THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 278, No. 7, Issue of February 14, pp. 5285–5291, 2003
Printed in U.S.A.

Mismatch Repair in Methylated DNA

STRUCTURE AND ACTIVITY OF THE MISMATCH-SPECIFIC THYMINE GLYCOSYLASE DOMAIN OF METHYL-CpG-BINDING PROTEIN MBD4*

Received for publication, October 23, 2002
Published, JBC Papers in Press, November 26, DOI 10.1074/jbc.M210884200

Peiying Wu‡, Chen Qiu‡§¶, Anjum Sohail¶‖, Xing Zhang‡, Ashok S. Bhagwat‖, and Xiaodong Cheng‡**

From the Departments of ‡Biochemistry and §Chemistry, Emory University, Atlanta, Georgia 30322 and the ‖Department of Chemistry, Wayne State University, Detroit, Michigan 48202-3489

MBD4 is a member of the methyl-CpG-binding protein family. It contains two DNA binding domains, an amino-proximal methyl-CpG binding domain (MBD) and a C-terminal mismatch-specific glycosylase domain. Limited *in vitro* proteolysis of mouse MBD4 yields two stable fragments: a 139-residue fragment including the MBD, and the other 155-residue fragment including the glycosylase domain. Here we show that the latter fragment is active as a glycosylase on a DNA duplex containing a G:T mismatch within a CpG sequence context. The crystal structure confirmed the C-terminal domain is a member of the helix-hairpin-helix DNA glycosylase superfamily. The MBD4 active site is situated in a cleft that likely orients and binds DNA. Modeling studies suggest the mismatched target nucleotide will be flipped out into the active site where candidate residues for catalysis and substrate specificity are present.

MBD4 is a mammalian DNA glycosylase that excises thymines from G:T mispairs and contains both a methyl-CpG binding domain (MBD)[1] and a domain found in the *Escherichia coli* endonuclease III class of DNA glycosylases (1). It has preference for G:T mismatches within a CpG sequence context (1), and hence this enzyme can act upon G:T mismatches that result from the deamination of 5-methylcytosines (5mC) at CpG sites. The importance of this enzyme for mutation avoidance in mammals is confirmed by an increase in 5mC to T mutations in *Mbd4*$^{-/-}$ Big Blue mouse and by increased occurrence of colon carcinoma in *Mbd4*$^{-/-}$ *Apc*$^{Min/+}$ mice (2). Additionally, studies of MBD4 (also called MED1) using the yeast two-hybrid system have shown that it interacts with MLH1 (a protein implicated in mismatch repair) and suggest a role for this enzyme in maintaining genome stability (3). Consistent with this observation, it is found that MBD4 is mutated in 26–43% of human colorectal tumors that show microsatellite instability (4).

MBD4 is not the only DNA glycosylase reported to excise thymines from G:T mismatches. Another enzyme, named thymine-DNA glycosylase (TDG), was identified earlier to have this ability (5). However, TDG is unrelated to MBD4 and belongs to the same structural superfamily as the uracil-excising enzymes UDG (6) and SMUG1 (7). MBD4 also differs from TDG in its substrate preference. Whereas the preferred substrates for TDG are N^4-ethenocytosine or uracil paired with a G (8), MBD4 prefers thymine over N^4-ethenocytosine (9). Recombinant MBD4 can also remove uracil, 5-fluorouracil, and 5mC at a low rate, particularly when these bases are opposite a guanine within CpG dinucleotides (1, 9, 10).

The MBD domain of MBD4 is similar to domains within four other mammalian proteins, MeCP2, MBD1, MBD2, and MBD3 (reviewed in Refs. 11 and 12). The latter proteins are involved in suppressing transcription in regions of heavy CpG methylation, but no such role has been ascribed to MBD4. Whereas the NMR structures of the MBD domains from MBD1 (13, 14) and MeCP2 (15) have been elucidated, no structural information regarding the glycosylase domain of MBD4 is available.

Here we present the crystal structure of the C-terminal glycosylase domain of MBD4, and we show that it belongs to the helix-hairpin-helix DNA glycosylase superfamily. The glycosylase domain alone is active on DNA duplex containing a G:T mismatch within a CpG sequence context.

EXPERIMENTAL PROCEDURES

Overexpression and Purification—The full-length mouse MBD4 was expressed as a His-tagged fusion protein in vector pET6H (16). The four fragments of MBD4 (Fig. 1*A*), amino acids 49–187 (MBD domain), 400–554 (glycosylase domain or Δ399), 49–554 (Δ48), and 429–554 (Δ428), were cloned into a modified pET28b (Novagen) vector, which contains an N-terminal tag of MGHHHHHH and accepts an *Nde*I-*Eco*RI insert. The Δ428 fragment was also expressed as a GST fusion in pGEX2T vector (Amersham Biosciences). *E. coli* strain BL21(DE3) carrying respective plasmid was grown in LB media supplemented with appropriate antibiotics at 37 °C to A_{600} = 0.6, shifted to 22 °C, and induced with 0.4 mM isopropyl-1-thio-β-D-galactopyranoside overnight at 22 °C, except that the full-length MBD4 was induced at 37 °C for 1 h.

The full-length and Δ48 proteins were purified from cleared lysates using three successive chromatography steps as follows: a nickel chelate column, a HiTrap heparin column, and Superdex 200 (Amersham Biosciences). The proteins were stored in 20 mM potassium phosphate, pH 7.5, 1 mM EDTA, 0.1% 2-mercaptoethanol, and 0.2 M NaCl.

The MBD domain and the glycosylase domain were purified using nickel chelate, HiTrap Q, and Superdex 75 columns. The proteins were stored in a high salt buffer for crystallization (20 mM Tris-HCl, pH 7.5, 1 mM EDTA, 1 mM DTT, 5% glycerol, and 0.5 M NaCl) or in a low salt buffer for activity assay (20 mM Tris-HCl, pH 8.0, 1 mM EDTA, 1 mM DTT, 50% glycerol, and 50 mM NaCl).

The GST-Δ428 was purified using glutathione-Sepharose 4B (Amer-

* This work was supported in part by National Institutes of Health Grants GM49245 (to X. C.) and GM57200 (to A. S. B) and the Georgia Research Alliance. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The atomic coordinates and structure factors (code 1NGN) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/).

¶ Both authors contributed equally to this work.

** To whom correspondence should be addressed: Dept. of Biochemistry, Emory University, 1510 Clifton Rd., Atlanta, GA 30322. Tel.: 404-727-8491; Fax: 404-727-3746; E-mail: xcheng@emory.edu.

[1] The abbreviations used are: MBD, methyl-CpG binding domain; 5mC, 5-methylcytosines; TDG, thymine-DNA glycosylase; GST, glutathione *S*-transferase; DTT, dithiothreitol.

