The EMBO Journal vol.12 no.4 pp.1311-1319, 1993

Structure of the HMG box motif in the B-domain of HMG1

Hazel M.Weir, Per J.Kraulis, Caroline S.Hill[1], Andrew R.C.Raine, Ernest D.Laue and Jean O.Thomas[2]

Centre for Molecular Recognition, Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, UK
[1]Present address: Imperial Cancer Research Fund Laboratories, PO Box 123, 44 Lincoln's Inn Fields, London WC2A 3PX, UK
[2]Corresponding author

Communicated by J.O.Thomas

The conserved, abundant chromosomal protein HMG1 consists of two highly homologous, folded, basic DNA-binding domains, each of ~80 amino acid residues, and an acidic C-terminal tail. Each folded domain represents an 'HMG box', a sequence motif recently recognized in certain sequence-specific DNA-binding proteins and which also occurs in abundant HMG1-like proteins that bind to DNA without sequence specificity. The HMG box is defined by a set of highly conserved residues (most distinctively aromatic and basic) and appears to define a novel DNA-binding structural motif. We have expressed the HMG box region of the B-domain of rat HMG1 (residues 88-164 of the intact protein) in *Escherichia coli* and we describe here the determination of its structure by 2D ^{1}H-NMR spectroscopy. There are three α-helices (residues 13-29, 34-48 and 50-74), which together account for ~75% of the total residues and contain many of the conserved basic and aromatic residues. Strikingly, the molecule is L-shaped, the angle of ~80° between the two arms being defined by a cluster of conserved, predominantly aromatic, residues. The distinctive shape of the HMG box motif, which is distinct from hitherto characterized DNA-binding motifs, may be significant in relation to its recognition of four-way DNA junctions.
Key words: 2D-NMR/DNA-binding motif/HMG1/HMG box/non-histone protein

Introduction

HMG1 and HMG2 belong to the 'high mobility group' of abundant non-histone chromosomal proteins and occur at a level of approximately one copy per 10-15 nucleosomes on average (Johns, 1982). They have been variously implicated in replication and transcription but their role remains elusive (reviewed by Bustin *et al.*, 1990). Their relative abundance, sequence conservation between species and apparent lack of sequence specificity in binding to DNA, suggests that they might perform some general function in chromatin, for example a structural role or possibly as general transcription factors.

HMG1 and HMG2 have a tripartite structure initially defined by limited proteolysis in 'structuring conditions' at high ionic strength (Reeck *et al.*, 1982; Carballo *et al.*, 1983; Cary *et al.*, 1983; Abdul-Razzak *et al.*, 1989). The N-terminal A-domain and central B-domain, each of 80-90 amino acid residues, are basic and ~30% identical (~43% homologous) in amino acid sequence. The highly acidic C-terminal C-domain, which is slightly shorter in HMG2 than in HMG1 (Shirakawa *et al.*, 1990), contains ~30 consecutive aspartic or glutamic acid residues. The proteins recognize both histones and DNA *in vitro*, through their acidic and basic regions respectively (Carballo *et al.*, 1983). They have been reported to show a preference for single-stranded DNA, to bind selectively to regions that are low-melting, in some studies to cause unwinding of double-stranded DNA (for references see Bustin *et al.*, 1990), and most recently to bind preferentially to DNA four-way junctions (Bianchi *et al.*, 1989). Very little is known about their binding sites in chromatin, except that there appears to be no obvious preferential association with either active or inactive genes (Postnikov *et al.*, 1991).

Regions of ~80 amino acid residues homologous to the A and B domains of HMG1 (the so called 'HMG box motif') have recently been recognized in other proteins, first in a human nucleolar transcription factor, hUBF (Jantzen *et al.*, 1990), which has several HMG boxes, then in the testis-determining factor SRY (Sinclair *et al.*, 1990) and subsequently in a number of other sequence-specific transcription factors including the lymphoid enhancer factor LEF-1 (for references see Ner, 1992). The HMG boxes in LEF-1 and SRY are able to bend DNA *in vitro* through a large angle (130° and 85°, respectively) (Ferrari *et al.*, 1992; Giese *et al.*, 1992), and this may also turn out to be a property of HMG1 (see Lilley, 1992). The SRY and HMG1 boxes both bind preferentially to four-way DNA junctions (Bianchi *et al.*, 1992; Ferrari *et al.*, 1992).

The most distinctive feature of the HMG box is a set of conserved aromatic and basic amino acid residues (Figure 1); the HMG box of abundant HMG1-like proteins (Ner, 1992), which bind to DNA without sequence specificity, contains additional conserved amino acids relative to the box of transcription factors, including two proline residues. No structural information has been reported for the HMG box motif and neither is there any clue as to the basis of sequence-specific versus non-specific binding. We describe here the determination, by 2D ^{1}H-NMR spectroscopy, of the 3D structure of the HMG box of the B domain of rat HMG1.

Results

Expression and characterization of the HMG box of the HMG1 B-domain

Based on sequence alignment, a fragment containing residues 84-184 of rat HMG1 was initially expressed in *Escherichia coli*. However, preliminary 2D-NMR studies showed many overlapping resonances in the NH-C_αH region. Tryptic digestion resulted in the stable fragment Phe88-Lys164,

