The EMBO Journal Vol.18 No.23 pp.6653–6661, 1999

Solution structure of the methyl-CpG-binding domain of the methylation-dependent transcriptional repressor MBD1

Izuru Ohki, Nobuya Shimotake, Naoyuki Fujita[1], Mitsuyoshi Nakao[1,2] and Masahiro Shirakawa[2]

Graduate School of Biological Sciences, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0101 and [1]Department of Tumor Genetics and Biology, Kumamoto University School of Medicine, 2-2-1 Honjo, Kumamoto 860-0811, Japan

[2]Corresponding authors
e-mail: shira@bs.aist-nara.ac.jp or mnakao@gpo.kumamoto-u.ac.jp

CpG methylation in vertebrates is important for gene silencing, alterations in chromatin structure and genomic stability, and differences in the DNA-methylation status are correlated with imprinting phenomena, carcinogenesis and embryonic development. Methylation signals are interpreted by protein factors that contain shared methyl-CpG-binding domains (MBDs). We have determined the solution structure of the MBD of the human methylation-dependent transcriptional repressor MBD1 by multi-dimensional heteronuclear NMR spectroscopy. It folds into an α/β-sandwich structure with characteristic loops. Basic residues conserved in the MBD family are largely confined to one face of this fold and a flexible loop, which together form a large positively charged surface. Site-directed mutagenesis and chemical shift changes upon complexing with a methylated DNA facilitated identification of this surface as the DNA interaction site. In addition to three basic residues, conserved Tyr34 and Asp32 were shown to be important for the DNA binding.

Keywords: DNA methylation/methyl-CpG-binding domain/NMR/transcription

Introduction

CpG methylation in vertebrates is important for control of gene activity either through effects on a single promoter region, or through global mechanisms that affect many genes (Bird, 1992; Kass *et al*., 1997a; Siegfried and Cedar, 1997; Eden *et al*., 1998; Razin, 1998; Tajima and Suetake, 1998). Differences in the DNA-methylation status correlate with a wide range of biological phenomena, including genomic imprinting, carcinogenesis, X inactivation and embryonic development (Ballabio and Willard, 1992; Li *et al*., 1993; Tate *et al*., 1996; R.Z.Chen *et al*., 1998; Surani, 1998). The methylation of promoter-associated CpG islands is involved in the transcriptional repression of tumor suppressor genes in tumor cells and imprinted genes (Sutcliffe *et al*., 1994; Feinberg *et al*., 1995; Lalande, 1996; Baylin *et al*., 1998). On the other hand, genome-wide demethylation has been suggested to be an early step in carcinogenesis. DNA hypomethylation was shown to lead to elevated mutation rates, and is associated with the occurrence of chromosomal abnormalities (Fearon and Vogelstein, 1990; R.Z.Chen *et al*., 1998).

The signals that DNA methylation represents are interpreted by protein factors that contain common MBD sequences; there are at present five known family members: MeCP2, MBD1 (formerly PCM1), MBD4/MED1, MBD2 and MBD3 (Figure 1) (Hendrich and Bird, 1998; Bhattacharya *et al*., 1999). Mammalian cells have the potential to encode both long and short forms of MBD2, and these have been termed MBD2a and 2b, respectively (Hendrich and Bird, 1998). One of them, MBD2b, was proposed to have methyl-CpG-specific demethylase activity (Bhattacharya *et al*., 1999), but more recent publications reported that MBD2b and its related proteins had failed to demethylate DNA, and instead implicate MBD2 in transcriptional repression (Ng *et al*., 1999; Wade *et al*., 1999). MBD4/MED1, another member of the MBD family, was shown to be an endonuclease that forms a complex with the DNA mismatch repair protein, MLH1, which suggests that MBD4/MED1 may be involved in mismatch repair and be responsible for maintenance of genome stability (Bellacosa *et al*., 1999). Recently, Hendrich *et al.* (1999) showed that MBD4/MED1 has DNA glycosylase activity and can remove thymine from a mismatch CpG site *in vitro*, and that its MBD binds preferentially to methyl-CpG·TpG mismatches. MBDs have also been found in *Xenopus laevis*. Wade *et al.* (1999) showed that *X.laevis* Mi-2 complex contains MBD3 and MBD3 LF, whose sequences are substantially similar to mammalian MBD3. MBD3 LF is a product of an alternative splicing variant, and contains an insertion of 20 amino acid residues. Sequences of all the MBD family are highly conserved. All but mammalian MBD3 and *X.laevis* MBD3 LF have been shown to bind specifically to symmetric methyl-CpG, and the binding appears to be as a monomer and be independent of the sequence context outside of the CG sequence (Nan *et al*., 1993; Hendrich and Bird, 1998; Chandler *et al*., 1999).

Methyl-CpG binding proteins (MeCPs) 1 and 2 contribute to methylation-dependent gene silencing (Meehan *et al*., 1989; Boyes and Bird, 1991; Lewis *et al*., 1992; Kass *et al*., 1997b; Nan *et al*., 1997; Bergman and Mostoslavsky, 1998). MeCP2 was shown to bind to a methylated DNA and to recruit the Sin3–histone deacetylase complex to promoters, resulting in deacetylation of core histones and subsequent transcriptional silencing (Jones *et al*., 1998; Nan *et al*., 1998). In a like manner, MeCP1 is a histone deacetylase complex that includes MBD2, RbA p48/p46, histone deacetylases (HDACs) 1 and 2, Sin3A and its related proteins (Ng *et al*., 1999; Wade *et al*., 1999). This complex is capable of repressing transcription from closely and sparsely methylated promoters, and its activity is ubiquitous in somatic cells and tissues but is notably absent in ES cells

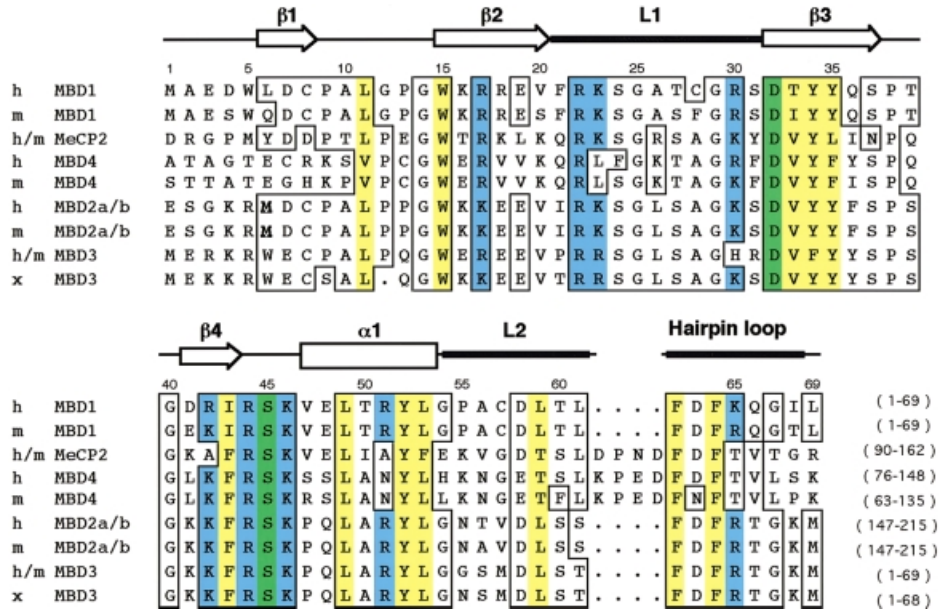

Fig. 1. Sequence alignment of the MBD family. The numbering is shown for human MBD1. Conserved residues are boxed. The secondary structure of human MBD1 is indicated at the top. Important residues discussed in the text are colored (blue, basic; yellow, hydrophobic; green, acidic or polar). The sequences aligned are those of MBD1, MeCP2, MBD4 (also called MED1), MBD2 and MBD3 of human and mouse, and *X.laevis* MBD3 (Hendrich and Bird, 1998; Bellacosa *et al*., 1999; Bhattacharya *et al*., 1999; Wade *et al*., 1999). MBD2 of human and mouse has a poteintial to encode both long (MBD2a) and short (MBD2b) forms. MBD2b starts at methionine 152 (underlined) of MBD2a.