TABLE I
X-ray data collection and structural refinement statistics

Crystal	Native	SeMet		
Wavelength (Å)	1.54	0.97890	0.97865	0.95
Resolution (Å)	20-2.1/2.14-2.1	20-2.1/2.15-2.1	20-2.1/2.15-2.1	20-2.1/2.15-2.1
Completeness (%)	99.5/99.8	80.4/29.1	80.2/28.7	85.1/39.6
R-merge	0.046/0.098	0.028/0.061	0.026/0.061	0.025/0.066
I/σ(I)	42.7	51.2	50.7	47.9
Unique reflections	12,345/585	9,965/230	9,944/228	10,531/314
Total reflections	107,778	58,654	55,029	63,278
Ramachandran plot				
Most favored regions (%)		93.4		
Additionally allowed regions (%)		6.6		
Root mean square deviation from ideality				
Bond lengths (Å)		0.01		
Bond angles (°)		1.5		
Dihedrals (°)		23.4		
Improper (°)		1.4		
G-factor				
Dihedrals		0.21		
Covalent		0.45		
Overall		0.31		
Average thermal factor (Å^2)				
Main chain		19.7		
Side chain		20.9		

sham Biosciences) and HiTrap Q column. The protein was stored in the low salt buffer.

Limited Proteolysis of Full-length MBD4 Protein—Protease V8 was added, as described previously (17), to MBD4 for 15 min incubation at room temperature. Addition of the protease inhibitor 1-chloro-3-tosylamido-7-amino-2-heptanone stopped the reaction, and two major MBD4 fragments were observed via SDS-PAGE. The mass of the two fragments was determined by electrospray ionization mass spectrometry to be 15,353.5 and 18,537.2 Da. The N-terminal sequences of the two fragments were also determined. Combining these results allowed us to deduce that the fragments represent residues 49–187 (the MBD domain) and residues 400–554 (the C-terminal glycosylase domain) (Fig. 1*A*).

Crystallography—Crystals of the MBD4 glycosylase domain stored in the high salt buffer were obtained in hanging drops under the conditions of 22–27% polyethylene glycol 2000 monomethyl ether, 190–230 mM ammonium sulfate, 10–15% ethylene glycol, 100 mM sodium citrate, pH 5.26. The initial drops were set up at 16 °C and moved to 4 °C after overnight incubation. Two crystal forms were observed, good looking diamond-shaped crystals appeared earlier at 16 °C but diffracted x-rays poorly. Rather unpromising colorless crystals appeared after 2–3 weeks only at 4 °C, but these diffracted x-rays strongly and belonged to space group $P3_121$, based on the systematic absence of reflections along the *z* axis. There is one molecule per asymmetric unit, and unit cell dimensions were a = b = 48.58 Å and c = 146.57 Å.

Selenomethionine-containing glycosylase domain Δ399 was expressed in a methionine auxotrophic *E. coli* strain B834(DE3) grown in LeMaster medium (18) supplemented with 25 μg/ml Se-methionine, and the protein was purified similarly to the native protein. X-ray diffraction data for the native and the single-site selenomethionine substitution crystals were collected (Table I), respectively, by an RAXIS-IV imaging plate detector equipped with a Rigaku rotating anode generator (50 kV, 100 mA) and a Brandeis B2 (1 × 1) CCD-based detector at the National Synchrotron Light Source beamline X12-C. The resulting images were processed using HKL (19). One surface selenium site (SeMet^{447} of αC) was determined by SOLVE (20). RESOLVE (21) was then used to modify the electron density map at 2.5-Å resolution (overall figure of merit 0.54). The modified map was of excellent quality to place amino acids 411–554 of MBD4 into the recognizable densities by using the graphic program O (22). Electron density was not observed for the first 11 residues (400–410 of the full-length protein). The resulting model was refined to 2.09 Å resolution using X-PLOR (23) (Table I), with a final crystallographic R-factor of 0.213 and R-free of 0.263 (for 9% of total 21,552 reflections).

Preparation of E. coli Extract—The His-tagged full-length MBD4, Δ48, MBD domain (amino acids 49–187), glycosylase domain Δ399, and Δ428 were all transformed into an *E. coli* strain BH161, which carries *ung*$^-$ and a copy of T7 RNA polymerase (24). Ten milliliters of culture was grown from each clone in LB media supplemented with 100 μg/ml ampicillin (full-length MBD4) or 50 μg/ml kanamycin (all MBD4 fragments) at 37 °C until the A_{600} reached 0.6. The cultures were induced by adding isopropyl-1-thio-β-D-galactopyranoside to a final concentration of 0.5 mM. After incubation at 37 °C for 3 h, the cells were recovered by centrifugation, and the cell pellets were washed with a buffer containing 20 mM Tris-HCl, pH 7.6, and 0.1 mM EDTA. Cells were resuspended in 0.5 ml of extraction buffer (20 mM Tris-HCl, pH 7.8, 0.1 mM EDTA, 5 mM 2-mercaptoethanol) containing 1 mg/ml lysozyme and incubated on ice for 15 min. Finally, the cells were broken by sonication on ice, and cell-free lysate was recovered following centrifugation at 12,000 × *g* for 15 min at 4 °C. Protein concentration in the extracts was determined using the Bradford Reagent (Bio-Rad).

Preparation of the Labeled DNA Substrate—The following oligonucleotides were obtained from Invitrogen:

T-oligo, 5′-GACTGGCTGCTCC*TG*GGCGAAGTGCC-3′;
G-oligo, 5′-GGGCACTTCGCC*CG*GGAGCAGCCAGTC-3′.

The oligonucleotides were gel-purified prior to their use. The T-oligo was labeled at the 5′ end using T4 polynucleotide kinase (New England Biolabs) in the presence of [γ-^{32}P]ATP (specific activity 6000 Ci/mmol, PerkinElmer Life Sciences). The reaction was terminated by heating it to 65 °C for 20 min. The labeled T-oligo was mixed with 3-fold molar excess of the unlabeled G-oligo in the STE buffer (150 mM NaCl, 10 mM Tris-HCl, pH 8.0, and 1 mM EDTA). The mixture was heated to 95 °C for 3 min and then slowly cooled to room temperature over a period of 2–3 h to promote duplex formation. The unincorporated [γ-^{32}P]ATP was removed from the labeled duplex by passage through a G-50 micro column (Amersham Biosciences).

DNA Glycosylase Assay—Twenty nM labeled duplex was equilibrated with nicking buffer (10 mM Tris-HCl, pH 8.0, 5 mM EDTA, 1 mM DTT, and 0.1 mg/ml bovine serum albumin), and the reactions were initiated by adding 100 ng of purified MBD4 variants (Fig. 5*A*) or 2 μg of cell-free extract (Fig. 5*B*). Following incubation at 37 °C for 1 h, the reaction was stopped by heating to 95 °C for 7 min in the presence of 0.1 M NaOH. Subsequently, 8 μl of gel loading dye (80% formamide, 10 mM EDTA, 1 mg/ml each of xylene cyanol and bromphenol blue) was added to the samples which were then heated to 95 °C and electrophoresed in 20% sequencing gel. The gel was exposed to a PhosphorImager screen (Amersham Biosciences), and the reaction products were quantified using ImageQuant software.

