Published online 29 March 2007

Nucleic Acids Research, 2007, Vol. 35, No. 7 ***2451–2459***
doi:10.1093/nar/gkm039

Structural insight into repair of alkylated DNA by a new superfamily of DNA glycosylases comprising HEAT-like repeats

Bjørn Dalhus[1,2], Ina Høydal Helle[1,2], Paul H. Backe[1,2], Ingrun Alseth[1], Torbjørn Rognes[1,3], Magnar Bjørås[1,2] and Jon K. Laerdahl[1,*]

[1]Centre for Molecular Biology and Neuroscience (CMBN) and Institute of Medical Microbiology, Rikshospitalet-Radiumhospitalet Medical Centre, N-0027 Oslo, Norway, [2]Institute of Clinical Biochemistry, University of Oslo, N-0027 Oslo, Norway and [3]Department of Informatics, University of Oslo, PO Box 1080 Blindern, N-0316 Oslo, Norway

Received November 14, 2006; Revised December 20, 2006; Accepted January 8, 2007

ABSTRACT

3-methyladenine DNA glycosylases initiate repair of cytotoxic and promutagenic alkylated bases in DNA. We demonstrate by comparative modelling that *Bacillus cereus* AlkD belongs to a new, fifth, structural superfamily of DNA glycosylases with an alpha–alpha superhelix fold comprising six HEAT-like repeats. The structure reveals a wide, positively charged groove, including a putative base recognition pocket. This groove appears to be suitable for the accommodation of double-stranded DNA with a flipped-out alkylated base. Site-specific mutagenesis within the recognition pocket identified several residues essential for enzyme activity. The results suggest that the aromatic side chain of a tryptophan residue recognizes electron-deficient alkylated bases through stacking interactions, while an interacting aspartate–arginine pair is essential for removal of the damaged base. A structural model of AlkD bound to DNA with a flipped-out purine moiety gives insight into the catalytic machinery for this new class of DNA glycosylases.

INTRODUCTION

The integrity of DNA in all living cells is constantly challenged by damaging agents of exogenous and endogenous origin. Damage to DNA can lead to mutations or may interfere with normal cellular processes such as DNA replication and transcription. To avoid the deleterious effects of DNA base damage, there has been a strong driving force for the evolution of enzymatic pathways that recognize and repair DNA base lesions. Damage to single DNA bases due to processes such as oxidation, alkylation or deamination are principally handled by the highly conserved multistep base excision repair (BER) pathway (1–3). BER is initiated by DNA glycosylases (4) which locate damaged and, in some cases, mispaired bases and excise them by hydrolyzing the *N*-glycosylic bond between the 2′-deoxyribose and the base. Monofunctional DNA glycosylases use an activated water molecule to attack the anomeric carbon of the damaged nucleotide, while bifunctional DNA glycosylases use an active site amine nucleophile and, in a concerted fashion, also nick the DNA backbone 3′ to the nascent abasic site (5). In the subsequent steps, abasic site endonucleases cleave 5′ to the abasic nucleotide and the DNA ends are processed by phosphodiesterases. Finally, DNA polymerase and DNA ligase activities result in the complete restoration of the DNA sequence.

All organisms express several DNA glycosylases, each with unique substrate specificities. In human cells 10 distinct DNA glycosylases have now been characterized (6). Despite the wide range of base lesion targets for the DNA glycosylases, nature has employed a limited number of protein folds for these enzymes (7,8). Four structural superfamilies of DNA glycosylases are currently recognized (7); the helix-hairpin-helix (HhH) and helix-two-turn-helix (H2TH) superfamilies termed after characteristic active site motifs, and the uracil DNA glycosylase (UDG) and alkyladenine DNA glycosylase (AAG) superfamilies named after structural similarity to two well-characterized human DNA glycosylases.

Members of the H2TH and HhH superfamilies contain multiple protein domains, and the active site and DNA-binding region is located at the junction between these domains. Examples of H2TH DNA glycosylases are *E. coli* formamidopyrimidine DNA glycosylase (Fpg) and endonuclease VIII (Nei), as well as the recently discovered vertebrate enzymes NEIL1 and NEIL2 (Nei-like), which all recognize and excise various

*To whom correspondence should be addressed. Tel: +47 22844784; Fax: +47 22844782; Email: j.k.lardahl@medisin.uio.no

oxidized base lesions. The HhH DNA glycosylase superfamily comprises a wide range of enzymes that excise oxidized bases, e.g. 8-oxoguanine DNA glycosylase (OGG1) and endonuclease III (Nth), alkylated bases, e.g. bacterial 3-methyladenine (3mA) DNA glycosylase II (AlkA) and mismatches, e.g. the adenine DNA glycosylase (MutY) that removes A from A/G mismatches. DNA glycosylases of the UDG and AAG structural superfamilies are compact single-domain enzymes. The members of the widely studied UDG superfamily contain a core of a four-stranded parallel twisted β-sheet flanked by α-helices. AAG consists of a single domain with an antiparallel β-sheet surrounded by α-helices. The major repair enzymes for removal of uracil in DNA belong to the UDG superfamily, while AAG is the only known mammalian DNA glycosylase specific to alkylated bases.

Recently, Alseth *et al.* (9) characterized two novel 3mA DNA glycosylases from *Bacillus cereus*, AlkC and AlkD, which are without sequence similarity to any protein of known function. Both proteins are monofunctional DNA glycosylases specific for removal of *N*-alkylated bases and without affinity for other base lesions such as deamination and oxidation products. In the present work we demonstrate that *B. cereus* AlkD and AlkC are members of a new and uncharacterized structural superfamily of DNA glycosylases with an alpha–alpha superhelix fold. A structural model of AlkD based on homology modelling has been generated, and the location of a putative active site pocket has been determined by investigating residue conservation. Residues participating in alkylated DNA recognition and catalysis were detected by functional and biochemical analysis of mutant proteins generated on the basis of the structural model. Finally, a model of AlkD in complex with DNA containing a flipped-out purine moiety is analysed.

MATERIALS AND METHODS

Computational structure prediction and analysis

The accession numbers for the *B. cereus* AlkC and AlkD proteins are CAJ31884 and CAJ31885, respectively. Single iteration sequence similarity searching in the NCBI non-redundant protein database were performed with PARALIGN (10) and NCBI BLAST (11) and gave a set of approximately 80 significant hits (*E*-value <0.01) for *B. cereus* AlkD. After removal of erroneous and redundant sequences, the remaining 43 AlkD homologs were aligned with the multiple sequence alignment (MSA) program T-Coffee (12). The MSAs were viewed and manipulated with Jalview (13). Protein secondary structure predictions were generated with PSIPRED (14,15), PROF (16), Jnet (17) and SSpro (18), and protein structural disorder predictions were performed with DISOPRED2 (19) and VSL1 (20).

