REVIEW ESSAY
Prospects & Overviews
BioEssays
www.bioessays-journal.com

Structural Biology of the HEAT-Like Repeat Family of DNA Glycosylases

*Rongxin Shi, Xing-Xing Shen, Antonis Rokas, and Brandt F. Eichman**

DNA glycosylases remove aberrant DNA nucleobases as the first enzymatic step of the base excision repair (BER) pathway. The alkyl-DNA glycosylases AlkC and AlkD adopt a unique structure based on α-helical HEAT repeats. Both enzymes identify and excise their substrates without a base-flipping mechanism used by other glycosylases and nucleic acid processing proteins to access nucleobases that are otherwise stacked inside the double-helix. Consequently, these glycosylases act on a variety of cationic nucleobase modifications, including bulky adducts, not previously associated with BER. The related non-enzymatic HEAT-like repeat (HLR) proteins, AlkD2, and AlkF, have unique nucleic acid binding properties that expand the functions of this relatively new protein superfamily beyond DNA repair. Here, we review the phylogeny, biochemistry, and structures of the HLR proteins, which have helped broaden our understanding of the mechanisms by which DNA glycosylases locate and excise chemically modified DNA nucleobases.

1. Introduction

Cellular metabolites and environmental toxins generate a diverse spectrum of chemical modifications to DNA that impair normal cellular function.[1,2] Alkylation of DNA nucleobases is one of the most common forms of DNA damage. The cytotoxic and mutagenic effects of alkylation damage can lead to genomic instability and disease, while also making some genotoxic DNA alkylating agents effective anticancer and antimicrobial chemotherapeutics.[3–5] In the cell, alkylation damage is repaired by several modification-specific pathways. Bulky, helix-distorting lesions typically are processed by nucleotide excision repair (NER), whereas small modifications are removed by direct reversal and base excision repair (BER) mechanisms. BER is initiated by a lesion-specific DNA glycosylase that liberates the modified nucleobase from the DNA backbone via hydrolysis of the N-glycosidic bond (**Figure 1**a). This reaction generates an apurinic/apyrimidinic (AP) site that is subsequently excised by an AP endonuclease and replaced by a gap-filling DNA polymerase.

Alkylation specific DNA glycosylases are widely distributed across all domains of life. These enzymes can be classified into one of three distinct structural superfamilies with overlapping substrate specificities.[6,7] Human AAG adopts a unique α + β fold,[8] while those from unicellular organisms, including yeast MAG and Mag1 and *Escherichia coli* AlkA and Tag, adopt a conserved helix-hairpin-helix (HhH) fold observed in some oxidation-specific glycosylases (Figure 1b).[6,7,9–11] Like most other glycosylase superfamilies, AAG, and HhH architectures scaffold a general DNA binding surface and an active site pocket, which helps capture the modified nucleobase extruded from the DNA helix in a process known as base flipping.[12] The remodeled DNA helix is stabilized by intercalation of one or more side chains into the void generated by the everted base. The nucleobase binding pocket in large part defines the enzyme's substrate specificity and places a limit on the size of the chemical modification that can be accommodated. For example, bacterial Tag and MagIII enzymes have a tight specificity and an active site perfectly shaped for 3-methyladenine (3 mA), whereas AlkA and Mag have a larger active site that allows recognition of a wide spectrum of alkylated and even deaminated nucleobases [reviewed in ref. [7]].

A third superfamily of DNA glycosylases was recently defined by bacterial AlkC and AlkD, which have a distinct specificity for cationic N3- and N7-alkylguanines and a unique construction of tandem α-helical repeats related to HEAT motifs (Huntington/Elongation/A-subunit/Target-of-rapamycin).[13,14] HEAT repeats are pairs of antiparallel α-helices that stack in parallel arrays to form extended superhelical or C-shaped structures, which often form scaffolds for large multiprotein assemblies, including those involved in chromatin maintenance and remodeling.[15–17] HEAT domains are also involved in binding and transporting protein ligands within the inner channel of the superhelix.[18–20] More recently, this channel has been identified as a nucleic acid binding surface in a set of proteins with diverse functions, including RNA nuclear export, regulation of mRNA stability, and chromosome segregation.[21–23]

The HEAT-like repeat (HLR) glycosylases AlkC and AlkD have illustrated how this versatile helical repeat architecture can serve as a DNA damage sensor in addition to a non-specific DNA binding platform,[24,25] and have provided new insight into how glycosylases can repair a wider variety of alkylated DNA lesions than previously known. In contrast to the base-flipping AAG and HhH glycosylases, AlkC and AlkD are unique

Dr. R. Shi, Dr. X.-X. Shen, Prof. A. Rokas, Prof. B. F. Eichman
Department of Biological Sciences
Vanderbilt University
Nashville, TN 37232, USA
E-mail: brandt.eichman@vanderbilt.edu

DOI: 10.1002/bies.201800133

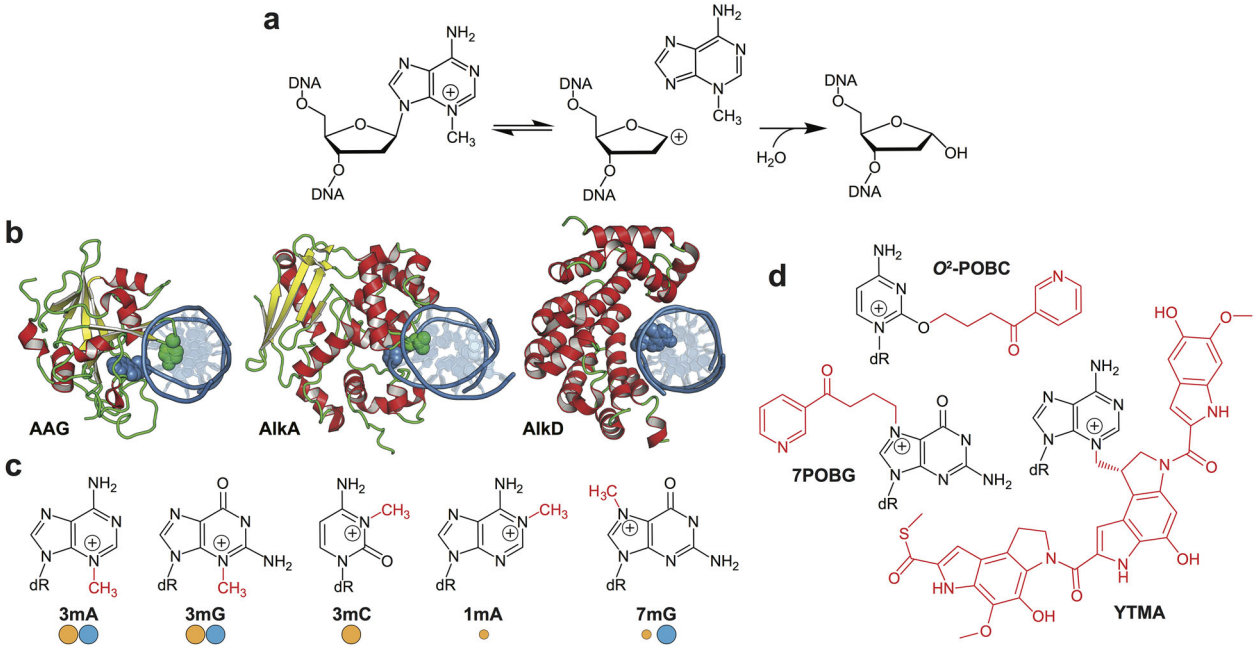

Figure 1. Function and specificity of HLR glycosylases. a) Base excision reaction catalyzed by DNA glycosylases. b) Crystal structures of the three alkylpurine DNA glycosylase families. Human AAG/1,N^6-ethenoadenine-DNA complex (PDB ID 1EWN), *E. coli* AlkA/1aR-DNA complex (1DIZ); *B. cereus* AlkD/3d3mA-DNA complex (5CL3). Proteins are colored according to secondary structure and DNA is blue. Damaged nucleobases and intercalating amino acids are shown as spheres. c) Methylated nucleobase substrates of AlkC (orange spheres) and AlkD (blue spheres). The smaller spheres indicate the lower activity of AlkC for 1 mA and 7 mG relative to the other nucleobases. 3 mA, 3-methyladenine; 3 mG, 3-methylguanine; 3 mC, 3-methylcytosine; 1 mA, 1-methyladenine; 7 mG, 7-methylguanine; dR, deoxyribose. d) Bulky alkylpurine lesions recognized by AlkD. POBC, pyridyloxobutylcytosine; POBG, pyridyloxobutylguanine; YTMA, yatakemycinyladenine.

