Uracil in DNA – occurrence, consequences and repair

Hans E Krokan*[,1], Finn Drabløs[1] and Geir Slupphaug[1]

[1]*Institute of Cancer Research and Molecular Biology, Norwegian University of Science and Technology, N-7489 Trondheim, Norway*

Uracil in DNA results from deamination of cytosine, resulting in mutagenic U : G mispairs, and misincorporation of dUMP, which gives a less harmful U : A pair. At least four different human DNA glycosylases may remove uracil and thus generate an abasic site, which is itself cytotoxic and potentially mutagenic. These enzymes are UNG, SMUG1, TDG and MBD4. The base excision repair process is completed either by a short patch- or long patch pathway, which largely use different proteins. UNG2 is a major nuclear uracil-DNA glycosylase central in removal of misincorporated dUMP in replication foci, but recent evidence also indicates an important role in repair of U : G mispairs and possibly U in single-stranded DNA. SMUG1 has broader specificity than UNG2 and may serve as a relatively efficient backup for UNG in repair of U : G mismatches and single-stranded DNA. TDG and MBD4 may have specialized roles in the repair of U and T in mismatches in CpG contexts. Recently, a role for UNG2, together with activation induced deaminase (AID) which generates uracil, has been demonstrated in immunoglobulin diversification. Studies are now underway to examine whether mice deficient in Ung develop lymphoproliferative malignancies and have a different life span.
Oncogene (2002) **21,** 8935–8948. doi:10.1038/sj.onc. 1205996

Keywords: uracil; uracil-DNA glycosylase; DNA repair

Introduction

The most common mutation in human cells is C→T transitions and this type of mutation is also found very frequently in human tumours. A major fraction of these transitions occurs in a CpG context and they have therefore been suggested to result from deamination of 5-methylcytosine (5-meC) to thymine, or cytosine to uracil. DNA replication would then insert A in the place of G (reviewed by Waters and Swann, 2000). However, C→T transitions are common in other sequences as well. Mismatches caused by deamination (T : G or U : G) are generally repaired by the base excision repair (BER) pathway. The BER pathway probably repairs the largest fraction of DNA damage, but its significance to human health remains unclear.

The present review will focus on the repair of one of the most common lesions in DNA, uracil. However, since several of the known uracil-DNA glycosylases (UDGs) are not strictly uracil-specific, the repair of some uracil analogues (Figure 1), namely 5-hydroxy-methyluracil (5-hmU), 3,N^4-ethenocytosine (εC) and 5-fluorouracil (5-FU) will also be discussed. In addition, repair of T : G mismatches by DNA glycosylases will be addressed. It has become clear in recent years that BER for removal of uracil from DNA involves several DNA glycosylases, which may have different functional roles. The steps downstream of the initial glycosylase step may also differ, depending on origin and localisation of uracil, the specific DNA glycosylase involved, and timing in the cell cycle. In this review the emphasis will be on the enzymes that initiate the repair process, the DNA glycosylases.

Uracil in DNA results from deamination of cytosine to uracil, creating a premutagenic U : G mispair, or from misincorporation of dUMP instead of dTMP during replication, creating a U : A pair (reviewed in Krokan *et al.*, 1997). The inappropriate uracil is recognised by a UDG activity, which cleaves the N-glycosidic bond and leaves an abasic site in the DNA. The first UDG was discovered in *Escherichia coli* in 1974 in a search for activities that would repair uracil in DNA. This also represented the discovery of the BER pathway. The activity that was discovered released uracil as a free base, leaving an intact abasic site in DNA (Lindahl, 1974). Ung from the *ung*-gene later proved to be a representative of a highly conserved family of UDGs present in most living organisms examined (Olsen *et al.*, 1989). However, mammalian cells contain at least three additional human glycosylases (TDG, SMUG1 and MBD4), that have the capacity to remove uracil from DNA (reviewed in Krokan *et al.*, 2001). UNG and SMUG1 prefer single-stranded DNA as substrate, but also remove U from double-stranded DNA. In contrast, TDG and MBD4 are strictly specific for double-stranded DNA, and have very low turnover numbers (Table 1). The functional roles of the different enzymes are only starting to become understood. It seems clear that UNG has a central function in removal of U from misincorporated dUMP-residues (Nilsen *et al.*, 2000; Otterlei *et al.*, 1999). SMUG1 has tentatively been proposed to have a specific role in removal of deaminated cytosine residues (Nilsen *et al.*, 2001), but this role needs more experimental confirmation. The
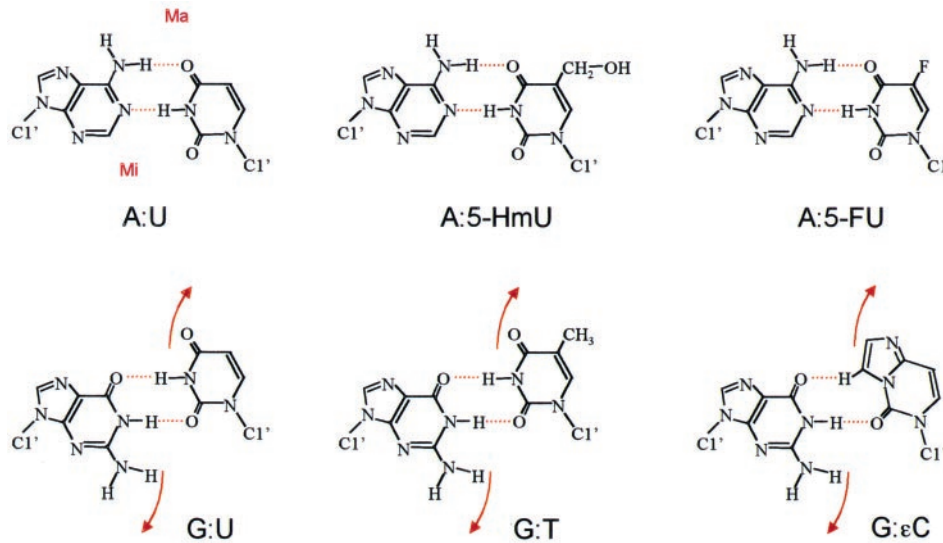
Figure 1 Pairs and mispairs of uracil or uracil analogs with guanine or adenine. Ma = major grove; Mi = minor groove. Red arrows indicate how bases in mispairs have moved relative to normal base pairs