(Meehan *et al*., 1989; Boyes and Bird, 1991, 1992). Besides these two complexes of MeCPs, other histone deacetylase core complexes were found to contain members of the MBD family. NuRD, a multisubunit complex having nucleosome remodeling and histone deacetylase activities, contains MBD3 and Mi-2 (Wade *et al*., 1999; Zhang *et al*., 1999). Mi-2 is known to be a member of the Swi2/Snf2 helicase/ATPase family (Wade *et al*., 1998, 1999). These observations illustrate a close link between CpG-methylation and histone deacetylation.

MBD1 was proposed to be a component of MeCP1, but this conclusion has been questioned (Cross *et al*., 1997; Ng *et al*., 1999). Recently, Fujita *et al*. (1999) showed that MBD1 can bind to methylated CpG islands from the tumor suppressors *p16*, *VHL* and *E-cadherin*, and imprinted *SNRPN* genes, and inhibit promoter activities of these genes in a methylation-dependent manner *in vitro* and in the cells. Here we describe the solution structure of MBD1 MBD. Based on the structure and results of chemical shift perturbation experiments and site-directed mutagenesis, we identified residues that are critical for binding to methylated CpG.

Results

Structure determination

Structure of the MBD of human MBD1 (residues 1–75) was determined from a total of 1329 NMR-derived distance and torsion angle restraints. The statistics for the 25 final structures are shown in Table I. The backbone and hydrophobic side chains have been well defined, except for a loop region (L1; residues 21–31), and N- and C-terminal residues. For loop L1, few long-range NOEs were observed. The local sequential NOE patterns and ^{1}H-^{15}N heteronuclear NOE values of this region are characteristic of an unstructured and flexible loop. In addition to this region, the N-terminal residues, including Met-1 and Ala-2, and the six C-terminal residues have not been defined regarding structure. Figure 2A depicts the backbones (N, C_α, C′) of the 25 final structures derived

Table I. Structural statistics for the methyl-CpG-binding domain of MBD1[a]

Total No. of distance constraints	1285
Intraresidual	491
Sequential (|i–j| = 1)	302
Medium range (1 <|i–j| ≤ 4)	165
Long range (|i–j| >4)	312
Hydrogen bonds	15
No. of dihedral angle constraints	44
R.m.s. deviations from experimental constraints[b]	
Distances (Å)	0.0070 ± 0.0007
Angles (°)	0.23 ± 0.08
R.m.s. deviations from idealized covalent geometry	
Bonds (Å)	0.0024 ± 0.00004
Angles (°)	0.67 ± 0.002
Impropers (°)	0.36 ± 0.005
X-PLOR potential energy	
E_{total} (kcal/mol)	179.0 ± 1.6
PROCHECK Ramachandran plot statistics (residues 3–20, 32–69)	
Most favored regions (%)	75.8
Additional allowed regions (%)	20.7
Generously allowed regions (%)	3.1
Disallowed regions (%)	0.4
R.m.s. deviations from the mean structure (residues 3–20, 32–69)	
Backbone heavy atoms (Å)	0.45
All heavy atoms (Å)	0.98

[a]These statistics comprise the ensemble of the 25 structures of lowest energy obtained from 125 starting structures. Structure calculations were performed following simulated annealing protocols using X-PLOR 3.8.
[b]None of these structures exhibited distance violations >0.3 Å or dihedral angle violations >5°.