in that they do not use a base flipping mechanism to recognize and cleave their substrates, and instead interact with damaged nucleotides that are stacked in the DNA helix, mainly through deoxyribose and phosphodiester backbone contacts. Consequently, these enzymes recognize a chemically different set of alkyl substituents previously associated with direct reversal and NER (Figure 1c,d).[5,26,27] AlkC is capable of excising 3-methylcytosine (3mC) and 1-methyladenine (1 mA),[28] which are also repaired directly by oxidative demethylation by the AlkB family of dioxygenases[29,30] (Figure 1c). The non-base flipping mechanism of AlkD enables it to remove bulky adducts, including pyridyloxobutyl (POB) adducts of guanine and cytosine[24] (Figure 1d). Likewise, *Bacillus cereus* AlkD and its ortholog in *Streptomyces* sp. TP-A0356, YtkR2, excise adenine that has been modified by yatakemycin (YTM), a bulky and highly toxic *Streptomyces* secondary metabolite that belongs to the duocarmycin/CC-1065 family of natural products (Figure 1d).[31–33] YTM-adenine (YTMA) adducts are known NER substrates.[34–36]

In addition to AlkC and AlkD enzymes, two distinct HLR proteins lacking base excision activity – AlkD2 and AlkF – have been identified. AlkD2 has very little detectable DNA binding activity, whereas AlkF has an affinity for branched DNA structures.[37,38] In this review, we will discuss the taxonomic distribution of members of the HLR superfamily and the structural basis for the unique specificities of the non-base-flipping glycosylases AlkC and AlkD and the non-enzymatic proteins AlkD2 and AlkF.

1.1. Phylogeny of the HLR Superfamily Shows Distinct Clustering of Enzymatic and Non-Enzymatic Proteins

The HLR superfamily consists of four distinct clades typified by AlkC, AlkD, AlkD2, and AlkF (**Figure 2**). We searched for AlkC, AlkD, AlkD2, and AlkF homologs from a PSI-BLAST search against the NCBI non-redundant protein sequence database using *Pseudomonas fluorescens* AlkC, *Bacillus cereus* AlkD, *Streptococcus mutans* AlkD2, and *Bacillus cereus* AlkF as queries. Following exclusion of redundant sequences and multiple sequence alignment, the phylogenetic tree shown in Figure 2a was constructed using the neighbor-joining method.[28] The AlkC and AlkD clades comprised the majority of HLR sequences, with 43% and 32% of the total, respectively (Figure 2a). HLR superfamily members were primarily confined to Bacteria, although a small percentage of AlkD and AlkC homologs was found in Eukaryota and Archaea (Figure 2b). Over 32% of AlkD orthologs were found in *Actinobacteria*, 57% distributed among *Firmicutes*, *Proteobacteria*, and *Bacteroidetes*, 5% in other bacterial species, and 6% distributed between Archaea and Eukaryota (Figure 2b and Figure S1 and Table S1, Supporting Information). Of these 69 non-bacterial AlkD proteins, 43 were found in Archaea, and 26 in Eukaryota. Eukaryotic AlkD proteins were primarily found in amoebozoans (e.g., *Entamoeba histolytica*, *Dictyostelium discoideum*), the unicellular ciliate *Paramecium tetraurelia*, unicellular relatives of metazoans (e.g., *Capsaspora owczarzaki* and *Salpingoeca rosetta*) and a few metazoans, such as a sea anemone *Nematostella*

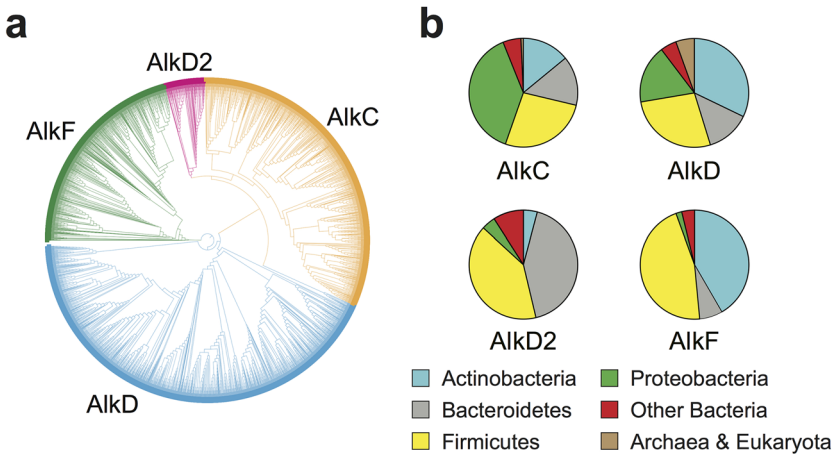

Figure 2. Phylogeny of the HLR superfamily. a) Phylogenetic tree constructed from 968 AlkC (orange), 1,299 AlkD (blue), 629 AlkF (green), and 123 AlkD2 (magenta) non-redundant protein sequences, generated, and visualized using Clustal Ω and iTOL [53,54]. b) Distribution of HLR proteins in diverse phyla.

vectensis, mollusks (*Lottia gigantea*), lancelets (e.g., *Branchiostoma floridae*), and a placozoan (*Trichoplax adhaerens*); they were not found in any fungal species (Table S1, Supporting Information). Sequence alignment of these proteins shows that they contain all the residues important for base excision in the *Bacillus* enzyme, and thus the eukaryotic AlkD orthologs are expected to have enzymatic activity.[24]

AlkC orthologs were primarily found in *Proteobacteria* (39%), 69% distributed among *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, and other bacterial species. Only 7 AlkC proteins (1% of the total) were found in Archaea and Eukaryota. All three eukaryotic AlkC proteins were found in the ciliate *Tetrahymena thermophila* (Table S1, Supporting Information). AlkF was the third most populated clade, comprising 21% of the total HLR sequences. The original identification of AlkF also defined a separate family of AlkG proteins.[37] AlkF and AlkG are closely related (the two paralogs from *Bacillus cereus* share 35% identity and 52.7% similarity), and thus were grouped together in the AlkF clade in our analysis. The vast majority (94%) of AlkF/AlkG orthologs were distributed among *Firmicutes* and *Actinobacteria* (Figure 2b). In contrast to the other groups, AlkD2 represented only 4% of the total HLR sequences and was primarily found in *Bacteroidetes* and *Firmicutes* (Figure 2b). Thus, the enzymes AlkD and AlkC are the most abundant HLR proteins and are distributed among bacterial phyla, with a few proteins found in uni- and multi-cellular eukaryotic organisms, whereas the non-enzymatic AlkF and AlkD2 sequences represent ~25% of the total HLR family and are distributed exclusively among bacterial phyla and rarely found in *Proteobacteria*.