Table 1 Substrates of uracil-releasing enzymes

Substrate*	UNG	SMUG1	TDG	MBD4
T : G	no	no	yes	yes
U in ssDNA	yes	yes	no	no
U in ds DNA	yes	yes	yes	yes
5hmeU in ssDNA	no	yes	no	no
5hmeU in dsDNA	no	yes	yes	no?
εC in ssDNA	no	yes	no	no?
εC in dsDNA	no	yes	yes	yes (weak)
5-FU in ssDNA	yes	yes	yes!	no
5-FU in dsDNA	yes	yes	yes	yes

*No quantification has been attempted because the efficiency depends on assay conditions. In mispairs, all the enzymes are generally more active when G is present in the opposite of the lesion as compared with other mismatches

roles of TDG (Hardeland et al., 2001) and MBD4 (Bellacosa, 2001) may be limited to repair of mismatched uracil, thymine and some damaged pyrimidines in double-stranded DNA, particularly in CpG and 5-methylCpG contexts. In addition, MBD4 apparently cross-talks with the mismatch repair (MMR) system through an interaction with MMR protein MLH1 (reviewed in Bellacosa, 2001). In the present paper, we focus on central and recent developments in the field. For more comprehensive overviews, the reader should consult previous reviews (Friedberg et al., 1995).

Generation and consequences of uracil in DNA

The chemical instability of cytosine in DNA (Shapiro, 1980) results in slow hydrolytic deamination that generates premutagenic G : U mispairs. These may cause G : C to A : T transitions unless repaired prior to the next round of replication, because DNA polymerases efficiently incorporate A opposite U in

the template. Subsequent removal of U and insertion of T generate the A : T transition. The number of cytosine deaminations has been calculated to be in the order of 60–500 per genome per day. The uncertainty depends on the average fraction of DNA present in single stranded form, since deamination is 200–300-fold faster from single stranded DNA than from double stranded DNA (reviewed in Lindahl, 1993; Krokan et al., 1997). Cytosine in cyclobutane dimers deaminates some 7–8 orders of magnitude faster than other cytosines (Krokan et al., 1997) and has been implicated in mutagenesis both in bacteria (Horsfall et al., 1997) and in human cells (Tu et al., 1998). Whether uracil in dimers is subject to repair is apparently not known. Enzymatic deamination of cytosine by (cytosine-5)-methyltransferase (MT) may happen as an aberrant process when the cellular content of S-adenosylmethionine (SAM) is low, or when MT is very high, as observed in several tumours (Krokan et al., 1997). The significance of this type of enzymatic deamination of cytosine remains unclear. However, activation-induced deaminase (AID) was recently suggested to directly target C : G base pairs in DNA. AID was initially described as a B-cell specific protein proposed to be involved in RNA editing (Muramatsu et al., 1999) and also potentiates changes in immunoglobulin gene DNA (Muramatsu et al., 2000). Interestingly, expression of AID in E. coli gave a mutator phenotype that yielded nucleotide transitions at dC : dG in a context-dependent manner, and the frequency of these mutations was enhanced in cells lacking Ung (Petersen-Mahrt et al., 2002). It was thus proposed that AID-triggered deamination of dC to dU and subsequent removal of U by UNG was involved in the first phase of hypermutation driving immunoglobulin gene diversification (Petersen-Mahrt et al., 2002). Furthermore, inhibition of chicken UNG by expression of Ugi in DT40 B-cells alters the pathway of

immunoglobulin hypermutation from transversion dominance (due to replication over sites of base loss) to transition dominance, as expected if the deaminated C is copied in replication (di Noia and Neuberger, 2002). How such a mechanism is regulated and targeted specifically to immunoglobulin genes in mammalian cells remains to be clarified. The proposed function of AID and UNG in immunoglobulin gene diversification (Petersen-Mahrt *et al.*, 2002) makes studies on possible B-cell abnormalities particularly relevant in an Ung-deficient mouse model. Such abnormalities would probably occur late in life since Ung-deficient mice do not show obvious signs of phenotypic changes in the first 12 months (Nilsen *et al.*, 2000). Studies on the possible occurrence of lymphomas in this model are now underway. The mechanism is still unclear, but one possibility would be that the delayed processing of AID-induced U:G mispairs results in abnormal recombinations due to strand breaks induced by DNA glycosylases other than Ung. It was recently shown that the CEM15 gene, which is similar to the AID gene, is inhibited by the HIV-1 Vif (virion infectivity factor) protein. The activity of CEM15 is still unknown (Sheehy *et al.*, 2002).