RESULTS

MBD4 Glycosylase Domain Structure—The overall structure of MBD4 glycosylase domain consists of 11 helices (αA to αK) (Fig. 1*B*) forming a single domain with a cleft in the middle (Fig. 2). Structural comparison with other DNA glycosylases (Fig. 3, *A* and *B*) reveals that the MBD4 glycosylase domain belongs to the helix-hairpin-helix (HhH) DNA glycosylase superfamily (25), named after a conserved structural motif αH-hairpin loop-αI (shown in *red* in Fig. 2*A*). The six helices before

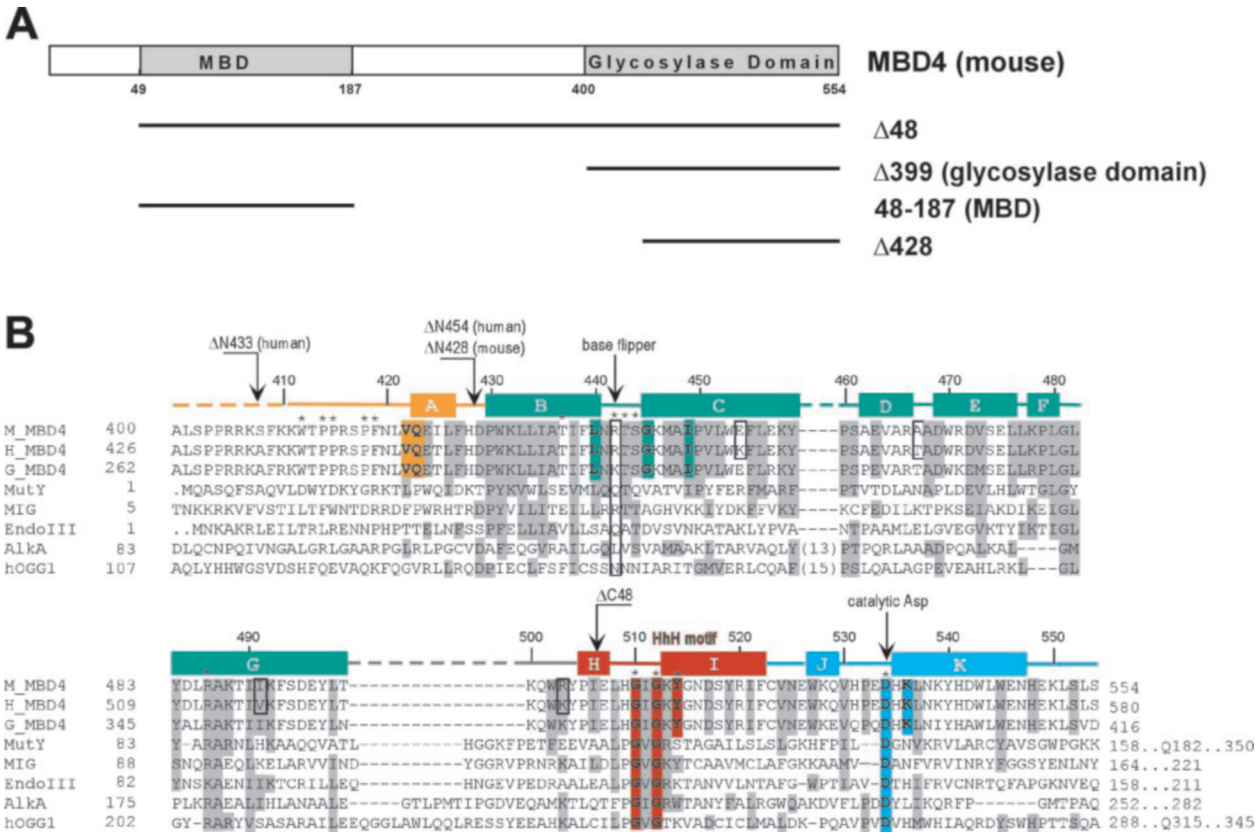

FIG. 1. *A*, schematic representation of mouse MBD4. Two stable fragments (*shaded*) were identified by limited proteolysis. Below the full-length mouse MBD4 are the depicted protein fragments (Δ48, MBD, glycosylase domain Δ399, and Δ428) used in this study. *B*, structure-guided sequence alignment of three MBD4 glycosylase domains (mouse, AAC68878; human, AAC68879; and Gallus, AAF68981) and five HhH glycosylases. We note that the reported MBD4 homolog in chickens only contains the glycosylase domain and has no consensus sequence for the N-terminal MBD (10). The secondary structure of MBD4 is indicated *above* the sequence and colored the same as in Fig. 2*A*. The *dashed lines* indicate gaps introduced to optimize alignments, and MBD4 has the shortest loop prior to the HhH motif. The *dots* indicate extra residues outside the glycosylase domain. The residues in the active site, proposed to interact with the extrahelical target nucleotide, are colored to match their associated region. The residues marked by * *above* the sequence are discussed in the text. The deletion mutants made in human MBD4 (10, 33) and mouse MBD4 (this study) are indicated by *arrows*. The four differences between mouse and human MBD4 sequences are *boxed*.

the HhH motif (αB to αG in green) are highly conserved structural elements among family members, forming the bottom of the cleft in the orientation shown (Fig. 2*A*). Among the known HhH enzymes, MBD4 has the shortest sequence following the HhH motif (Fig. 1*B*). The C-terminal helices αJ and αK, the short N-terminal helix αA, and its 12-residue preceding loop, the HhH motif, come together to form a hydrophobic core (Fig. 2*C*), forming the top of the cleft.

Model of the MBD4-DNA Complex—The high degree of structural similarity among HhH glycosylases allowed us to create a model of the MBD4 glycosylase domain bound to DNA. By using the coordinates of the AlkA-DNA (26) or hOGG1-DNA (27) complexes, we superimposed the protein components, and then the DNA was positioned over the surface of MBD4 with the cleft. Previous modeling studies of other HhH glycosylases MutY and EndoIII suggested that they bind to DNA in a manner similar to that of AlkA (26). Our modeling suggests that the MBD4 glycosylase domain also binds DNA similarly to AlkA and hOGG1, which bind DNA via the minor groove and bend it ~70° at the damaged base (26, 27).

The residues that contact the DNA backbone in the hOGG1 and AlkA structures occupy similar positions in the free MBD4 structure (Fig. 3*C*), and the MBD4 glycosylase domain could contact bent DNA without major physical distortion of the protein component (Fig. 3*D*). Two important DNA-binding loops are superimposed, the loop between helices αB and αC and the Gly-rich hairpin loop of HhH motif (Fig. 3*C*). Arg^{442} of MBD4, as well as Arg^{47} of MIG (28), is in the same position as Leu^{125} of AlkA (or Asn^{149} of hOGG1) that fills the space in the DNA duplex vacated by the flipped nucleotide. Thr^{443} of MBD4 is in the same position as Asn^{150} of hOGG1 that makes main chain contacts to the phosphate groups 3′ to the flipped nucleotide. Ser^{444} of MBD4 is in the position of Asn^{151} of hOGG1 that forms hydrogen bonds with the base 5′ immediate to the flipped nucleotide. It seems that the loop between helices αB and αC contains residues (Arg^{442}–Thr^{443}–Ser^{444}) important for DNA binding and base flipping.

