bending by sequence-specific and non-sequence-specific HMG proteins (Yen *et al*., 1998; Allain *et al*., 1999; Ohndorf *et al*., 1999). The non-sequence-specific HMG domain proteins can intercalate the DNA at two major sites (residues 13 and 32/33) at non-adjacent base steps, whereas the sequence-specific proteins intercalate at one site (residue 13), making direct DNA base hydrogen bonds with the hydrophilic residue at the other site (residues 32/33). Interestingly, the predicted sites and residues involved in partial intercalation in the model structure of HMG-D bound to DNA (Balaeff *et al*., 1998) are correct for Met13 and Val32. However, Leu9 does not intercalate in this crystal structure, as was predicted in the model. The modeling study of NHP6A (Allain *et al*., 1999) also correctly predicted intercalation by Met13 and Phe32 (Val32 in HMG-D), but the relative base steps of intercalation are different from those observed in the crystal structure; Val32 of HMG-D partially intercalates two base steps away from Met13, but the intercalations occur at adjacent base steps in the NHP6A model structure. The structure-specific HMG1 box A–DNA complex has only one intercalating residue, which binds at the kinked base step. Intercalation by residue 13 is not observed because the DNA structure is different (as described above), and residue 13 in HMG1 box A (Figure 1A) is an alanine. The specificity of HMG domain proteins for cisplatin-modified DNA appears to be related to the size of the residue at position 32, because HMG1 box A has a greater preference for cisplatinated DNA than HMG1 box B and HMG-D, which have smaller residues at position 32.

Direct and water-mediated hydrogen bonds to DNA bases are generally considered to be sequence-specific interactions (Seeman *et al*., 1976; Pabo and Sauer, 1992). However, it is well known that the DNA minor groove has more degeneracy in hydrogen bonding donor and acceptor positions among possible base pairs than the major groove (Seeman *et al*., 1976). It appears that the chromosomal HMG domain proteins have capitalized on this feature of the minor groove by making water-mediated hydrogen bonds (or possibly direct hydrogen bonds, as predicted in the NHP6A model) to the base N3 position. This position is a hydrogen bond acceptor, and is well suited to direct or water-mediated interactions. Such a sequence-neutral hydrogen bond still provides binding energy, but does not restrict the sequences to which the protein can bind, unlike the base-specific hydrogen bonds made by the equivalent residue, Asn10 of LEF-1.

Structures of unmodified DNA bound to two other non-sequence-specific chromosomal protein complexes, the nucleosome core particle (Luger *et al*., 1997) and the Sac7d archaeal chromosomal proteins (Robinson *et al*., 1998), have been determined recently using X-ray crystallography. These structures revealed two different general modes of non-sequence-specific DNA binding. Electrostatic interactions dominate the core histone DNA interaction on the inside of the DNA bend, whereas hydrophobic and intercalation interactions dominate the Sac7d DNA interaction on the outside of the DNA bend. The general mode of non-sequence-specific binding observed in this structure is similar to Sac7d. The multiple partially intercalating residues and DNA contacts mediated by water seen in the HMG-D structure are also observed in Sac7d–DNA complexes, suggesting that these mechanisms will be generally observed for chromosomal proteins that bend DNA by binding to the minor groove.

Materials and methods

Structure determination

Co-crystallization of the HMG-D–DNA complex with the duplex DNA decamer, GCGATATCGC, and native data collection are described elsewhere (Murphy *et al*., 1999). The complex crystallized in the space group $P2_12_12_1$ with a = 43.74 Å; b = 53.797 Å; c = 86.843 Å, with one DNA duplex and two protein molecules in the asymmetric unit. One protein makes canonical contacts with the DNA, and the other binds at the DNA junction formed by the DNA and a symmetry-related DNA duplex. The initial phases for model building were obtained using MAD methods with co-crystals that contained a 5-bromocytosine at position 8 in the DNA sequence. Four wavelengths of MAD data were collected at beamline X-12C at the National Synchrotron Light Source (Brookhaven, NY). The data were processed using DENZO and SCALEPACK (Otwinowski and Minor, 1997) (Table I). Intensities were processed and initial MAD phases were calculated using the program SOLVE (Terwilliger and Berendzen, 1999) followed by phase refinement using the program SHARP (Fortelle and Bricogne, 1997). Solvent flattening using a solvent content of 45% with the program Solomon (Abrahams and Leslie, 1996) yielded interpretable maps of the main core of the complex. The initial model was built into the 2.4 Å MAD experimental maps using O (Jones *et al*., 1991) and refined using CNS 0.1 (Brünger *et al*., 1998). Cycles of model building with O and refinement with CNS were then performed against native data using