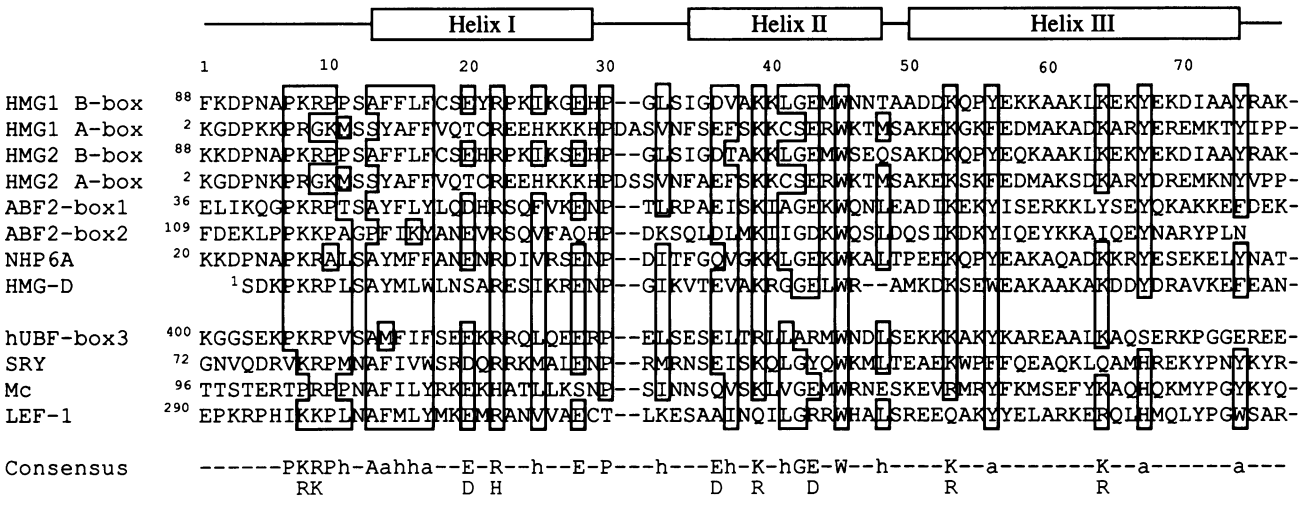

Fig. 1. Protein sequence alignment of representative HMG boxes. The top eight represent HMG1 and HMG1-like proteins that bind DNA non-specifically and the bottom four members of the sequence-specific HMG box transcription factors (Ner, 1992). The numbering at the top refers to the HMG1 B-domain box. Residue 1 of the box corresponds to residue 88 in the intact protein (Bianchi *et al.*, 1989). Amino acid positions within the parent proteins are shown at the beginning of each line. Gaps have been introduced for optimal alignment. Boxed residues in the sequence are identical or conserved in at least eight of the 12 sequences; the consensus sequence is shown below. Aromatics (a) include Y, F, W and H; hydrophobics (h) include V, I, A, L and F. For ease of reference, the positions of the three helices in the HMG1 B-domain box structure determined here by NMR (see Figure 4) are shown above the sequence. Abbreviations: HMG1 and HMG2, high mobility group proteins 1 (from rat) (Bianchi *et al.*, 1989) and 2 (from pig) (Shirakawa *et al.*, 1990); ABF2, *Saccharomyces cerevisiae* ARS binding factor (Diffley and Stillman, 1991); NHP6A, *S.cerevisiae* non-histone protein A (Kolodrubetz and Burgum, 1990); HMG-D, *Drosophila melanogaster* high mobility group protein D (Wagner *et al.*, 1992) also called HMG-N (Ner, 1992); hUBF, human upstream binding factor for RNA polymerase I (Jantzen *et al.*, 1990); SRY, human testis-determining factor (Sinclair *et al.*, 1990): Mc, mating type protein of *Schizosaccharomyces pombe* (Kelly *et al.*, 1988); LEF-1, murine lymphoid enhancer factor 1 (Travis *et al.*, 1991).

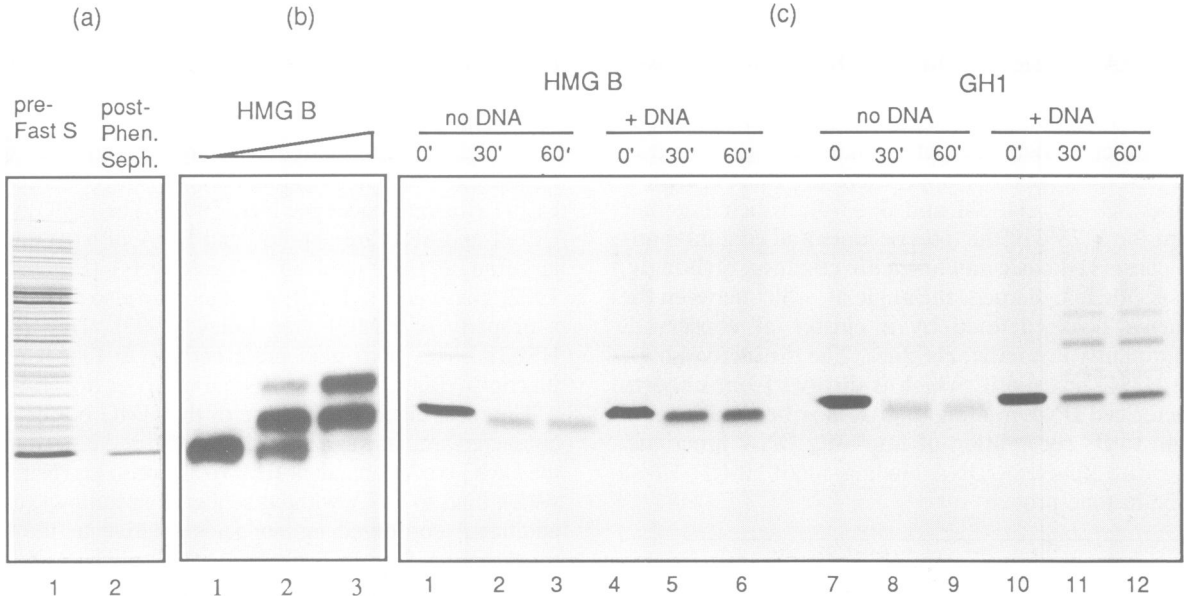

Fig. 2. Purification and characterization of the HMG box fragment (residues 88–164) of the HMG1 B-domain. (**a**) Lane 1, whole cell extract before application to a Fast S Sepharose column; lane 2, purified HMG box fragment after Phenyl Sepharose chromatography. (**b**) Gel retardation assay. Different concentrations of purified fragment (88–164) of the HMG B-domain were incubated with 5′ ^{32}P-labelled four-way junction DNA (2.5 nM) in a final volume of 10 μl and the mixtures analysed in a 5% polyacrylamide gel. The autoradiograph is shown. Lane 1, DNA alone; lanes 2 and 3, 2.8 and 5.6 μM HMG B-domain fragment, respectively. (**c**) Cross-linking assay. Treatment of the HMG1 B-domain fragment and GH1 with suberic acid *bis*(*N*-hydroxysuccinimide ester) in 20 mM sodium phosphate, pH 8. Lanes 1–3, B-domain (no DNA) in the presence of 0.15 M NaCl treated for 0, 30 and 60 min; lanes 4–6, B-domain (plus DNA, no NaCl) treated for the same times. Lanes 7–9, GH1 but otherwise as lanes 1–3; lanes 10–12, GH1 but otherwise as lanes 4–6. (The slightly increased mobility of the HMG B-domain fragment in lanes 2 and 3, and 5 and 6, compared with 1 and 4, is due to lysine modification; the effect is less in the presence of DNA due to lysine protection. A similar effect is apparent for GH1 in lanes 8 and 9, but is substantially less in lanes 11 and 12 due to DNA binding.)

whose NMR spectra were far less crowded. This well defined fragment represents a 'minimal' B-domain containing the HMG box of the HMG1 B-domain; its sequence is shown in Figure 1 (top line). The fragment was expressed in *E.coli* in milligram quantities using a pT7-7 expression system and purified to homogeneity (Figure 2a) by cation exchange chromatography followed by hydrophobic chromatography. The N-terminal sequence was MFKDPNA.... , as expected

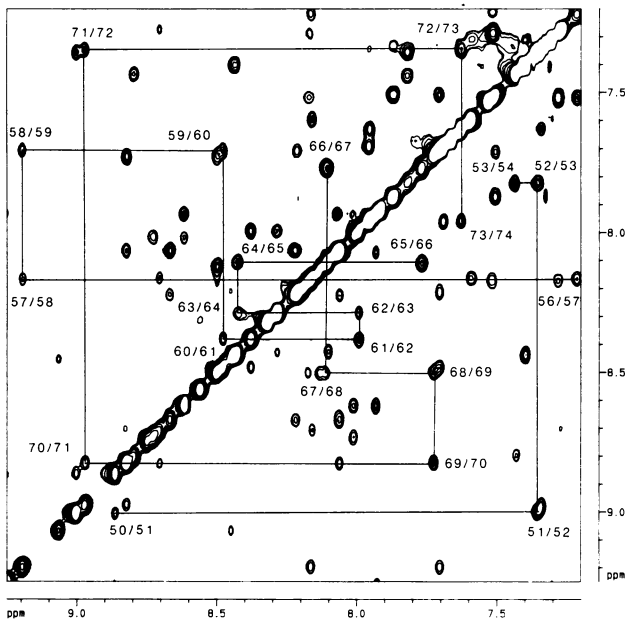

Fig. 3. Part of a 600 MHz NOESY ^{1}H-NMR spectrum of a 2.7 mM solution of the HMG box of the B-domain of HMG1 (residues 88–164) in a 90% H_2O–10% D_2O solution, containing 10 mM sodium phosphate pH 5.0, 0.15 M NaCl and 0.2 mM DTT, recorded at 293 K with a 200 ms mixing time. The spectral region displayed contains cross-peaks resulting from the d_{NN} NOE interactions. The d_{NN} connectivities used in the sequential assignment of helix III (residues 50 to 74; see Figure 4) are identified by lines that connect the appropriate cross-peaks. The cross-peaks due to the d_{NN} NOE interactions between residues 65 and 66 and between residues 66 and 67 are exactly coincident under these conditions.

with the addition of methionine encoded by the start codon; electrospray ionization mass spectra recorded before and after NMR data collection showed one major species, with a molecular mass of 8903.5 ± 0.7 Da (before) and 8904.5 ± 2.3 Da (after), (calculated molecular mass for residues 88–164 of the B-domain plus the additional N-terminal methionine is 8904.3 Da). The large negative ellipticity at 222 nm in the CD spectrum (not shown) recorded at 296 K in 10 mM sodium phosphate pH 5, 0.15 M NaCl, 0.2 mM dithiothreitol (DTT), (see below) gave an estimate of ~50% α-helix for the HMG box. Like intact HMG1, a larger B-domain fragment containing residues 91–176 (Bianchi *et al.*, 1992) and a similar A-domain fragment (Ferrari *et al.*, 1992), our minimal B-domain bound to four-way DNA junctions (Figure 2b). The existence of two complexes at the higher protein concentration is consistent with the binding of two protein molecules to structurally equivalent sites in the four-way junction.

