Excision of thymine and 5-hydroxymethyluracil by the MBD4 DNA glycosylase domain: structural basis and implications for active DNA demethylation

Hideharu Hashimoto, Xing Zhang and Xiaodong Cheng*

Department of Biochemistry, Emory University School of Medicine, 1510 Clifton Road, Atlanta, GA 30322, USA

ABSTRACT

The mammalian DNA glycosylase—methyl-CpG binding domain protein 4 (MBD4)—is involved in active DNA demethylation via the base excision repair pathway. MBD4 contains an N-terminal MBD and a C-terminal DNA glycosylase domain. MBD4 can excise the mismatched base paired with a guanine (G:X), where X is uracil, thymine or 5-hydroxymethyluracil (5hmU). These are, respectively, the deamination products of cytosine, 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC). Here, we present three structures of the MBD4 C-terminal glycosylase domain (wild-type and its catalytic mutant D534N), in complex with DNA containing a G:T or G:5hmU mismatch. MBD4 flips the target nucleotide from the double-stranded DNA. The catalytic mutant D534N captures the intact target nucleotide in the active site binding pocket. MBD4 specifically recognizes the Watson–Crick polar edge of thymine or 5hmU via the O2,N 3 and O4 atoms, thus restricting its activity to thymine/uracil-based modifications while excluding cytosine and its derivatives. The wild-type enzyme cleaves the N-glycosidic bond, leaving the ribose ring in the flipped state, while the cleaved base is released. Unexpectedly, the C10 of the sugar has yet to be hydrolyzed and appears to form a stable intermediate with one of the side chain carboxyl oxygen atoms of D534, via either electrostatic or covalent interaction, suggesting a different catalytic mechanism from those of other DNA glycosylases.

INTRODUCTION

Mammalian DNA glycosylases have been proposed to be involved in active DNA demethylation via the base excision repair pathway (1–4). MBD4 contains both an N-terminal methyl-CpG binding domain (MBD) and a C-terminal DNA glycosylase domain (Supplementary Figure S1a) that acts on G:T and G:U mismatches (5). In zebrafish, the activation-induced cytidine deaminase (AID) and MBD4 cooperate to demethylate DNA (1). Consistent with a role in DNA demethylation in mammals, AID is required to demethylate pluripotency genes during reprogramming of the somatic genome in embryos (6), and AID-deficient animals are less efficient in erasure of DNA methylation in primordial germ cells (7). It is noteworthy that AID promotes 5-methylcytosine (5mC or M) deamination, resulting in thymine (1,8), as well as 5-hydroxymethylcytosine (5hmC or H) deamination, which would produce 5-hydroxymethyluracil (5hmU) (3) (Figure 1). There are three mammalian ten eleven translocation (Tet) proteins that convert 5-methylcytosine (5mC or M) to 5hmC (9). 5hmC is a constituent of nuclear DNA, present in many tissues and cell types (9–11). The genomic content of 5hmU (<3.5 pmol) is orders of magnitude lower than that of 5hmC (hundreds of pmols) (10), suggesting that modification products of 5hmC are probably short lived and possibly removed by subsequent enzymatic reactions.

The importance of MBD4 for mutation avoidance in mammals and in maintaining genome stability is confirmed by an increase in 5mC to T mutations and increased occurrence of colon carcinoma in Mbd4−/− mice (13). Consistent with this observation, MBD4 is mutated in 26–43% of human colorectal tumors that show microsatellite instability (14). Here we show, by means of X-ray crystallography, that the mouse MBD4 glycosylase domain (residues 411–554; Supplementary Figure S1b) binds to a G:T or G:5hmU mismatch in the context of a CpG dinucleotide. The mismatched nucleotide (T or 5hmU) is flipped completely out of the DNA helix and is positioned in a binding pocket with Watson–Crick polar hydrogen bonds specific for thymine/uracil-based modifications.