The other main source of uracil in DNA is by incorporation of dUMP instead of dTMP during DNA replication. dUTP is a normally occurring intermediate in nucleotide metabolism, but the level is kept very low due to an efficient dUTPase which prevents incorporation of dUMP. Although incorporation of dUMP into DNA was demonstrated both in bacteria (Tye *et al.*, 1977; Wist *et al.*, 1978) and isolated mammalian nuclei 25 years ago (Wist *et al.*, 1978), the extent and significance remain unclear. However, *Ung*-deficient mice have several-fold increased content of uracil in their genome even though the capacity to remove incorporated dUMP is not completely abolished (Nilsen *et al.*, 2000). In addition, inhibition of nuclear UNG2 by antibodies reduced the ability of the nuclei to remove misincorporated dUMP (Otterlei *et al.*, 1999). Together, these results indicate that dUMP is misincorporated to a significant extent and that UNG-proteins have a central role in removal of uracil in a U:A context. The full physiological significance of misincorporated dUMP remains unclear.

Human uracil-DNA glycosylase (UNG) encoded by the *UNG* gene

Mitochondrial UNG1 and nuclear UNG2 are major uracil-DNA glycosylases in human cells (Slupphaug *et al.*, 1995). UNG2 has a specific role in removal of misincorporated uracil in an immediate post-replicative process, (Otterlei *et al.*, 1999; Nilsen *et al.*, 2000). However, it may also have an important role in removal of deaminated cytosine residues (Kavli *et al.*, 2002).

A human uracil-DNA glycosylase activity was reported in 1976 (Sekiguchi *et al.*, 1976), and shown to have a function in removal of incorporated dUMP

from nuclear DNA (Wist *et al.*, 1978). cDNA for human UNG was cloned (Olsen *et al.*, 1989) based upon information on the amino acid sequence of the N-terminal end of human placental UNG (Wittwer *et al.*, 1989). The sequence encoded by the cDNA demonstrated that the major human UNG was highly conserved in herpes viruses, bacteria, yeast and man (Olsen *et al.*, 1989). The human *UNG* gene spans approximately 13.8 kb and is located at 12q24.1 (Krokan *et al.*, 2001). The gene encodes both the mitochondrial form UNG1 and nuclear form UNG2, which have a common catalytic domain, but different N-terminal sequences (Nilsen *et al.*, 1997). Promoters P_B and P_A direct synthesis of mRNA for UNG1 and UNG2, respectively and they are differentially regulated, with UNG2 mRNA as well as enzyme activity peaking in the S phase (Haug *et al.*, 1998). Exon 1A encodes 44 amino acids unique to UNG2 and is spliced into codon 35 of exon 1B. All of exon1B is used to form mRNA for the unique N-terminal end of UNG1, which therefore contains 35 N-terminal amino acids not found in UNG2. Amino acids downstream of position 44 in UNG2 and 35 in UNG1 are identical in the two forms. The N-terminal end of UNG1 forms a classical and very strong mitochondrial targeting signal (MTS), most of which is removed by mitochondrial processing peptidase upon entry in the mitochondria (Bharati *et al.*, 1998). Complete nuclear targeting requires approximately 100 of the UNG2 N-terminal amino acids, although the unique N-terminal 44 amino acids contain the most important residues in the nuclear localization signal (NLS). When fused to the N-terminal end of UNG2, the MTS overrides the NLS and directs mitochondrial targeting (Otterlei *et al.*, 1998). The N-terminal 90 amino acids of UNG2 are readily lost due to proteolysis by cellular proteases during purification, but the compact and fully active catalytic domain is even resistant to Proteinase K treatment. The N-terminal part of UNG2 therefore probably forms an independent structural domain, or may be unstructured. UNG2 is located in the nucleoplasm and replication foci. The mechanism for entry of UNG2 into replication foci is not known, but it may be co-transported with proliferating cell nuclear antigen (PCNA), similar to the targeting of DNA ligase I (Rossi *et al.*, 1999) and DNA pol *ι* (Haracska *et al.*, 2001). The N-terminal part contains a classical motif (Q_4xxLxxFF$_{11}$) for binding of PCNA. In addition, the N-terminal part binds trimeric replication factor A (RPA) through a motif that overlaps with the PCNA-binding motif, and binds the 34 kD monomer of RPA by a second motif in residues 73–84. These interactions take place in replication foci (Otterlei *et al.*, 1999). The interactions of UNG2 with PCNA and RPA, both putatively required for long patch BER as well as replication, and the presence of several other replication factors used in long patch repair, suggest that UNG2 is engaged in long patch repair in replication foci. However, UNG2 is also present in

the nucleoplasm and may participate in repair of U:G mispairs resulting from cytosine deamination (Kavli *et al*., 2002). The specific role of UNG2 in removal of misincorporated uracil was demonstrated by the inhibition of immediate post-replicative removal of incorporated uracil in isolated nuclei by neutralizing anti-UNG antibodies (Otterlei *et al*., 1999), and by the slow removal of incorporated uracil in nuclei from $Ung^{-/-}$ mice (Nilsen *et al*., 2000). Incorporated uracil was eventually removed, however, indicating the presence in nuclei of a less efficient back-up system for removal of uracil (Nilsen *et al*., 2000). UNG2 is ideally suited for removal of misincorporated uracil close to the rapidly moving replication fork because it has a turnover number (600–1000 per min) which is orders of magnitude higher than that of other uracil-DNA glycosylases. In $Ung^{-/-}$ cells the steady state level of uracil in genomic DNA is approximately 2000 per cell, while it was below detection level, and at least 10-fold lower, in wild-type cells (Nilsen *et al*., 2000). Surprisingly, the spontaneous mutation level in $Ung^{-/-}$ mice was increased only 1.3-fold (thymus) and 1.4-fold (spleen) as compared with wild type. This is likely due to the presence of back-up activities, among which SMUG1 may have a dominant role (Nilsen *et al*., 2000, 2001). While these results are seemingly at odds with previous results in bacteria and yeast (reviewed in Krokan *et al*., 1997), a direct comparison between rapidly growing microorganisms and a mixture of cells in complex multicellular organisms may not be entirely informative. It is possible that mutation rates in a homogenous fraction of rapidly growing cells would be different. In fact, inhibition of UNG2 in human glioma cells by expression of the protein inhibitor Ugi resulted in a threefold increase in spontaneous mutation rates in a shuttle vector, which is comparable to the average mutation rates in microorganisms (Radany *et al*., 2000).