A longer B-domain fragment (91–176) was recently reported to be dimeric free in solution (Bianchi *et al.*, 1992). However, treatment of our expressed protein (88–164) with the lysine-specific bifunctional reagent, suberic acid *bis*(*N*-hydroxysuccinimide ester), showed no evidence of cross-linked products in the presence or absence of double-stranded DNA at pH 7.5–8 (Figure 2c, lanes 1–6), suggesting that the protein was monomeric both free in solution and when bound to DNA under these conditions. As a 'positive control' for the method, the central globular domain of histone H1 (GH1), which is similar in size and lysine content to the B-domain of HMG1, was treated with cross-linking reagent under identical conditions. As expected this was monomeric when free in solution (lanes 7–9), but gave cross-linked oligomers when bound to DNA (lanes 10–12), due to cooperative binding (Thomas *et al.*, 1992). Dimerization of the HMG1 fragment (88–164) occurs only through thiol oxidation on storage without reducing agents or through prolonged exposure to trichloroacetic acid during sample preparation for SDS–polyacrylamide gel electrophoresis.

Proton NMR assignments and location of secondary structure elements

The first set of 2D spectra were recorded at 293 K in 10 mM sodium phosphate pH 5.0, 0.15 M NaCl, 0.2 mM DTT (see Materials and methods). Signals in the ^{1}H-NMR spectrum were assigned to specific protons within the protein using well established procedures (Wüthrich, 1986). Identification of as many spin systems (representing protons within individual amino acid residues) as possible was made using 2D-double quantum filtered correlation (DQF-COSY) spectra (Piantini *et al.*, 1982) and total correlation (TOCSY) spectra (Braunschweiler and Ernst, 1983) recorded with a 65 ms mixing time. Sequential assignments (Wagner and Wüthrich, 1982) were made using nuclear Overhauser enhancement (NOESY) spectra (Jeener *et al.*, 1979; Kumar *et al.*, 1980) recorded with a 200 ms mixing time. A region of the NOESY spectrum illustrating part of the sequential assignment is shown in Figure 3. All these spectra were recorded in an H_2O solution. At this stage ~75% of the resonances in the molecule had been assigned.

To complete the assignment, further DQF-COSY, TOCSY and NOESY spectra were recorded in H_2O at 298 and 303 K, and a NOESY spectrum of a D_2O solution of the protein was recorded at 303 K. By superposition and systematic comparison of these spectra, particularly those recorded at 293 and 303 K, using the graphics display and assignment program ANSIG (Kraulis, 1989), the assignment of the N-terminal residues (1–14), the proline side-chains and the connection of the aromatic rings to their respective $C_\alpha H$-$C_\beta H$ spin systems was completed. A summary of the identified sequential NOEs is shown in Figure 4. With the exception of the two N-terminal residues, at least one sequential connection (d_{NN}, $d_{\alpha N}$ or $d_{\beta N}$) was identified between all adjacent amino acid residues. For proline residues 7, 10, 11, 23, 30 and 55, sequential NOEs between the $C_\alpha H$ protons in the preceding residue and the proline $C_\delta H$ protons indicates that these X-Pro peptide bonds have a *trans* conformation. Due to overlap in the spectra, the conformation of the peptide bond to the remaining proline residue (residue 4) can only tentatively be assigned as *trans*.

Figure 4 also summarizes the medium range NOE connectivities used to identify secondary structure elements. The presence of $d_{\alpha N}$ (*i*, *i* + *3*) and $d_{\alpha\beta}$ (*i*, *i* + *3*) NOEs, as well as successive strong d_{NN} and weak $d_{\alpha N}$ NOEs (Wüthrich *et al.*, 1984), together with the slow exchange of successive amide protons, identifies three helices containing approximately residues 13–29 (helix I), 34–48 (helix II) and 50–74 (helix III) (see Figure 4). Helical wheel representations show that all three helices are amphipathic. The three helical segments together account for ~75% of the total residues in the expressed fragment (88–164). The lower helical content (~ 50%) estimated from the CD spectrum is possibly due (Bradley *et al.*, 1990) to the presence of tyrosine, which has a positive ellipticity band at 225 nm (there are three tyrosine residues in helix III and