cross-validated maximum likelihood techniques throughout, with the R_{free} calculations performed on 5% of the reflections. The canonical protein model includes amino acids 2–74, with S2 and K6 built as alanines, the second protein includes residues 3–73 (r.m.s.d. between the two proteins is 0.8 Å), and there is an alternative conformer for the nucleoside Cyt10.

Structure analysis

DNA parameters were calculated using the program CURVES (Lavery and Sklenar, 1989). Intercalation distances and roll angles were calculated using customized XPLOR (Brünger, 1992) scripts (Balaeff *et al*., 1998). Interatomic distances and hydrogen bonds were determined using CONTACT and accessible surface area calculations were performed using AREAIMOL, both distributed in the CCP4 package (Bailey, 1994). R.m.s.d. values were calculated using LSQKAB (Kabsch, 1976), with values reported for the most similar model when comparing the structure with an NMR ensemble. The proteins were aligned along the three helices of the HMG domain as follows: HMG-D: 5–28, 30–46, 49–69; HMG-D free (1hma): 5–28, 30–46, 49–69; HMG1 box A (1aab): 8–31, 35–51, 56–76; HMG1 box B (1hmf): 7–30, 32–48, 53–73; SRY (1hrz): 5–28, 30–46, 51–71; HMG1 box A bound to cisplatin-modified DNA (1ckt): 8–31, 37–49, 56–76; and LEF-1 (2lef): 5–25, 28–44, 51–71 (LEF-1 had to be compared with HMG-D residues 7–27, 30–46, 51–69). The DNA segments to be compared were selected by spatial proximity to the bound DNA of HMG-D following alignment of bound proteins. The regions of the DNA selected for comparisons are: LEF-1 B6–12 and C4–10 with HMG-D C4–10 and D11–17; SRY 1–8 and 9–16 with HMG-D C3–10 and D11–16; HMG1 box A B107–116 and C201–210 with HMG-D C1–10 and D11–20.

Figures

Figures were constructed using O (Jones *et al*., 1991) (Figure 1B), Molscript (Kraulis, 1991) and Raster3D (Merritt and Bacon, 1997) (Figures 1C, 2A and 3A); VMD (Humphrey *et al*., 1996) and Raster3D (Merritt and Bacon, 1997) (Figures 2C, 3B and 3C).

Coordinate deposition

Atomic coordinates have been deposited in the Protein Data Bank under accession No. 1qrv.

Acknowledgements

We appreciate the helpful suggestions and contributions of Drs Hillary Nelson and Venki Ramakrishnan during the phasing and refinement of this structure, Dr David Jones during preparation of the manuscript, and members of the Churchill laboratory, including Dr Linda Dow for protein biochemistry and Ms Heidi Wanaski for technical assistance. The HKL software package is supported by NIH grant GM-53163 to Z.Otwinowski and W.Minor. Beamline X12-C in the National Synchrotron Light Source is supported by the US DOE, the NSF and the NIH (RR12408-01A1). We are grateful for support from the NIH (Shannon Award and RO1-GM59456), American Cancer Society and the American Heart Association for a Grant-In-Aid to M.E.A.C., and from the NIH Molecular Biophysics pre-doctoral training grant and a University of Illinois Graduate College Travel Grant to F.V.M.

References

Abrahams,J.P. and Leslie,A.G.W. (1996) Methods used in the structure determination of bovine mitochondrial F1 ATPase. *Acta Crystallogr. D*, **52**, 30–42.

Allain,F.H.-T., Yen,Y.-M., Masse,J.E., Schultze,P., Dieckmann,T., Johnson,R.C. and Feigon,J. (1999) Solution structure of the HMG protein NHP6A and its interaction with DNA reveals the structural determinants for non-sequence-specific binding. *EMBO J.*, **18**, 2563–2579.

Bailey,S. (1994) The CCP4 suite—programs for protein crystallography. *Acta Crystallogr. D*, **50**, 760–763.