1.2. HLR Proteins Share a Common Structural Architecture

Crystal structures of each member of the HLR superfamily (Table S2, Supporting Information) show that they are all composed of an N-terminal helical bundle (NTB) followed by 5 HLR motifs (**Figure 3**a), which together form a C-shaped structure (Figure 3b). The C-terminal helices of each HLR line the concave surface, which is the most conserved region (Figures 3b and S2, Supporting Information). In AlkC and AlkD this surface is highly positively charged as a result of the distribution of lysine and arginine side chains that are important for DNA binding.[39] In contrast, the positive charge on the surfaces of AlkD2 and AlkF is not concentrated within the concave cleft (Figure 3b). Each of the five HLR motifs (HLR1-5) has a distinct structure[39] that is largely conserved in each of the four HLR proteins. The HLR families mainly differ in the relative positions and stacking of adjacent HLR motifs that affect the overall solenoid structure. For example, AlkD2 is a hybrid of AlkC and AlkD families in that HLR1-2 shows a high degree of structural and sequence similarity to AlkD, while HLR3-5 are more similar to AlkC (Figures 3b and S2, Supporting Information). In addition to differences in HLR1-5 stacking, HLR families can be defined by unique motifs inserted within the HLRs. The AlkC family is largely defined by the presence of an inserted DNA binding loop (DBL) between HLR3 and HLR4 (Figure 3a). This insertion is present in both AlkCα and AlkCβ subtypes, which differ in the presence of an additional 100-residue immunoglobulin (Ig)-like domain at the C-terminus. Similarly, AlkF contains an inserted β-hairpin motif between helices αL and αM of HLR5 that affects its ability to bind branched DNA structures.[37]

The largest structural differences among the four HLR family members are found in the NTB, which also plays a role in DNA binding in different ways. The NTB of AlkD and AlkF are composed of three helices (αA, αB, αC), in which antiparallel helices αA and αC stack onto HLR1, leaving helix αB to protrude into the concave surface (Figure 3a). In AlkD, helix αB contacts DNA.[24,32,40] The NTB of AlkD2 has the same antiparallel arrangement and packing of αA and αC, but lacks the αB helix, which greatly reduces DNA binding affinity of AlkD2.[38] In contrast, AlkC's NTB is highly divergent from the other HLR members, largely in the orientation of the αA and αB helices.[28] In AlkC, αA and αB are each split into two co-linear helices that pack against helix αC. Interestingly, AlkC's αA helix resides in the same location as those in the other three families, but with opposite polarity. Consequently, the N-termini of both helices αA and αC point directly toward DNA, forming favorable helix

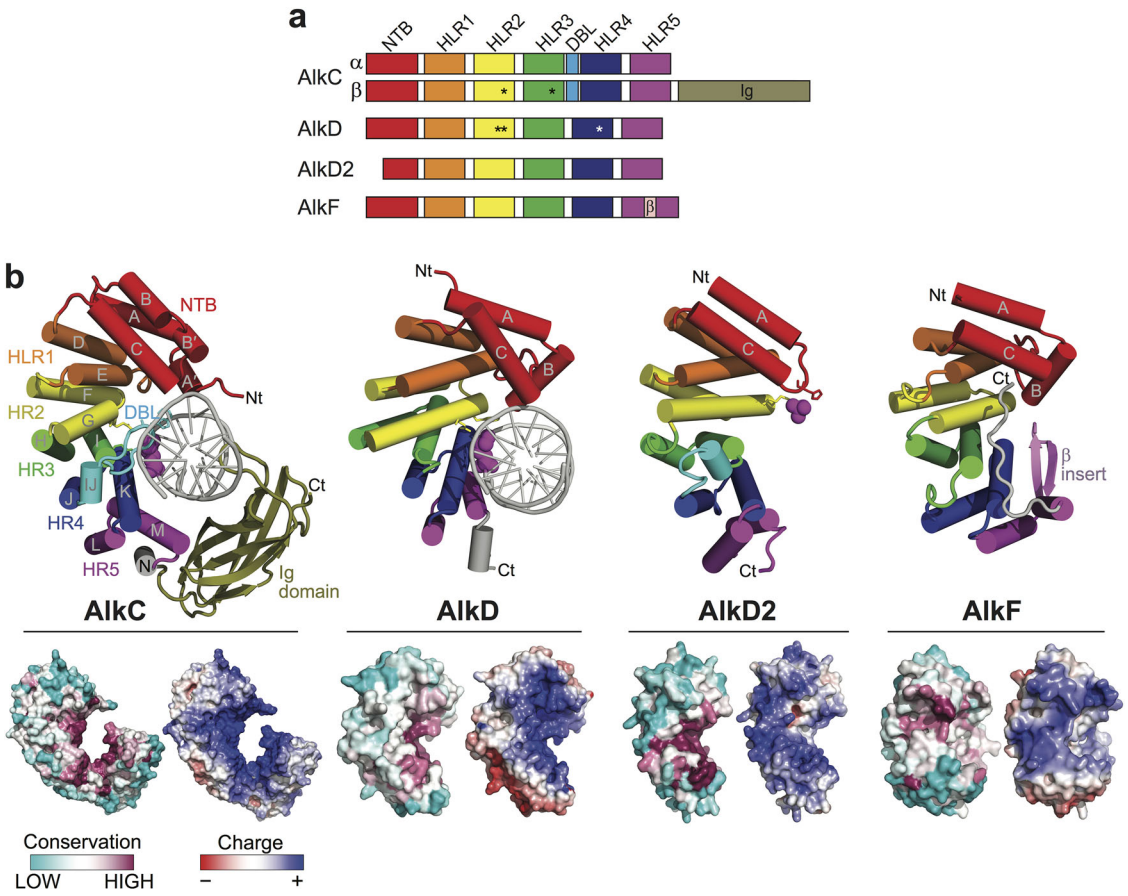

Figure 3. Structural comparison of HLR proteins. a) Protein sequence schematic, colored by helical repeat motif: N-terminal bundle (NTB), red; HEAT-like repeats, orange, yellow, green, navy blue, and purple; Ig domain, olive; DNA binding loop (DBL), cyan; inserted β-hairpin, pink. Asterisks denote the locations of catalytic residues. b) Crystal structures of *Pseudomonas fluorescens* AlkC in complex with DNA containing 1aR-DNA (PDB ID 5VHV), *Bacillus cereus* AlkD in complex with 1aR-DNA (5CLD), *Streptococcus mutans* AlkD2 (4 × 8Q), and *Bacillus cereus* AlkF (3ZBO). Proteins are colored as in panel a. DNA is shown as a silver cartoon, and bound 1aR nucleotide (AlkC, AlkD) and inorganic phosphate (AlkD2) are shown as magenta spheres. Solvent accessible surface representations are shown at the bottom, colored by sequence conservation within each ortholog (left) and electrostatic potential (right).

dipole interactions with the phosphate backbone.[28] These structural differences endow each HLR protein with a particular substrate specificity, which we describe in more detail below.

1.3. AlkD Uses a Non-Base-Flipping Mechanism to Excise Bulky Lesions

AlkD excises a variety of cationic alkylated DNA lesions. The *Bacillus cereus* enzyme (BcAlkD) was characterized initially to remove 3 mA, 3-methylguanine (3 mG), and 7-methylguanine (7 mG), and to provide Tag and AlkA deficient *E. coli* with resistance to methylating agents.[13] BcAlkD was subsequently shown to also remove bulky, cationic POB adducts.[24] More recent work revealed that AlkD proteins likely evolved specifically to repair DNA adducts of the duocarmycin/CC-1065/YTM family of bacterial toxins.[31] The *Streptomyces* AlkD homolog, YtkR2, excises YTM-adenine adducts, and protects *E. coli* challenged with either YTM or the methylating agent methyl methane sulfonate (MMS). Deletion of YtkR2 from *Streptomyces* sp. TP-A0356 reduces YTM production by 40%.[33] BcAlkD also excises YTM-adenine lesions in vitro, and deletion of AlkD from *Bacillus anthracis* (*ΔalkD*) sensitized cells to low concentrations of YTM.[34] In contrast, *ΔalkD* cells were not sensitive to MMS, likely because of the AAG, AlkA, and AlkC orthologs present in *Bacillus*.[13]

A homology model of BcAlkD, constructed from an uncharacterized, structural genomics crystal structure of a putative AlkD ortholog from *Enterococcus faecalis* (PDB ID 2B6C), enabled identification of functionally important residues that subsequently were corroborated by a BcAlkD crystal structure.[14,39] The active site consists of electrostatically paired Asp113-Arg148 side chains surrounded by a several aromatic residues (Tyr27, Trp109, Phe179, Phe180, and Trp187) clustered on the concave surface of the protein (**Figure 4**). Mutational analysis confirmed roles for Asp113, Arg148, Trp109, and Trp187 in glycosylase activity in vitro and in cells, providing the first direct evidence for catalytic activity by an HLR protein.[14,39] Structures of AlkD bound to DNA containing either a 3-deaza-3-methyladenine (3d3 mA) substrate analog (Figure 4a), a