Several lines of evidence indicate that human UNG may also play a role in viral infection. Several DNA viruses such as the poxviruses and the herpesviruses encode their own UDG, whereas RNA viruses such as retroviruses do not. It could, however, be hypothesized that the latter utilizes cellular UNG to avoid misincorporation of uracil during synthesis of the replicative DNA intermediate. Interestingly, the HIV-1 proteins Vpr and integrase IN have both been shown to bind UNG (Bouhamdan *et al*., 1996; Willetts *et al*., 1999). Furthermore, human UNG was efficiently incorporated into wild-type HIV-1 virions. IN-defective viruses fail to incorporate UNG, indicating that the integrase is required for packaging of UNG into virions (Willetts *et al*., 1999). These findings indicate that the presence of UNG is important near the site of nascent viral DNA synthesis. Alternatively, the association of viral proteins with UNG might be important for their nuclear translocation. Further studies of the association between UNG and viral proteins may provide important insight into the biology of HIV-1 and other viruses.

The catalytic domain of UNG proteins

Our understanding of the individual molecular steps involved in enzymatic removal of uracil by the Ung proteins has largely evolved from analysis of the human, *E. coli* and the Herpes simplex virus type-1 UDGs. Crystal structures of HSV-1 (Savva *et al*., 1995), human (Slupphaug *et al*., 1996) and *E. coli* (Xiao *et al*., 1999) enzyme-ligand complexes provided decisive information on both the substrate recognition and the catalytic mechanism. The three enzymes are structurally essentially identical, and substrate binding occurs in a highly conserved pocket providing shape and electrostatic complementarity to uracil and which is too narrow to accommodate purines. Numbering of amino acids in human UNG below is with reference to UNG1 (Mol *et al*., 1995). UNG2 is nine amino acids longer at the unique N-terminal end (reviewed in Krokan *et al*., 2001). Selectivity against other pyrimidines is provided by human UNG Y147 (Y90 in HSV-1 and Y66 in *E. coli*), which exclude bulky 5-substitutions (such as thymine), and N204 (N147 in HSV-1 and N123 in *E. coli*) which specifically hydrogen bonds to N3 and O4 of uracil. Correct orientation of the latter amide group is fixed by a cluster of water molecules at the base of the uracil-binding pocket (Pearl, 2000). The contribution of both N204 as well as Y147 to specificity was also verified experimentally, as replacement of Asn204 with Asp in human UNG shifted the specificity of the mutant towards cytosine, while replacement of Y147 by the smaller Ala, Cys or Ser shifted the specificity towards thymine (Kavli *et al*., 1996).

An important conclusion from the crystal structures was that uracil within the helical context of DNA could not be accommodated within the buried uracil-binding pocket. Instead, uracil binding was likely mediated by 'base flipping'. A conserved leucine (human L272), positioned directly above the uracil-binding pocket, was suggested as a candidate to assist the local melting of the DNA helix (Mol *et al*., 1995). Subsequently, base flipping by human UNG was directly visualized by cocrystallization of hUNG with a uracil-containing DNA duplex (Slupphaug *et al*., 1996). In this complex uracil, deoxyribose and 5′-phosphate were rotated 180° relative to their starting position in DNA, and the cleaved uracil was found in the specificity pocket. We thus suggested that the term 'nucleotide flipping' provided a better description of the process. Furthermore, compression of the backbone flanking uracil was for the first time implicated in catalysis, assisted by extensive conformational changes in the enzyme upon formation of the productive complex. This established a 'push–pull' hypothesis in which the L272 side chain penetrates into the minor groove expelling uracil (push), and complementary interactions from the specificity pocket facilitate productive binding (pull). The mechanistics of this was further refined by high resolution co-crystal structures of wild type and mutant UNG bound to U:A and U:G oligonucleotides (Parikh *et al*., 1998).

In the L272A mutant structure the uracil had dissociated, and the enzyme rebound to the product, an extrahelically positioned AP-site. This implied that the extrahelical conformation could be achieved even in the absence of the insertion of a hydrophobic side chain (push). A mechanism was proposed in which serine-proline rich loops compress the phosphates flanking uracil by a 'pinching' mechanism and thereby stabilize the extrahelical conformation (Parikh et al., 1998) without contributing significant energy to the cleavage reaction itself (Werner et al., 2000). Such a 'pinch–push–pull' mechanism was also hypothesised for E. coli Ung (Stivers et al., 1999). More recent data, however, indicate that the function of the inserting leucine side chain may be more complex than merely pushing the uracil out of the double helix. When analysing kinetic parameters of E. coli Ung L191A against single-stranded substrates, a 15-fold reduction in k_{cat}/K_m was observed compared to the wild type (Handa et al., 2001; Jiang and Stivers, 2001), which would not be expected in the absence of stabilizing base pairing. A likely explanation for this could be that the leucine side chain rather functions as a 'doorstop' to prevent return of the flipped out uracil residue, as recently suggested by Wong et al. (2002). Using stopped-flow experiments of E. coli Ung and monitoring fluorescence of 2-aminopurine adjacent to uracil, they demonstrated that Leu191 insertion occurred after nucleotide-flipping but before excision, thus indicating a 'pinch–pull–push' rather than a 'pinch–push–pull' mechanism.