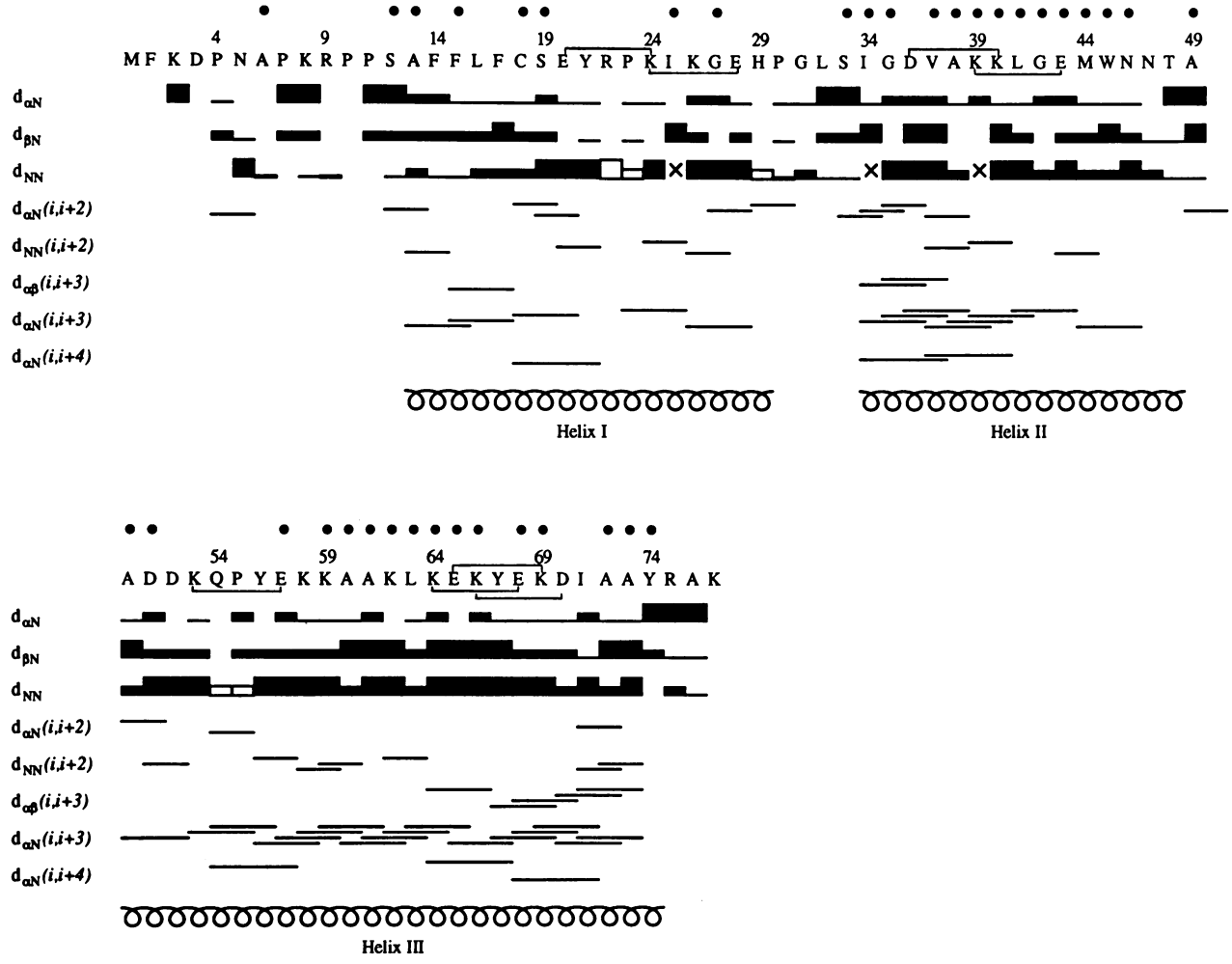

Fig. 4. Summary of sequential and medium range NOE connectivities observed in the HMG box of the B-domain of HMG1. Filled circles above the amino acid sequence (single letter code; numbering starts after the initiating methionine from the expression system) represent the location of slowly exchanging amide protons, which were observed in a NOESY spectrum recorded at 303 K after dissolving the protein in a D_2O solution. Below the sequence, the three rows of solid bars represent the observed sequential $d_{\alpha N}$, $d_{\beta N}$ and d_{NN} NOE connectivities; the thickness of the bars indicate the intensities of the cross-peaks in a NOESY spectrum recorded with a 90 ms mixing time. The open bars represent connectivities involving $C_\delta H$ protons of proline residues. Crosses represent potential NOE connectivities that are obscured by overlap with the spectrum diagonal. Lines below the sequential connectivities represent the $d_{\alpha N}(i, i + 2)$, $d_{NN}(i, i + 2)$, $d_{\alpha N}(i, i + 3)$, $d_{\alpha\beta}(i, i + 3)$, $d_{NN}(i, i + 3)$ and $d_{\alpha N}(i, i + 4)$ NOE connectivities, observed in a NOESY spectrum recorded with a 200 ms mixing time. A complete set of assignments is available from the authors. The bottom line of the figure shows the location of the three helices identified from the data collated here. The brackets in the amino acid sequence connecting acidic (D and E) and basic (K and R) residues in the relationship (i, i + 4) indicate possible salt-bridges; ⌊—⌋ and ⌈—⌉ indicate basic → acidic and acidic → basic (N→C) respectively.

one in helix I), and helix distortion (helices I and III both contain a proline residue).

Extended regions of polypeptide chain, characterized by successive strong $d_{\alpha N}$ and weak d_{NN} connectivities (Billeter *et al.*, 1982), are found between residues 7 and 13 (which includes three proline residues), 32 and 34 (turn between helices I and II), and 48 and 50 (turn between helices II and III). The lack of sequential NOEs in the N-terminus [residues 1–4; (M)FKDP] is probably due to flexibility, which is not surprising since this region is almost certainly part of the A-domain–B-domain hinge in the intact protein.

Determination of tertiary structure

The medium and long-range NOEs were assigned using the ANSIG program (Kraulis, 1989). It was possible to assign most of them by careful matching of the chemical shifts of cross-peaks at three different temperatures (293, 298 and 303 K) with the sequence-specific assignments. These assignments resulted in an initial set of 750 NOE distance restraints. The remaining cross-peaks were assigned by reference to a set of initial structures computed using only the unambiguous NOE data. The intensities of the cross-peaks in a NOESY spectrum (303 K; mixing time 90 ms) were classified as strong, medium, weak and very weak, corresponding to distance restraints of $\leq$2.5, $\leq$3.2, $\leq$5.0 and $\leq$7.0 Å, respectively. There were 1183 identifiable NOE restraints, including 308 sequential (neighbouring residues), 274 medium range [i to $\leq(i + 4)$], and 227 long range [i to $\geq (i + 5)$], the latter containing information about the tertiary structure of the protein. Stereospecific assignments and χ^1 torsion angle restraints were obtained for only four residues because of severe overlap in the $C_\alpha H$-$C_\beta H$ region of the 2D spectra. A total of 22 dihedral angle restraints derived from $^3J_{N\alpha}$ coupling constants were used (Pardi *et al.*, 1984). Further restraints were derived from the presence of slowly exchanging amide protons (a

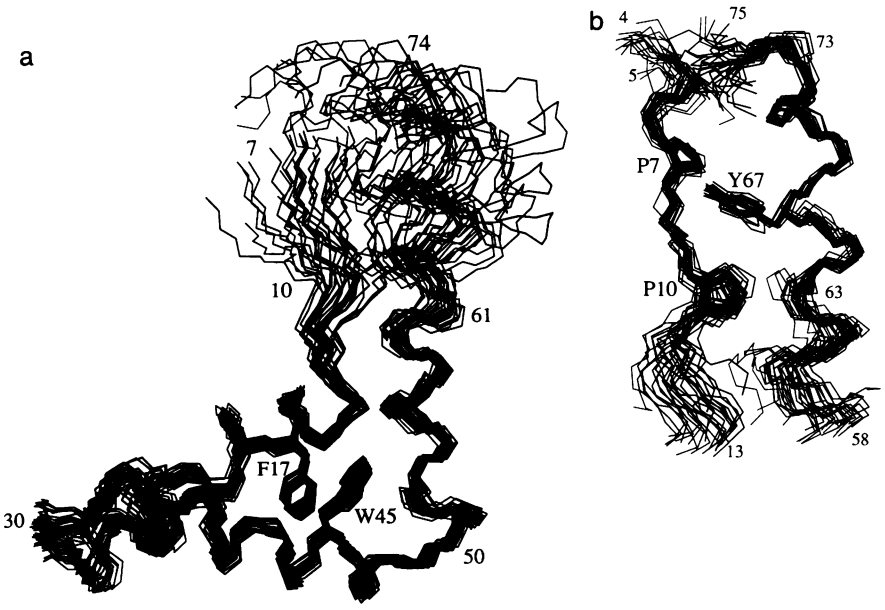

Fig. 5. Structures computed from the NMR data for the HMG box of the HMG1 B-domain. (**a**) Superposition of 30 structures showing the backbone of residues 7–74, after a least-squares fit of the backbone of the region 10–61. (**b**) Superposition of the 30 structures showing the backbone of residues 4–13 and 58–75 after a least-squares fit of the regions 5–10 and 63–73, illustrating that these regions are well-defined locally, although not globally (see text). The side-chains of some conserved residues are shown. The coordinates will be deposited in the Brookhaven Data Bank.

total of 38; Figure 4); 22 hydrogen bonds were unambiguously identified using the set of initial structures.

Thirty structures with no restraint violations >0.5 Å were computed from the NMR data (Figure 5). These are well defined, with an atomic root-mean-square deviation about the average structure of 0.69 Å for the backbone atoms and 0.94 Å for all heavy atoms, for residues 10–61; for residues 5–10/63–73, the corresponding values were 0.67 and 1.10 Å, respectively. The stereochemistry of the structures, as judged by the ϕ and ψ angles in a Ramachandran plot and by energy calculations, is good. A Lennard-Jones van der Waals energy (E_{L-J}) of −268 ± 11 kcal/mol was calculated using the CHARMm force field (Brooks *et al.*, 1983). The consistency of the structure with the NMR data was checked by creating a simulated NOESY spectrum based on the coordinates of the structure and the experimentally determined chemical shifts (Arseniev *et al.*, 1988). Superposition of the simulated spectrum and the experimental NOESY spectrum recorded at 303 K showed excellent agreement.