Fold recognition protein structure predictions were obtained with mGenTHREADER (14,21), Phyre (22) and SAM-T02 (23). Experimental structures for proteins homologous to *B. cereus* AlkD were obtained from the Protein Data Bank (PDB) (24), including the structure of the hypothetical protein EF3068 of *Enterococcus faecalis* V583 (PDB identifier 2B6C) determined by J. Osipiuk, C. Hatzos, S. Moy, F. Collart and A. Joachimiak. 2B6C is a 2.1 Å resolution X-ray structure generated as a part of the structural genomics effort of the Midwest Center for Structural Genomics (MCSG; http://www.mcsg.anl.gov). No accompanying publication or details about the structure have been released yet, apart from the experimental procedure and the coordinates in the PDB. A structural model for the *B. cereus* AlkD target was generated from the 2B6C template employing standard comparative modelling with SwissModel (25), and the model was analysed with WhatCheck (26).

The SCOP structural classification of proteins database (27) was used to investigate the relationships between the various repeat-rich structural protein families. Electrostatic potentials were calculated with APBS (28) using the structure prepared for continuum electrostatistics calculations with PDB2PQR (29) and employing PROPKA (30) for the pK_a calculations. Amino acid conservation mapped onto the protein surface was generated with ConSurf (31,32). All illustrations of protein structure were generated with PyMOL (33).

Modelling of the AlkD–DNA complex

DNA coordinates were extracted from the structure of AlkA in complex with a 15-mer double-stranded DNA incorporating the abasic site (AP-site) analogue 1-azaribose (PDB structure 1DIZ (34)). A normal adenine nucleotide was superimposed onto the ribose of the flipped-out AP-site analogue, and the base was oriented to approximate a proposed model of 3mA inserted into the AlkA active site pocket (34). Subsequently, this DNA model with a flipped-out adenine, was manually docked onto the surface of AlkD making sure that the widened DNA minor groove was facing the protein and with complementary surfaces. Various orientations involving the flipped-out purine stacked with either Trp109 or Trp187 and with the anomeric carbon atom in reasonable distance to Asp113 were examined. The solution with the extrahelical adenine stacking with Trp109 appeared superior to the other orientations. This initial model complex was solvated using SYBYL (Tripos Inc.) and the protein–DNA interface was optimized, removing a few steric clashes, using 500 steps of conjugated gradient minimization in CNS (35). Initially, DNA from several DNA glycosylases as well as AP endonucleases were examined, but only DNA from AlkA and human OGG1 gave sensible, essentially equivalent, models without suffering from substantial steric conflicts or limited protein–DNA contacts.

Site-directed mutagenesis and cloning

The Quick Change site-directed mutagenesis kit (Stratagene) was used to introduce mutations with the pUC18-AlkD construct as template (9). Eight different oligonucleotides were designed using the manufacturer's protocol (from 5′ to 3′, new codon in bold): gcagaaccgatg gcacgt**gct**atgaaaaatcacttccta (Y27A), attgtaacaaaatcttgg **gcg**gacactgtcgatagcatc (W109A), tcttggtgggacactgtc**aat**ag

catcgtccctacattt (D113N), attgcatcagataacata**gcg**ttacaacg ggccgctatt (W145A), gataacatatggttacaa**gcg**gccgctattttat tccag (R148A), ctacattcttcgaaagaa**gct**ttcattcaaaaa gcgatt (F179A), cattcttcgaaagaatttt**gcc**attcaaaaagcg attgga (F180A), attcaaaaagcgattggc**gcg**gtccttcgtgaatatgca (W187A). The AlkD mutant W187A was designed using a pT7-SCII AlkD construct as template. Mutant constructs were verified by DNA sequencing.

Preparation of crude cell extracts

AlkD mutant and wild-type constructs, as well as the pUC18 vector (control) were transformed into the 3mA DNA glycosylase deficient strain *E. coli* BK2118 (*tag alkA*), described by Clarke *et al.* (36), grown in 1 l LB medium containing 100 μg ml^{-1} ampicillin to an OD_{600} of ~1.0 and induced with 0.3 mM isopropyl-β-D-thiogalactopyranoside for 2 h at 37°C. Cells were harvested by centrifugation and protein extracts were prepared by sonication of the suspended cell pellets in 50 mM NaCl, 10 mM Tris–HCl, pH 7.0 and 10 mM β-mercaptoethanol. The cell debris was separated from the protein extract by centrifugation at 15 000 rpm for 20 min at 4°C. Extracts prepared from the strains expressing AlkD showed a strong band corresponding to the molecular weight of AlkD on denaturing protein gels stained with Coomassie blue. This protein band was absent in extracts of the strain transformed with empty vector. The AlkD protein concentrations of each extract were quantified from the band intensity and showed that all mutant constructs and wild-type construct expressed equal amounts of soluble AlkD protein. The activities of the crude extracts were assessed by the DNA glycosylase assay.

DNA glycosylase assay

Calf thymus DNA alkylated with *N*-[^{3}H]-methyl-*N*-nitrosourea ([^{3}H]-MNU) (1.5 Ci mmol^{-1}) was incubated with different amounts of cell extracts within the linear range, for 30 min at 37°C as described previously (37). DNA was precipitated with ethanol and the radioactivity in the supernatant was quantified in a liquid scintillation counter (Tri-Carb 2900TR, Packard).

Alkylation survival of *E. coli* BK2118 (*tag alkA*) and transformed derivatives

BK2118 cells transformed with the expression constructs pUC18-AlkD, pUC18-Y27A, pUC18-W109A, pUC18-D113N, pUC18-W145A, pUC18-R148A, pUC18-F179A, pUC18-F180A and pT7-W187A were grown in LB medium with ampicillin (100 μg ml^{-1}) to an OD_{600} of ~1.0 and diluted in M9-buffer to form a density gradient from 3×10^8 to 3×10^5 cells ml^{-1}. One microlitre aliquots of the diluted cultures were spotted onto LB agar plates without (control) and with 1 mM methyl methanesulfonate (MMS) as well as 100 μg ml^{-1} ampicillin. BK2118 cells transformed with empty pUC18-vector were used as a control. Following incubation for 1 day at 37°C, the plates were inspected for growth of colonies.