Balaeff,A., Churchill,M.E.A. and Schulten,K. (1998) Structure prediction of a complex between the chromosomal protein HMG-D and DNA. *Proteins Struct. Funct. Genet.*, **30**, 113–135.

Baxevanis,A.D. and Landsman,D. (1995) The HMG-1 box protein family: classification and functional relationships. *Nucleic Acids Res.*, **23**, 1604–1613.

Bianchi,M.E., Falciola,L., Ferrari,S. and Lilley,D.M.J. (1992) The DNA binding site of HMG1 protein is composed of 2 similar segments (HMG boxes), both of which have counterparts in other eukaryotic regulatory proteins. *EMBO J.*, **11**, 1055–1063.

Boonyaratanakornkit,V. *et al*. (1998) High-mobility group chromatin proteins 1 and 2 functionally interact with steroid hormone receptors to enhance their DNA binding *in vitro* and transcriptional activity in mammalian cells. *Mol. Cell. Biol.*, **18**, 4471–4487.

Brünger,A.T. (1992) *X-PLOR Manual Version 3*.1. Yale University, New Haven, CT.

Brünger,A.T. *et al*. (1998) Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D*, **54**, 905–921.

Bustin,M. and Reeves,R. (1996) High-mobility-group chromosomal proteins: architectural components that facilitate chromatin function. *Prog. Nucleic Acid Res. Mol. Biol.*, **54**, 35–100.

Churchill,M.E.A., Jones,D.N.M., Glaser,T., Hefner,H., Searles,M.A. and Travers,A.A. (1995) HMG-D is an architecture-specific protein that preferentially binds to DNA containing the dinucleotide TG. *EMBO J.*, **14**, 1264–1275.

Churchill,M.E.A., Changela,A., Dow,L.K. and Krieg,A.J. (1999) Interactions of high mobility group box proteins with DNA and chromatin. *Methods Enzymol.*, **304**, 99–133.

Clark,K.L., Halay,E.D., Lai,E.S. and Burley,S.K. (1993) Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone-H5. *Nature*, **364**, 412–420.

David-Cordonnier,M.H., Payet,D., D'Halluin,J.C., Waring,M.J., Travers,A.A. and Bailly,C. (1999) The DNA-binding domain of human c-Abl tyrosine kinase promotes the interaction of a HMG chromosomal protein with DNA. *Nucleic Acids Res.*, **27**, 2265–2270.

Fortelle,E. and Bricogne,G. (1997) Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.*, **276**, 472–494.

Ge,H. and Roeder,R.G. (1994) The high mobility group protein HMG1 can reversibly inhibit class II gene transcription by interaction with the TATA-binding protein. *J. Biol. Chem.*, **269**, 17136–17140.

Giese,K., Amsterdam,A. and Grosschedl,R. (1991) DNA-binding properties of the HMG domain of the lymphoid-specific transcriptional regulator LEF-1. *Genes Dev.*, **5**, 2567–2578.

Giese,K., Cox,J. and Grosschedl,R. (1992) The HMG domain of lymphoid enhancer factor-1 bends DNA and facilitates assembly of functional nucleoprotein structures. *Cell*, **69**, 185–195.

Goodfellow,P.N. and Lovellbadge,R. (1993) Sry and sex determination in mammals. *Annu. Rev. Genet.*, **27**, 71–92.

Grosschedl,R., Giese,K. and Pagel,J. (1994) HMG domain proteins—architectural elements in the assembly of nucleoprotein structures. *Trends Genet.*, **10**, 94–100.

Haqq,C.M., King,C.Y., Ukiyama,E., Falsafi,S., Haqq,T.N., Donahoe,P.K. and Weiss,M.A. (1994) Molecular basis of mammalian sexual determination: activation of Müllerian inhibiting substance gene expression by SRY. *Science*, **266**, 1494–1500.

Hardman,C.H., Broadhurst,R.W., Raine,A.R.C., Grasser,K.D., Thomas,J.O. and Laue,E.D. (1995) Structure of the A-domain of HMG1 and its interactions with DNA as studied by heteronuclear three- and four-dimensional NMR spectroscopy. *Biochemistry*, **34**, 16596–16607.

Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD—visual molecular dynamics. *J. Mol. Graphics*, **14**, 33–38.