MATERIALS AND METHODS

Expression and purification of MBD4 glycosylase domain

Hexahistidine-SUMO tagged mouse MBD4 residues 411–554 (pXC1064) and its mutants D534N (pXC1088), K536A (pXC1160) and Y514F (pXC1162) were expressed...
in *Escherichia coli* BL21(DE3)-Gold cells from the RIL-Codon plus plasmid (Stratagene). Expression cultures were grown at 37°C until OD600 reached 0.5, shifted to 16°C and then induced by adding 0.4 mM isopropyl β-D-1-thiogalactopyranoside. Cell pellets were suspended with 4-fold volume of Ni buffer (300 mM NaCl, 20 mM sodium phosphate pH 7.4, 20 mM imidazole, 1 mM dithiothreitol and 0.25 mM phenylmethylsulfonyl fluoride) and sonicated for 5 min (1 s on and 2 s off). The lysate was clarified by centrifugation twice at 38,000 g for 30 min. His6-SUMO fusion proteins were isolated on a nickel-charged chelating column (GE Healthcare). The His6–SUMO tag was cleaved by ULP-1 protease (16 h at 4°C), leaving two extraneous N-terminal amino acids (HisMet). The cleaved protein was loaded onto tandem ion exchange HiTrap-Q and HiTrap-SP columns (GE-Healthcare), eluted from the SP column and further purified by Superdex 75 (16/60) in the presence of 500 mM NaCl.

Crystallography of MBD4 and DNA complexes

For co-crystallization, the MBD4 proteins (WT or D534N mutant) and annealed G:X mismatch oligonucleotides labeled with 6-carboxy-fluorescein (FAM), and the excision of the target base was monitored by denaturing gel electrophoresis following NaOH hydrolysis (Supplementary Figure S1c). In Figure 2a, MBD4 protein (0.5 μM) and equal amount of

DNA glycosylase activity assay

MBD4 glycosylase activity assays were performed using various oligonucleotides labeled with 6-carboxy-fluorescein (FAM), and the excision of the target base was monitored by denaturing gel electrophoresis following NaOH hydrolysis (Supplementary Figure S1c). In Figure 2a, MBD4 protein (0.5 μM) and equal amount of

![Figure 1. A putative pathway of DNA demethylation involving DNA methylation by DNMTs, hydroxylation by Tet proteins, deamination by AID or members of APOBEC superfamily (12) and base excision by MBD4 linked to base excision repair (BER). DNA major groove and minor groove sides are indicated. Small arrows in G:M pair and G:T mismatch indicate the hydrogen bond donors and acceptors for 5mC and thymine bases.](image-url)
double-stranded FAM labeled 32-bp duplexes were mixed in 20μl nicking buffer (10 mM Tris–HCl, pH 8.0, 1 mM EDTA, 0.1% BSA) and incubated at 37°C for 30min. Reactions were stopped by adding 2μl of 1 N NaOH, boiled for 10min, and 20μl of loading buffer (98% formamide, 1 mM EDTA and 1 mg/ml of Bromphenol Blue and Xylene Cyanole) was added. The samples were boiled for 10min and then immediately cooled in ice water and loaded onto a 10 cm x 10 cm 15% denaturing gel containing 7 M urea, 24% formamide, 15% acrylamide and 1× TBE. The gels were run in 1× TBE buffer for 60 min at 200 V. FAM-labeled single-stranded DNA was visualized under UV exposure. The following 32-bp oligonucleotides were synthesized at the New England Biolabs:

(FAM)-5'-TCGGATTTGTGGTTCGACGTAGTGTGA-3' (where X = C, 5mC, 5hmC, U, T, 5hmU or 5caC) and 5'-TACACTATCATGCAGCTGACCCCACAACATCCGA-3'.

For enzymatic reactions under single turnover conditions (Figure 2b), the FAM-labeled 32-bp duplexes (0.25 μM) and 10-fold excess of MBD4 catalytic domain (2.5 μM) were incubated for 0–10 min at room temperature (~22°C) and processed as described earlier. The intensities of the FAM-labeled DNA were measured by Typhoon Trio+ (GE Healthcare) and quantified by the image-processing program ImageJ (NIH). The data were fitted by non-linear regression using software GraphPad PRISM 5.0d (GraphPad Software Inc.): [Product] = \[ P_{max}(1 - e^{-kt}) \]

where \( P_{max} \) is the product plateau level, \( k \) is the observed rate constant, and \( t \) is the reaction time.