RESULTS

AlkD and AlkC have an alpha–alpha superhelix fold

Several PSI-BLAST (11) search iterations with *B. cereus* AlkD in the NCBI non-redundant protein database pick up many homologous proteins and demonstrate that AlkD belongs to a large family of prokaryotic proteins with currently more than 200 members in the sequence databases. *B. cereus* AlkC is also in this family, but it is a remote homolog of AlkD with sequence identity below 20%. Eukaryotic homologs have been identified in *Entamoeba histolytica* and *Dictyostelium discoideum* (9).

Structural disorder predictions suggest that *B. cereus* AlkD (237 residues) folds into a regular, well-defined tertiary structure, with only the C-terminal 15–20 amino acids being structurally flexible and disordered. *B. cereus* AlkC (256 residues) is predicted to have no regions with significant structural disorder. Several secondary structure prediction tools concurrently show that AlkD consists of 13 α-helices connected by short loops (Supplementary Figure 1). The AlkC secondary structure appears to comprise 13 or 14 α-helices although there is less consensus between the various predictions compared with those for AlkD (Supplementary Figure 2).

The mGenTHREADER (14,21) and Phyre (22) fold recognition servers predict with high confidence *B. cereus* AlkD to have the same overall structure as a family of all-alpha proteins—e.g. PDB (24) structures 1T06, 1OYZ and 1BK5—categorized in SCOP (27) as belonging to the alpha–alpha superhelix fold. They are further classified as being members of a superfamily with a fold similar to ARM- and HEAT-repeat-containing proteins (38). The results are less clear-cut for AlkC, but even so, 9 out of the 10 best similarity hits are for proteins belonging to the ARM/HEAT-repeat containing superfamily.

The SAM-T02 (23) fold recognition server also predicts both AlkD and AlkC to be ARM- or HEAT-repeat-containing proteins. In addition, the SAM-T02 server shows with very high confidence (SAM-T02 *E*-value below 10^{-20}) that AlkD has the same fold as the *E. faecalis* V583 hypothetical protein EF3068 (PDB structure 2B6C). The experimental structure 2B6C was released in November 2005, and was not detected by mGenTHREADER and Phyre due to its absence in the data set of these programs. 2B6C is also the best SAM-T02 hit for the AlkC structure (SAM-T02 *E*-value is 6×10^{-5}). The amino acid sequences of *E. faecalis* EF3068 and *B. cereus* AlkD have 35% sequence identity for a length of more than 200 residues, and with just one single insertion/deletion (Supplementary Figure 3). Proteins with this level of similarity generally have the same overall structure and protein fold (39). These predictions demonstrate that AlkD and AlkC are proteins comprising ARM/HEAT-like repeats and that a reliable model of AlkD can be generated with standard comparative modelling based on the structure of EF3068 from *E. faecalis*.

AlkD has a groove lined with positively charged residues, ideally shaped for accommodating DNA

Based on the sequence alignment between *E. faecalis* EF3068 and *B. cereus* AlkD (Supplementary Figure 3),

a structural model for AlkD was generated from the 2B6C chain B template employing standard comparative modelling with SwissModel (25), including backbone generation, loop building and side-chain optimization (Supplementary data file in PDB format). EF3068 is shorter than AlkD at both ends, and no reliable structure could be built for the 10–15 N- and C-terminal residues. However, the C-terminal region is predicted to be structurally disordered, and the N-terminal region, including α1, most likely packs against α3 and α4 as in another recently solved ARM/HEAT-repeat structure (PDB identifier 1T06). No abnormalities, not to be expected in a homology-based model, were detected by the WhatCheck (26) standard evaluation checks.

A cartoon representation of the *B. cereus* AlkD model is given in Figure 1a, with the APBS (28) calculated electrostatic potential mapped onto the protein surface in Figure 1b (see also Supplementary Video 1 online). A multiple sequence alignment of AlkD and homologs (Supplementary Figure 4) highlights conserved residues that probably are involved in catalysis and substrate recognition and binding (Supplementary Table 1). The dense packing of the AlkD α-helices results in a scoop-shaped overall structure containing a wide groove with a diameter of ~20–25 Å (Figure 1 and Supplementary Video 1). The groove, or front side of the protein, is positively charged and has a number of conserved Lys and Arg residues along the edges of this feature (Supplementary Figure 4 and Table 1). The groove region of the protein is ideally shaped and charged for accommodating double-stranded DNA through a salt-bridge network involving some of the conserved Lys/Arg residues and the DNA phosphate groups. The rear side of the protein is mainly negatively charged, making the protein overall neutral.

A structural alignment of the tandem repeats of *B. cereus* AlkD was generated manually from the model (Figure 2). Each repeat of AlkD, as in *E. faecalis* hypothetical protein EF3068, consists of only two anti-parallel helices (Figures 1a and 2) and is clearly more closely related to the standard HEAT repeat than to the ARM repeat (38). AlkD comprises six HEAT-like repeats, each with ~35 residues (Figure 2). The similarity is very low between the repeats and is mainly limited to the conserved location of hydrophobic residues.

Site-specific mutagenesis identifies active site residues

The degree of conservation of the various amino acid residues among the AlkD family members was calculated from a multiple sequence alignment of 43 AlkD homologs. Amino acid conservation was mapped onto the protein surface by employing the ConSurf program (31) in order to identify patches of conservation that might be functionally important (Figure 1c). A nest of conserved residues in *B. cereus* AlkD is located at one end of the putative DNA binding groove, near the C-terminus of α2 and the N-termini of α7, α9 and α11 (Figure 1c and d and Supplementary Table 1). Eight of these residues were selected for mutational studies in order to characterize their role in catalysis (Figure 1d).

Mutant AlkD proteins were expressed in the *E. coli tag alkA* mutant strain BK2118 (36) in order to avoid interference from endogenous 3mA DNA glycosylase activity. Crude extracts loaded on denaturing SDS–PAGE gels stained with Coomassie blue showed that all mutant proteins and wild-type AlkD were equally expressed (data not shown). The crude extracts were examined for the ability to remove alkylated bases from DNA treated with [^{3}H]-MNU. Extracts of mutants Y27A and F180A displayed DNA glycosylase activities comparable to the wild-type, whereas the activity of the W187A extract was reduced to approximately one-third of the wild type (Figure 3a). Extracts expressing AlkD mutants W109A, D113N and R148A had lost all DNA glycosylase activity for alkylated bases (Figure 3a), suggesting that these three residues are essential for the catalytic mechanism and/or recognition of the modified base. The three residues Trp109, Asp113 and Trp187 form a structural motif in AlkD with a configuration comparable to Trp272, Asp 238 and Trp218, respectively, in AlkA (34) (Figure 4). It is tempting to speculate that these three residues play similar roles in AlkD as in AlkA, which is, π–π and/or π-cation stacking with the extra-helical alkylated base (Trp109), a function in the nucleophilic attack on the glycosylic bond (Asp113) and coordination and stabilization of the DNA backbone around the alkylated nucleotide (Trp187).