Jayaraman,L., Moorthy,N.C., Murthy,K.G.K., Manley,J.L., Bustin,M. and Prives,C. (1998) High mobility group protein-1 (HMG-1) is a unique activator of p53. *Genes Dev.*, **12**, 462–472.

Jones,D.N.M., Searles,A., Shaw,G.L., Churchill,M.E.A., Ner,S.S., Keeler,J., Travers,A.A. and Neuhaus,D. (1994) The solution structure and dynamics of the DNA-binding domain of HMG-D from *Drosophila melanogaster. Structure*, **2**, 609–627.

Jones,T.A., Zou,J.Y., Cowan,S.W. and Kjelgaard,M. (1991) Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A*, **47**, 110–119.

Kabsch,W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, **A32**, 922–923.

Kim,J.L., Nikolov,D.B. and Burley,S.K. (1993a) Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature*, **365**, 520–527.

Kim,Y.C., Geiger,J.H., Hahn,S. and Sigler,P.B. (1993b) Crystal structure of a yeast TBP TATA-box complex. *Nature*, **365**, 512–520.

King,C.-Y. and Weiss,M.A. (1993) The SRY high-mobility-group box

recognizes DNA by partial intercalation in the minor groove: a topological mechanism of sequence specificity. *Proc. Natl Acad. Sci. USA*, **90**, 11990–11994.
Kornberg,R.D. and Lorch,Y. (1992) Chromatin structure and transcription. *Annu. Rev. Cell Biol*., **8**, 563–687.
Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, **24**, 946–950.
Lavery,R. and Sklenar,H. (1989) Defining the structure of irregular nucleic acids: conventions and principles. *J. Biomol. Struct. Dynam*., **6**, 655–667.
Love,J.J., Li,X., Case,D.A., Giese,K., Grosschedl,R. and Wright,P.E. (1995) Structural basis for DNA bending by the architectural transcription factor LEF-1. *Nature*, **376**, 791–795.
Luger,K., Mader,A.W., Richmond,R.K., Sargent,D.F. and Richmond,T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, **389**, 251–260.
Merritt,E.A. and Bacon,D.J. (1997) Raster3D: photorealistic molecular graphics. *Methods Enzymol.*, **277**, 505–524.
Murphy,F.V., Sehy,J.V., Gao,Y.-G., Dow,L.K. and Churchill,M.E.A. (1999) Co-crystallization and preliminary crystallographic analysis of the high mobility group domain of HMG-D bound to DNA. *Acta Crystallogr. D*, **55**, 1594–1597.
Ner,S.S. and Travers,A.A. (1994) HMG-D, the *Drosophila melanogaster* homologue of HMG 1 protein is associated with early embryonic chromatin in the absence of histone H1. *EMBO J.*, **13**, 1817–1822.
Nightingale,K., Dimitrov,S., Reeves,R. and Wolffe,A. (1996) Evidence for a shared structural role for HMG1 and linker histones B4 and H1 in organizing chromatin. *EMBO J.*, **15**, 548–561.
Ohndorf,U.-M., Rould,M.A., He,Q., Pabo,C.O. and Lippard,S.J. (1999) Basis for recognition of cisplatin-modified DNA by high-mobility-group proteins. *Nature*, **399**, 708–712.
Otwinowski,Z. and Minor,W. (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.*, **276**, 307–326.
Pabo,C. and Sauer,R. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem*., **61**, 1053–1095.
Paull,T.T., Haykinson,M.J. and Johnson,R.C. (1993) The nonspecific DNA-binding and -bending proteins HMG1 and HMG2 promote the assembly of complex nucleoprotein structures. *Genes Dev.*, **7**, 1521–1534.
Payet,D. and Travers,A.A. (1997) The acidic tail of the high mobility group protein HMG-D modulates the structural selectivity of DNA binding. *J. Mol. Biol.*, **266**, 66–75.
Pil,P. and Lippard,S.J. (1992) Specific binding of chromosomal protein HMG-1 to DNA damaged by the anticancer drug cisplatin. *Science*, **256**, 234–237.
Ramakrishnan,V., Finch,J.T., Graziano,V., Lee,P.L. and Sweet,R.M. (1993) Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature*, **362**, 219–223.