When considering the energy contribution of each discrete event above to the overall catalytic reaction, one should bear in mind that DNA is a very heterogeneous substrate. This is also reflected by the different efficiency whereby uracil is excized from different sequence contexts (Eftedal et al., 1993; Nilsen et al., 1995). Recently the sequence specificity was re-examined using both single-stranded and duplex DNA substrates (Bellamy and Baldwin, 2001), and the authors conclude that the observed variations were not due to stability of the uracil itself within the DNA structure. Rather, local structure perturbations could affect uracil recognition, e.g. by affecting the phosphate recognition by the serine pinch.

Uracil binding induces considerable conformational changes in UNG, bringing key residues in optimal distances to favour catalysis (Slupphaug et al., 1996). This is accompanied by large conformational strain induced upon the deoxyuridine (Parikh et al., 2000b; Werner and Stivers, 2000). U is both pulled relative to the sugar and rotated $\sim 90°$ around the N-glycosidic bond. This increases the orbital overlap and favours electron flow from O4′ of the deoxyribose to O2 of uracil. The developing negative charge at O2 is enzymatically stabilized by a neutral histidine (E. coli H187, human H268) (Drohat et al., 1999; Shroyer et al., 1999). Hydrolysis is completed by a weakly nucleophilic water molecule positioned directly below the deoxyribose C1′ that becomes the 1′-α-OH, and protonation of the uracil enolate anion N1 to give the

more stable amide (Parikh et al., 2000b; Werner and Stivers, 2000). Moreover, recent quantum- and molecular-mechanical calculations indicate that negative phosphate charges in the substrate itself may repel the anionic leaving group, and thus make a major contribution to the catalytic rate (Dinner et al., 2001). The authors suggest that such substrate autocatalysis may emerge as a general feature of DNA glycosylases.

The observation that UNG had an higher affinity for the product AP-site than the actual substrate itself (Parikh et al., 1998), led to the assumption that subsequent to product release, UNG quickly rebound to the AP site to protect the cell until the AP site was conveyed to the next enzyme in the BER-pathway. Such rebinding has subsequently been observed for several DNA glycosylases (Vidal et al., 2001; Waters et al., 1999). Recent studies have, however, demonstrated that the human nuclear UNG2 does not bind AP-sites (Kavli et al., 2002), and the in vivo significance of AP-site shielding thus remains elusive, at least for UNG2.

Perhaps the least understood stage in the processing of uracil-DNA is how the glycosylases recognize these subtle lesions within vast stretches of DNA. This is further complicated by the fact that eukaryotic DNA is organized in complex nucleoprotein structures. In vitro, the UNG-proteins appear to function in both a processive and distributive fashion, depending on the salt concentration (Bennett et al., 1995; Slupphaug et al., 1995). When a uracil residue is encountered, the mechanism of initial recognition is not obvious. U : A base pairs only differ from T : A by the lack of a 5-methyl in the minor groove. U : G mispairs are somewhat more evident, as they exist in the wobble conformation with the uracil shifted towards the major groove leaving partially hydrophobic gaps within the base stack. Both U : A and U : G base pairs might conceivably be detected in the minor groove by the Leu272 loop (human) reading head. However, neither of these structural determinants appear necessary for base flipping to occur, since the UNG mutants TDG and CDG are able to recognize and cleave perfectly normal bases from the DNA (Kavli et al., 1996). Thus, the enzyme might instead flip every DNA base to probe against the specificity pocket. How the energetic cost of such a scanning mechanism is covered merits further investigation, however.

SMUG1 (Single strand-selective Monofunctional Uracil-DNA Glycosylase)

SMUG1 removes uracil, as well as 5-hydroxymethyluracil (5-hmeU), from single- and double stranded DNA and is proposed to have an important role in removal of uracil resulting from cytosine deamination (Nilsen et al., 2001). It also removes εC, although relatively inefficiently (Kavli et al., 2002). SMUG1 is not thought to have a role in removal of incorporated uracil and it does not accumulate in replication foci. SMUG1 was identified in extracts of Xenopus (xSMUG1) using a proteomics approach. This procedure comprised in

vitro expression from a library of cDNA, and electrophoretic mobility shift upon binding of damage-recognising protein to DNA containing modified nucleotides designed to target the active site of the glycosylases. The human counterpart was identified from EST databases (Haushalter *et al.*, 1999). Interestingly, the gene for human SMUG1, like TDG and UNG-proteins, is located in the long arm of chromosome 12, more specifically in position 12q13.1–14, centromeric of *UNG* and *TDG* which are located in positions 12q24.1 and 12q22–24.1, respectively (reviewed in Krokan *et al.*, 2001). Thus, genes for three out of four uracil-removing activities are located on chromosome 12. These genes are thought to have evolved from a common ancestral gene by duplications and divergence and although the amino acid sequence conservation is <10%, sequence analysis suggests that the enzymes have retained a common fold (Pearl, 2000). The phylogeny of the uracil-DNA glycosylase genes will be discussed in more detail below.