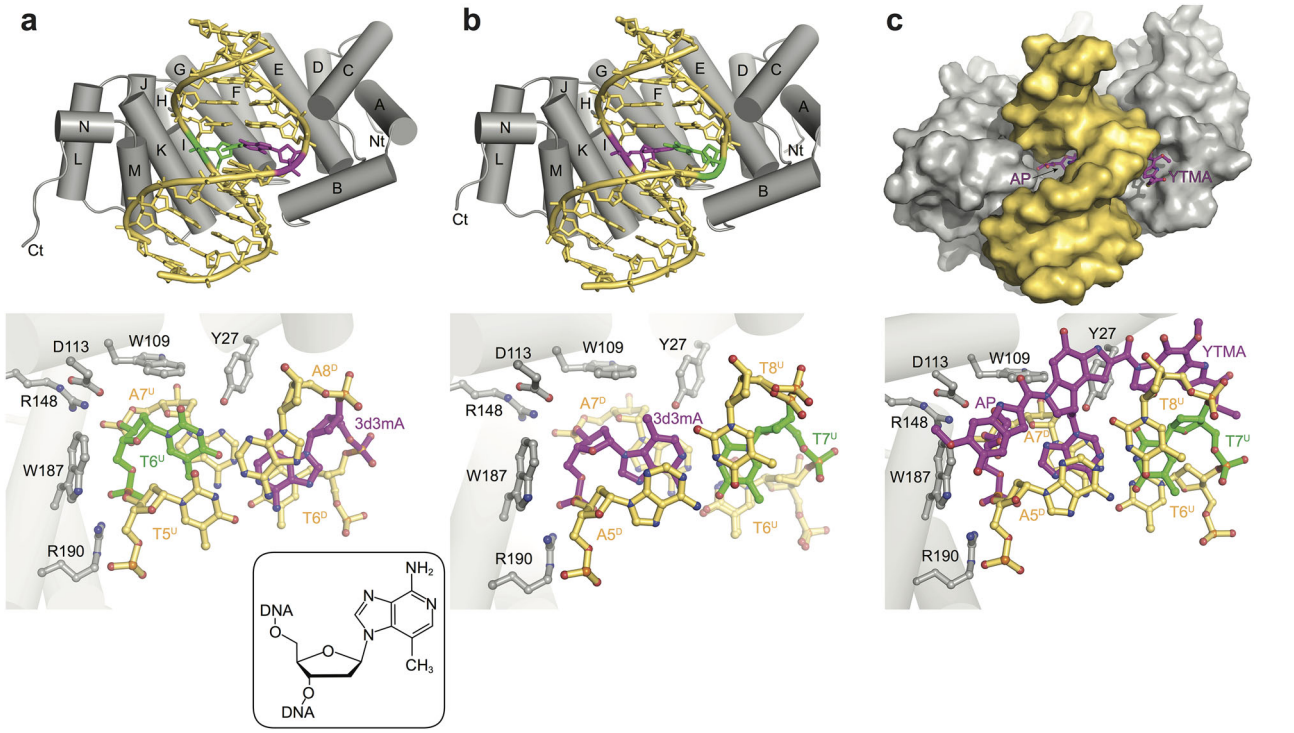

Figure 4. DNA recognition by AlkD. a and b) Crystal structures of AlkD bound to DNA containing 3d3 mA in non-catalytic (a, PDB ID 3JX7) and catalytic (b, PDB ID 5CL3) orientations. A chemical schematic of 3d3 mA is shown in the inset. c) Structure of the AlkD/AP-DNA/YTMA ternary product complex (PDB ID 5UUF). Overall structures are shown at the top, and details of the active site interactions are shown looking along the DNA helical axis. AlkD is colored grey, DNA gold, modified nucleotide (3d3 mA, YTMAde, AP site) magenta, and opposing thymine green.

tetrahydrofuran (THF) abasic site product analog, or a G•T mismatch illustrated how AlkD encircles half of the DNA duplex to place the shallow active site surface against the DNA backbone of the aberrant base pair.[24] These initial structures revealed that the AlkD active site is distinctly different from traditional base flipping glycosylases in that there was neither a putative extrahelical nucleobase binding pocket nor an obvious helix intercalating residue. Instead, the enzyme was able to form a specific complex with a variety of non-Watson-Crick base pairs without contacting the damage itself. However, the protein-DNA contacts important for catalysis were not discernable because the damaged nucleotides resided on the solvent exposed strand and not adjacent to the active site.

In order to trap a catalytically relevant complex containing the damaged nucleotide in contact with the protein, we took advantage of a nearest neighbor effect on AlkD catalysis in which 7 mG excision activity was weakest with thymine 3′ to the lesion (E.H. Rubinson and B.F. Eichman, unpublished). Modeling thymine immediately 3′ to the aberrant base pair in the crystal structures ($A7^U$ in Figure 4a) generated a potential steric clash between the thymine methyl group and Lys183, providing a structural rationale for the sequence effect and a strategy for preventing binding of the enzyme to the strand opposite the lesion. Indeed, altering the original $^{5'}$T(3d3 mA)A$^{3'}$/$^{5'}$TTA$^{3'}$ sequence to $^{5'}$A(3d3 mA)A$^{3'}$/$^{5'}$TTT$^{3'}$ resulted in a structure with the modified nucleotide positioned against the active site (Figure 4b).[32] Although the uncharged 3d3 mA was refractory to excision by AlkD at neutral pH, it was clear from the electron density that hydrolysis of the N-glycosidic bond occurred in crystals grown under acidic conditions, producing a mixture of intact 3d3mA-DNA substrate and AP-DNA + 3d3 mA nucleobase product, thus confirming the catalytic relevance of this new DNA orientation. As in the previous mismatch structures, the 3d3 mA nucleotide remains stacked within the DNA helix in a sheared orientation, with the 3d3 mA displaced toward the minor groove (Figure 4b). As a consequence, the 3d3 mA deoxyribose is cradled by active site residues Asp113, Trp109, and Trp187, along with a water molecule positioned by the Asp113 carboxylate for in-line attack of the deoxyribose C1′ carbon. In addition to substrate and product complexes generated by harvesting crystals at various times after protein-DNA mixing, an AlkD-DNA complex mimicking the oxocarbenium intermediate was determined using 1-azaribose (1aR)-DNA and 3 mA nucleobase. Strikingly, the positions of the protein, DNA, and free nucleobase did not differ significantly among the substrate, intermediate, and product complexes, indicating that the protein structure is tailor-made to recognize a sheared non-Watson-Crick base pair for N-glycosidic bond hydrolysis without base flipping.