**DNA binding assay**

As shown in Figure 2c, MBD4 protein (1.0 μM) and 0.5 μM of the 32-bp FAM-labeled DNA were mixed in 20 μl nicking buffer and incubated at 37°C for 15 min. As shown in Figure 2d, the 32-bp FAM-labeled DNA

![Figure 2. Base excision and binding activities of MBD4 glycosylase domain on double-stranded DNA containing various forms of CpG dinucleotide.](image-url)
RESULTS

MBD4 specifically binds and cleaves thymine/uracil-based mismatches

We analyzed the specificity of the glycosylase and binding activities of MBD4 glycosylase domain using various 32-bp DNA oligonucleotides, with each containing a single G:X modification within the CpG sequence, where X = C, M, H, U, T, 5hmU or 5-carboxyketosine (5caC). As expected, substrates bearing G:U, G:T and G:5hmU mismatches were efficiently cleaved (Figure 2a), but glycosylase activity was not detected with MBD4 on oligonucleotides bearing the ‘natural’ G:C and G:M base pairs, or G:H and G:5caC, which presumably preserve Watson–Crick base pair hydrogen bonds (Supplementary Figure S2a). Therefore, MBD4 is capable of acting on deamination-linked mismatches, with the observed rate constants ($k_{\text{cat}}$) of 1.7, 1.0 or 0.45 min$^{-1}$ for G:U, G:5hmU and G:T substrates, respectively, under single turnover conditions ($[E_{\text{MBD4}}] >> [SDNA]$) (Figure 2b). The differences in the MBD4 glycosylase activity on G:U and G:T substrates are in agreement with previous findings (5,19). A non-catalytic amount) were incubated at 20 nM and MBD4 D534N mutant (with an indicated structure containing 11 helices, is highly similar to that of AlkA (17). The protein component of MBD4, an all-helical DNA glycosylases including hOGG1, MutY, EndoIII and NTH (23). The polar groups of the thymine base along the Watson–Crick edge are all hydrogen bonded with the protein (Figure 3i). The O$_2$ oxygen atom accepts hydrogen bonds from the side chains of Tyr514 (hydroxyl group) and Gln423 (amino group), the N$_3$ nitrogen atom donates a hydrogen bond to the carbonyl oxygen of Gln423 side chain and the O$_4$ oxygen atom accepts a hydrogen bond from the main-chain amide nitrogen of Val422. It is noteworthy that a thymine base has the opposite hydrogen bond potential at the ring positions of 3 and 4 compared to a cytosine base (or its C$_5$ derivatives), of which the ring N$_3$ nitrogen can only be a hydrogen-bond acceptor and the exocyclic N$_4$ amino group can only donate a hydrogen bond (Figure 1). Therefore, the network of hydrogen bonds in the MBD4 active site will not accommodate the Watson–Crick edge of a cytosine and exclusion of cytosine base by the active site allows MBD4 to discriminate against all cytosine derivatives, including 5caC, which is a substrate for mammalian thymine DNA glycosylase (TDG) (24). We do note, however, that the discrimination of a mismatched thymine/uracil from that of G:C pair could also occur prior to base flipping.