The selectivity of SMUG1 for single-stranded DNA is more pronounced for the *Xenopus* enzyme than for the human enzyme, which in fact removes uracil relatively efficiently both from U:G and U:A contexts (Haushalter *et al.*, 1999; Kavli *et al.*, 2002). Thus, the term single-strand selective is not entirely appropriate for the human enzyme. The primary sequence of the SMUG1 proteins is very different from the UNG-proteins. Furthermore, the xSMUG1 activity was not inhibited by the peptide inhibitor Ugi that efficiently inhibits both prokaryotic and eukaryotic uracil-DNA glycosylases belonging to the *Ung*-family. Although hSMUG1 is weakly inhibited by Ugi (Kavli *et al.*, 2002, data not shown), the DNA-binding region in SMUG1 is probably substantially different from the Ung-family of proteins. Human SMUG1 (hSMUG1) is a protein of predicted size 270 amino acids (30 kDa) and is 60% identical and 71% similar to the *Xenopus* enzyme. xSMUG1 acted upon U:G and U:A essentially without preference for either, but showed no enzymatic activity against T:G, in contrast to the T(U)-mismatch glycosylase TDG (Hardeland *et al.*, 2001). Recently, it was found that SMUG1 is probably the major enzyme for removal of 5-hmeU (Boorstein *et al.*, 1987) from DNA (Boorstein *et al.*, 2001). 5-hmeU is a result of oxidation of the 5-methyl group of thymine in DNA due to ionizing radiation and other forms of oxidative stress (Faure *et al.*, 1998). However, it may also be formed in a two-step reaction; first the methyl-group of 5-meC in CpG-contexts is oxidised to 5-hmeC, and subsequently this residue is deaminated to yield 5-hmeU (Cannon Carlson *et al.*, 1989). The latter pathway might result in G:C to A:T transitions, but based upon information on rate constants for deamination and oxidation, the amount of 5-hmeU generated spontaneously by this scheme would be extremely low (Rusmintratip and Sowers, 2000). In *Ung*$^{-/-}$ mice mSMUG1 is apparently the major uracil-DNA glycosylase activity, both in various organs and in mouse embryonal fibroblasts from knock-out mice (Nilsen *et al.*, 2001). xSMUG1 has a low Km for U:G (0.035 μM), which would make it suitable for recognising U:G mismatches resulting from cytosine-deamination (Haushalter *et al.*, 1999). Using a double-stranded DNA containing a single U:G mismatch at low substrate concentration to mimic the *in vivo* situation, mSMUG1 was found to be the major UDG activity in extracts of cells from *Ung*$^{-/-}$ mice. Even in extracts from wild-type mice, mSMUG1 contributed a substantial fraction of the total UDG-activity under these assay conditions. In addition, mSMUG1 activity is stimulated several-fold by APE1. These results would be compatible with mSMUG1 as the major enzyme responsible for removal of U from U:G mismatches (Nilsen *et al.*, 2001). However, recent data from our laboratory on hSMUG1 as compared with hUNG2 are at variance with these results (Kavli *et al.*, 2002). Thus, the Km values for hSMUG1 on double-stranded DNA containing U:G mismatches is *not* lower than the Km for hUNG2 tested under different conditions, including different concentrations of Mg^{2+}. Mg^{2+}, present in relatively high concentrations intracellularly, substantially enhances both hSMUG1 and hUNG2 activities. In the presence of 7.5 mM Mg^{2+}, the Km values for hUNG2 and hSMUG1 against U:G are 0.4$\pm$0.1 μM and 1.3$\pm$0.2 μM, respectively. In fact, under all conditions tested with U:A, U:G and ssU, the Km for hUNG2 was either lower (most frequently), or similar to the Km values for hSMUG1. Furthermore, hUNG2 was also by far the most efficient enzyme in extracts from human tumor cells (CX-1, from a colorectal carcinoma) in removal of U from U:G mismatches, even when using the same oligonucleotide substrate described in the mSMUG1-study by Nilsen *et al.* (2001). Thus, at least in human cells both hUNG2 and hSMUG1 are candidates for removal of U from U:G mismatches, and if anything, hUNG2 is the major candidate with hSMUG1 as a back-up enzyme (Kavli *et al.*, 2002). The situation may be different in mouse. hSMUG1 is a nuclear enzyme, but it is not cell cycle-regulated and in contrast to UNG2, it does not accumulate in replication foci (Kavli *et al.*, 2002). This would agree with a function of hSMUG1 in removal of U from U:G mismatches independent of DNA replication. The efficiency of hSMUG1 in removal of 5hmU (k$_{cat}$/K$_m$) was 5hmU in ssDNA > 5hmeU:G > 5hmU:A (Kavli *et al.*, 2002).

In conclusion, hSMUG1 is a non-abundant enzyme present in the nucleoplasm. It efficiently removes U and 5hmUra from ssDNA and double-stranded DNA. It also inefficiently removes εC. The major function may be in removal of 5hmU and deaminated cytosines, although it may be less important than UNG2 in the latter process, at least in human cells.

Other activities for removal of 5-hmU from DNA

In addition to SMUG1, one well defined enzyme, TDG which has a strict requirement for dsDNA, removes 5-hmeU, but less efficiently than U and T (Rusmintratip

and Sowers, 2000). In addition, an activity that removes 5-hmeU and that is different from SMUG1 and TDG has recently been partially purified. This enzyme has a strict requirement for double-stranded substrates (Baker et al., 2002). Thus, there are apparently at least three different human enzymatic activities for removal of 5-hmeU from DNA. While 5-hmeU-removing activity is relatively abundant in extracts from human nuclei, especially activity acting on 5-hmeU:G mismatches (Rusmintratip and Sowers, 2000), 5-hmeU-removing activity is apparently absent in bacteria and yeast (Boorstein et al., 1987). The 60-fold increased activity towards 5-hmeU:G over 5-hmeU:A may seem surprising, since by far the major route of formation of 5-hmeU is thought to be by oxidation of the methyl-group in thymine (Rusmintratip and Sowers, 2000). However, 5-hmeU:G mispairs may cause G:C to A:T transition mutations, while 5-hmeU:G may not, since 5-hmeU is not miscoding (Levy and Teebor, 1991), does not block polymerases (Vilpo and Vilpo, 1995), and does not perturb DNA structure (Mellac et al., 1993; Pasternack et al., 1996). These results may indicate that 5-hmeU:G mispairs may be generated more frequently than assumed, although decisive experimental evidence supporting this view is still missing.