In another set of experiments, we tested the mutant constructs by functional complementation of the alkylation sensitive phenotype of the *E. coli* strain BK2118 (*tag alkA*) upon exposure to the alkylating agent MMS. No rescue was observed with plasmids expressing mutants W109A, D113N and R148A, while constructs expressing Y27A, F180A and W187A complemented the MMS sensitivity of BK2118 in the same way as the wild-type construct (Figure 3b). Finally, cells transformed with mutant constructs W145A and F179A had a growth comparable to cells expressing wild-type AlkD (data not shown). These survival studies support our biochemical analysis and computational predictions that Trp109, Asp113 and Arg148 are active site residues.

Molecular modelling of AlkD in complex with DNA proposes a partially open recognition pocket with specificity for electron deficient alkylated bases

Based on our AlkD model and the experimental structures of several DNA glycosylases in complex with DNA, an approximate model of AlkD in complex with DNA incorporating an extrahelical purine was constructed. The model (Figure 5 and Supplementary Video 2) suggests that the DNA backbone interacts with several strongly conserved Lys and Arg residues (Supplementary Figure 4 and Table 1) along the edges of the wide AlkD groove. The lesion-containing strand is partially buried in the AlkD groove with contacts primarily on the 3′ side of the lesion. The flipped-out base fits nicely into the partially open recognition pocket with Trp109 forming the base of a platform onto which the electron deficient alkylated bases may stack. The side-chain of Trp187 is positioned in the vicinity

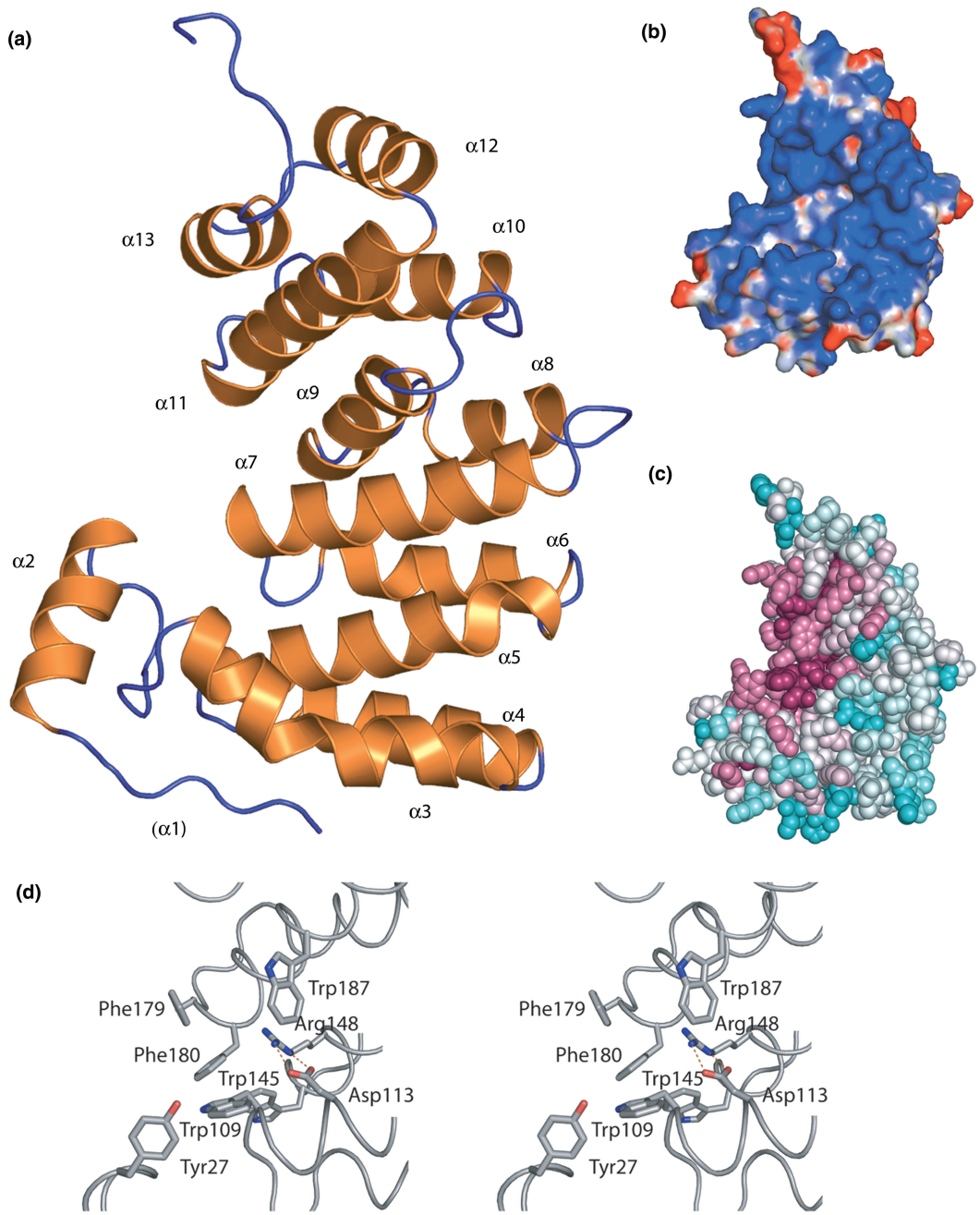

Figure 1. Structural model with proposed active site and lesion recognition pocket of *B. cereus* AlkD. The model contains residues 11 through to 226 and is lacking α-helix α1 and the 11 C-terminal residues (predicted to be disordered). (**a**) Cartoon rendering of the protein which comprises 13 α-helices contributing to the six repeats in Figure 2. (**b**) APBS (28) calculated electrostatic potential mapped onto the protein surface (red = negative, white = neutral and blue = positive) showing the 20–25 Å wide, positively charged, putative DNA binding groove. (**c**) Amino acid residue conservation in 43 AlkD homologs mapped onto the space filling representation of the model generated with ConSurf (31). The scale extends from magenta (highly conserved), through white to cyan (highly variable). There is a nest of conserved residues in the putative DNA binding groove, and several conserved basic amino acid residues (Arg and Lys) are sited along the upper and lower edge of the groove. (**d**) Stereo view of a close-up of the highly conserved nest shows the eight conserved residues that were mutated by site-directed mutagenesis. The catalytic activity and MMS sensitivity of the resulting mutants were determined (Figure 3). The eight conserved residues have identical geometry in the experimental structure 2B6C. The orientation of the protein in space is identical in all panels. The model is also available as Supplementary Video 1 online.

of the ribose ring of the flipped-out base. Finally, the rough model shows that Asp113 is positioned within a reasonable distance to facilitate hydrolytic cleavage of the *N*-glycosylic bond in the extrahelical nucleotide.