The structure determined for the HMG box has an unusual L-shape and consists of two arms ~31 and 36 Å long at an angle of ~80° (Figures 5 and 6a). The shorter arm consists of helices I and II (residues 13–29 and 34–48, respectively; see Figures 1 and 4). The longer arm consists of the extended N-terminal region (residues 5–12) packed against helix III (residues 50–74), the relative positions of the two arms being maintained by a cluster of conserved residues (Figure 6b). Helices I and III are slightly bent, due to the presence of proline residues at positions 23 and 55. Residues 10–61, consisting of helices I and II and the N-terminal half of helix III, are well defined (Figure 5a). Likewise residues 5–10 and 63–73 together form a region that is well defined locally (Figure 5b). However, the mutual disposition of the two regions (10–61 and 5–10/63–73) is less well defined. On superimposing the 30 structures after least-squares fit of residues 10–61, the end of the longer arm (residues 5–10/63–73) fans out in a variety of directions (Figure 5a). The reason for this is that the longer arm displays varying degrees of long-range bending between the different structures. Analysis of the different structures shows that this structural imprecision cannot be attributed to any single residue or small set of residues (data not shown). Rather, the divergence builds up gradually along the arm with increasing distance from the junction with the shorter arm. Although this effect may be due to insufficient experimental data in this region (some NOEs are unidentifiable in the spectra because of overlap) or to mobility within the longer arm, it is also possible that the elongated shape of the molecule precludes better definition of the structure using essentially short-range and conservatively quantified NMR data. [In the case of elongated molecules such as oligonucleotide duplexes, more precise distance estimates than those usually used for protein structure determination are necessary in order to achieve precision and accuracy (reviewed by James, 1991).] Test calculations using a complete set of restraints derived from one of the 30 HMG box structures, classified in the usual way, showed that even if we had been able to determine every possible restraint, NMR data of this precision cannot uniquely define the overall conformation of the longer arm in the structure (data not shown). We are currently developing approaches for the refinement of protein structures using methods similar to those used for DNA duplexes.

Structural features of the HMG box

The amino acid sequence element that constitutes the HMG box is defined by a number of conserved residues (Figure 1) (Ner, 1992), the role of many of which is evident from the tertiary structure (Figures 5 and 6). Phe14, Phe17, Trp45, Lys53 and Tyr56 form a cluster at the junction of the two arms in the structure (Figure 6b), thereby maintaining the angle between the arms, and are thus probably conserved for structural reasons. The rings of Phe17, Trp45 and Tyr56 pack at right angles to each other, while Phe14 lies between

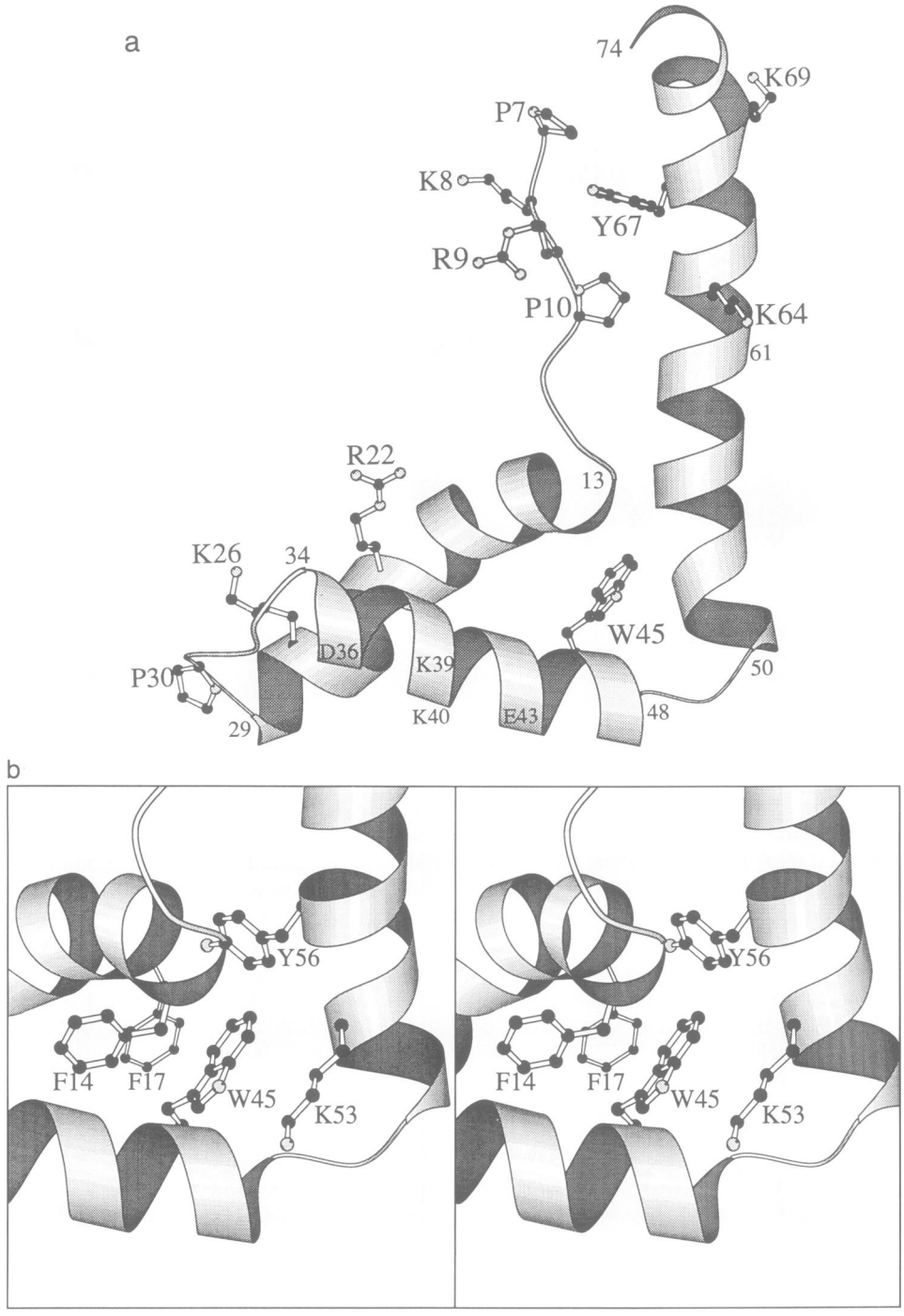

Fig. 6. Schematic representations of one of the 30 structures of the B-domain HMG box, showing the location of some of the conserved residues discussed in the text. (**a**) The structure in the same orientation as that in Figure 5. (**b**) A stereo view of the cluster of conserved residues at the junction of the two arms of the structure. The conformations of all the side-chains shown are well defined. These drawings and those in Figure 5 were generated using MOLSCRIPT (Kraulis, 1991).

helix I and helix II. The amino–aromatic interaction between Lys53 and Trp45 may account for differences in quenching of tryptophan fluorescence by reagents of different charge (Cs^+ and I^-) (Butler *et al.*, 1985). Pro residues at positions 7 and 10 in the N-terminal strand interact with Tyr67 in helix III, whereas Pro 30 is important for the conformation of the loop between helices I and II (Figure 6a).