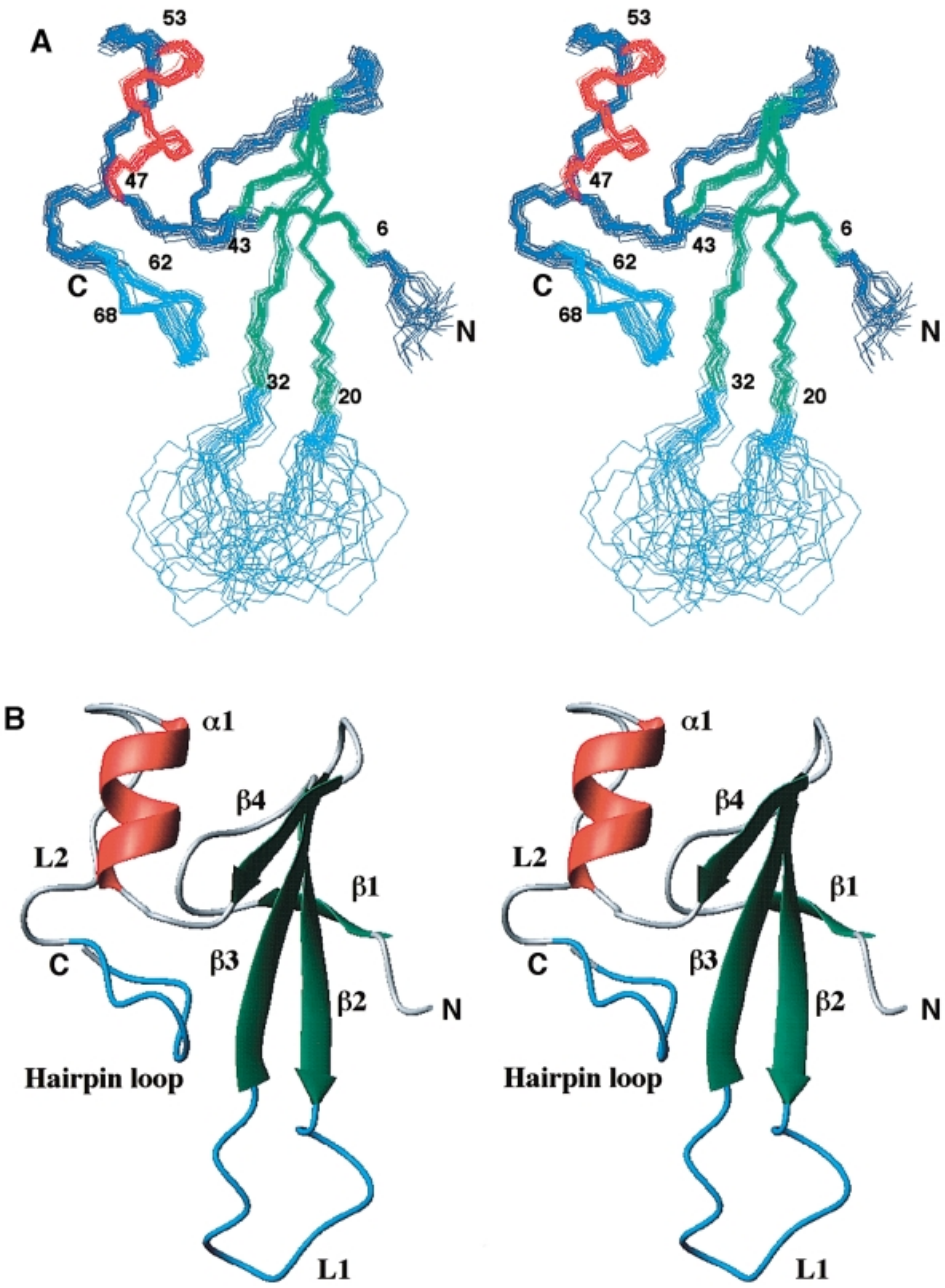

Fig. 2. Solution structure of MBD1 MBD. (**A**) Best-fit backbone superpositions of the 25 final structures in stereo. The backbone atoms of residues 3–20 and 32–69 are superimposed. Helix α1 (residues 47–53) is colored in red, and strands in green. Loop L1 (residues 21–31) and the hairpin loop (residues 62–68) are colored in light blue. (**B**) Schematic ribbon drawing of the NMR structure of the MBD in stereo, drawn with MOLMOL (Koradi *et al*., 1996). Secondary structure elements are indicated. The disordered two N-terminal and six C-terminal residues are omitted from the figures for clarity.

from NMR data. The r.m.s. deviation (RMSD) of residues 3–69 from the mean coordinate positions, excluding loop L1, is 0.45 Å for the backbone heavy atoms, and 0.98 Å for all non-hydrogen atoms.

Structure description

The three-dimensional structure of the MBD has a compact fold, the N- and C-termini being on opposite faces of the molecule. It assumes an α/β-sandwich fold: one layer consists of a four-stranded twisted β-sheet (strand β1, residues 6–8; β2, 15–20; β3, 32–37; β4, 41–43), and the other layer comprises a helix (α1, residues 47–53) with a characteristic hairpin loop at the C-terminus (Figure 2B). Both helix α1 and the β-sheet have an amphipathic character, and their hydrophobic faces are tightly packed against each other, so that helix α1 and strand β4 are arranged in a roughly antiparallel manner.

The MBD contains three long loops, L1, L2 and the C-terminal hairpin loop. Loop L2 (residues 54–61), which connects helix α1 and the hairpin loop, is well defined. It starts at a characteristic short turn at Gly54–Pro55, followed by a helical coil arranged antiparallel to α1. The C-terminal hairpin loop (residues 62–68) resembles a β hairpin, with a main chain hydrogen bond between Gly67 H_N and Asp63 C′=O in its stem region. This loop has structurally vital hydrophobic residues, Phe62 and

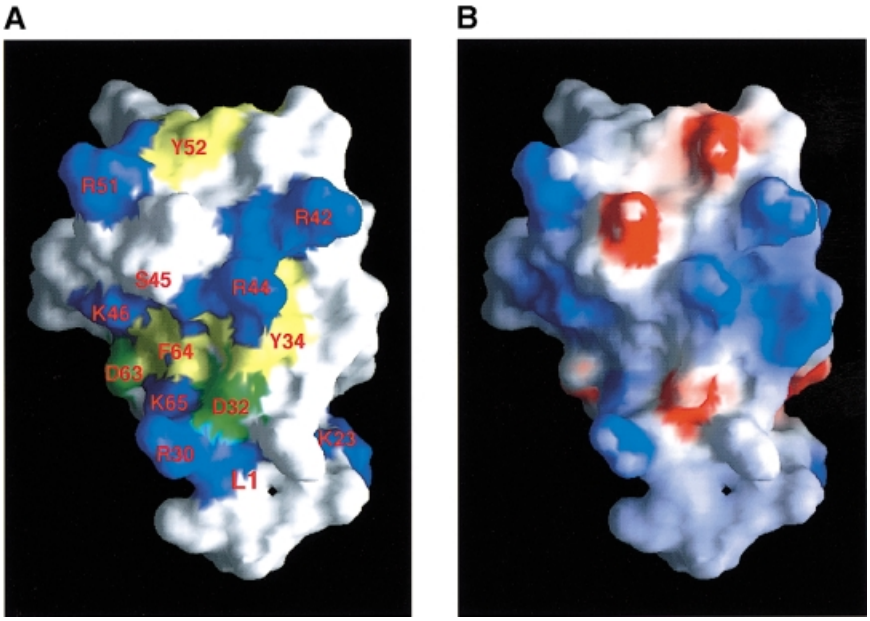

Fig. 3. Surface diagrams of MBD1 MBD. (**A**) The protein surface (Nicholls *et al*., 1991) viewed in the same orientation as in Figure 2. The conserved basic and acidic residues are colored in blue and green, respectively. The hydrophobic superficial patches made up of the conserved residues are colored in yellow. The location of Ser45 is also indicated. (**B**) Distribution of the electrostatic potential on the solvent-accessible surface (Nicholls *et al*., 1991). Blue corresponds to positive potential and red to negative potential.

Phe64, which stabilize orientation of the hairpin relative to the rest of the molecule, and are identical in all MBD family members. Substitution of Phe64 by alanine caused disruption of the native tertiary structure and a consequent loss of the DNA binding (see Mutagenesis and DNA binding assays, below). Deletion of residues 157–162 from MeCP2, which corresponds to most of the hairpin loop of MeCP2, resulted in a total loss of methyl-CpG binding activity, further indicating the structural significance of the hairpin loop (Nan *et al*., 1993). With knowledge of the tertiary structure of the MBD and structurally vital residues, we could align sequences of the MBD family, as shown in Figure 1, finding the insertion of four residues at the junction between L2 and the hairpin loop in both MeCP2 and MBD4. Considering the insertion position in the tertiary structure, these residues would not be expected to have a large impact on the overall folding of these proteins. In contrast to L2, L1 (residues 21–31) apparently does not form a single definite structure judging from observation of the few long-range NOEs, although its residues are highly conserved in the MBD family. ^{1}H-^{15}N heteronuclear NOE measurements (Cavanagh *et al*., 1996) showed smaller NOE values for residues in this loop (0.34 ± 0.17) than for the structurally well defined region (0.70 ± 0.08), further indicating that L1 is highly mobile in solution.

This MBD fold generates a well defined hydrophobic core that includes the entire polypeptide domain, and is formed by Leu11, Trp15, Thr33, Tyr35, Ile43, Leu49, Leu53, Leu59, Phe62 and Phe64. The high degree of conservation of these residues in mammalian and *X.laevis* MBDs, which are responsible for structural integrity of the domain, suggests that the folding is essentially identical throughout the MBD family (Figure 1). Side chains of most of the conserved hydrophobic residues point to the interior of the protein. Exceptions are Tyr34 and Tyr52. These residues are highly conserved and their aromatic rings largely exposed to the solvent.

In addition to hydrophobic residues that form the core, several basic residues are well conserved in the mammalian and *X.laevis* MBD family. Side chains of these residues are exposed to the solvent, are largely confined to one side of the molecule and cause positive charges on the surface (Figure 3A and B), while Arg17 is located on the other side. Well conserved Arg30, Arg42, Arg44, Lys46 and Lys65 are lined up in the middle of this basic surface, and Arg22 and Lys23 are clustered in loop L1.

DNA binding site

To identify the DNA binding site, we did chemical shift perturbation experiments. A ^{15}N-labeled MBD was titrated with an unlabeled double-stranded oligonucleotide carrying a methyl-CpG pair. Selective chemical shift changes were observed for signals in the (^{15}N, ^{1}H) HSQC spectrum upon complex formation (Figure 4A). As is often seen for strong interactions, a slow exchange mode was exhibited at the NMR time scale, whereas titration with a non-methylated DNA revealed smaller chemical shift changes and a fast exchange mode (data not shown).