The mutated catalytic residue D534N is roughly perpendicular to the sugar ring of the flipped thymine with the side chain nitrogen atom forming a hydrogen bond with the O$_{3'}$ atom (2.7 Å), while the side chain oxygen atom is $\approx 3.7$ Å from the C$_{4'}$ atom and 4.6 Å from the N$_1$ atom—the two atoms that form the $\gamma$-glycosidic bond (Figure 3i). No water molecule is found nearby. Superimposition with DNA containing a normal intrahelical thymine reveals that the everted yet uncleaved thymine ring rotates around the glycosidic bond as well as bends slightly relative to the sugar ring (Figure 3j). This is reminiscent of the structures of uracil DNA glycosylase–DNA complex with a flipped uracil (25) and Methanobacterium thermoformicum mismatch glycosylase MIG–DNA complex with an extrahelical thymine (26), where the enzyme-induced distortion was suggested to facilitate the cleavage of the glycosidic bond.
The C5 methyl group of the flipped thymine is 3.6 Å away from one of the 5' phosphate oxygen atoms, and 4.0 Å away from the C2 atom of Gly445 (Figure 3i), suggesting that an additional hydroxyl group attached to the C5 methyl could fit into the space, consistent with 5hmU being a substrate of MBD4. Indeed, 5hmU flips completely out of the DNA helix by MBD4 and can be well overlaid with thymine in the same binding pocket (Figure 4a). However, the hydroxyl group points away from Gly445 and forms an inter-molecular polar interaction with the terminal ε-amino group of Lys536 (Figure 4b), whose aliphatic portion of side chain stacked with the ring of the extrahelical nucleotide.
The structure of a stalled intermediate formed by WT MBD4 glycosylase domain

To obtain a product complex structure, WT MBD4 protein were mixed with substrate DNA containing a G:T mismatch and co-concentrated (~3 h at 4°C) before setting up crystallization trials. As expected, the thymine base was already cleaved and possibly released via the open-active-site cleft connected to the solvent (Figure 3g and Supplementary Figure S1d). Superimposition of the protein components of the D534N mutant and WT structures (with rms deviation of ~0.5 Å when compared 144 pairs of Cα atoms) reveals a conformational change of side chain of Lys536, moving toward the base and possibly pushing out the cleaved base (Figure 4c). Mutating Lys536 to alanine (K536A) reduces the activity by a factor of ~5–10 (Figure 4d, left panel), comparing to that of WT enzyme (Figure 2b). While K536A mutation is expected to affect product release, it also reduces the reaction rate probably due to the involvement of Lys536 side chain in stabilizing the flipped base in a twisted state (Figure 4b).

Unexpectedly, the electron density clearly indicates that the sugar ring of the cleaved nucleotide does not yet contain a hydroxyl oxygen atom at the C1′ carbon atom (Figure 5a). Rather, the ribose ring C1′ carbon is in close contact with one of the carboxylate oxygen atoms of the catalytic residue Asp534 (refined to ~2.6 Å) and the hydroxyl oxygen atom of Tyr514 (~3.1 Å) (Figure 5b), leaving no room for a water nucleophile between the aspartic acid and the substrate nucleotide. However, a water molecule is observed, coordinated by the second carboxylate oxygen atom of Asp534 and the hydroxyl oxygen atom of Tyr514 and is ~3.1 Å away from C1′ (Figure 5b), a position approximately corresponding to the O2 position of the uncleaved thymine. Perhaps this water molecule will be able to serve as the nucleophile to attack the C1′ and eventually complete the reaction (Supplementary Figure S3).

MBD4 is known for exhibiting single-turnover, pre-steady state kinetics (27). It appears that the crystal structure of the MBD4 product complex represents a stalled reaction intermediate after the cleavage of the base but before the final nucleophilic attack of the C1′ atom (Supplementary Figure S3). To investigate whether this is due to the crystallization condition being unfavorable for MBD4 reaction, we incubated the protein–DNA complex (in 20 mM HEPES pH 7.0 and 100 mM NaCl) overnight at room temperature, to assure that the first round of base excision reaction was completed (under 1 h, Supplementary Figure S1e), before setting up crystallization. We then repeated the crystallographic study and the resulting structure is essentially identical to the structure without the extended pre-incubation, showing that the C1′ lacks the hydroxyl group while the base has already been cleaved (data not shown).

The time course of an MBD4 glycosylase reaction with a 4:1 DNA to protein molar ratio, under the same condition as pre-crystallization incubation (100 mM NaCl), shows that minimal enzyme turnover occurred during the 20-h incubation period, while the first round of base excision was completed before 1 h, suggesting that indeed the reaction was stalled for a significant amount of time after base excision (Supplementary Figure S1e). When no salt was added to reaction, MBD4 has a turnover rate of ~4 h (Supplementary Figure S1f).