Arg or Lys residues are conserved at positions 8, 9, 22, 39 and 64 in both classes of HMG box (Figure 1) and additionally at positions 26, 40 and 69 in the HMG1/2-like proteins. Interestingly, most of these basic residues are on or close to the concave surface formed between the two arms of the structure (Figure 6a), possibly implicating this region in DNA binding. However, residues 39 and 40 in helix II might instead be involved in salt-bridges with Glu43 and Asp36; these and the other two conserved acidic residues at positions 20 and 28 are all in the short arm of the structure. [In fact, all three helices have acidic and basic amino acid residues in the relationship (i, i + 4), suggesting possible salt-bridges (as many as eight in total; Figure 4) which could,

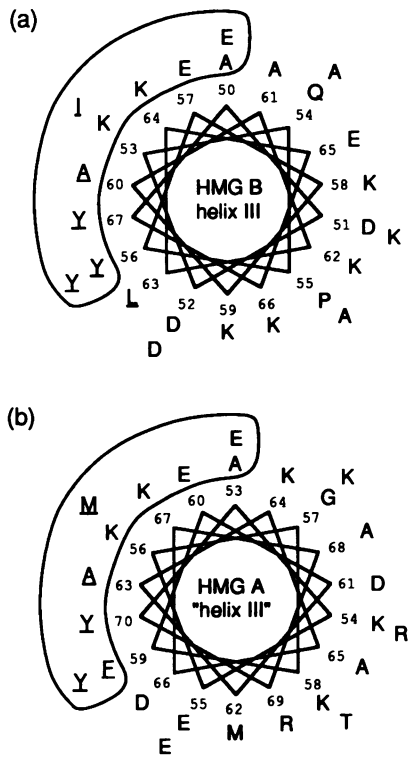

Fig. 7. Helical wheel projections for helix III of the B-domain and the homologous sequence in the A-domain of rat HMG1. (**a**) Helix III (residues A50–Y74) of the B-domain. Numbering (inner numbers) refers to the expressed fragment as shown in Figure 1; only the residues that constitute one complete turn of the 18-point helical wheel are numbered. (**b**) Helical wheel for the region of the A-domain of HMG1 (residues 53–77 in the intact protein; denoted 'helix III') that corresponds to the helix III region of the B-domain. Underlining emphasizes hydrophobic (including aromatic) amino acid residues which are clustered in one segment of the wheels. Note the similarity in amphipathic character between the helices in (a) and (b), and in addition the striking conservation of ten amino acids (two acidic, two basic, three aliphatic hydrophobic and three aromatic; shown boxed), as well as other similarities.

in principle, lead to helix stabilization (Marqusee and Baldwin, 1987).] The turns between helices I and II, and II and III can evidently accommodate structural variation since they are major sites of difference between HMG1 and the closely related but distinct protein HMG2 (turn between helices II and III), and between the A and B domains of HMG1 (turn between helices I/II and II/III) (Figure 1).

Discussion

The HMG box region of the B-domain of rat HMG1 (residues 88–164 of the intact protein), like intact HMG1, binds four-way DNA junctions. Its structure, determined by 2D ^{1}H-NMR spectroscopy, is L-shaped and the angle of ~80° between the two arms is defined by a cluster of conserved residues. 75% of the total residues are contained in three helical segments that contain many of the conserved basic and aromatic residues that define the HMG box. The sequence conservation suggests that the abundant HMG1-like proteins and the sequence-specific class of HMG box proteins (Ner, 1992) indeed share a common structural fold.

The presence of helices linked by short turns in the HMG box prompted us to compare it with the helix–turn–helix (HTH) motifs of other DNA-binding proteins (Harrison, 1991). Neither helices I/II, nor helices II/III in the HMG box match the HTH motif of, for example, the 434 repressor (Mondragon *et al.*, 1989) or the *engrailed* homeodomain (Kissinger *et al.*, 1990) in their relative orientation and, moreover, the HMG box in TCF-1, SRY and LEF-1 interacts primarily with the minor rather than major groove of DNA (Giese *et al.*, 1991, 1992; van de Wetering and Clevers, 1992). The HMG box thus appears to be a novel DNA-binding motif. The HMG box regions of LEF-1 and SRY are able to bend DNA *in vitro* (Ferrari *et al.*, 1992; Giese *et al.*, 1992) and this may also turn out to be a property of HMG1 (Lilley, 1992). Indeed HMG1 has been shown to recognize a DNA kink of ~34° (Pil and Lippard, 1992). It is tempting to speculate that the L-shape of the HMG box might be relevant here and in binding to four-way junctions.

The structure of the HMG box sheds light on some of the mutations in SRY and LEF-1 that reduce or abolish DNA-binding activity (Giese *et al.*, 1991; Harley *et al.*, 1992). Conserved basic residues at positions 8 and 9 in the box may interact with DNA (see above); mutation of these to Glu in LEF-1 indeed abolishes DNA binding. The conserved Lys53, which interacts with Trp45 (Figure 6b), appears to have a structural role, but may also be involved in DNA binding; mutation to Ile in the SRY box abolishes DNA-binding activity. In LEF-1, mutation of Tyr56 to Ser also impairs DNA binding, probably for structural reasons (Figure 6b). The residue at position 42 is nearly always small (Gly, Ala or Ser); in SRY, mutation to Arg abolishes DNA binding. A small side-chain in this position does not appear to be required to maintain the structure, implying that a larger side-chain may interfere sterically with DNA binding.

The A-domain of HMG1 is homologous with the B-domain and also has a high α-helical content (Abdul-Razzak *et al.*, 1989). A helical wheel projection (Figure 7) of the segment of the A-domain (residues 53–77 in the intact protein) that corresponds exactly with the longest helix (25 residues) of the B-domain identified here (helix III) reveals that the amphipathic character of the helix is clearly conserved and moreover, a patch of 10 residues, including hydrophobic/aromatic, acidic and basic residues, is identical apart from conservative changes of F for Y and M for I. Strikingly, this region is similar or even identical in the helical wheel projections of the corresponding regions of other abundant HMG box proteins such as yeast NHP6 (Kolodrubetz and Burgum, 1990) and *Drosophila* HMGD(N) (Ner, 1992; Wagner *et al.*, 1992), based on sequence alignment (Figure 1; Ner, 1992). Conservation not only of hydrophobic but also of basic and acidic residues in different HMG boxes suggests not only common structural features, but also common surface features involved in recognition of DNA (or possibly other proteins). Some, but not all, of these features are retained in the transcription factor class of HMG box proteins, which is entirely consistent with individual functional differences being superimposed upon a common structural fold.

Materials and methods

Plasmids and bacterial strains

pT7-7 and pGP1-2 are as described by Tabor and Richardson (1985). Plasmid pRNHMG1 [from R.Cortese (Bianchi *et al.*, 1989)] contains nucleotides −30 to 787 of rat HMG1 cDNA (Paonessa *et al.*, 1987) cloned into the *Eco*RI and *Sma*I restriction sites of pTZ18R (Pharmacia). Bacterial strains

were *E.coli* TG1 [genotype K12 Δ(*lac-pro*) *supE thi hsd* Ds/F' *tra*D36 *pro*A$^+$B$^+$ *lac*Iq *lac* Z ΔM15 *rec*O] (constructed by P.Oliver) and K38 [HfrC(l)] (Russel and Model, 1984).

Construction of pT7-7 HMG1-B

PCR was used to subclone the region of plasmid pRNHMG1 corresponding to amino acid residues 88−164 of the rat HMG1. Primer 1 (5'-GGATCC-ATATGACCAAAAAGAAGTTCAA-3') introduced an *Nde*I restriction site encompassing the ATG start codon at the 5' end of the coding sequence; primer 2 (5'-TCTAGAATTCATCACTTCTTTTTCTTGCTCT-3') introduced two in-frame TGA stop codons and an *Eco*RI restriction site at the 3' end of the gene. The PCR products were digested with *Nde*I and *Eco*RI and cloned into pT7-7. The ligation products were used to transform *E.coli* TG1 to ampicillin resistance. The DNA sequence of the insert in the resulting plasmid, designated pT7-7 HMG1-B, was verified.

Expression, purification and characterization of HMG1 fragment (88 − 164)

E.coli K38 cells (containing plasmid pGP1-2, which contains the T7 RNA polymerase gene under the control of the λ P_L promoter and the temperature-sensitive *c*I857^{ts} λ repressor) were transformed to ampicillin resistance with pT7-7 HMG1-B. Cultures grown at 30°C in 2×YT medium containing 50 μg/ml each of ampicillin and kanamycin, were induced at 42°C for 30 min and then incubated at 37°C for 2 h.