Mechanisms for Recognition of Flipped Bases and Catalysis—First, where is the active site? In analogy to the AlkA-DNA (26) and hOGG1-DNA (27) complexes, the MBD4 cleft defines the location of the active site (Fig. 4*A*). The target nucleotide is likely to be flipped out from the DNA helix into the active-site cleft of the enzyme, in a similar manner to AlkA or hOGG1. The structural superimposition of the HhH glycosylase-DNA complexes and the unbound MBD4 reveals several informative features. Interestingly, the flipped base can only be docked into the active site by stacking the base between the side chains of Leu^{440} of αB and Lys^{536} of αK (Fig. 4*B*). Although these residues are not conserved in HhH glycosylases, similar stacking appears to be conserved: in hOGG1 8-oxoguanine is between Cys^{253} and Phe^{319} (27) and in MutY adenine soaked into the crystal lies between Leu^{40} and Met^{185} (29). Leu^{440} of MBD4 corresponds to Leu^{40} of MutY (Figs. 1*B* and 4*E*), whereas MutY Met^{185} corresponds to Phe^{319} of hOGG1.

A second question is where the key catalytic residues are located. Asp^{534}, the last residue prior to helix αK (Fig. 1*B*), is in

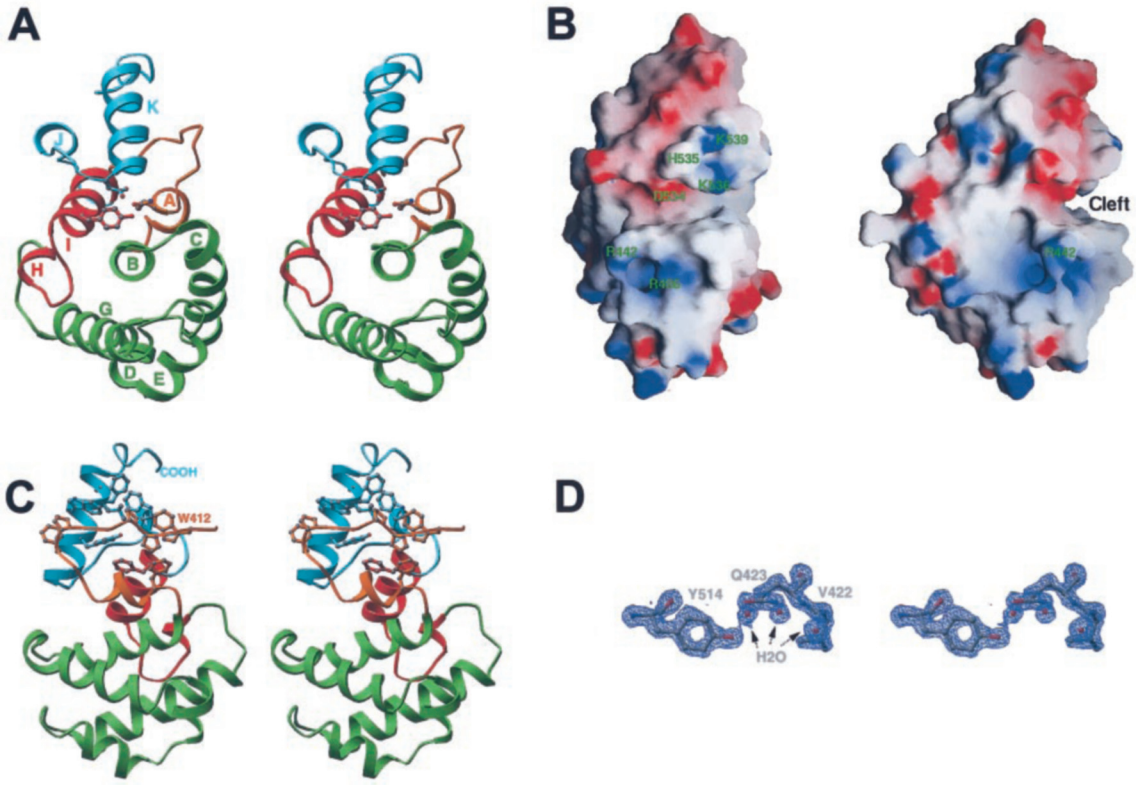

FIG. 2. **Structure of MBD4 glycosylase domain.** *A*, stereo view of Ribbon (36) diagram. The protein is colored according to Fig. 1*B*. The proposed thymine recognition residues Gln^{423} (*orange*) and Tyr^{514} (*red*), and the catalytic Asp^{534} (*cyan*) are shown in *ball-and-stick* model. *B*, two orthogonal views of solvent-accessible Grasp (37) surface with charge distribution (*blue* for positive charge, and *red* for negative charge). *Left*, the protein is shown in a similar orientation as in *A*. *Right*, the cleft predicted to hold the active site is visible after rotation ~90° on the vertical axis. *C*, stereo view of the hydrophobic core formed by the N-terminal loop and helix αA (*orange*), helix αI of HhH motif (*red*), and the C-terminal helices αK and αJ (*cyan*). The hydrophobic side chains are shown in *ball-and-stick model*, colored to match their associated region. *D*, stereo view of difference electron density (*blue*) contoured at 3.0σ above the mean, superimposed with Val^{422}, Gln^{423}, and Tyr^{514}, and three associated water molecules.

a position structurally equivalent to the catalytically important Asp^{238} of AlkA (26), Asp^{268} of hOGG1 (27), Asp^{138} of MutY (29), and Asp^{138} of EndoIII (30). Two mechanisms have been suggested for the function of this structurally conserved aspartic acid in HhH glycosylases: (i) it activates a catalytic nucleophile, which is either a water (29) or the ϵ-amino group of a lysine (27), for the attack on the deoxyribose C1′ carbon atom of the target nucleotide; or (ii) it directly assists base removal by protonating the leaving group of the substrate sugar (26). In the docking model of MBD4-thymine (Fig. 4*C*), the C1′ position of a modeled substrate is in direct contact (~3.0 Å) with the carboxylate of Asp^{534}, which would favor the second (protonation) mechanism.

A third question regarding the MBD4 action is how it distinguishes an A:T pair from a G:T. Although it is possible that the protein distinguishes G:T from an A:T because of their differing geometries, it is also possible that it may make specific contacts with the guanine in a manner similar to *E. coli* MUG (31) or hOGG1 (27); Arg^{486} of MBD4 is in the same position as Arg^{204} of hOGG1 that forms hydrogen bonds in the minor groove side with the G on the opposite strand of the flipped nucleotide. A detailed answer to this question must await the availability of a MBD4-DNA co-crystal structure.