DISCUSSION

Initiation of base excision by MBD4

MBD4 does not have the equivalent of Lys249 of hOGG1, which is proposed to attack the deoxyribose C1′ atom (22). Instead, MBD4 has Tyr514 in the corresponding position, whose hydroxyl group makes a hydrogen bond to the O2 of thymine or 5hmU (Figures 3i and 4a). Together with the O2-interacting amino group of Gln423, Tyr514 could serve to activate the leaving group by providing a hydrogen bond to the developing negative charge on a pyrimidine O2 in the transition state (Figure 4a), as suggested for the interaction between uracil DNA glycosylase His157 and uracil O2 (28). Indeed, mutating Try514 to phenylalanine (Y514F) nearly abolishes MBD4 activity (Figure 4d, right panel).

Potential reaction mechanism(s)

The stalled complex shows that MBD reaction is a stepwise S_N1-type cleavage reaction where the base

---

**Figure 4.** Comparison of the 5hmU conformation with that of thymine in MBD4. (a) Superimposition of 5hmU-bound (in gray) and thymine-bound (in color) structures. (b) The hydroxyl group of 5hmU interacts with Lys536. We note that the interaction does not form an ideal hydrogen bond. Rotation of the C–C bond between the ring C5 and the methyl hydroxyl (CH2–OH) could position the hydroxyl group in several alternative conformations (Supplementary Figure S2b). (c) Superimposition of the WT (in gray) and mutant MBD4 (in color) structures indicates a conformational change of Lys536 in conjunction with the release of the cleaved base. (d) The activity of MBD4 mutants, K536A (left) and Y514F (right), on G:U (top panel), G:T (middle panel) and G:5hmU (bottom panel) substrates under the same single turnover conditions as that of the wild-type enzyme (see Figure 2b).
excision and nucleophilic attack of $C_1'$ are clearly separated (29). In this mechanism, the $C_1'$ should exist as a positively charged carbocation or an oxacarbenium ion intermediate following base removal (29). The half-life of a highly reactive oxacarbenium ion is estimated to be only $\sim 10^{-10} - 10^{-12}$ s (30,31), too short-lived to be observed in the time scale of crystallization. However, it is potentially possible that the oxacarbenium ion is stabilized by negatively charged Asp534, forming a long-lived ion-pair intermediate at the active site of MBD4. A similar mechanism has been proposed for *Bacillus stearothermophilus* MutY, where Asp144 of MutY was in the vicinity of $C_1'$ of a non-cleavable substrate analogue and proposed to stabilize positive charge accumulation on an oxacarbenium ion as well as coordinate a water nucleophile (32,33) (Supplementary Figure S4a). In addition to Asp144, a second acidic residue Glu43 of MutY interacts with the substrate adenine-$N_7$ that was thought to facilitate glycosidic bond scission (32). The equivalent functional group in MBD4 might be the $O_2$-interacting Tyr514.

An alternative mechanism would be a covalent glycosyl-enzyme intermediate, like those used by retaining O-glycosylases such as hen egg-white lysozyme (HEWL) (34). HEWL uses two carboxylate residues (Glu35 and Asp52) during the catalysis. When the catalytic acid/base Glu35 is mutated to the corresponding amide (E35Q), the covalent glycosyl-Asp52 intermediate is observed because hydrolysis of this intermediate is slowed enormously as the attack of water no longer benefits from the base catalysis ordinarily provided by Glu35. MBD4 only has the Asp52 equivalent carboxylate (Asp534 in MBD4) and lacks the Glu35 equivalent residue in the active site, which might have allowed us to capture the covalent intermediate between Asp534 and the $C_1'$ atom in the absence of a base activated water.
nucleophile. A similar interaction has been observed in the structure of *E. coli* AlkA complexed with DNA containing 1-azaribose abasic inhibitor, where Asp238 directly interacts with the positively charged N1' (equivalent to C1' of the 1-azaribose ring) (21) (Supplementary Figure S4b). It was proposed that Asp238 of AlkA directly assists the removal of positively charged bases, natural substrates for AlkA. However, MBD4 only removes uncharged bases and it remains to be determined whether Asp534 carboxylate initiates the attack on C1' that breaks the N-glycosidic linkage.