Strongly affected residues were observed in loop L1, the linker between strand β4 and helix α1, and the tip of the hairpin loop, as well as in their neighboring regions (Figure 4A). These putative DNA-binding residues form a continuous surface on one side of the molecule, which greatly overlaps the area covered by the conserved basic residues, suggesting that they are involved in the interaction with DNA (Figure 4B). In addition to these basic ones, three conserved residues, Asp32, Tyr34 and Ser45, are located in the middle of the putative DNA site. Of them, Tyr34 forms a superficial hydrophobic patch together with another conserved residue, Phe64 (Figure 3A).

$^{13}C_{\alpha}$, $^{13}C_{\beta}$ and $^{13}C'$ chemical shifts depend on the main

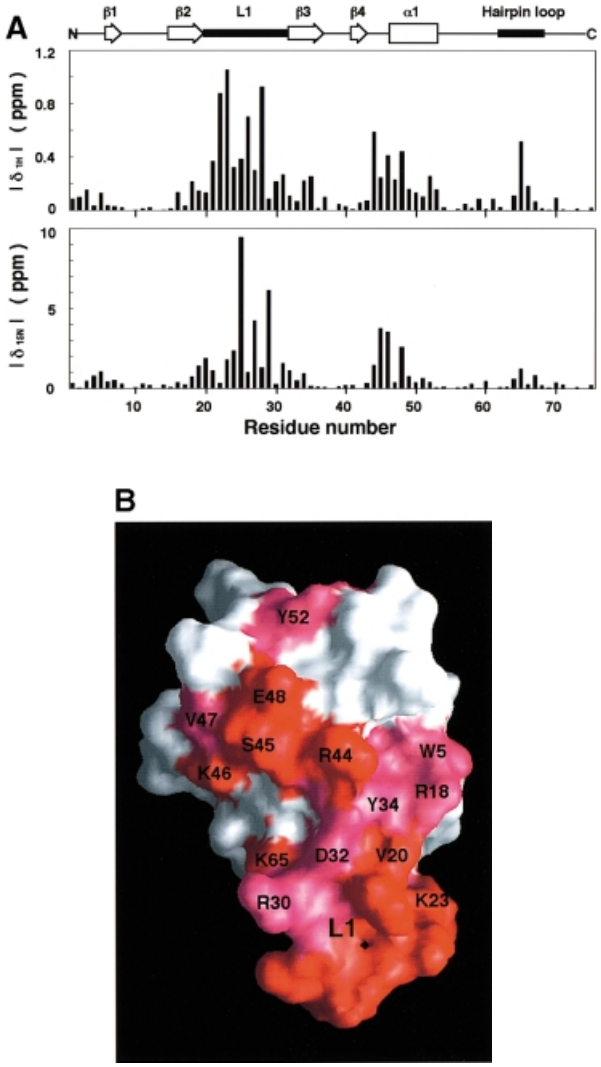

Fig. 4. Effects of titration of methyl-CpG DNA on backbone ^{15}N and ^{1}H chemical shifts of the MBD. (**A**) Chemical shift changes, $|\delta_{^1H}|$ (upper panel) and $|\delta_{^{15}N}|$ (lower panel), are plotted versus residue number, where $\delta_{^1H}$ and $\delta_{^{15}N}$ are the differences in p.p.m. between the free and bound chemical shifts. The secondary structure is indicated at the top. (**B**) The protein surface representation with color coding to show strongly affected residues viewed in the same orientation as in Figures 2 and 3. The color coding reflects the normalized weighted average shift differences, δ_{av}/δ_{max} (see Materials and methods; Garrett *et al*., 1997; Foster *et al*., 1998). The residues strongly affected with addition of the methyl-CpG DNA are in red ($\delta_{av}/\delta_{max} > 0.2$) or magenta ($0.1 < \delta_{av}/\delta_{max} \leq 0.2$).

chain dihedral angles, and are good indicators of the secondary structure (Wishart and Sykes, 1994). The chemical shifts exhibited by MBD1 MBD in its complex with the methyl-CpG DNA indicated that elements of the secondary structure of the MBD in the complex are maintained essentially as those of the protein alone (data not shown). On the other hand, a large conformational change in the conserved L1 loop was suggested by the observation of much larger changes in the L1 loop ($^{13}C_\alpha$, 0.95 ± 0.82 p.p.m.; $^{13}C_\beta$, 0.83 ± 0.67 p.p.m.; and $^{13}C'$, 1.44 ± 1.15 p.p.m.) than those in the α/β structured region ($^{13}C_\alpha$, 0.18 ± 0.15 p.p.m.; $^{13}C_\beta$, 0.37 ± 0.37 p.p.m.; and $^{13}C'$, 0.27 ± 0.25 p.p.m.). Furthermore, the ^{1}H-^{15}N NOE values for residues in loop L1 are markedly greater in the complex (0.70 ± 0.05) than in the free form (0.34 ± 0.17). These observations indicate that the DNA binding induces a well defined structure in loop L1.

Mutagenesis and DNA binding assay

In order to identify amino acids that are important for DNA binding activity of MBD, site-directed mutants were prepared in a GST-fusion form, and each was examined as to the potential to bind to a double-stranded oligonucleotide carrying a methyl-CpG pair, by means of gel-shift assays. The following eight residues for mutation were chosen based on results of chemical shift perturbation experiments, the sequence homology and the NMR structure: Arg22, Arg30, Asp32, Tyr34, Arg44, Ser45, Tyr52 and Phe64. We only chose surface-exposed residues, except for Phe64, whose side chain aromatic ring is 50% exposed in the ensemble of the 25 NMR structures. All point mutations were to alanine.

For each of the mutants except S45A, a ^{15}N-labeled MBD was prepared and the ^{1}H-^{15}N HSQC spectrum was compared with that of the wild-type MBD. If mutants are structurally similar to wild type, ^{1}H-^{15}N HSQC spectra with chemical shift differences only for residues in the vicinity of the mutation site and in the flexible part are expected. On the other hand, mutant proteins that disrupt the native fold show differences throughout the spectra. Only one of the eight mutants, F64A, caused changes indicative of disruption of the native tertiary structure. For Y52A, the spectrum showed a minor conformational change. All the other mutants (R22A, R30A, D32A, Y34A and R44A) showed no evidence for mutant-induced disruptions of the native fold. We could not obtain a sufficient amout of the S45A MBD for ^{1}H-^{15}N HSQC measurements. However, a ^{1}H NMR spectrum of the S45A MBD showed no evidence for disruption of the native overall fold.

Six of the eight mutant proteins (R22A, R30A, D32A, Y34A, R44A and F64A) exhibited a considerably decreased ability to bind to methylated CpG in the gel-shift assays (Figure 5A). Methyl-DNA bound fractions for each protein, obtained from the averaged value of four experiments, were: 0.95 for the wild protein, 0.0 for R22A, 0.04 for R30A, 0.02 for D32A, 0.15 for Y34A, 0.0 for R44A, 0.48 for S45A, 0.83 for Y52A and 0.0 for F64A. For F64A, the loss of the binding activity is apparently due to disruption of the native tertiary structure. Thus, the mutagenesis indicated that the five residues (Arg22, Arg30, Asp32, Tyr34 and Arg44) are important for the binding. As shown in Figure 5B, these important residues are all clustered in the middle of the putative DNA binding site suggested by evidence in the chemical shift perturbation experiments. Of the remaining two mutants, Y52A could bind to the DNA with a similar affinity to that seen with the wild protein, and the mutation of Ser45 caused a moderate loss of DNA binding activity.