Thymine and Uracil—How does the flipped base specifically bound in the active site? In MutY the adenine soaked into the crystal are recognized by Glu^{37} and Gln^{182} (29) (Fig. 4*D*). Structural superimposition between MutY and MBD4 (Fig. 3*A*) indicates the side chains of Gln-423 and Tyr^{514} of MBD4 are in the vicinity of the adenine-specific interacting side chains of MutY (Fig. 4*E*).

In MBD4, the two polar residues (Gln^{423} of αA and Tyr^{514} of αI) and three hydrophobic residues (Val^{422} prior to αA, Gly^{445}, and Ile^{449} of αC) line in the cleft next to the catalytic Asp^{534} (Fig. 4*A*). We suggest that these amino acids are the major determinants of specificity after docking the flipped thymine into the binding pocket. In the absence of the target nucleotide, the active site is occupied by ordered water molecules (Fig. 2*D*), which lie almost in a plane and directly interact with Tyr^{514}, Gln^{423}, and Val^{422} (Fig. 4*C*). We docked a thymine with its Watson-Crick pairing edge (O-2, N-3, and O-4) occupying three water sites (Fig. 4*C*). The OH group of Tyr^{514} can make one hydrogen bond with the O-2 atom, the side chain carbonyl C = O of Gln^{423} can make a hydrogen bond to the protonated N-3-H. In addition, the main chain N-H group of Val^{422} can make a hydrogen bond to the O-4 atom. Gly^{445} and Ile^{449} form a surface hydrophobic patch near the end of the cleft, in a perfect position to accommodate the methyl group of thymine. Of all contacts made to the thymine base (Fig. 4*F*), the hydrophobic-methyl interaction will be absent for a uracil base.

Interestingly, Glu, Gln, or Tyr are often found in the active site of the HhH glycosylases. A glutamate is found in MIG (Glu^{42}; Ref. 28) and TAG (Glu^{38}; Ref. 32) in the equivalent position as Glu^{37} of MutY; the corresponding main chain position in MBD4 is Thr^{437} (helix αB), whose side chain points opposite as that of Glu^{37} in MutY (Fig. 4*E*) and makes a hydrogen bond with main chain carbonyl oxygen atom of Leu^{434}. A tyrosine is proposed for interacting target thymine in MIG (28) (Tyr^{126} occupying equivalent position as Tyr^{514} of MBD4 in helix αI), for interacting target 3-methyladenine in AlkA (26) (Tyr^{222} from an αI-equivalent helix, *i.e.* the second

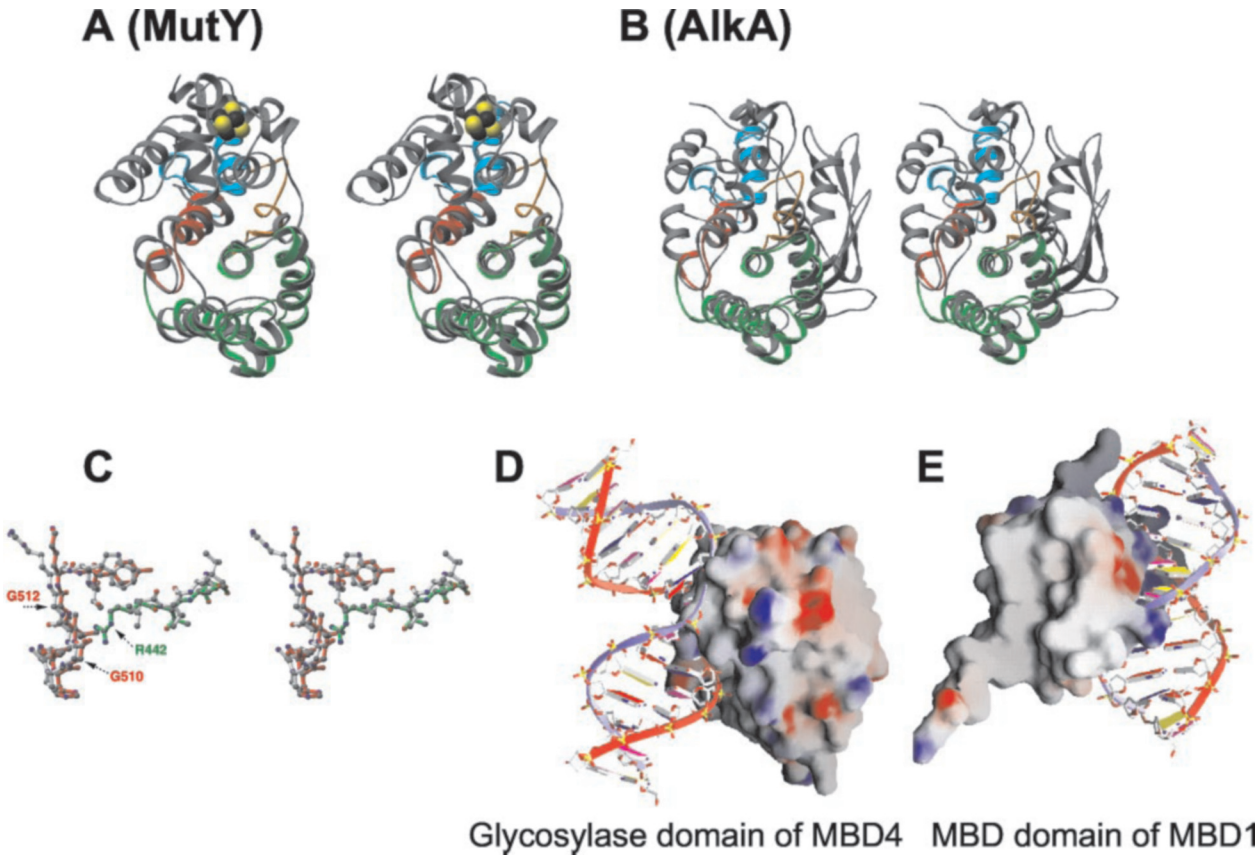

FIG. 3. **Structural comparison of HhH glycosylases.** *A*, superimposition of MBD4 glycosylase domain (colored according to Fig. 1*B*) and MutY (*gray*; Protein Data Bank code 1MUD). MutY (29), as well as EndoIII (30) and MIG (28), contains a C-terminal [4Fe-4S] cluster (shown in space-filling model). *B*, superimposition of MBD4 glycosylase domain and AlkA (*gray*; Protein Data Bank code 1DIZ). AlkA (26) and hOGG1 (27) contain an additional N-terminal β-sheet domain, and a 13- or 15-residue insertion between αC and αD (see Fig. 1*B*). *C*, superimposition of two DNA binding loops between MBD4 (*red* and *green*) and AlkA (*gray*). The Gly-rich hairpin loop of HhH motif is indicated by conserved Gly^{510} and Gly^{512} of MBD4. The minor groove wedge (Leu^{125} in AlkA), which assists in base flipping, superimposed on Arg^{442} of MBD4 in the loop between helices αB and αC (see Fig. 1*B*). *D*, based on the superimposition shown in *C*, the MBD4 glycosylase domain is docked to DNA from the minor groove side. *E*, the MBD domain of MBD4 has not yet been structurally characterized; however, the NMR solution structure of the MBD domain of MBD1 was shown to bind DNA from the major groove side (13).