DISCUSSION

B. cereus AlkD and AlkC are 3mA DNA glycosylases specific for excision of *N*-alkylated DNA bases and without significant sequence similarity to any other

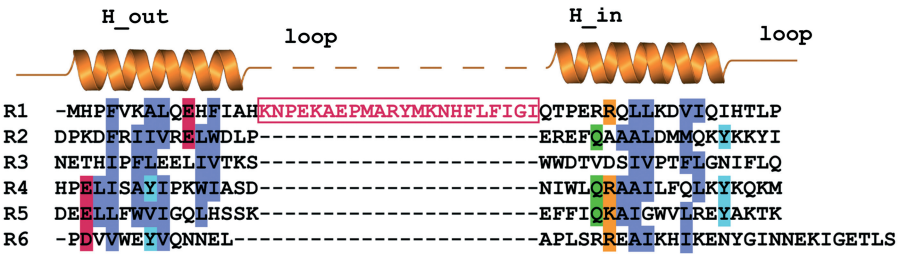

Figure 2. *B. cereus* AlkD has six HEAT-like tandem repeats. Structure-based multiple alignment of repeats R1 (α1 and α3), R2 (α4 and α5), R3 (α6 and α7), R4 (α8 and α9), R5 (α10 and α11) and R6 (α12 and α13) of AlkD with conserved residues shown on a coloured background; hydrophobic residues blue, basic residues orange, acidic residues magenta and polar Tyr and Gln residues light blue and green, respectively. Each α-helix (H_in) of the putative DNA binding surface of AlkD is linked to the two rear side α-helices (H_out), the one preceding it (same repeat) and the other succeeding it (next repeat), through short loops. There is an additional inserted α-helix, α2 (boxed, magenta), between the first and second α-helices of R1.

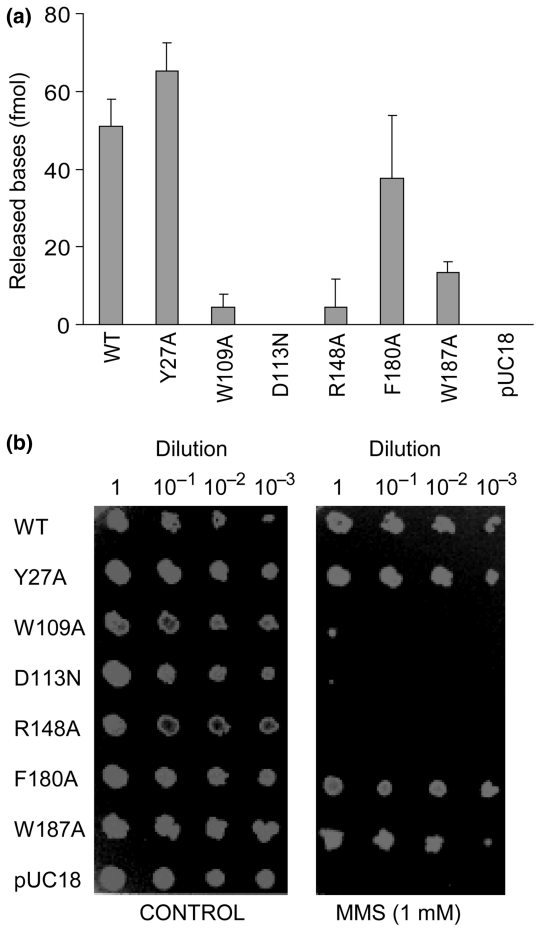

Figure 3. Identification of Trp109, Asp113 and Arg148 as active site residues. (**a**) AlkD active site mutants have strongly reduced 3mA DNA glycosylase activity. Release of alkylated bases was measured in protein extracts from *E. coli tag alkA* cells expressing wild-type AlkD, mutants Y27A, W109A, D113N, R148A, F180A and W187A, and empty pUC18 (negative control). One microgram extract was incubated with [^{3}H]-MNU-treated calf thymus DNA for 30 min, the DNA was ethanol precipitated and the supernatant subjected to scintillation counting. (**b**) The MMS sensitive phenotype of the *E. coli* BK2118 *tag alkA* strain is not complemented by expression of AlkD active site mutants. One microlitre of serially diluted mid-log phase cultures of *E. coli* BK2118 transformed with pUC18-AlkD wild-type, mutants Y27A, W109A, D113N, R148A, F180A and W187A, or empty pUC18 vector, were spotted onto LB agar plates with or without MMS as indicated. The undiluted samples correspond to 3×10^8 cells ml^{-1}.

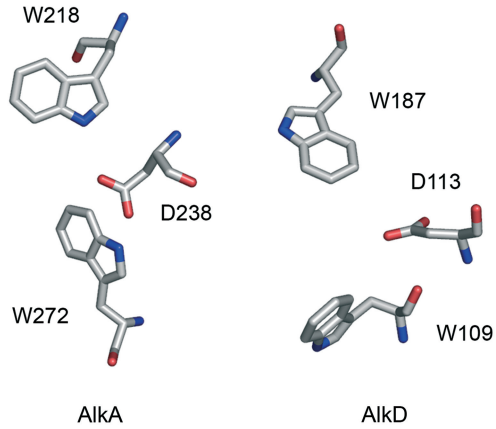

Figure 4. Coordination around the catalytic Asp residue in AlkA and AlkD. View of the Trp218/Asp238/Trp272 triad in the active site of *E. coli* AlkA (34) and their structural counterparts Trp187/Asp113/Trp109 in *B. cereus* AlkD.

protein of known function (9). The present study was initiated in order to determine the structure of proteins belonging to this novel family of DNA glycosylases. Our attempts at crystallizing the proteins for X-ray crystallographic structure determination have so far been unsuccessful. However, the release of the atomic coordinates for the crystal structure of the hypothetical protein EF3068 from *E. faecalis* provided the necessary input to determine the overall structure of *B. cereus* AlkD with a high level of confidence by computational methods. The scoop-shaped overall structure of AlkD consists of an arrangement of α-helices forming a wide groove with a diameter of ~20–25 Å, lined with positively charged residues (Figure 1a and b). The shape and electrostatic environment of the groove seem suitable to accommodate double-stranded DNA. Identification of a single nest of highly conserved residues at one end of the putative DNA binding cavity (Figure 1c and d) strongly suggests that the active site of AlkD is located in this region. Functional analysis of site-specific mutant proteins support the structural model of AlkD, and elucidates residues involved in DNA binding and alkylated base recognition (Trp109) and catalysis (Asp113 and Arg148). The mechanistic and catalytic features are further