Discussion

Comparison of MBD with other DNA-binding domains

The secondary structure organization of the MBD fold has a limited similarity to those of several known DNA binding proteins. For example, both the N-terminal domain

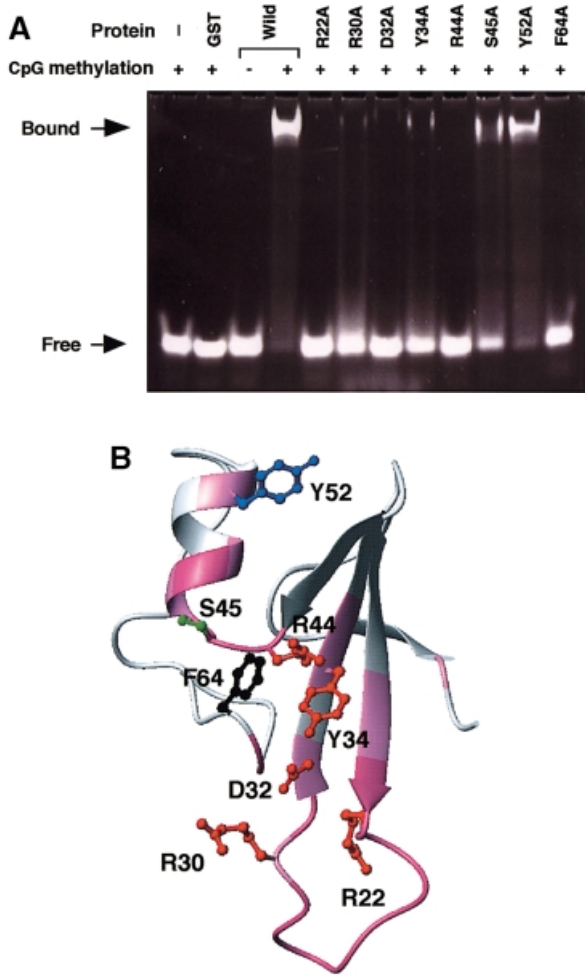

Fig. 5. Mutational analysis of MBD. (**A**) Gel-shift assay to examine the DNA binding of different mutants of MBD1 MBD. The purified proteins in their GST-fusion form and the 12mer oligonucleotide duplex carrying a methyl-CpG pair were used. The wild-type and mutated residues are indicated above each lane. (**B**) Stick representation of residues, the mutations of which led to a large reduction in DNA binding (red), viewed in the same orientation as for Figures 2–4. The side chains of Ser45 and Tyr52, whose mutations had moderate or little impact, are shown in green and blue, respectively. The mutation of Phe64 (shown in black) caused disruption of the native tertiary structure and a consequent loss of the DNA binding activity. Main chains of residues strongly affected by addition of methyl-CpG DNA are in dark magenta ($\delta_{av}/\delta_{max} > 0.2$) or light magenta ($0.1 < \delta_{av}/\delta_{max} \leq 0.2$).

of the Tn916 integrase and the GCC-box binding domain of AtERF1 have a three-stranded antiparallel β-sheet and a C-terminal helix, which are packed against each other, thus resembling the MBD (Allen *et al*., 1998; Wojciak *et al*., 1999), although there is no apparent sequence similarity between the MBD and these proteins. RMSDs between MBD and the integrase (PDB accession code 2bb8) (1.97 ± 0.08 Å over the 22 C_α coordinates for residues in β2, β3, β4 and α1 of MBD), and MBD and the GCC-box binding domain (PDB accession code 2gcc) (3.42 ± 0.04 Å over the 22 C_α coordinates for residues in β2, β3, β4 and α1 of MBD), indicate a distant similarity in topology. However, neither the integrase nor the GCC-box binding domain has a hairpin loop similar to that of the MBD. To check the uniqueness of the MBD fold, the coordinate of the lowest energy structure was compared against a database of known structures, using the DALI server version 2.0 (Holm and Sander, 1993). This analysis showed that there is no previously determined structure with the Z score ≥2.0, suggesting that MBD is like no other structure in the database. The comparison with the nucleic acid binding domain of transcriptional elongation factor TFIIS (Qian *et al*., 1993) gave the highest Z score (1.99 on the average for the 25 calculated structures of the MBD). The structure of this domain exhibits a three-stranded antiparallel β-sheet and a Cys-4 Zn^{2+} binding site and contains no α helix or loop similar to those of MBD; therefore, the similarity is limited to the β-sheet. The Z scores between the lowest energy structure of MBD and the structue of the Tn916 integrase, and the lowest energy structure of MBD and the structure of GCC-binding domain are 1.0 and 0.8, respectively, thereby suggesting the novelty of the MBD fold.

Important residues for DNA binding

The results of site-directed mutagenesis experiments are in good agreement with the DNA binding site suggested from chemical shift perturbation experiments. The putative DNA binding site is made up of several conserved residues. Of these, the following five residues were shown to be important for the DNA binding by the mutagenesis: Arg22, Arg30, Asp32, Tyr34 and Arg44.

One of them, Tyr34, forms a continuous hydrophobic patch in the middle of the putative DNA binding site, together with Phe64 (Figure 3A). The aromatic ring of Tyr34 is largely exposed to the solvent (>80% in the ensemble of the 25 NMR structures), while that of Phe64 on one face projects inward to establish the protein core and on the other side is exposed. There are no other exposed methyl or aromatic groups of conserved residues in or adjacent to the DNA binding site, except for Tyr52 that is located at the far edge of the site (Figure 4B). However, the mutation of Tyr52 had little impact on the DNA binding activity. In previous structure determinations of the complexes of sequence-specific DNA binding proteins with DNAs, the 5-methyl groups of thymidines were commonly recognized by hydrophobic groups of the proteins, such as methyl groups, aromatic rings or long aliphatic chains, through hydrophobic interactions (Chuprina *et al*., 1993; Suzuki *et al*., 1995a,b; König *et al*., 1996; L.Chen *et al*., 1998). As the 5-methyl groups of methyl-cytidines in the major groove of a B-form DNA are considered to be located at a similar position to the 5-methyl groups of thymidines, it can be assumed that they are mainly recognized through similar hydrophobic contacts. These observations suggest that the aromatic side chain of Tyr34 serves as a part of the contact site for methyl groups in methyl-CpG.