helix of HhH motif), and in TAG (32) (Tyr^{16} of an N-terminal helix). A glutamine is common to MutY (Gln^{182}) and hOGG1 (Gln^{315}) in recognizing their substrate base, adenine and 8-oxoguanine, respectively; both Gln^{182} of MutY and Gln^{315} of hOGG1 are located in a C-terminal helix outside of the structurally homologous regions among the HhH glycosylases shown in Fig. 1*B*. Although MBD4 does not have an equivalent C-terminal helix, the N-terminal and C-terminal regions of all structurally characterized HhH glycosylases are folded together, above the cleft as shown in Fig. 3; and in the case of MBD4, Gln^{423} is from the N-terminal helix αA and its side chain occupies a similar position as that of Gln from the C-terminal helix.

DNA Glycosylase Activity of MBD4 N-terminal Truncations—Among the known HhH enzymes, MBD4 has the longest N-terminal sequence before the glycosylase domain (for examples, see Fig. 1*B*). Zhu *et al.* (10) analyzed a series of N-terminal deletion mutants of human MBD4, and the results are consistent with our glycosylase domain structure presented here. In that study, N-terminal deletions of up to 65% of the total length of MBD4 retain the DNA glycosylase activity. The smallest fragment that retained activity, ΔN433 (10), is very similar in size to our glycosylase domain determined by proteolysis (see Fig. 1*B*).

We used a DNA duplex containing a G:T within a CpG sequence context as the substrate to test the glycosylase activities of purified full-length MBD4 and several of its deletion derivatives. The T-containing strand was radiolabeled, and the excision of this base was monitored by gel electrophoresis. Typical results are presented in Fig. 5*A* and show that in addition to the full-length MBD4, the Δ399 mutant used for crystallography is an active thymine DNA glycosylase. A construct missing the first 48 amino acids of the full-length protein (Δ48) has less activity, but the construct containing only the MBD segment of the protein (amino acids 48–187) has no detectable activity (Fig. 5*A*). We also measured the activity of all MBD4 constructs in crude cell extract by expressing the proteins in a strain lacking the endogenous uracil glycosylase gene (*ung*$^-$) to minimize background. Full-length MBD4, Δ48, and Δ399 all have detectable activity in this assay (Fig. 5*B*).

Petronzelli *et al.* (33) have reported that a deletion of the first 454 amino acids of the human MBD4 still retained its enzymatic activity. The murine MBD4 equivalent to this deletion would be missing 428 N-terminal residues (Fig. 1*B*), which include helix αA and its preceding loop that provides part of the hydrophobic core above the cleft (Trp^{412}, Pro^{414}, Pro^{415}, Pro^{418}, and Phe^{419}; Fig. 2*C*) and Val^{422} and Gln^{423} that are proposed to contact the target thymine (Fig. 4*C*). Thus the results reported by Petronzelli *et al.* (33) are not compatible with the crystal structure and are surprising.

To resolve these discrepancies, we attempted to duplicate the result of Petronzelli *et al.* (33) by making the equivalent murine MBD4 truncation (Δ428; Fig. 1) and fusing it to a six histidine tag or GST tag. We were unable to detect any expression of the

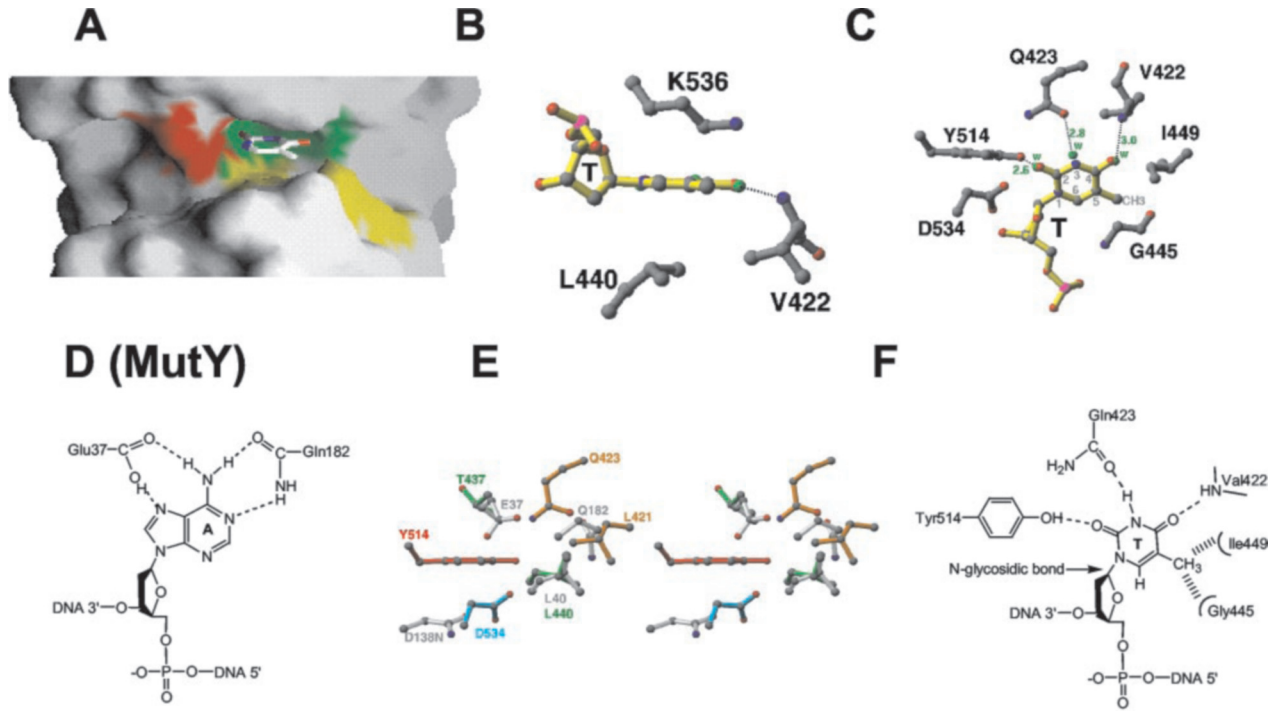

FIG. 4. **The active-site cleft and where the flipped, mismatched thymine will bind.** *A*, close-up view of the cleft, where Asp^{534} is colored in *red*; Tyr^{514}, Gln^{423}, and Val^{422} are colored in *green*, and Ile^{449}, Gly^{445}, and Leu^{440} are in *yellow*. *B*, the thymine (*yellow*) is docked in between Leu^{440} and Lys^{536}. *C*, the Watson-Crick pairing edge of thymine (O-2, N-3, and O-4) is aligned with three ordered water molecules (*green*) that make hydrogen bonds with Tyr^{514}, Gln^{423}, and Val^{422}. The C-5 methyl group would make van der Waals contact with Ile^{449} and Gly^{445}. The C1′ of the target sugar would be in a direct contact with the carboxylate group of Asp^{534}. The atom spheres are colored in *red* (oxygen), *blue* (nitrogen), and *gray* (carbon); the chemical bonds are colored in gray for the protein residues and *yellow* for thymine. *D*, schematic drawing of adenine-specific interactions in MutY (29). *E*, stereo view of superimposition of active site residues of MutY (*gray*) and the proposed MBD4 active site residues (colored according to Fig. 1*B*). *F*, schematic drawing of proposed thymine-specific interactions in MBD4.