TDG – a T(U) mismatch glycosylase that may have multiple functions

TDG removes T or U from mismatches in double-stranded DNA, with U:G being preferred over T:G. TDG also efficiently removes εC in mismatches, and removes 5hmU with lower efficiency. It is thought to have an important role in correcting mismatches resulting from deamination of C or 5-meC in CpG contexts, which are major sources of G→A transition mutations in man (Brown and Jiricny, 1987; Hardeland et al., 2001). However, the major physiological role of TDG remains elusive. Interestingly, TDG may also function as a transcription factor (Hardeland et al., 2001).

Identification of TDG and biochemical properties

TDG was discovered during a search for mammalian activities correcting T:G mismatches. After transfection of SV40 DNA containing this mismatch, C was found to be preferentially inserted in place of T in monkey cells (Brown and Jiricny, 1987). Later a protein of apparent size 55 kDa necessary for this repair was identified, and it was demonstrated to be a G:T mismatch glycosylase (TDG). Cloning of the cDNA revealed that the correct size of TDG was 46 kDa. The enzyme can also remove thymine from C:T and T:T mispairs. The order of efficiency is G:T>>C:T>T:T. It does not cleave the DNA backbone, and thus contains no lyase activity (Neddermann et al., 1996; Neddermann and Jiricny, 1993). However, TDG removes U mispaired to G with some

10-fold higher k_{cat} than T from G:T mispairs (reviewed in Hardeland et al., 2001). The cyclic alkylation product εC is also a good substrate both for the bacterial homologue MUG and human TDG and removal of this lesion may be a major function of TDG (Barrett et al., 1999; Hang et al., 1998; Hardeland et al., 2001; Saparbaev and Laval, 1998). Surprisingly, TDG is able to process 5-fluorouracil (5-FU) in DNA with a higher efficiency than any other substrate that has been tested. The enzyme even removes 5-FU from single stranded DNA (Hardeland et al., 2001). This is unexpected since TDG is generally very double-strand-specific, and since the substrate recognition, as deduced from data on the homologous bacterial protein MUG, involves interactions with both strands (Barrett et al., 1999). Although different assay systems may influence the relative activities of TDG towards different substrates, 5-FU and εC appear to be the best substrates (Hardeland et al., 2001; Saparbaev and Laval, 1998), but in contrast to 5-FU, εC in single stranded oligonucleotides is not a substrate for TDG (Saparbaev and Laval, 1998). εC may result from exposure to environmental pollutants such as vinylchloride, but may also be caused by lipid peroxidation (reviewed in Bartsch and Nair, 2000). Since εC has lost the specific coding potential it is mutagenic, as well as cytotoxic, and causes both transitions and transversions (Basu et al., 1993; Moriya et al., 1994). It has been reported that TDG may also remove 5-meC from CpG contexts (Zhu et al., 2000b), but this has not been verified in a separate study (Hardeland et al., 2001).

In the cases described above, the substrate is a base that has been modified. However, TDG is also able to remove T from T:O^{6}meG mismatches arising from miscoding O^{6}meG residues in the template (reviewed in Hardeland et al., 2001). Repair of O^{6}meG in mispairs by O^{6}methylG-DNA methyltransferase would result in a G:T mispair that could either be repaired by TDG or the mismatch repair system. O^{6}meG in the template may direct incorporation of either C or T. Thus, TDG may remove T and give the BER system a second chance to incorporate C opposite of the miscoding O^{6}meG-residue.

Unlike UNG-proteins, isolated TDG has an extremely low catalytic turnover number. The rate-limiting step in the mechanism in vitro is the release of the product (e.g. G:AP-site) due to rigid hydrogen bonding interactions between the enzyme and the Watson-Crick face of the guanine opposite the AP-site (Hardeland et al., 2000). This shields the cytotoxic AP-site until the next enzyme in the pathway, AP-endonuclease APE1 (also called HAP1) enters. No evidence for direct interaction between TDG and APE1 has been found. Therefore, APE1 most likely displaces TDG because it has even higher affinity for the AP-site than TDG, and in addition APE1 is more abundant. In agreement with this, APE1 strongly stimulates the activity of TDG (Waters et al., 1999). Recent findings indicate, however, that the AP-site binding capacity of TDG might be subject to

regulation. Hardeland et al. (2002) demonstrated that a substantial fraction of cellular TDG is covalently modified by the ubiquitin-like proteins SUMO-1 and SUMO-2/3. Furthermore, in vitro sumoylation of TDG dramatically reduced its DNA-substrate and AP-site affinity. This increased the enzymatic turnover of G:U substrates, with a concomitant loss of G:T-processing activity. The authors propose that TDG binds its substrate in the unmodified state, and that subsequent to catalysis, sumoylation allows detachment from the product AP-site. Interestingly, a sumoylation consensus site is also present in MBD4 (MED1), but the functional implications of this remain to be established.