Read,C.M., Cary,P.D., Crane-Robinson,C., Driscoll,P.C. and Norman,D.G. (1993) Solution structure of a DNA-binding domain from HMG1. *Nucleic Acids Res.*, **21**, 3427–3436.
Read,C.M., Cary,P.D., Crane-Robinson,C., Driscoll,P.C., Carrillo, M.O.M. and Norman,D.G. (1995) The structure of the HMG box and its interactions with DNA. In Ekstein,F. and Lilley,D.M.J. (eds), *Nucleic Acids and Molecular Biology*. Springer-Verlag, Berlin, Vol. 9, pp. 222–250.
Read,C.M., Cary,P.D., Preston,N.S., Lnenicek-Allen,M. and Crane-Robinson,C. (1994) The DNA sequence-specificity of HMG boxes lies in the minor wing of the structure. *EMBO J.*, **13**, 5639–5646.
Revzin,A. (1990) *Non-specific DNA–protein Interactions*. CRC Press, Boca Raton, FL.
Rice,P.A., Yang,S.W., Mizuuchi,K. and Nash,H.A. (1996) Crystal structure of an IHF–DNA complex: a protein-induced DNA U-turn. *Cell*, **87**, 1295–1306.
Robinson,H., Gao,Y.-G., McCrary,B.S., Edmondson,S.P., Shriver,J.W. and Wang,A.H.-J. (1998) The hyperthermophile chromosomal protein Sac7d sharply kinks DNA. *Nature*, **392**, 202–205.
Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA*, **73**, 804–805.
Shykind,B.M., Kim,J. and Sharp,P.A. (1995) Activation of the TFIID–TFIIA complex with HMG-2. *Genes Dev.*, **9**, 1354–1365.
Tanaka,I., Appelt,K., Dijk,J., White,S.W. and Wilson,K.S. (1984) 3-Å resolution structure of a protein with histone-like properties in prokaryotes. *Nature*, **310**, 376–381.
Teo,S.H., Grasser,K.D. and Thomas,J.O. (1995) Differences in the DNA binding properties of the HMG domains of HMG1 and the sex determining factor SRY. *Eur. J. Biochem*., **230**, 943–950.
Terwilliger,T.C. and Berendzen,J. (1999) Automated structure solution for MIR and MAD. *Acta Crystallogr. D*, **55**, 849–861.
van Gent,D.C., Hiom,K., Paull,T.T. and Gellert,M. (1997) Stimulation of V(D)J cleavage by high mobility group proteins. *EMBO J.*, **16**, 2665–2670.
Wagner,C.R., Hamana,K. and Elgin,S. (1992) A high-mobility-group protein and its cDNAs from *Drosophila melanogaster. Mol. Cell. Biol.*, **12**, 1915–1923.
Weir,H.M., Kraulis,P.J., Hill,C.S., Raine,A., Laue,E.D. and Thomas,J.O. (1993) Structure of the HMG box motif in the B-domain of HMG1. *EMBO J.*, **12**, 1311–1319.
Werner,M.H., Bianchi,M.E., Gronenborn,A.M. and Clore,G.M. (1995a) NMR spectroscopic analysis of the DNA conformation induced by the human testis determining factor SRY. *Biochemistry*, **34**, 11998–12004.
Werner,M.H., Huth,J.R., Gronenborn,A.M. and Clore,G.M. (1995b) 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.
Wolfe,S.A., Ferentz,A.E., Grantcharova,V., Churchill,M.E.A. and Verdine,G.L. (1995) Modifying the helical structure of DNA by design: recruitment of an architecture-specific protein to an enforced DNA bend. *Chem. Biol.*, **2**, 213–221.
Yen,Y.M., Wong,B. and Johnson,R.C. (1998) Determinants of DNA binding and bending by the *Saccharomyces cerevisiae* high mobility group protein NHP6A that are important for its biological activities: role of the unique N-terminus and putative intercalating methionine. *J. Biol. Chem.*, **273**, 4424–4435.
Zappavigna,V., Falciola,L., Citterich,M., Mavilio,F. and Bianchi,M. (1996) HMG1 interacts with HOX proteins and enhances their DNA binding and transcriptional activation. *EMBO J*., **15**, 4981–4991.
Zwilling,S., Konig,H. and Wirth,T. (1995) High-mobility-group protein-2 functionally interacts with the POU domains of octamer transcription factors. *EMBO J*., **14**, 1198–1208.

Received August 31, 1999; revised and accepted October 11, 1999