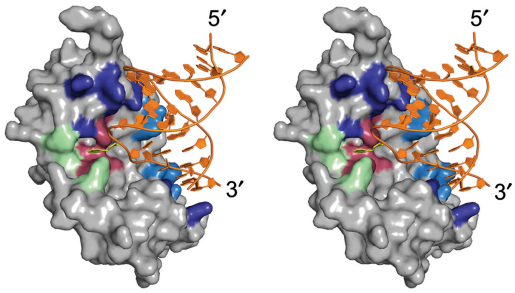

Figure 5. Stereo view of the proposed AlkD–DNA complex. Selected amino acid side chains are coloured as follows: Arg and Lys residues with a high degree of conservation in the AlkD family (dark blue), additional Lys residues lining the DNA binding groove (light blue), residues with no impact on DNA glycosylase activity as shown by alkylbase DNA assay (green), residues with reduced or impaired DNA glycosylase activity when mutated (dark red), docked DNA (orange) and flipped-out DNA base (yellow). The model is also available as Supplementary Video 2 online.

illuminated by a computational model of AlkD complexed with DNA (Figure 5).

Alseth *et al.* (9) found that AlkD was specific for the excision of the *N*-alkylated purine bases 3mA, 3-methylguanine (3mG) and 7-methylguanine (7mG). A common feature of these lesions is that the bases are electron deficient and positively charged. *N*-alkylation of a purine base also significantly destabilizes the *N*-glycosylic bond, making the base susceptible to spontaneous hydrolysis (40). Finally, normal Watson–Crick base pairing and base stacking ability is affected, which may lower the barrier for extrahelical flipping. Unlike the majority of DNA glycosylases, 3mA DNA glycosylases only need a catalytic rate enhancement of a few orders of magnitude (4) due to the destabilized *N*-glycosylic bond. Several 3mA DNA glycosylases have been subjected to extensive structural and biochemical studies, including 3-methyladenine DNA glycosylase I (Tag) (41,42) and AlkA (34,43–45) from *E. coli*, human AAG (40,46,47), as well as the 3mA DNA glycosylase MagIII from *Heliobacter pylori* (48,49). Like all previously studied DNA glycosylases (8), these enzymes employ a DNA-bending and base-flipping strategy that exposes the damaged base in a recognition pocket or cleft. A common feature of all the studied 3mA DNA glycosylases is that they have extensive aromatic networks in the respective substrate pockets (34,42,43,47,48). This may allow for alkylated base specificity due to favourable π–π and/or π–cation interactions between the positively charged base lesion and the aromatic residues lining the catalytic pockets. Alternatively, the activation barrier for glycosylic bond cleavage for alkylated bases may be lowered due to destabilization of the ground-state upon insertion of the charged base lesion into the non-polar active site pocket (4,42). Interestingly, the proposed pocket region in AlkD also contains several highly conserved aromatic residues (Figure 1d)—Trp109, Trp145, Phe179, Phe180 and Trp187. Mutants W145A, F179A and F180A showed catalytic activities comparable to the wild-type enzyme (Figure 3 and data not shown). This observation is in line with our model of AlkD bound to DNA, where these residues are not part of the protein-DNA interface despite the close proximity to the catalytic Asp113 and Arg148 (Figure 5).

Both AlkA and AAG can excise a variety of base lesions, including 3mA, 3mG, 7mG and other alkylated bases as well as uncharged lesions such as cyclic 1,N^6-ethenoadenine (εA) and deaminated adenine (hypoxanthine) (40,44). AlkA has a wide and open recognition pocket that forms few specific contacts with the substrate bases (34,43,45). AAG has a tight pocket that snugly accommodates neutral εA (47). The specificity of AlkA and AAG towards *N*-alkylated bases appears to be due to the weak *N*-glycosylic bonds and a recognition pocket rich in aromatic residues that can stack with electron-deficient bases (40,44), in addition to modified base stacking and Watson–Crick base pairing for these lesions. In contrast, Tag and MagIII are specific for removal of 3-methyl purines from DNA, and are unable to excise 7mG (41,48,49). Steric hindrance excludes 7mG from the active site pocket of MagIII (48). Unlike the other 3mA DNA glycosylases, Tag appears to specifically sense the purine 3-methyl group and undergoes an induced fitting conformational change that makes base excision possible (42). AlkD excises both 3- and 7-methylated purines and thus has a specificity that lies between AlkA/AAG and Tag/MagIII. It appears unlikely that AlkD is able to specifically recognize methyl groups both at N3 and N7 while excluding all other normal or damaged bases from the recognition pocket employing a similar strategy for base discrimination as does Tag for purines methylated at N3 only. More likely AlkD recognizes its targets in a fashion more similar to AlkA and AAG, with electron rich aromatic residues interacting with electron deficient *N*-alkylated bases. Mutational analysis of aromatic residues in AlkD indicates that Trp109 is important for recognition of the alkylated base (Figure 3). Trp272 of AlkA forms stacking interactions with the electron deficient alkylated bases (34,43) and we propose that Trp109 has a similar role in AlkD. The exact shape of the lesion recognition pocket and the details of the base orientation can not be determined from our approximate docking-model of the AlkD–DNA complex (Figure 5). However, it appears that AlkD primarily recognizes *N*-alkylated purines through interactions in an open catalytic pocket that is rich in aromatic residues.