All of the following procedures were carried out at 0−4°C and monitored by SDS−polyacrylamide (18%) gel electrophoresis (Thomas and Kornberg, 1978). Cell pellets were resuspended in buffer A [10 mM Na phosphate pH 7.0, 1 mM Na_2EDTA, 0.5 mM PMSF, 1 mM DTT, 100 μg/ml each of 1-chloro-3-tosylamido-7-amino-L-2-heptanone (TLCK) and benzamidine] and lysed with a French press at 1000 p.s.i. 0.25 vol of 5 M NaCl was added to give a final concentration of 1 M NaCl and the lysate was then centrifuged at 39 000 r.p.m. for 1 h in a Beckman SW40 rotor. The supernatant was extensively sonicated, diluted with buffer A to give a final NaCl concentration of 0.1 M, filtered through a 0.22 μm Millipore filter and then applied to a column of S Sepharose Fast Flow ('Fast S') equilibrated in buffer B (10 mM Na phosphate pH 7.0 and 1 mM DTT) containing 0.1 M NaCl. The column was washed with this buffer and then eluted with a linear gradient of 0.1−0.6 M NaCl in buffer B. To the pooled peak fractions containing the HMG1 B-domain, solid NH_4SO_4 was added to a final concentration of 2 M (51.3% saturation at 0°C). The mixture was centrifuged at 9000 r.p.m. for 20 min in a Sorvall SS-34 rotor and the supernatant applied to a Phenyl Sepharose FPLC column (Pharmacia) equilibrated with buffer B containing 2 M NH_4SO_4. The column was eluted with a linear gradient from 2 M NH_4SO_4 in buffer B to buffer B alone, and gave a single major peak containing the HMG box fragment of the HMG1 B-domain, which appeared homogeneous by SDS−polyacrylamide gel electrophoresis. The yield was ~4−6 mg per litre of culture.

N-terminal sequence analysis was carried out on ~ 500 pmol of the B-domain fragment using an Applied Biosystems 477 pulsed liquid sequencer. The molecular mass of the protein was determined on ~250 pmol by electrospray ionization mass spectrometry using a VG BioQ quadrupole instrument (50% aqueous methanol−2% acetic acid as solvent; one injection of 11 μl). CD spectra from 195 to 260 nm (1 mm pathlength) were recorded at room temperature (~23°C) using a Jobin-Yvon CD6 spectropolarimeter. Samples were at 0.1 mg/ml in 10 mM Na phosphate pH 5.0, 0.15 M NaCl and 0.2 mM DTT.

Cross-linking assay

Treatment with the cross-linking reagent, suberic acid *bis*(*N*-hydroxy-succinimide ester; Sigma), was carried out at 23°C in 20 mM Na phosphate, pH 8, with the protein [the minimal HMG1 B-domain or GH1 prepared as described (Thomas *et al.*, 1992)] at 25 μg/ml and the reagent at 0.2 mg/ml (added from a freshly prepared stock solution at 20 mg/ml in dry dimethyl-formamide). For the 'protein alone' the buffer also contained 0.15 M NaCl; for the 'plus DNA' samples, NaCl was omitted and sonicated salmon sperm DNA (average ~600 bp) was present at 125 μg/ml. Samples taken after 30 min and 60 min and untreated proteins were precipitated for 10 min on ice with an equal volume of 50% trichloroacetic acid, washed with acetone−0.1 M HCl and then acetone, and dissolved in SDS gel sample buffer containing 2-mercaptoethanol at 100°C (Thomas, 1989). To ensure complete reduction the samples were incubated with 1 mM DTT at 37°C for 1 h and then analysed in an SDS−polyacrylamide (18%) gel, which was stained with Coomassie brilliant blue R250 (Thomas, 1989).

Gel retardation assay

DNA-binding assays (final volume typically 10 μl) contained 10 mM HEPES pH 7.9, 8% Ficoll, 0.2 M NaCl, 10 mM $MgCl_2$, 100 μg/ml bovine serum albumin, 5 mM KCl, 1 mM Na_2EDTA, 1 mM spermidine and 0.5 mM DTT (Bianchi *et al.*, 1992). Various concentrations of protein were added and finally 5' ^{32}P-labelled four-way junction DNA (with the core sequence of 'substrate 1' of Elborough and West (1988) but with longer arms: formed by annealing four 35mers). After incubation on ice for 20 min the samples were analysed by electrophoresis through a 5% polyacrylamide gel containing 0.5×TBE at 150 V for 3 h at 23°C. The gel was fixed in 10% acetic acid, dried under vacuum and autoradiographed at −70°C for ~18 h with intensifying screens.

NMR sample preparation and NMR spectroscopy

Pooled Phenyl Sepharose fractions were desalted by dialysis against 2 mM Na phosphate, pH 5.0, 0.04 mM DTT and vacuum-concentrated 5-fold in a Speedvac concentrator (Savant), giving a final buffer concentration of 10 mM Na phosphate, pH 5.0, 0.2 mM DTT (buffer C). Preliminary 1D ^{1}H-NMR spectra of the minimal HMG1 B-domain (residues 88−164 of HMG1) were recorded in buffer C containing 10% D_2O at pH values between 5.0 and 7.3, at added NaCl concentrations of up to 0.15 M and at temperatures between 283 and 303 K. 2D ^{1}H-NMR spectra were recorded at pH 5.0 in the presence of 0.15 M NaCl (which stabilized the protein) at 293, 298 and 303 K on an AM500 or an AMX 600 Bruker spectrometer and data-processing was carried out using the manufacturer's software. For the identification of slowly exchanging amide protons, the protein was dialysed against 0.1% (w/v) ammonium hydrogen carbonate, lyophilized, dissolved in buffer C containing 99.8% D_2O and spectra were immediately recorded. The sample concentration was typically 2−3 mM.

All spectra were acquired in the phase-sensitive mode with quadrature detection in the t_1 dimension using time-proportional phase incrementation (TPPI) (Bodenhausen and Ruben, 1980; Marion and Wüthrich,1983). In all cases the carrier was placed on the ^1H_2O resonance, which was suppressed by presaturation. The receiver reference phase and the delay between the opening of the receiver gate and acquisition of the first data point were optimized to obtain a flat baseline (Hoult *et al.*, 1983; Marion and Bax, 1988). During data processing, baseline correction in ω_2 was carried out using a polynomial of order 3.

Double-quantum- and triple-quantum-filtered COSY (DQF and TQF-COSY) spectra, in either H_2O or D_2O solution, were acquired using the standard phase cycle (Piantini *et al.*, 1982); spectra were recorded using either 32 or 48 scans for each t_1 point (DQF and TQF-COSY respectively) and the maximum acquisition times were 70.66 ms and 0.283 s in t_1 and t_2 respectively. $^3J_{N\alpha}$ couplings were measured by fitting in-phase and anti-phase Gaussian peak shapes to TOCSY and COSY traces, respectively, for each cross-peak. NOESY spectra in H_2O solution were recorded with mixing times of 90 and 200 ms. TOCSY spectra in H_2O were recorded with mixing times of between 54 and 68 ms using a spin locking field of 5 kHz. The TOCSY experiments employed the DIPSI-2 mixing sequence (Shaka *et al.*, 1988) and were acquired using the method of Rance (1987). The NOESY and TOCSY spectra were acquired with 32 or 64 scans for each t_1 point and maximum acquisition times were 35.33 ms and 0.283 s in t_1 and t_2 respectively.

Structures were calculated from the NMR data using the program X-PLOR and the simulated annealing protocol YASAP starting from initial structures with random ϕ and ψ values (Brünger, 1990).

Acknowledgements

We thank Dr R.Cortese for plasmid pRNHMG1, Dr P.Oliver for the TG1 *rec*O strain of *E.coli*, Dr M.Nilges for protocols for the simulated annealing calculations and Christine Rees for assistance with the cross-linking experiment. Protein microsequencing, oligonucleotide synthesis and amino acid analysis were carried out in the Protein and Nucleic Acid Chemistry Facility in this Department. H.M.W. thanks S.Brocklehurst, C.Hardman and Dr Y.Kalia for help and discussion. We acknowledge the support of the Science and Engineering Research Council through the Molecular Recognition Initiative and of The Wellcome Trust. P.J.K. is a Postdoctoral Fellow of the Swedish Natural Science Research Council and C.S.H. was a Research Fellow at New Hall, Cambridge, supported by the Gateway Corporation.