Perhaps the most remarkable feature of the structures of AlkD in complex with methylated DNA is that the active site residues do not directly contact the target nucleobase. Instead, the protein facilitates base excision exclusively through interactions with the deoxyribose of the alkylated nucleotide. The position of the Asp113 carboxylate is consistent with a role of this side chain in positioning the water nucleophile for in-line attack of the

anomeric C1′ carbon, and in stabilization of the developing positive charge on the oxocarbenium reaction intermediate, as proposed for other DNA glycosylases.[41–43] In addition, the CH-π interactions between Trp109 and Trp187 indole side chains with the lesion C2′ and C4′/C5′, respectively, suggested that these residues also play a role in stabilization of positive charge that develops on the deoxyribose as the N-glycosidic bond is broken.[32,44] Aromatic residues in base flipping glycosylases either bind the extrahelical base or plug the gap through π–π stacking, and thus mutation of these residues interferes with substrate binding.[8,45,46] In contrast, mutation of the highly conserved Trp109 and Trp187 in AlkD resulted in a more than 100-fold decrease in catalysis (k_{cat}) without compromising DNA binding.[44]

1.4. AlkD Is Uniquely Suited for Bulky YTM-DNA Adducts

The structures and non-base-flipping mechanism of BcAlkD helped explain the ability of the orthologous YtkR2 protein to excise bulky YTMA lesions as a means of providing *Streptomyces* sp. TP-A0356 resistance against the YTM natural product.[33] YTM covalently modifies DNA at the N3 nitrogen of adenine to produce minor groove adducts. Extensive CH-π interactions between the indole subunits of YTM and five deoxyribose rings along each DNA strand dramatically stabilize the double-helix by non-covalently tethering the opposing strands.[34] As a consequence, the YTM-adenosine N-glycosidic bond is protected from hydrolysis, as evidenced by its 1.2-year half-life of spontaneous depurination (compared to 1.1-week half-life of 3-methyladenosine).[34] The AlkD/3d3mA-DNA structures showed a large solvent filled cavity between the DNA minor groove and the AlkD active site surface that is perfectly shaped to accommodate a YTMA adduct.[32] Subsequent crystal structures of AlkD bound to the products of the YTM-DNA excision reaction (AP−DNA + YTMA nucleobase) confirmed this position of the toxin and further detailed how AlkD overcomes the YTM-DNA stability to enhance its rate of depurination by at least 10^7-fold[34] (Figure 4c). Sixteen residues that contact the DNA and/or YTM moiety help to pry open the minor groove of the DNA, allowing access to the N-glycosidic bond by a water nucleophile and catalytic residues Trp109, Asp113, and Trp187. The collective power of the three catalytic AlkD residues for YTMA hydrolysis is highlighted by the fact that either single alanine mutants or addition of an electron-withdrawing fluorine atom onto the deoxyribose C2′ carbon still allowed for a 10^4 rate enhancement of base excision activity.[34]

The BER pathway is typically not considered to play a role in excision of bulky lesions, and in fact in human cells NER has been shown to protect against CC-1065, a member of the same family of natural products as YTM.[35,36,47] Using *alkD* and *uvrA* deletion strains of *Bacillus anthracis*, we found that both BER and NER provided modest protection against YTM toxicity in bacteria. However, even very low (nM) concentrations of YTM were sufficient to kill wild-type cells, indicating that neither pathway was particularly effective at conferring YTM resistance.[34] We would not expect YTM-DNA lesions to be substrates for the NER pathway given the enhanced stability of the YTM-DNA lesions. However, the inefficiency of AlkD-mediated BER against YTM toxicity in cells is somewhat paradoxical given the high activity of AlkD at YTMA excision in vitro. One explanation for this paradox was provided by the fact that the AlkD/AP-DNA product inhibited incision of the AP site by either of two canonical bacterial AP endonucleases, ExoIII, and EndoIV.[34] Thus, although AlkD is highly efficient at excising YTMA, the product of that reaction inhibits the remaining steps of the pathway. Presumably, this inhibition is not found in the host strain, since *ytkR2* and several other putative BER genes – *ytkR4* and *ytkR5* – are embedded within the YTM synthesis cluster and likely constitute a self-resistance mechanism against YTM toxicity.[48] YtkR5 is a putative AP endonuclease distantly related to EndoIV that may recognize the AP-DNA product of the YtkR2 reaction. Similarly, YtkR4 is distantly related to the TatD exonuclease and thus conceivably also could help displace a product-inhibited glycosylase from the AP-site to allow completion of the BER pathway. The divergence of *ytkR4* and *ytkR5* from other bacterial BER genes, and the fact that more canonical TatD and EndoIV homologs are found in the host strain, outside of the YTM gene cluster, raises the possibility that a specialized BER pathway exists for YTM repair in the host.[34]

1.5. AlkC Uses a Non-Base-Flipping Mechanism to Select for Small Alkyl-Adducts

Compared to AlkD, AlkC cleaves a narrow range of positively charged alkylpurines. AlkC displays only weak activity for 7 mG and no activity for bulky lesions.[13,28] Originally found to remove minor groove methylated adducts 3 mA and 3 mG,[13] AlkC also has robust excision activity for 3 mC and weaker activity for 1 mA.[28] Both 3 mC and 1 mA are also substrates of AlkB-catalyzed oxidative demethylation. Of all bacterial genomes that contain *alkB* or *alkC*, only 6% contain both genes, suggesting that AlkC acts to remove 3 mC and 1 mA in bacteria that lack AlkB. However, the two proteins do not have analogous repair activities. Although methylation of cytosine N3 and adenine N1 disrupts Watson-Crick base pairing, AlkC only excises 3 mC and 1 mA in double-stranded DNA, and has no activity toward these lesions in single-stranded (ss) DNA or RNA.[28] In contrast, AlkB has a preference for 1 mA and 3 mC in ssDNA and RNA,[49] suggesting that AlkC and AlkB are not functionally homologous.

The structure of *Pseudomonas fluorescens* (Pf) AlkC in complex with DNA containing a 1aR oxocarbenium intermediate analog (**Figure 5**a) showed that like AlkD, AlkC is also a non-base-flipping glycosylase.[28] However, despite their similar HLR architectures, AlkC, and AlkD have different strategies for lesion recognition and excision. Whereas AlkD recognizes and catalyzes base excision with only modest perturbation of the DNA helical axis and base stacking, AlkC kinks the DNA by 60–80°, opening up the minor groove at the lesion (Figure 5). As a consequence, the lesion is exposed to the active site on the concave cleft of the protein without the need for a flipping mechanism. Instead of the catalytic Trp-Asp-Trp motif found in AlkD, AlkC utilizes two glutamate residues (Glu121 and Glu156) to position a water nucleophile for attack at the C1′ carbon and to stabilize the developing positive charge on the deoxyribose (Figure 5a). Substitution of either of the two glutamate residues with alanine in PfAlkC impairs 3 mA excision and abolishes

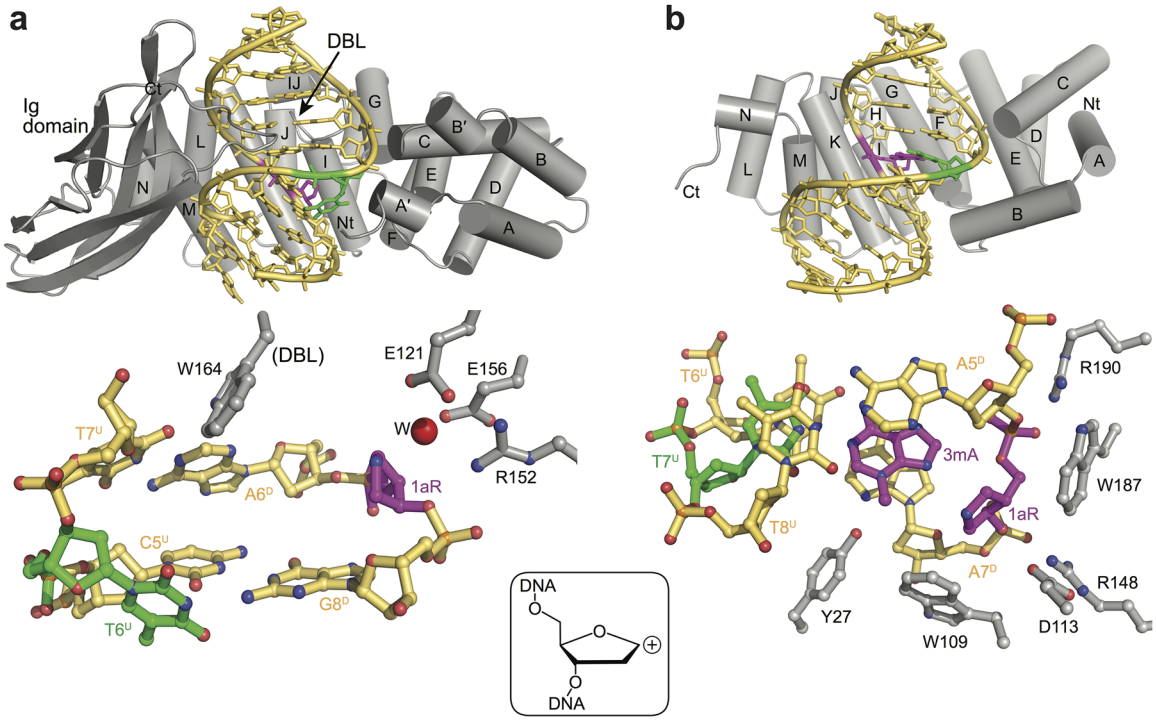

Figure 5. Comparison of AlkC and AlkD active sites. a) AlkC/1aR-DNA binary product complex (PDB ID 5VHV). b) AlkD/1aR-DNA/3mAde ternary product complex (PDB ID 5CLD). Structures are colored as in **Figure 4**. A chemical schematic of 1aR is shown in the inset.