His-tagged Δ428 protein, either by Coomassie staining or anti-His tag antibody (data not shown), whereas all other MBD4 fragments were expressed and soluble under the same conditions. Not surprisingly, no glycosylase activity was detected in the extract of Δ428 construct using the *ung*$^-$ strain (Fig. 5*B*). The GST-tagged Δ428 was expressed to high level, but most of the protein was insoluble (data not shown). However, we did manage to partially purify some GST-Δ428 fusion protein using a glutathione affinity column and a HiTrap Q column. The protein was heavily associated with Hsp60 (data not shown), an indication that the protein may not be folded properly. When the GST-Δ428 protein was tested for glycosylase activity, none was detected (Fig. 5*A*). The observation that Δ428 mutant does not fold properly is consistent with the important structural roles of the missing residues. In addition, although sequence similarity of MBD4 to other glycosylases starts at helix αB, MutY, MIG, EndoIII, and TAG all have N-terminal extensions similar in size to Δ399 of MBD4 (Fig. 1*B*). We do not know the origin of the discrepancy between our data and that of Petronzelli *et al.* (33), as the sequences of human Δ454 and mouse Δ428 deletions are almost 100% identical except 4 residues (see Fig. 1*B*). One possibility is that the pET28b vector (Novagen) used for the human Δ454 construct would add at least 10 additional residues besides the 6 histidines at the N terminus. These residues may fortuitously substitute the natural MBD4 residues and allow folding and enzymatic activity.

The activity of the Δ399 deletion was easiest to detect in the extracts, whereas the full-length MBD4 and the Δ48 construct displayed relatively poor activity (Fig. 5*B*). The lower activity of the full-length MBD4 in cell-free extracts was surprising but reproducible. It is noted that the MBD domain of MBD4 binds DNA with G:T mismatches (1), and it is possible that both the MBD and the glycosylase domains compete for the DNA substrate. Regardless, it is clear from these data that the Δ399 construct of the murine MBD4, which has almost the same N-terminal extension as the MutY, MIG, EndoIII, and TAG, is a stable protein fragment with substantial glycosylase activity.

DISCUSSION

We have described the crystallographic structure of the glycosylase domain of the methyl-CpG-binding protein MBD4. The structure reveals that the MBD4 glycosylase domain belongs to the HhH DNA glycosylase superfamily. Modeling studies suggest that MBD4 glycosylase domain, similar to that of AlkA and hOGG1 HhH glycosylases, binds DNA from the minor groove side (Fig. 3*D*).

Unlike other HhH glycosylases, MBD4 contains an additional DNA binding domain, the MBD, near its N terminus. An NMR solution structure of the MBD domain from human MBD1, in complex with methylated DNA, revealed that the MBD domain contacts both methyl groups of methyl-CpG site via the major groove of B-form DNA (13) (Fig. 3*E*). This is consistent with the observation that of the DNA sequence tested, only the fully methylated CpG or the methylated mismatch 5mCpG/TpG (both contain two methyl groups in the major groove) is bound by the MBD of MBD4 (1). Because all structurally characterized HhH glycosylases in complex with DNA appear to bind DNA exclusively via the minor groove, it is attractive to think that the MBD and the glycosylase domains of MBD4 would come together at 5mCpG/TpG mismatches to engulf DNA from opposite directions (28). However, because the MBD domain does not bend DNA (13), whereas all HhH glycosylases appear to significantly bend DNA and flip the target, it is not clear how DNA would be bent when both domains bind together. Alternatively, perhaps the two domains separated by ~200 residues bind DNA at adjacent but non-

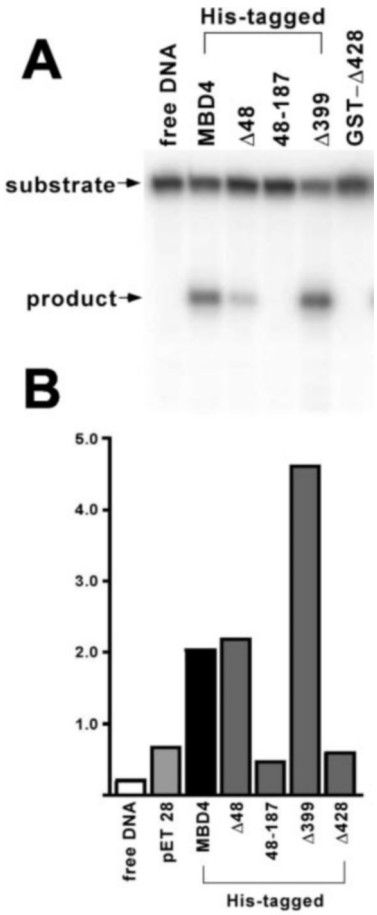

FIG. 5. **Glycosylase activities of MBD4 N-terminal truncations.** *A*, thymine excision activities of purified full-length MBD4 and its deletion derivatives are shown. One hundred nanograms of the His-tagged or GST-tagged (Δ428) versions of the proteins were used with 20 nM radiolabeled duplex containing a G:T. The products of the reaction were separated on a sequencing gel, and the gel was scanned with a PhosphorImager. The PhosphorImager scan is shown. *B*, glycosylase activities in cell extracts prepared from overproducers of MBD4 variants. Thymine excision activities in the cell extracts of the overproducers of full-length MBD4 and its deletion derivatives are shown. Cell extracts containing 2 μg of total proteins were used with 20 nM radiolabeled duplex containing a G:T. The products of the reaction were separated on a sequencing gel, and the gel was scanned with a PhosphorImager. The released product as a percentage of total labeled DNA is shown.

overlapping sites. The function of the MBD domain in MBD4 may be to target the glycosylase activity to regions of heavily methylated DNA as methyl-CpG dinucleotides tend to occur in clusters (reviewed in Ref. 34), so the tethered glycosylase domain could sample nearby sites for G:T mismatches. This would raise the local concentration of glycosylase activity in regions where methylated mismatch 5mCpG/TpG is most likely to occur.