The mutation of another conserved residue, Asp32, located adjacent to Tyr34, also led to a large reduction in DNA binding activity, further suggesting the importance of Tyr34 (Figures 3A and 5A). This residue is invariant in the family, and is strongly affected by DNA titration (Figure 4B). Because of its negatively charged side chain, it is unlikely that this residue interacts with the negatively charged phosphate backbone of a DNA. The side chain carboxyl group of Asp is capable of forming hydrogen bonds with a cytidine base (Suzuki *et al*., 1995a), suggesting that Asp32 may be responsible for recognition of

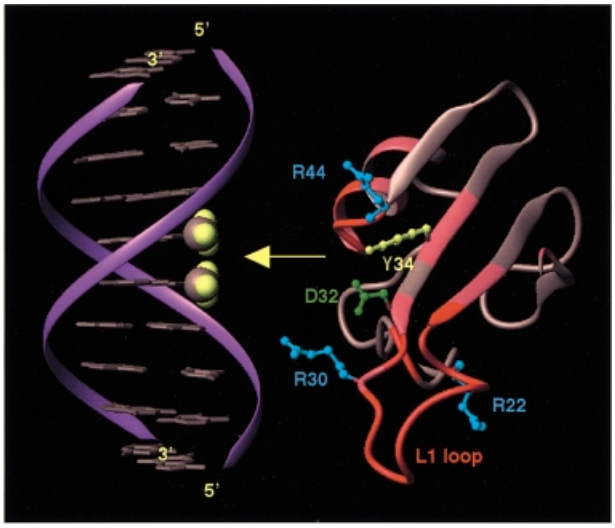

Fig. 6. Putative DNA binding site of MBD. Stick representation (Koradi *et al*., 1996) of the MBD with selected conserved residues in the proposed DNA binding site. Basic residues are colored in blue; aromatic residues, yellow; an acidic residue, green. Main chains of residues strongly affected by addition of methyl-CpG DNA are colored in red ($\delta_{av}/\delta_{max} > 0.1$). The molecule is rotated approximately –90° about the vertical axis relative to that shown in Figures 2–5. B-form DNA is also shown in the left-hand figure, with methyl groups in the symmetric methyl-CpG highlighted in yellow. With the DNA binding site placed in the major groove of the B-form DNA, loop L1 and the linker between strand β4 and helix α1 are located close to the phosphate backbone. The side chains of Tyr34 and Asp32 can come into contact with the methyl-CpG.

the cytosine in the methyl-CpG sequence. The mutation of another residue neighboring the hydrophobic patch, Ser45, led to a moderate loss of the binding activity (Figure 5A).

The mutation of either of two basic residues, Arg22 or Arg30, in flexible loop L1 led to a reduction in DNA binding activity. These mutations, as well as results of chemical shift perturbation experiments, indicated the significance of loop L1. However, because there are no conserved hydrophobic residues in loop L1, it probably does not serve as a contact site with the methyl groups of the methylated DNA. It is of interest that Arg30 was seen to be needed for sufficient DNA binding. While all the other functional MBD family members have Arg or Lys at position 30, mammalian MBD3 has a histidine at this position (Figure 1). This MBD3 does not bind to methyl-CpG, despite a high sequence homology with other functional MBD family members (Hendrich and Bird, 1998; Zhang *et al.*,1999). The present mutagenesis related results suggest that the loss of binding of the MBD3 may be due to this substitution. Interestingly, unlike its mammalian counterpart, *X.laevis* MBD3 has a lysine at this position, and exhibits selectivity for methylated DNA (Wade *et al*., 1999), further suggesting the importance of basic residues at this position in the MBD family. *Xenopus laevis* MBD3 LF has an insertion of 20 amino acid residues at the junction between strand β3 and the following loop (Wade *et al*., 1999), and the presence of this insertion probably causes a loss of the DNA binding of MBD3 LF. *Drosophila* MBD-like (Wade *et al*., 1999) lacks a sequence corresponding to strand β4, helix α1 and the hairpin loop; therefore, the folding of this protein is probably dissimilar to that of MBD.

Model for the interaction between MBD and methyl-CpG

Visual inspection revealed that size and curvature of the putative DNA binding surface of the MBD fits the major groove of a standard B-form DNA. Using the chemical shift perturbation results and locations of the important residues determined through mutagenesis, we present a putative model for the complex formation between MBD1 MBD and methyl-CpG DNA, as shown in Figure 6. Assuming that Tyr34 and Asp32 are important for recognition of the methyl-CpG sequence, the side chains of Tyr34 and Asp32 are placed in close proximity to the CG sequence. The residues of flexible loop L1 and the linker between β4 and α1 could have large interfaces with the phosphate backbone in a methylation-independent manner, as seen in winged helix–turn–helix protein–DNA interactions (Clark *et al*., 1993; Furui *et al*., 1998). This is particularly the case for two basic residues, Arg30 and Arg44, the mutation of which had effects.

The structure of the MBD is compact, and its putative DNA binding site is confined to the limited surface of the fold. This structural feature seems to be consistent with a recent observation that no major rearrangement of core-histone DNA contacts occurs upon binding of an MBD protein, MeCP2, to exposed methyl-CpG sites in nucleosomal DNA (Chandler *et al*., 1999).

Conclusion

CpG methylation, a major modification which occurs in the mammalian genome, is important in differential control of gene expression, and is involved in alteration of chromatin structure and genome stability. These effects are mainly the result of interactions between methlyated sites and methyl-CpG binding proteins. Methyl-CpG binding proteins are something of an enigma: the recognition site is short and can only be efficiently distinguished by the presence of two methyl groups. Our data provide the basis for a model of the unique interaction of the MBD with methyl-CpG, and a framework for the structural and functional analysis of the MBD family, in particular members such as MBD2 and MBD4/MED1.

Materials and methods

Sample preparation

^{15}N-labeled and ^{15}N/^{13}C-labeled GST–MBD1$_{1–84}$ fusion proteins were expressed in *Escherichia coli* strain DH5α grown in synthetic media containing $^{15}NH_4Cl$ and $^{15}NH_4Cl$/[^{13}C]glucose, respectively, and then purified chromatographically. The nine C-terminal residues were degraded in the bacterial cells and during steps of purification. After cleavage by thrombin, MBD1$_{1–75}$ with an extra Gly–Ser sequence at its N-terminus, referred to as MBD1 MBD, was further purified chromatographically, and analyzed by TOF mass spectroscopy and N-terminal sequencing, or TOF mass spectroscopy. A palindromic 12mer oligonucleotide with methyl-CpG, 5′-GTATCmCGGATAC-3′, was chemically synthesized, annealed and used for chemical shift perturbation experiments. Samples for NMR measurements typically comprised 1.3 mM MBD1 MBD in 20 mM potassium phosphate buffer pH 6.5, 50 mM KCl, 5mM DTT.

Structure determination

NMR spectra were acquired at 303 K with a Bruker DMX500, DRX500 or DRX800 NMR spectrometer. For assignments of the ^{1}H, ^{15}N and ^{13}C resonances, a series of three-dimensional experiments (HNCA, HNCO, HN(CA)CO, CBCA(CO)NH, CBCANH, C(CO)NH, H(CCO)NH and HCCH-TOCSY) were done using the ^{15}N, ^{13}C-labeled protein (Cavanagh *et al*., 1996). The stereospecific assignment of the methyl groups of Leu

and Val residues was achieved using a 15% fractionally ^{13}C-labeled protein as described (Hu and Zuiderweg, 1996). Distance information was obtained by means of ^{15}N- or ^{13}C-resolved 3D, or ^{15}N, ^{13}C- or ^{13}C, ^{13}C-resolved 4D NOESY experiments with a mixing time of 100 ms (Cavanagh *et al*., 1996). For torsion angle constraints, the backbone vicinal coupling constants ($^3J_{H_NH_\alpha}$) were determined by means of HMQC-J (Cavanagh *et al*., 1996). The torsion angles, χ_1, of Tyr35, Tyr52, Phe62 and Phe64 were estimated from $^3J_{C'C_\gamma}$ and $^3J_{NC_\gamma}$ coupling constants (Hu *et al*., 1997). The structure of MBD1 MBD was calculated with simulated annealing protocols, using X-PLOR (Brunger, 1993). The NOE connectivities from strong, medium and weak cross peaks were categorized, and assumed to correspond to the upper limits for proton–proton distances of 3.0, 4.0 and 5.0 Å, respectively. In the final steps of structure calculation, restraints were included for 15 slowly exchanging backbone amides obtained at 283 K, pH 5.0 [2.8–3.4 Å (N–O), 1.8–2.4 Å (H–O)]. In total, 1270 NOE constraints (491 intraresidual, 302 sequential, 165 medium range and 312 long range) and 44 dihedral angle constraints were used. A total of 125 structures was calculated. Of these, the 25 structures that showed the lowest energy were selected, and analyzed using MOLMOL (Koradi *et al*., 1996), AQUA and PROCHECK-NMR (Laskowski *et al*., 1996) software.