1 mA and 3 mC excision activities. These catalytic glutamates are part of an AlkC-specific network of alternating charged residues Glu121-Arg152-Glu156-Arg159-Asp203 that interact electrostatically with the DNA phosphate backbone.

The specificity of AlkC for small methylated bases can be explained by the presence of the highly conserved, AlkC-specific DNA binding loop (DBL) located between HLR3 and HLR4. This loop protrudes into the exposed minor groove and presumably blocks binding to DNA that contains a bulky minor groove lesion. At the tip of the DBL is Trp164, which forms one face of the active site (Figure 5a). This residue is predicted to stabilize cationic nucleobases in the active site through a cation-π interaction, while also sterically preventing binding by 1 mA.[28] Widening of the minor groove to allow for access by Trp164 and active site residues is augmented by the helix dipolar interactions from AlkC's unique NTB to the backbone of the undamaged strand (Figures 3a and 5a).

AlkC is also distinct in that it is the only HLR glycosylase to contain a second, non-HLR domain. The majority (70%) of AlkC orthologs (AlkCβ) contain an additional C-terminal 100-residue Ig-like domain, which is essential for DNA binding in this subfamily (Figure 5a).[28] The Ig domain contacts DNA through backbone and side chain interactions from the extended EF-loop (following Ig domain nomenclature)[50] and the short turn between strands βC and βC′ (what we refer to as the CC′-turn). Interestingly, this mode of DNA binding by AlkC's Ig domain differs from those found in eukaryotic transcription factors.[28,51] AlkC's EF-loop makes nucleobase contacts in the major groove (directly across the helix from the DBL-minor groove interaction), whereas in transcription factors the EF-loop forms DNA backbone contacts. Because deletion of the Ig domain abrogates DNA binding by PfAlkC, we postulate that AlkCα proteins, which lack the Ig domain, contain a compensating DNA binding motif that helps stabilize the sharp kink in the DNA. In support of this, the sequences of the NTB and HLR1, both of which contribute to DNA binding in the PfAlkC structure, differ between AlkCα and AlkCβ subtypes.[28] It is clear from the structure of PfAlkC (an AlkCβ type) that the NTB plays an important role in stabilizing a particular DNA conformation by contacting the backbone of the undamaged strand in the immediate vicinity of the lesion.

1.6. AlkD2 Highlights the Importance of the N-Terminal Bundle of the HLR Family

AlkD2 is a non-enzymatic AlkC/AlkD paralog in a small number of bacteria and exhibits only very weak DNA binding affinity.[38] Its overall structure resembles a hybrid of the N-terminal half of AlkD and the C-terminal half of AlkC's HLR domain. The most notable structural feature of AlkD2 is the absence of the αB-helix, an essential DNA binding element in AlkD that provides the only nucleobase contact from the protein.[24,32] Absence of the αB-helix from AlkD2 is responsible for its weak DNA binding.[38] While the role of AlkD2 is not clear, it is noteworthy that the *S. mutans alkD2* gene is clustered with those predicted to be involved in purine biosynthesis, suggesting a potential role of AlkD2 in purine metabolism. Consistent with this hypothesis, we found that AlkD2 specifically binds inosine-5′-monophosphate (IMP), an intermediate in purine metabolism, in AlkD2

crystals soaked with the nucleotide (Table S3, Supporting Information). The new 1.4-Å structure (PDB ID 6M9M) shows that IMP is bound to AlkD2 by the loop that spans helices αA and αC, and by the N-terminal end of helix αG, both of which are locations that contribute to DNA binding by AlkD (**Figure 6**a). Similar to their role in binding inorganic phosphate in the unliganded AlkD2 structure[38] (Figure 3b), His17 and Gly18 in the αA-αC loop and Arg85 on αG contribute electrostatic interactions to the phosphate, while the inosine nucleobase forms π-stacking interactions with the indole side chain of Trp84 (Figure 6b). Interestingly, this tryptophan corresponds to the catalytic Trp109 in AlkD (Figure 6c). Although this manner of IMP binding is consistent with a role for AlkD2 in nucleotide metabolism, the biological significance remains to be determined.

The AlkD2 structure provided a useful tool in evaluating the importance of the NTB in DNA binding by AlkD. We now know that this unique region of AlkC also contributes to DNA binding and may impart specific DNA binding functions to AlkCα and AlkCβ subtypes.[28] This motif has been observed to play a similar binding role in other helical repeat proteins. For example, a Dali search[52] of the AlkD2 structure showed a strong structural similarity to the N-terminal six HEAT repeats of exportin-5, which forms a complex with the guanine triphosphatase Ran to transport pre-miRNA from the nucleus. Exportin-5 forms a highly curved solenoid structure from over 20 HEAT repeats that cradle both nucleotide bound Ran•GTP and pre-miRNA in the concave surface.[21] There is a striking similarity to the position of Ran•GTP, DNA, and nucleotide bound to the NTB and HEAT repeats of exportin-5, AlkD, and AlkD2, respectively (Figure 6a). Thus, the NTB is an important ligand specificity determinant in these and possibly other helical repeat proteins.

1.7. AlkF Binds Branched DNA Structures

The cellular functions of AlkF and its close ortholog AlkG have not been well characterized. They do not display glycosylase activity toward alkylated, oxidized, or other modified DNA nucleotides. Single- and double *alkF*$^-$ and *alkG*$^-$ knockout cells show no or only very low sensitivity to genotoxic stress agents, and neither protein is essential for cell growth under normal conditions. Thus, there is so far no evidence that AlkF or AlkG participates in repair of DNA damage. Interestingly, however, AlkF and AlkG show a specific DNA binding activity toward branched structures, including three- and four-way (Holliday) junctions,[37] although the affinity for branched DNA is relatively weak (low μM) compared the nM binding affinities of AlkC and AlkD for damaged DNA.[28,38]

As the first HLR protein deposited into the Protein Data Bank, the structure of *Bacillus cereus* AlkF originally appeared as a