The active-site cleft of the glycosylase domain suggests a base flipping mechanism for accessing the damaged or mispaired base (reviewed in Ref. 35), the mismatched base should be swung completely out of the DNA helix by torsional rotation of its flanking sugar-phosphate backbones so as to occupy the active-site cleft of MBD4. The structure also reveals candidate residues for catalysis (Asp^{534}), for thymine (or uracil)-specific recognition hydrogen bonding (Tyr^{514}, Gln^{423}, and Val^{422}), for the methyl group of thymine (Ile^{449} and Gly^{445}), and for the stacking stabilization of the flipped base (Leu^{440} and Lys^{536}). With this information, our structure provides useful starting points for more detailed studies of this interesting enzyme.

Acknowledgments—We thank Dr. Adrian Bird (University of Edinburgh) for providing the constructs to overexpress mouse MBD4; Drs. Anand Saxena and Dieter Schneider (Brookhaven National Laboratory) and John R. Horton (Emory University) for help with x-ray data collection at beamlines X12-C and X26-C in the National Synchrotron Light Source; and Drs. Robert Blumenthal (Medical College of Ohio) and Paul Wade (Emory University) for critical comments on the manuscript.

REFERENCES

1. Hendrich, B., Hardeland, U., Ng, H.-H., Jiricny, J., and Bird, A. (1999) *Nature* **401,** 301–304
2. Millar, C. B., Guy, J., Sansom, O. J., Selfridge, J., MacDougall, E., Hendrich, B., Keightley, P. D., Bishop, S. M., Clarke, A. R., and Bird, A. (2002) *Science* **297,** 403–405
3. Bellacosa, A., Cicchillitti, L., Schepis, F., Riccio, A., Yeung, A. T., Matsumoto, Y., Golemis, E. A., Genuardi, M., and Neri, G. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 3969–3974
4. Riccio, A., Aaltonen, L. A., Godwin, A. K., Loukola, A., Percesepe, A., Salovaara, R., Masciullo, V., Genuardi, M., Paravatou-Petsotas, M., Bassi, D. E., Ruggeri, B. A., Klein-Szanto, A. J., Testa, J. R., Neri, G., and Bellacosa, A. (1999) *Nat. Genet.* **23,** 266–268
5. Neddermann, P., Gallinari, P., Lettieri, T., Schmid, D., Trung, O., Hsuan, J. J., Wiebauer, K., and Jiricny, J. (1996) *J. Biol. Chem.* **271,** 12767–12774
6. Krokan, H. E., Nilsen, H., Skorpen, F., Otterlei, M., and Slupphaug, G. (2000) *FEBS Lett.* **476,** 73–77
7. Nilsen, H., Haushalter, K. A., Robins, P., Barnes, D. E., Verdine, G. L., and Lindahl, T. (2001) *EMBO J.* **20,** 4278–4286
8. Saparbaev, M., and Laval, J. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95,** 8508–8513
9. Petronzelli, F. Riccio, A. Markham, G. D., Seeholzer, S. H., Stoerker, J., Genuardi, M., Yeung, A. T., Matsumoto, Y., and Bellacosa, A. (2000) *J. Biol. Chem.* **275,** 32422–32429
10. Zhu, B., Zheng, Y., Angliker, H., Schwarz, S., Thiry, S., Siegmann, M., and Jost, J.-P. (2000) *Nucleic Acids Res.* **28,** 4157–4165
11. Bird, A. P., and Wolffe, A. P. (1999) *Cell* **99,** 451–454
12. Wade, P. A. (2001) *Oncogene* **20,** 3166–3173
13. Ohki, I., Shimotake, N., Fujita, N., Jee, J.-G., Ikegami, T., Nakao, M., and Shirakawa, M. (2001) *Cell* **105,** 487–497
14. Ohki, I., Shimotake, N., Fujita, N., Nakao, M., and Shirakawa, M. (1999) *EMBO J.* **18,** 6653–6661
15. 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
16. Hendrich, B., and Bird, A. (1998) *Mol. Cell. Biol.* **18,** 6538–6547
17. Dong, A. Yoder, J. A., Zhang, X., Zhou, L., Bestor, T. H., and Cheng, X. (2001) *Nucleic Acids Res.* **29,** 439–448
18. Hendrickson, W. A., Horton, J. R., and LeMaster, D. M. (1990) *EMBO J.* **9,** 1665–1672
19. Otwinowski, Z., and Minor, W. (1997) *Methods Enzymol.* **276,** 307–326
20. Terwilliger, T. C., and Berendzen, J. (1999) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **55,** 849–861
21. Terwilliger, T. C. (2000) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **56,** 965–972
22. Jones, T. A., and Kjeldgard, M. (1997) *Methods Enzymol.* **277,** 173–208
23. Brünger, A. T. (1992) *X-PLOR: A System for X-ray Crystallography and NMR*, version 3.1, Yale University, New Haven, CT
24. Beletskii, A., Grigoriev, A., Joyce, S., and Bhagwat, A. S. (2000) *J. Mol. Biol.* **300,** 1057–1065
25. Nash, H. M., Bruner, S. D., Scharer, O. D., Kawate, T., Addona, T. A., Spooner, E., Lane, W. S., and Verdine, G. L. (1996) *Curr. Biol.* **6,** 968–980
26. Hollis, T., Ichikawa, Y., and Ellenberger, T. (2000) *EMBO J.* **19,** 758–766
27. Bruner, S. D., Norman, D. P. G., and Verdine, G. L. (2000) *Nature* **403,** 859–866
28. Mol, C. D., Arvai, A. S., Begley, T. J., Cunningham, R. P., and Tainer, J. A. (2002) *J. Mol. Biol.* **315,** 373–384
29. Guan, Y., Manuel, R. C., Arvai, A. S., Parikh, S. S., Mol, C. D., Miller, J. H., Lloyd, S., and Tainer, J. A. (1998) *Nat. Struct. Biol.* **5,** 1058–1064
30. Thayer, M. M., Ahern, H., Xing, D., Cunningham, R. P., and Tainer, J. A. (1995) *EMBO J.* **14,** 4108–4120
31. Barrett, T. E., Scharer, O. D., Savva, R., Brown, T., Jiricny, J., Verdine, G. L., and Pearl, L. H. (1999) *EMBO J.* **18,** 6599–6609
32. Drohat, A. C., Kwon, K., Krosky, D. J., and Stivers, J. T. (2002) *Nat. Struct. Biol.* **9,** 659–664
33. Petronzelli, F., Riccio, A., Markham, G. D., Seeholzer, S. H., Genuardi, M., Karbowski, M., Yeung, A. T., Matsumoto, Y., and Bellacosa, A. (2000) *J. Cell. Physiol.* **185,** 473–480
34. Jones, P. A., and Takai, D. (2001) *Science* **293,** 1068–1070
35. Roberts, R. J., and Cheng, X. (1998) *Annu. Rev. Biochem.* **67,** 181–198
36. Carson, M. (1997) *Methods Enzymol.* **277,** 493–505
37. Nicholls, A., Sharp, K. A., and Honig, B. (1991) *Protein Struct. Funct. Genet.* **11,** 281–296