Chemical shift perturbation

The (^{15}N, ^1H) HSQC spectrum of MBD1 MBD complexed with a 12mer oligonucleotide duplex with methyl-CpG was compared with that of the free MBD. Since the interaction exhibited a slow exchange mode, the backbone assignment of the MBD in the complex was determined from results of CBCA(CO)NH, CBCANH, HN(CA)CO and HNCO experiments, using $^{15}N/^{13}C$-labeled MBD1 MBD complexed with the oligonucleotide. To evaluate the effects, the normalized weighted average shift difference of the ^1H and ^{15}N resonances, δ_{av}/δ_{max}, for each residue is calculated. The weighted average shift difference, δ_{av}, is calculated as $\{[\delta_{^1H}{}^2 + (\delta_{^{15}N}/5)^2]/2\}^{1/2}$, where $\delta_{^1H}$ and $\delta_{^{15}N}$ are the differences in p.p.m. between the free and bound chemical shifts. δ_{max} is the maximum observed weighted average shift difference (Garrett *et al*., 1997; Foster *et al*., 1998).

Mutational analyses

Mutant constructs were prepared with a GeneEditor *in vitro* site-directed mutagenesis system (Promega), according to the manufacturer's instructions. The coding regions were sequenced. GST–MBD fusion proteins were purified. To examine DNA binding, DNA binding reactions were carried out in 10 µl of 10 mM Tris–HCl pH 8.0, 5 mM $MgCl_2$, 5 mM DTT, 5% glycerol. The mixtures contained 1.7 µg of each purified protein and 74 ng of the methylated or unmethylated 12mer oligonucleotide duplex. The reactions were carried out for 30 min at room temperature. The binding reaction mixtures were loaded on a 10% acrylamide gel and run in TBE. The gels were stained with SYBR green I (Molecular Probes).

Structure data submission

The atomic coordinates have been deposited in the Protein Data Bank (accession code 1D9N).

Acknowledgements

We thank Y.Kyogoku, S.Tajima, H.Matsuo, M.Wälchli and T.Ikegami for helpful discussions. We thank M.Ohara for careful reading of the manuscript. This study was supported by grants to M.S. and M.N. from the Ministry of Education, Science, Sports and Culture of Japan.

References

Allen,M.D., Yamasaki,K., Ohme-Takagi,M., Tateno,M. and Suzuki,M. (1998) A novel mode of DNA recognition by a β-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA. *EMBO J.*, **15**, 5484–5496.

Ballabio,A. and Willard,H.F. (1992) Mammalian X-chromosome inactivation and the XIST gene. *Curr. Opin. Genet. Dev.*, **2**, 439–447.

Baylin,S.B., Herman,J.G., Graff,J.R., Vertino,P.M. and Issa,J.P. (1998) Alterations in DNA methylation: a fundamental aspect of neoplasia. *Adv. Cancer. Res.*, **72**, 141–196.

Bhattacharya,S.K., Ramchandani,S., Cervoni,N. and Szyf,M. (1999) A mammalian protein with specific demethylase activity for mCpG DNA. *Nature*, **397**, 579–583.

Bellacosa,A., Cicchillitti,L., Schepis,F., Riccio,A., Yeung,A.T., Matsumoto,Y., Golemis,E.A., Genuardi,M. and Neri,G. (1999) MED1, a novel human methyl-CpG-binding endonuclease, interacts with DNA mismatch repair protein MLH1. *Proc. Natl Acad. Sci. USA*, **96**, 3969–3974.

Bergman,Y. and Mostoslavsky,R. (1998) DNA demethylation: turning genes on. *Biol. Chem.*, **379**, 401–407.

Bird,A. (1992) The essentials of DNA methylation. *Cell*, **70**, 5–8.

Boyes,J. and Bird,A. (1991) DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. *Cell*, **64**, 1123–1134.

Boyes,J. and Bird,A. (1992) Repression of genes by DNA methylation depends upon CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. *EMBO J.*, **11**, 327–333.

Brunger,A.T. (1993) *X-PLOR 3.1: A System for X-ray Crystallography and NMR*. Yale University Press, New Haven, CT.

Cavanagh,J., Fairbrother,W.J., Palmer,A.G.,III and Skelton,N.J. (1996) *Protein NMR Spectroscopy.* Academic Press, San Diego, CA.

Chandler,S.P., Guschin,D., Landsberger,N. and Wolffe,A.P. (1999) The methyl-CpG binding transcriptional repressor MeCP2 stably associates with nucleosomal DNA. *Biochemistry*, **38**, 7008–7018.

Chen,L., Glover,J.N.M., Hogan,P.G., Rao,A. and Harrison,S.C. (1998a) Structure of the DNA-binding domains from NFAT, Fos and Jun bound specifically to DNA. *Nature*, **392**, 42–48.

Chen,R.Z., Pettersson,U., Beard,C., Jackson-Grusby,L. and Jaenisch,R. (1998b) DNA hypomethylation leads to elevated mutation rates. *Nature*, **395**, 89–93.

Chuprina,V.P., Rullmann,J.A.C., Lamerichs,R.M.J.N., van Boom,J.H., Boelens,R. and Kaptein,R. (1993) Structure of the complex of *lac* repressor headpiece and an 11 base-pair half-operator determined by nuclear magnetic resonance spectroscopy and restrained molecular dynamics. *J. Mol. Biol.*, **234**, 446–462.

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

Cross,S.H., Meehan,R.R., Nan,S. and Bird,A. (1997) A component of the transcriptional repressor MeCP1 shares a motif with DNA methyltransferase and HRX proteins. *Nature Genet.*, **16**, 256–259.

Eden,S., Hashimshony,T., Keshet,I. and Cedar,H. (1998) DNA methylation models histone acetylation. *Nature*, **394**, 842.

Fearon,E.R. and Vogelstein,B. (1990) A genetic model for colorectal tumorgenesis. *Cell*, **61**, 759–767.

Feinberg,A.P., Rainier,S. and DeBaun,M.R. (1995) Genomic imprinting, DNA methylation and cancer. *J. Natl Cancer Inst. Monogr.*, **17**, 21–26.

Foster,M.P., Wuttke,D.S., Clemens,K.R., Jahnke,W., Radhakrishnan,I., Tennant,L., Reymond,M., Chung,J. And Wright,P.E. (1998) Chemical shift as a probe of molecular interfaces: NMR studies of DNA binding by the three amino-terminal zinc finger domains from transcription factor IIIA. *J. Biomol. NMR*, **12**, 51–71.

Fujita,N., Takebayashi,S., Okumura,K., Kudo,S., Chiba,T., Saya,H. and Nakao,M. (1999) Methylation-mediated transcriptional silencing in euchromatin by methyl-CpG binding protein MBD1 isoforms. *Mol. Cell. Biol.*, **19**, 6415–26.