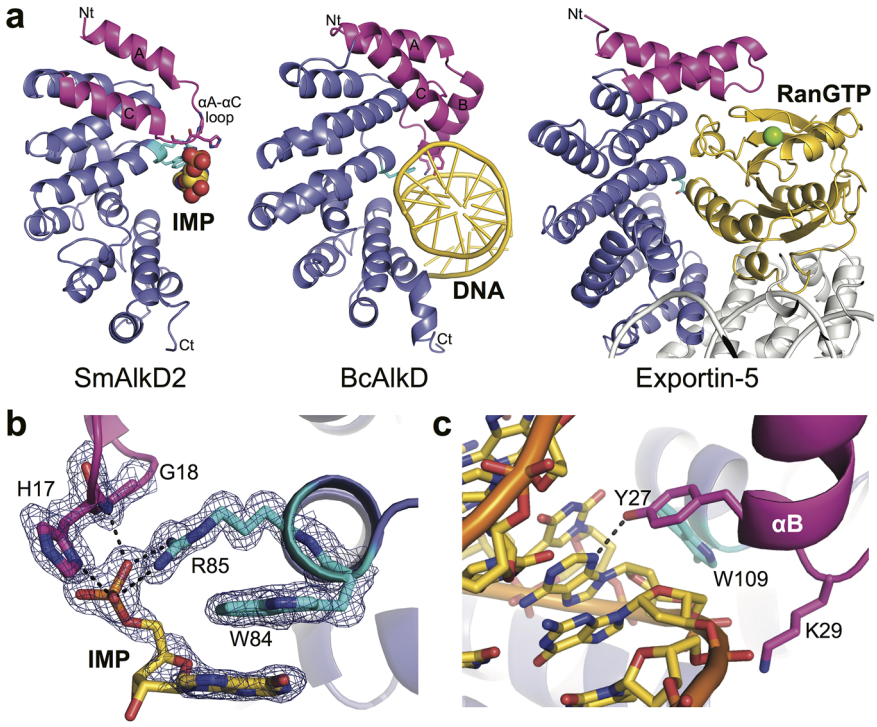

Figure 6. The N-terminal helical bundle is important for substrate recognition in HLR and related proteins. a) Structures of SmAlkD2 bound to IMP (PDB ID 6M9M), BcAlkC bound to G•T mismatch-containing DNA (3JXY), and exportin-5 bound to RanGTP and pre-miRNA (3A6P). The N-terminal bundle (NTB) is colored magenta, the binding partner in the concave cleft is colored gold, and HLR residues at the N-terminal end of helix αG that interact with the binding partner are colored cyan. b) Contacts between SmAlkD2 and IMP. Residues are colored as in panel A. 2Fo-Fc electron density is contoured at 1σ. Hydrogen bonds are shown as black dashed lines. c) AlkD-DNA contacts from helix αB (magenta) and αB (cyan) in the BcAlkD/G•T-DNA mismatch structure (PDB ID 3JXY).

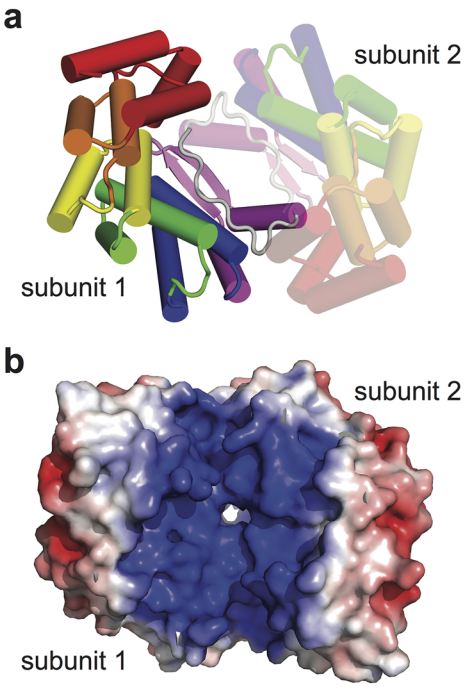

Figure 7. Dimerization of AlkF. Dimer structure of AlkF as observed in the crystal structure (PDB ID 3ZBO). a) Secondary structure representation colored as in **Figure 3**. b) Electrostatic potential (blue, positive; red, negative) mapped onto the dimer solvent accessible surface. The saturation of the colors is proportional to the degree of electrostatic charge from -7 to $+7$ k_BT/e_C. The orientation of the dimer is identical to that shown in panel a.

hypothetical protein and structural genomics target (PDB ID 1T06), and later was determined to higher resolution after its characterization as a distinct evolutionary homolog of AlkC and AlkD (PDB ID 3ZBO)[37] (Table S2, Supporting Information). The concave surface of AlkF is not as uniformly positively charged as other DNA binding HLR proteins (Figure 3b), consistent with the protein's weak DNA affinity. In addition, the normal DNA surface along the concave surface is disrupted by a unique 12-amino acid β-hairpin insertion between helix αL and αM of HLR5 at the C-terminus. This element may explain the preference of the protein for branched DNA structures, as alanine substitution of Arg203, Lys206, and Lys207 within the β-hairpin greatly reduces preference of AlkF to branched DNA.[37] However, the precise role of the β-hairpin is still not well defined. Interestingly, AlkF crystallized as a dimer with helix αM of the disrupted HLR5 facilitating a significant intermolecular contact (**Figure 7**a). An extended random coil at the extreme C-terminus folded back on the concave surface to form a highly positively charged groove at the dimer interface (Figure 7b). It is intriguing to speculate that this dimeric surface may endow the protein with the ability to bind branched nucleic acids or other ligands.

2. Conclusion

The structures of AlkC and AlkD have revealed a number of novel principles regarding DNA damage recognition. Unlike other DNA glycosylases, AlkC and AlkD do not flip their substrate nucleobases into an active site pocket, but rather engage the lesion deoxyribose using the concave surface of the HLR solenoid structure. DNA binding by AlkD widens the minor groove, allowing its catalytic Trp-Asp-Trp triad to stabilize the developing positive charge on the deoxyribose through CH-π interactions. There is enough room between the AlkD surface and the DNA base stack to accommodate adenine adducts of YTM, a bulky and highly toxic secondary metabolite also recognized by NER machinery. In contrast, AlkC imposes a more dramatic perturbation to the DNA helix, exposing the damage site to a pair of catalytic glutamate residues that likely play the same function as AlkD's Trp-Asp-Trp triad. The AlkC family is largely defined by an inserted DNA binding loop (DBL), which blocks access to bulky DNA adducts and helps provide specificity for 3 mC, a known substrate for direct reversal by the oxidative demethylase AlkB. The majority of AlkC proteins are of the AlkCβ type, which have an Ig domain appended to the C-terminus of the HLR domain that clamps the DNA by interacting with the major groove directly across helix from where the DBL binds the minor groove at the lesion. DNA binding by the AlkC Ig domain expands the role of these domains in bacteria and is distinct from the mode of DNA binding observed eukaryotic Ig-containing transcription factors. The non-enzymatic AlkD2 protein is a hybrid of AlkC and AlkD architectures and has very low DNA binding affinity as a result of the αB-helix important for DNA binding in AlkD. Similarly, AlkF contains a β-hairpin insertion in the DNA binding cleft that may endow the protein with its ability to bind branched DNA structures.

The novel DNA binding and enzymatic functions of the proteins within the HLR superfamily have expanded the role of the HEAT repeat architecture beyond protein-protein scaffold to now include nucleic acid binding to mediate DNA repair and likely other functions. Despite advances in our understanding of their structures and biochemistry, however, important biological questions remain. Mostly glaringly, we do not yet know the cellular functions of non-enzymatic AlkD2 and AlkF. Current evidence for both proteins suggests roles outside of DNA repair, and AlkD2's putative nucleotide binding function and genetic context may indicate a role in nucleotide metabolism. Second, the significance of AlkC's 3 mC activity and whether AlkC protects against this lesion in bacteria lacking AlkB remains to be determined. Finally, it seems clear that AlkD proteins evolved to protect bacteria from YTM/CC-1065/duocarmycin toxins. However, we do not yet know if AlkD or its homologs within the YTM (YtkR2) or CC-1065 (C10R5) gene clusters have a specificity for one particular secondary metabolite, nor do we know the fates of the potentially inhibitory enzyme-product complexes in the toxin-producing strains.[34] More work is needed to answer these important questions related to a unique family of proteins.

Abbreviations

1aR, 1-azaribose; 3d3mA, 3-deaza-3-methyladenine; 3mA, 3-methyladenine; 7mG, 7-methylguanine; BER, base excision repair; NER, nucleotide excision repair; YTM, yatakemycin; YTMA, 3-yatakemycinyladenine.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The authors thank Dr. Elwood Mullins for comments on the manuscript. Research on DNA glycosylases in the Eichman laboratory is funded by the National Science Foundation (MCB-1517695).

Conflict of Interest

The authors declare no conflict of interest.