Most structures of monofunctional DNA glycosylases suggest a catalytic mechanism in which an Asp or Glu residue deprotonates and activates a water molecule. Following an S_N2-like substitution mechanism this leads to a simultaneous base release and attack by the activated nucleophile on the C1′ of the sugar group. This appears to be the mechanism, for example, of mammalian AAG (46,47). However, the structure of *E. coli* AlkA in complex with abasic DNA appears to better fit an alternative S_N1-like mechanism, in which the catalytic Asp238 assists base removal by direct stabilization of a oxacarbenium-like transition state or intermediate (34). There are two strong hydrogen bonds between Arg148 and the putative catalytic Asp113 in our model of *B. cereus* AlkD (Figure 1d) which also appear between the corresponding

residues in the experimental structure of *B. faecalis* EF3068. The estimated pK_a for Asp113 in wild-type AlkD is 1.0, while it is 4.0 in the R148A mutant, approximately the same as for free Asp, indicating that Arg148 regulates the catalytic performance of Asp113. The interaction between Arg148 and the carboxylate group of Asp113 stabilizes the charged form of both residues, thus the Asp residue cannot gain protons easily. In order to activate a water molecule for nucleophilic attack (S_N2 mechanism), the salt bridge between Asp113 and Arg148 probably have to be abrogated upon substrate binding. AlkD would in this case be catalytically activated only upon binding DNA. Interestingly, the catalytic site general base of AAG, Glu125, is located next to a conserved Arg182 in the AAG:εA-DNA structure (47) with both residues forming hydrogen bonds to the water molecule that is activated for S_N2-like nucleophilic attack. Alternatively, the reduced pK_a of AlkD Asp113 can ensure that the residue is not protonated and thus stabilizes the transition state in a fashion similar to the AlkA S_N1-like cleavage mechanism.

In AlkA and AAG the side chains of Leu125 and Tyr162, respectively, protrude from the protein surfaces and intercalate into the DNA at the position of the flipped-out nucleotide (34,47), thereby reducing the possibility of re-entry of the extrahelical base into the DNA duplex. A candidate residue for a similar role in AlkD could be Tyr27 (Figure 1d). However, mutant Y27A shows catalytic activity comparable to the wild-type AlkD (Figure 3). Indeed, the current model of the AlkD–DNA complex does not have any clear candidate for an intercalating residue. This may in part explain why AlkD can only excise positively charged alkylated bases that have weak *N*-glycosylic bonds.

The present study demonstrates that *B. cereus* AlkD and AlkC are single domain proteins belonging to a superfamily of proteins with a right-handed alpha–alpha superhelix fold related in structure to ARM/HEAT-repeat containing proteins (Figures 1a and 2). These proteins are built from imperfect tandem repeats of a sequence motif containing ~40 amino acid residues (38). The canonical ARM repeat (50) has three α-helices, while the HEAT repeat (51) has only two antiparallel α-helices which form a hairpin. The repeated units of two (HEAT) or three (ARM) helices are packed together to form a superhelix of helices. The number of repeats may be greater than 20, but can also be below 6, and it has been estimated that 1 in 500 eukaryotic proteins contains these repeats (38). ARM repeats are found, for example, in mammalian β-catenin and importin α, while the HEAT motif is found in—and takes its name from—proteins such as **h**untingtin, **e**longation factor 3, a protein phosphatase 2**A** subunit and the lipid kinase **T**OR (target of rapamycin), in addition to a diverse family of other proteins. These proteins are involved in many different cellular processes. A common theme appears to be interactions with other proteins (38), but HEAT-repeat containing enzymes have recently been reported (52). While AlkD and AlkC appears to be more similar to the HEAT than to the ARM family, they do not appear to be closely related to any of the eukaryotic HEAT subfamilies described by Andrade *et al.* (38).

In conclusion, AlkD and AlkC are DNA glycosylases built from HEAT-like repeats. To the authors' knowledge, no other DNA glycosylases, nor indeed any BER proteins, have been shown to belong to this structural class. Proteins of both the UDG and AAG superfamilies have a core built around a β-sheet, while the H2TH DNA glycosylases have two protein domains that both contain important β-sheet regions. Some of the HhH DNA glycosylases are all-alpha proteins (e.g. *E. coli* Nth), but they invariably consist of at least two domains and have no superhelix of α-helices forming the overall solenoid fold of AlkD and AlkC. Consequently, AlkD and AlkC are members of a new, fifth, structural superfamily of DNA glycosylases.

ACKNOWLEDGEMENTS

This work was supported by the Research Council of Norway, the Norwegian Cancer Society and the Consortium for Advanced Microbial Sciences and Technologies (FUGE-CAMST). We are grateful to the scientists of the Midwest Center for Structural Genomics for making their solved structures available to the community, and thereby making studies like the present work possible. Funding to pay the Open Access publication charge was provided by the Research Council of Norway.

REFERENCES

1. Seeberg,E., Eide,L. and Bjørås,M. (1995) The base excision repair pathway. *Trends Biochem. Sci.*, **20**, 391–397.
2. Krokan,H.E., Nilsen,H., Skorpen,F., Otterlei,M. and Slupphaug,G. (2000) Base excision repair of DNA in mammalian cells. *FEBS Lett.*, **476**, 73–77.
3. Barnes,D.E. and Lindahl,T. (2004) Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu. Rev. Genet.*, **38**, 445–476.
4. Stivers,J.T. and Jiang,Y.L. (2003) A mechanistic perspective on the chemistry of DNA repair glycosylases. *Chem. Rev.*, **103**, 2729–2759.
5. Dodson,M.L., Michaels,M.L. and Lloyd,R.S. (1994) Unified catalytic mechanism for DNA glycosylases. *J. Biol. Chem.*, **269**, 32709–32712.
6. Wood,R.D., Mitchell,M. and Lindahl,T. (2005) Human DNA repair genes, 2005. *Mutat. Res.*, **577**, 275–283.
7. Fromme,J.C., Banerjee,A. and Verdine,G.L. (2004) DNA glycosylase recognition and catalysis. *Curr. Opin. Struct. Biol.*, **14**, 43–49.
8. Huffman,J.L., Sundheim,O. and Tainer,J.A. (2005) DNA base damage recognition and removal: new twists and grooves. *Mutat. Res.*, **577**, 55–76.
9. Alseth,I., Rognes,T., Lindbäck,T., Solberg,I., Robertsen,K., Kristiansen,K.I., Mainieri,D., Lillehagen,L., Kolstø,A.-B. and Bjørås,M. (2006) A new protein superfamily includes two novel 3-methyladenine DNA glycosylases from *Bacillus cereus*, AlkC and AlkD. *Mol. Microbiol.*, **59**, 1602–1609.
10. Sæbø,P.E., Andersen,S.M., Myrseth,J., Laerdahl,J.K. and Rognes,T. (2005) PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology. *Nucleic Acids Res.*, **33**, W535–539.
11. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
12. Poirot,O., O'Toole,E. and Notredame,C. (2003) Tcoffee@igs: a web server for computing, evaluating and combining multiple sequence alignments. *Nucleic Acids Res.*, **31**, 3503–3506.

13. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
14. Bryson,K., McGuffin,L.J., Marsden,R.L., Ward,J.J., Sodhi,J.S. and Jones,D.T. (2005) Protein structure prediction servers at University College London. *Nucleic Acids Res.*, **33**, W36–38.
15. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
16. Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.
17. Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.
18. Pollastri,G., Przybylski,D., Rost,B. and Baldi,P. (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47**, 228–235.
19. Ward,J.J., Sodhi,J.S., McGuffin,L.J., Buxton,B.F. and Jones,D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
20. Obradovic,Z., Peng,K., Vucetic,S., Radivojac,P. and Dunker,A.K. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, **61**, 176–182.
21. McGuffin,L.J. and Jones,D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, **19**, 874–881.
22. Kelley,L.A., MacCallum,R.M. and Sternberg,M.J.E. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499–520.
23. Karplus,K., Karchin,R., Draper,J., Casper,J., Mandel-Gutfreund,Y., Diekhans,M. and Hughey,R. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins*, **53**, 491–496.
24. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
25. Schwede,T., Kopp,J., Guex,N. and Peitsch,M.C. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.*, **31**, 3381–3385.
26. Hooft,R.W.W., Vriend,G., Sander,C. and Abola,E.E. (1996) Errors in protein structures. *Nature*, **381**, 272.
27. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–229.
28. Baker,N.A., Sept,D., Joseph,S., Holst,M.J. and McCammon,J.A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA*, **98**, 10037–10041.
29. Dolinsky,T.J., Nielsen,J.E., McCammon,J.A. and Baker,N.A. (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.*, **32**, W665–667.
30. Li,H., Robertson,A.D. and Jensen,J.H. (2005) Very fast empirical prediction and rationalization of protein pK_a values. *Proteins*, **61**, 704–721.
31. Landau,M., Mayrose,I., Rosenberg,Y., Glaser,F., Martz,E., Pupko,T. and Ben-Tal,N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**, W299–302.
32. Mayrose,I., Graur,D., Ben-Tal,N. and Pupko,T. (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, **21**, 1781–1791.
33. Delano,W.L. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos, CA, USA (http://www.pymol.org).
34. Hollis,T., Ichikawa,Y. and Ellenberger,T. (2000) DNA bending and a flip-out mechanism for base excision by the helix-hairpin-helix DNA glycosylase, *Escherichia coli* AlkA. *EMBO J.*, **19**, 758–766.
35. Brunger,A.T., Adams,P.D., Clore,G.M., DeLano,W.L., Gros,P., Grosse-Kunstleve,R.W., Jiang,J.-S., Kuszewski,J., Nilges,M. *et al.* (1998) Crystallography & NMR System: a new software suite for macromolecular structure determination. *Acta Cryst. D*, **54**, 905–921.
36. Clarke,N.D., Kvaal,M. and Seeberg,E. (1984) Cloning of *Escherichia coli* genes encoding 3-methyladenine DNA glycosylases I and II. *Mol. Gen. Genet.*, **197**, 368–372.
37. Bjelland,S. and Seeberg,E. (1987) Purification and characterization of 3-methyladenine DNA glycosylase I from *Escherichia coli*. *Nucleic Acids Res.*, **15**, 2787–2801.
38. Andrade,M.A., Petosa,C., Donoghue,S.I., Muller,C.W. and Bork,P. (2001) Comparison of ARM and HEAT protein repeats. *J. Mol. Biol.*, **309**, 1–18.
39. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
40. O'Brien,P.J. and Ellenberger,T. (2004) Dissecting the broad substrate specificity of human 3-methyladenine-DNA glycosylase. *J. Biol. Chem.*, **279**, 9750–9757.
41. Bjelland,S. and Seeberg,E. (1996) Different efficiencies of the Tag and AlkA DNA glycosylases from *Escherichia coli* in the removal of 3-methyladenine from single-stranded DNA. *FEBS Lett.*, **397**, 127–129.
42. Cao,C., Kwon,K., Jiang,Y.L., Drohat,A.C. and Stivers,J.T. (2003) Solution structure and base perturbation studies reveal a novel mode of alkylated base recognition by 3-methyladenine DNA glycosylase I. *J. Biol. Chem.*, **278**, 48012–48020.
43. Labahn,J., Schaerer,O.D., Long,A., Ezaz-Nikpay,K., Verdine,G.L. and Ellenberger,T.E. (1996) Structural basis for the excision repair of alkylation-damaged DNA. *Cell*, **86**, 321–329.
44. O'Brien,P.J. and Ellenberger,T. (2004) The *Escherichia coli* 3-methyladenine DNA glycosylase AlkA has a remarkably versatile active site. *J. Biol. Chem.*, **279**, 26876–26884.
45. Yamagata,Y., Kato,M., Odawara,K., Tokuno,Y., Nakashima,Y., Matsushima,N., Yasumura,K., Tomita,K.-I., Ihara,K. *et al.* (1996) Three-dimensional structure of a DNA repair enzyme, 3-methyladenine DNA glycosylase II, from Escherichia coli. *Cell*, **86**, 311–319.
46. Lau,A.Y., Schärer,O.D., Samson,L., Verdine,G.L. and Ellenberger,T. (1998) Crystal structure of a human alkylbase-DNA repair enzyme complexed to DNA: mechanisms for nucleotide flipping and base excision. *Cell*, **95**, 249–258.
47. Lau,A.Y., Wyatt,M.D., Glassner,B.J., Samson,L.D. and Ellenberger,T. (2000) Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, AAG. *Proc. Natl Acad. Sci. USA*, **97**, 13573–13578.
48. Eichman,B.F., O'Rourke,E.J., Radicella,J.P. and Ellenberger,T. (2003) Crystal structures of 3-methyladenine DNA glycosylase MagIII and the recognition of alkylated bases. *EMBO J.*, **22**, 4898–4909.
49. O'Rourke,E.J., Chevalier,C., Boiteux,S., Labigne,A., Ielpi,L. and Radicella,J.P. (2000) A novel 3-methyladenine DNA glycosylase from *Helicobacter pylori* defines a new class within the endonuclease III family of base excision repair glycosylases. *J. Biol. Chem.*, **275**, 20077–20083.
50. Huber,A.H., Nelson,W.J. and Weis,W.I. (1997) Three-dimensional structure of the Armadillo repeat region of β-catenin. *Cell*, **90**, 871–882.
51. Andrade,M.A. and Bork,P. (1995) HEAT repeats in the Huntington's disease protein. *Nat. Genet.*, **11**, 115–116.
52. Park,J.-H., Aravind,L., Wolff,E.C., Kaevel,J., Kim,Y.S. and Park,M.H. (2006) Molecular cloning, expression, and structural prediction of deoxyhypusine hydroxylase: A HEAT-repeat-containing metalloenzyme. *Proc. Natl Acad. Sci. USA*, **103**, 51–56.