Furui,J., Uegaki,K., Yamazaki,T., Shirakawa,M., Swindells,M.B., Harada,H., Taniguchi,T. and Kyogoku,Y. (1998) Solution structure of the IRF-2 DNA binding domain: a novel subgroup of the winged helix-turn-helix family. *Structure*, **6**, 491–500.

Garrett,D.S., Seok,Y., Peterkofsky,A., Clore,G.M. and Gronenborn,A.M. (1997) Identification by NMR of the binding surface for the histidine-containing phosphocarrier protein Hpr on the N-terminal domain of enzyme I of the *Escherichia coli* phosphotransferase system. *Biochemistry*, **36**, 4393–4398.

Hendrich,B. and Bird,A. (1998) Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol. Cell. Biol.*, **18**, 6538–6547.

Hendrich,B., Hardeland,U., Ng,H., Jiricny,J. and Bird,A. (1999) The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature*, **401**, 301–304.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Hu,W. and Zuiderweg,E.R.P. (1996) Stereospecific assignments of Val and Leu methyl groups in a selectively ^{13}C-labeled 18 kDa polypeptide using 3D CT- (H)CCH-COSY and 2D $^1J_{C\text{-}C}$ edited heteronuclear correlation experiments. *J. Magn. Reson. B*, **113**, 70–75.

Hu,J.-S., Grzesiek,S. and Bax,A. (1997) Two-dimensional NMR methods for determining $\chi 1$ angles of aromatic residues in proteins from three-bond $J_{C'C\gamma}$ and $J_{NC\gamma}$ couplings. *J. Am. Chem. Soc.*, **119**, 1803–1804.

Jones,P.L., Veenstra,G.J.C., Wade,P.A., Vermaak,D., Kass,S.U., Landsberger,N., Strouboulis,H. and Wolffe,A.P. (1998) Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nature Genet.*, **19**, 187–191.

Kass,S.U., Landsberger,N. and Wolffe,A.P. (1997a) DNA methylation directs a time-dependent repression of transcription inititation. *Curr. Biol.*, **7**, 157–165.

Kass,S.U., Pruss,D. and Wolffe,A.P. (1997b) How does DNA methylation repress transcription? *Trends Genet.*, **13**, 444–449.

König,P., Giraldo,R., Chapman,L. and Rhodes,D. (1996) The crystal structure of the DNA-binding domain of yeast Rap1 in complex with telomeric DNA. *Cell*, **85**, 125–136.

Koradi,R., Billeter,M. and Wuethrich,K. (1996) MOLMOL: a program for the display and analysis of macromolecular structures. *J. Mol. Graph.*, **14**, 51–55.

Lalande,M. (1996) Parental imprinting and human disease. *Annu. Rev. Genet.*, **30**, 173–195.

Laskowski,R.A., Rullmann,J.A.C., MacArthur,M.W., Kaptein,R. and Thornton,J.M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR*, **8**, 477–486.

Lewis,J.D., Meehan,R.R., Henzel,W.J., Maurer-Fogy,I., Jeppesen,P., Klein,F. and Bird,A. (1992) Purification, sequence and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell*, **69**, 905–914.

Li,E., Beard,C. and Jaenisch,R. (1993) Role for DNA methylation in genomic imprinting. *Nature*, **366**, 362–365.

Meehan,R.R., Lewis,J.D., McKay,S., Kleiner,E.L. and Bird,A.P. (1989) Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell*, **69**, 905–914.

Nan,X., Meehan,R.R. and Bird,A. (1993) Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic Acids Res.*, **21**, 4886–4892.

Nan,X., Campoy,F.J. and Bird,A. (1997) MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell*, **88**, 471–481.

Nan,X., Ng,H.-H., Johnson,C.A., Laherty,C.D., Turner,B.M., Eisenman,R.N. and Bird,A. (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*, **393**, 386–389.

Ng,H.-H., Zhang,Y., Hendrich,B., Johnson,C.A., Turner,B.M., Erdjument-Bromage,H., Tempst,P., Reinberg,D. and Bird,A. (1999) MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nature Genet.*, **23**, 58–61.

Nicholls,A., Sharp,K.A. and Honig,B. (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins*, **11**, 281–296.

Qian,X., Gozani,S.N., Yoon,H., Jeon,C.J., Agarwal,K. and Weiss,M.A. (1993) Novel zinc finger motif in the basal transcriptional machinery: three-dimensional NMR studies of the nucleic acid binding domain of transcriptional elongation factor TFIIS. *Biochemistry*, **32**, 9944–9959.

Razin,A. (1998) CpG methylation, chromatin structure and gene silencing—a three-way connection. *EMBO J.*, **17**, 4905–4908.

Siegfried,Z. and Cedar,H. (1997) DNA methylation: a molecular lock. *Curr. Biol.*, **7**, 305–307.

Surani,M.A. (1998) Imprinting and the initiation of gene silencing in the germ line. *Cell*, **93**, 309–312.

Sutcliffe,J.S., Nakao,M., Christian,S., Orstavik,K.H., Tommerup,N., Ledbetter,D.H. and Beaudet,A.L. (1994) Deletions of a differentially methylated CpG island at the SNRPN gene define a putative imprinting control region. *Nature Genet.*, **8**, 52–58.

Suzuki,M., Brenner,S.E., Gerstein,M. and Yagi,N. (1995a) DNA recognition code of transcription factors. *Protein Eng.*, **8**, 319–328.

Suzuki,M., Yagi,N. and Gerstein,M. (1995b) DNA recognition and superstructure formation by helix-turn-helix proteins. *Protein Eng.*, **8**, 329–338.

Tajima,S. and Suetake,I. (1998) Regulation and function of DNA methylation in vertebrates. *J. Biochem.*, **123**, 993–999.

Tate,P., Skarnes,W. and Bird,A. (1996) The methyl-CpG binding protein MeCP2 is essential for embryonic development in the mouse. *Nature Genet.*, **12**, 205–208.

Wade,P.A., Jones,P.L., Vermaak,D. and Wolffe,A.P. (1998) A multiple subunit Mi-2 histone deacetylase from *Xenopus laevis* cofractionates with an associated Snf2 superfamily ATPase. *Curr. Biol.*, **8**, 843–846.

Wade,P.A., Gegonne,A., Jones,P.L., Ballestar,E., Aubry,F. and Wolffe,A.P. (1999) Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nature Genet.*, **23**, 62–66.

Wishart,D.S. and Sykes,B.D. (1994) Chemical shifts as a tool for structure determination. *Methods Enzymol.*, **239**, 363–392.

Wojciak,J.M., Connolly,K.M. and Clubb,R.T. (1999) NMR structure of the Tn916 integrase-DNA complex. *Nature Struct. Biol.*, **6**, 366–373.

Zhang,Y., Ng,H.-H., Erdjument-Bromage,H., Tempst,P., Bird,A. and Reinberg,D. (1999) Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev.*, **13**, 1924–1935.

Received August 23, 1999; revised and accepted October 6, 1999

Note added in proof

After submission of this manuscript, the solution structure of the MBD from MeCP2 was published [Wakefield,R.I.D., Smith,B.O., Nan,X., Free,A., Soteriou,A., Uhrin,D., Bird,A.P. and Barlow,P.N. (1999) *J. Mol. Biol.*, **291**, 1055–1065]. The structure and the result of chemical shift perturbation with DNA are similar to those of MBD1 shown in this paper.