References

Abdul-Razzak,K.K., Denton,M.L., Cox.D.J. and Reeck,G.R. (1989) *Biochim. Biophys. Acta*, **996**, 125−131.

Arseniev,A., Schultze,P., Wörgötter,E., Braun,W., Wagner,G., Vašák,M., Kägi,J.H.R. and Wüthrich,K. (1988) *J. Mol. Biol.*, **201**, 637−657.

Bianchi,M.E., Beltrame,M. and Paonessa,G. (1989) *Science*, **243**, 1056−1059.

Bianchi,M.E., Falciola,L., Ferrari,S. and Lilley,D.M.J. (1992) *EMBO J.*, **11**, 1055–1063.
Billeter,M., Braun,W. and Wüthrich,K. (1982) *J. Mol. Biol.*, **155**, 321–346.
Bodenhausen,G. and Ruben,D. (1980) *Chem. Phys. Lett.*, **69**, 185–187.
Bradley,E.K., Thomason,J.F., Cohen,F.E., Kosen,P.A. and Kuntz,I.D. (1990) *J. Mol. Biol.*, **215**, 607–622.
Braunschweiler,L. and Ernst,R.R. (1983) *J. Magn. Reson.*, **53**, 521–528.
Brooks,B.R., Bruccoleri,R.E., Olafson,B.D., States,D.J., Swaminathan,S. and Karplus,M. (1983) *J. Comput. Chem.*, **4**, 187–217.
Brünger,A.T. (1990) *X-PLOR Manual V2.1*, Yale University.
Bustin,M., Lehn,D.A. and Landsman,D. (1990) *Biochim. Biophys. Acta*, **1049**, 231–243.
Butler,A.P., Mardian,J.K.W. and Olins,D.E. (1985) *J. Biol. Chem.*, **260**, 10613–10620.
Carballo,M., Puigdomenech,P. and Palau,J. (1983) *EMBO J.*, **2**, 1759–1764.
Cary,P.D., Turner,C.H., Mayes,E. and Crane-Robinson,C. (1983) *Eur. J. Biochem.*, **131**, 367–374.
Diffley,J.F.X. and Stillman,B. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 7864–7868.
Elborough,K.M. and West,S.C. (1988) *Nucleic Acids Res.*, **16**, 3603–3616.
Ferrari,S., Harley,V.R., Pontiggia,A., Goodfellow,P.N., Lovell-Badge,R. and Bianchi,M.E. (1992) *EMBO J.*, **11**, 4497–4506.
Giese,K., Amsterdam,A. and Grosschedl,R. (1991) *Genes Dev.*, **5**, 2567–2578.
Giese,K., Cox,J. and Grosschedl,R. (1992) *Cell*, **69**, 185–195.
Harley,V.R., Jackson,D.I., Hextall,P.J., Hawkins,J.R., Berkowitz,G.D., Sockanathan,S., Lovell-Badge,R. and Goodfellow,P.N. (1992) *Science*, **255**, 453–456.
Harrison,S.C. (1991) *Nature*, **353**, 715–719.
Hoult,D.I., Chen,C.N., Eden,H. and Eden,M. (1983) *J. Magn. Reson.*, **51**, 110–114.
James,T.L. (1991) *Curr. Opin. Struct. Biol.*, **1**, 1042–1053.
Jantzen,H.-M., Admon,A., Bell,S.P. and Tjian,R. (1990) *Nature*, **344**, 830–836.
Jeener,J., Meier,B.H., Bachmann,P. and Ernst,R.R. (1979) *J. Chem. Phys.*, **71**, 4546–4553.
Johns,E.W. (1982) *The HMG Chromosomal Proteins*. Academic Press, London.
Kelly,M., Burke,J., Smith,M., Klar,A. and Beach,D. (1988) *EMBO J.*, **7**, 1537–1547.
Kissinger,C.R., Liu,B., Martin-Blanco,E., Kornberg,T.B. and Pabo,C.O. (1990) *Cell*, **63**, 579–590.
Kolodrubetz,D. and Burgum,A. (1990) *J. Biol. Chem.*, **265**, 3234–3239.
Kraulis,P.J. (1989) *J. Magn. Reson.*, **84**, 627–633.
Kraulis,P.J. (1991) *J. Appl. Cryst.*, **24**, 946–950.
Kumar,A., Ernst,R.R. and Wüthrich,K. (1980) *Biochem. Biophys. Res. Commun.*, **95**, 1–6.
Lilley,D.M.J. (1992) *Nature*, **357**, 282–283.
Marion,D. and Bax,A. (1988) *J. Magn. Reson.*, **79**, 352–356.
Marion,D. and Wüthrich,K. (1983) *Biochem. Biophys. Res. Commun.*, **113**, 967–974.
Marqusee,S. and Baldwin,R. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 8898–8902.
Mondragón,A., Subbiah,S., Almo,S.C., Drottar,M. and Harrison,S.C. (1989) *J. Mol. Biol.*, **205**, 189–200.
Ner,S. (1992) *Curr. Biol.*, **2**, 208–210.
Paonessa,G., Frank,R. and Cortese,R. (1987) *Nucleic Acids Res.*, **15**, 9077.
Pardi,A., Billeter,M. and Wüthrich,K. (1984) *J. Mol. Biol.*, **180**, 741–751.
Piantini,U., Sørensen,O.W. and Ernst,R.R. (1982) *J. Am. Chem. Soc.*, **104**, 6800–6801.
Pil,P.M. and Lippard,S.J. (1992) *Science*, **256**, 234–237.
Postnikov, Yu.,V., Schick,V.V., Belyavsky,A.V., Khrapko,K.R., Brodolin,K.L., Nicolskaya,T.A. and Mirzabekov,A.D. (1991) *Nucleic Acids Res.*, **19**, 717–725.
Rance,M. (1987) *J. Magn. Reson.*, **74**, 557–564.
Reeck,C.R., Isackson,P.J. and Teller,D.C. (1982) *Nature*, **300**, 76–78.
Russel,M. and Model,P. (1984) *J. Bacteriol.*, **159**, 1034–1039.
Shaka,A.J., Lee,C.J. and Pines,A. (1988) *J. Magn. Reson.*, **77**, 274–293.
Shirakawa.H., Tsuda,K. and Yoshida,M. (1990) *Biochemistry*, **29**, 4419–4423.
Sinclair,A.H. *et al.* (1990) *Nature*, **346**, 240–244.
Tabor,S. and Richardson,C.C. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 1074–1078.
Thomas,J.O. (1989) *Methods Enzymol.*, **170**, 549–571.
Thomas,J.O. and Kornberg,R.D. (1978) *Methods Cell Biol.*, **18**, 429–440.
Thomas,J.O., Rees,C.R. and Finch,J.T. (1992) *Nucleic Acids Res.*, **20**, 187–194.
Travis,A., Amsterdam,A., Belanger,C. and Grosschedl,R. (1991) *Genes Dev.*, **5**, 880–894.
van de Wetering,M. and Clevers,H. (1992) *EMBO J.*, **11**, 3039–3044.
Wagner,C.R., Hamana,K. and Elgin,S.C.R. (1992) *Mol. Cell. Biol.*, **12**, 1915–1923.
Wagner,G. and Wüthrich,K. (1982) *J. Mol. Biol.*, **155**, 347–366.
Wüthrich,K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York.
Wüthrich,K., Billeter,M. and Braun,W. (1984) *J. Mol. Biol.*, **180**, 715–740.

Received on December 8, 1992; revised on January 20, 1993