Unfortunately, our current structure at the resolution of 2.8Å is insufficient to determine whether a covalent bond (~1.45Å) formed between the C1' and the carboxylate oxygen of Asp534 or to determine the details of the sugar conformation. The distance between the two interacting atoms could be significantly reduced by a simple rotation around the side chain χ2 torsion angle of Asp534 (Figure 5c) and/or by relaxation of the sugar conformation. In the 1.64Å-resolution structure of HEWL E53Q mutant, such movement caused the C1 of the six-membered pyranose ring to be 1.75Å closer to the carboxylate oxygen of Asp52, resulting in a covalent bond (34). For MBD4, an atomic resolution structure will be required to settle whether the covalent bond forms.

The loss of activity by mutating Asp534 to the corresponding amide (D534N) (Figure 2a) suggests that the negative charge on Asp534 is critical for catalysis. In addition, the pH profile of MBD4 activity is also consistent with the requirement of a deprotonated carboxylate for catalysis: the activity is similar in the pH range of 6.5–9.4, but drops precipitously below that until total loss of activity at pH 4.0 (Supplementary Figure S1g).

In summary, we show that the mouse MBD4 glycosylase domain binds to G:X mismatched substrate DNA and flips out the target nucleotide into the active-site pocket. Many structural features are conserved among the HhH DNA glycosylases [including the human MBD4 glycosylase domain (35), Supplementary Figure S5]: the phosphodiester backbone pinching (36) caused by extensive protein–phosphate contacts surrounding the flipped nucleotide, the significant DNA bending due to hydrophobic intercalation of protein residues between the Cyt and Gua of the unmodified strand, and the use of an arginine finger to penetrate DNA from the minor groove and fill the space left by the flipped nucleotide. However, the recognition of the flipped nucleotide and catalytic mechanism in the active site could be very different among the HhH enzymes. MBD4 specifically recognizes the Watson–Crick polar edge of thymine and 5hmU (Figures 3i and 4a), thus restricting its activity to thymine/uracil-based modifications. The observation of a long-lived intermediate is unique for DNA N-glycosylases. If the involvement of a covalent bond between enzyme and substrate can be confirmed, it would offer significant new insight for the catalytic mechanism of DNA N-glycosylases. The tight binding of the glycosylase may have an important biological function for protection of the cleaved site and prevents its non-specific processing until subsequent repair activities are recruited to the site.

**ACCESSION NUMBERS**

Protein Data Bank: The coordinates and structure factors of the mouse MBD4 catalytic domain—DNA complexes have been deposited with accession numbers 4EVV (D534N and G:T mismatch), 4EW0 (D534N and G:5hmU mismatch) and 4EW4 (WT and G:T resulting in a stalled intermediate).

**SUPPLEMENTARY DATA**

Supplementary Data are available NAR Online: Supplementary Table 1 and Supplementary Figures 1–5.

**ACKNOWLEDGEMENTS**

We thank sincerely Brenda Baker of New England Biolabs for DNA oligo synthesis, Drs John R. Horton and John Chrzas for help with X-ray data collection, Dr Robert M. Blumenthal of the University of Toledo College of Medicine for critical comments and editing the manuscript and Dr Paul J. Berti of McMaster University for discussion. H.H. performed all experiments. X.Z. and X.C. organized and designed the scope of the study, and all were involved in analyzing data and writing and revising the manuscript.

**FUNDING**

U.S. National Institutes of Health [GM049245-18]; Georgia Research Alliance Eminent Scholar (to X.C.) and The Department of Biochemistry at the Emory University School of Medicine (use of the Southeast Regional Collaborative Access Team synchrotron beamlines at the Advanced Photon Source of Argonne National Laboratory). Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

**REFERENCES**