Keywords

base excision repair, base flipping, DNA alkylation, DNA glycosylase, DNA repair, HEAT repeat, yatakemycin

Received: July 24, 2018
Revised: August 27, 2018
Published online: September 28, 2018

[1] T. Lindahl, *Nature* **1993**, *362*, 709.
[2] E. C. Friedberg, *Cell Res.* **2008**, *18*, 3.
[3] M. D. Wyatt, D. L. Pittman, *Chem. Res. Toxicol.* **2006**, *19*, 1580.
[4] A. Tubbs, A. Nussenzweig, *Cell* **2017**, *168*, 644.
[5] B. Sedgwick, *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 148.
[6] E. H. Rubinson, S. Adhikary, B. F. Eichman, in: *ACS Symposium Series: Structural Biology of DNA Damage and Repair*, (Ed: M. P. Stone), Vol. *1041*. Oxford University Press, Washingon, D.C. **2009**, pp. 29–45.
[7] S. C. Brooks, S. Adhikary, E. H. Rubinson, B. F. Eichman, *Biochim. Biophys. Acta* **2013**, *1834*, 247.
[8] A. Y. Lau, O. D. Scharer, L. Samson, G. L. Verdine, T. Ellenberger, *Cell* **1998**, *95*, 249.
[9] S. Adhikary, B. F. Eichman, *EMBO Rep.* **2011**, *12*, 1286.
[10] A. C. Drohat, K. Kwon, D. J. Krosky, J. T. Stivers, *Nat. Struct. Biol.* **2002**, *9*, 659.
[11] J. Labahn, O. D. Scharer, A. Long, K. Ezaz-Nikpay, G. L. Verdine, T. E. Ellenberger, *Cell* **1996**, *86*, 321.
[12] R. J. Roberts, X. Cheng, *Annu. Rev. Biochem.* **1998**, *67*, 181.
[13] I. Alseth, T. Rognes, T. Lindback, I. Solberg, K. Robertsen, K. I. Kristiansen, D. Mainieri, L. Lillehagen, A. B. Kolsto, M. Bjoras, *Mol. Microbiol.* **2006**, *59*, 1602.
[14] B. Dalhus, I. H. Helle, P. H. Backe, I. Alseth, T. Rognes, M. Bjoras, J. K. Laerdahl, *Nucleic Acids Res.* **2007**, *35*, 2451.
[15] A. F. Neuwald, T. Hirano, *Genome Res.* **2000**, *10*, 1445.
[16] J. Perry, N. Kleckner, *Cell* **2003**, *112*, 151.
[17] B. L. Sibanda, D. Y. Chirgadze, D. B. Ascher, T. L. Blundell, *Science* **2017**, *355*, 520.
[18] M. A. Andrade, C. Petosa, S. I. O'Donoghue, C. W. Muller, P. Bork, *J. Mol. Biol.* **2001**, *309*, 1.
[19] T. Z. Grove, A. L. Cortajarena, L. Regan, *Curr. Opin. Struct. Biol.* **2008**, *18*, 507.
[20] S. H. Yoshimura, T. Hirano, *J. Cell Sci.* **2016**, *129*, 3963.
[21] C. Okada, E. Yamashita, S. J. Lee, S. Shibata, J. Katahira, A. Nakagawa, Y. Yoneda, T. Tsukihara, *Science* **2009**, *326*, 1275.
[22] R. M. Lahr, S. M. Mack, A. Heroux, S. P. Blagden, C. Bousquet-Antonelli, J. M. Deragon, A. J. Berman, *Nucleic Acids Res.* **2015**, *43*, 8077.
[23] M. Kschonsak, F. Merkel, S. Bisht, J. Metz, V. Rybin, M. Hassler, C. H. Haering, *Cell* **2017**, *171*, 588.
[24] E. H. Rubinson, A. S. Gowda, T. E. Spratt, B. Gold, B. F. Eichman, *Nature* **2010**, *468*, 406.
[25] E. H. Rubinson, B. F. Eichman, *Curr. Opin. Struct. Biol.* **2012**, *22*, 101.
[26] Y. Mishina, E. M. Duguid, C. He, *Chem. Rev.* **2006**, *106*, 215.
[27] L. C. Gillet, O. D. Scharer, *Chem. Rev.* **2006**, *106*, 253.
[28] R. Shi, E. A. Mullins, X. X. Shen, K. T. Lay, P. K. Yuen, S. S. David, A. Rokas, B. F. Eichman, *EMBO J.* **2018**, *37*, 63.
[29] S. C. Trewick, T. F. Henshaw, R. P. Hausinger, T. Lindahl, B. Sedgwick, *Nature* **2002**, *419*, 174.
[30] P. O. Falnes, R. F. Johansen, E. Seeberg, *Nature* **2002**, *419*, 178.
[31] D. L. Boger, D. S. Johnson, *Angew Chem. Int. Edit.* **1996**, *35*, 1438.
[32] E. A. Mullins, R. Shi, Z. D. Parsons, P. K. Yuen, S. S. David, Y. Igarashi, B. F. Eichman, *Nature* **2015**, *527*, 254.
[33] H. Xu, W. Huang, Q. L. He, Z. X. Zhao, F. Zhang, R. X. Wang, J. W. Kang, G. L. Tang, *Angew Chem. Int. Edit.* **2012**, *51*, 10532.
[34] E. A. Mullins, R. Shi, B. F. Eichman, *Nat. Chem. Biol.* **2017**, *13*, 1002.
[35] D. Gunz, M. T. Hess, H. Naegeli, *J. Biol. Chem.* **1996**, *271*, 25089.
[36] C. P. Selby, A. Sancar, *Biochemistry* **1988**, *27*, 7184.
[37] P. H. Backe, R. Simm, J. K. Laerdahl, B. Dalhus, A. Fagerlund, O. A. Okstad, T. Rognes, I. Alseth, A. B. Kolsto, M. Bjoras, *J. Struct. Biol.* **2013**, *183*, 66.
[38] E. A. Mullins, R. Shi, L. A. Kotsch, B. F. Eichman, *PLoS ONE* **2015**, *10*, e0127733.
[39] E. H. Rubinson, A. H. Metz, J. O'Quin, B. F. Eichman, *J. Mol. Biol.* **2008**, *381*, 13.
[40] E. A. Mullins, E. H. Rubinson, B. F. Eichman, *DNA Repair (Amst)* **2014**, *13*, 50.
[41] R. M. Werner, J. T. Stivers, *Biochemistry* **2000**, *39*, 14054.
[42] J. A. McCann, P. J. Berti, *J. Am. Chem. Soc.* **2008**, *130*, 5789.
[43] K. M. Schermerhorn, S. Delaney, *Accounts Chem. Res.* **2014**, *47*, 1238.
[44] Z. D. Parsons, J. M. Bland, E. A. Mullins, B. F. Eichman, *J. Am. Chem. Soc.* **2016**, *138*, 11485.
[45] T. Hollis, Y. Ichikawa, T. Ellenberger, *EMBO J.* **2000**, *19*, 758.
[46] B. Dalhus, J. K. Laerdahl, P. H. Backe, M. Bjørås, *FEMS Microbiol. Rev.* **2009**, *33*, 1044.
[47] S. G. Jin, J. H. Choi, B. Ahn, T. R. O'Connor, W. Mar, C. S. Lee, *Mol. Cells* **2001**, *11*, 41.
[48] W. Huang, H. Xu, Y. Li, F. Zhang, X. Y. Chen, Q. L. He, Y. Igarashi, G. L. Tang, *J. Am. Chem. Soc.* **2012**, *134*, 8831.
[49] P. O. Falnes, M. Bjoras, P. A. Aas, O. Sundheim, E. Seeberg, *Nucleic Acids Res.* **2004**, *32*, 3456.
[50] P. Bork, L. Holm, C. Sander, *J. Mol. Biol.* **1994**, *242*, 309.
[51] M. J. Rudolph, J. P. Gergen, *Nat. Struct. Biol.* **2001**, *8*, 384.
[52] L. Holm, C. Sander, *J. Mol. Biol.* **1993**, *233*, 123.
[53] I. Letunic, P. Bork, *Nucleic Acids Res.* **2011**, *39*, W475.
[54] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, *Mol. Syst. Biol.* **2011**, *7*, 539.