Structural aspects

UNG and TDG have <10% identical amino acid residues, and would therefore not immediately be considered homologous proteins. However, they appear to have very similar structures. The structure of the bacterial homologue MUG has been solved (Barrett et al., 1998, 1999; Pearl, 2000). MUG is closely related (37% identity) to human TDG, and their structures must be similar. Crystal structures of MUG and UNG reveal a striking homology, containing similarly positioned α-helices on the surface and a centrally located β-sheet. Both enzymes are traversed by a DNA-binding groove connecting to a uracil-binding pocket, which in the case of MUG is less tailor-made for uracil. Two active site motifs are conserved. In human UNG1 these are 143-GQDPY-147 and 268-HPSPLS-273. These are identical in human, E. coli, yeast and herpes viral proteins. The corresponding motifs in MUG are 16-GINPGL-20, which is identical to the corresponding motif in TDG, and 140-NPSGLS-145, which is related to the corresponding motif in human TDG (MPSSSS). While aspartate in the first motif of UNG has a function in activating water for attack on the C1′ of deoxyribose, asparagine in MUG/TDG can not activate the nucleophilic water, although it may bind and position it. The two conserved motifs in UNG and MUG/TDG have very similar conformations. The active sites of UNG and MUG/TDG are thus related, but specific differences also explain the broader substrate specificity of MUG/TDG. In both cases, the base can only be accommodated in the catalytic pocket after the nucleotide is flipped out of the helix. In UNG, binding of T is sterically hindered by the bulky side chain of Tyr147, which would clash with the 5′-methyl group of T. In the corresponding position in MUG, the amino acid is Gly20 which does not represent a major barrier to thymine. The preference for U:G over T:G is probably explained by position of the side chain in Ser23, the OH-group of which would clash with the 5′ methyl-group of thymine. However, the Ser side chain is free to rotate. Thus, unlike UNG where the Tyr side chain is fixed, there is no rigid structure in MUG preventing binding of pyrimidines with substitutions in the 5′ position. In mammalian TDG, Ala is in the

position corresponding to Ser23 in MUG. The smaller side chain of Ala represents an even smaller barrier to binding of T. This probably explains why MUG has a stronger preference for U:G over T:G than TDG (Barrett et al., 1999). Interestingly, substitution of Tyr147 in human UNG with Ala, Ser or Cys results in enzymes that release T and U, but not C, from DNA. Although the catalytic turnover is reduced compared with wild type UNG, it is still higher than for TDG (Kavli et al., 1996).

The discrimination between cytosine and thymine/uracil is less obvious. Possibly, the hydrophobic nature of the catalytic pocket, the specificity for G in the complementary strand, and a weak push that can only flip out unstable base pairs may explain the selectivity for U/T:G over the more stable C:G. The hydrophobic εC would be even better accommodated in the pocket, and it is weakly bonded to the complementary G, explaining why it is a good substrate for TDG (Barrett et al., 1999).

Like several other DNA glycosylases, TDG has N-terminal and C-terminal extensions not present in the bacterial homologues. These extensions may be involved in subcellular sorting and protein–protein interactions, although this has not yet been demonstrated. Yeast two-hybrid screening surprisingly revealed interaction between TDG and the retinoid acid receptor (RAR), as well as retinoid X receptor (RXR). RAR and RXR are intracellular receptors that after binding of ligand function as transcription factors. In fact, in vitro studies have demonstrated that TDG stimulated RAR/RXR-dependent transactivation (Um et al., 1998). Furthermore, TDG has been shown to be a strong repressor of thyroid transcription factor-1 (TTF-1) transcriptional activity (Missero et al., 2001), indicating that TDG may function both as a positive and negative regulator of transcription. Whether the transcription-associated activity of TDG is modulated by sumoylation is presently not known. This, however merits further investigation, since several transcription factors appear to be directly or indirectly regulated by sumoylation (reviewed by Muller et al., 2001).

In conclusion, TDG removes 5-FU, εC, U, T and 5-hmeU in this order from double stranded DNA, preferentially when mismatched to G. In addition, it may function as a transcription factor. Which of these functions is the most important remains unclear.

MBD4 (MED1) – A protein containing a methyl binding domain and a separate glycosylase domain

MBD4 (also called MED1) has central properties in common with TDG, it binds to T:G or U:G mismatches preferentially in CpG contexts and releases the pyrimidine base with low efficiency. Like TDG, it also removes 5-FU and, weakly, εC (reviewed in Bellacosa, 2001). TDG and MBD4 therefore apparently have very similar substrate specificities. Possibly MBD4 through its specific methyl-binding domain may

have a more important role in methylated CpG contexts than TDG. MBD4 may also have an additional role in mismatch repair through its interaction with MLH1 (reviewed in Bellacosa, 2001). Furthermore, there is significant evidence associating defects in the MBD4 gene with human cancer (reviewed in Hardeland *et al.*, 2001) while similar evidence is lacking for TDG. The gene for MBD4 has been mapped to position 3q21.22 (Bellacosa, 2001). It is thus the only one of the four uracil-removing proteins that is not located to human chromosome 12 and it is apparently unrelated to the other uracil-DNA glycosylases.

Human MBD4 was discovered as one member of a family of proteins that bind specifically to methylated DNA *in vitro* (Hendrich and Bird, 1998). MBD4 was also found to interact with the human mismatch repair protein MLH1 in a yeast two-hybrid system. It was also found to have a domain with homology to glycosylases/lyases and displayed an apparent endonuclease activity (Bellacosa *et al.*, 1999). However, MBD4 was later shown to be a monofunctional G:T and U:G mismatch glycosylase (Hendrich *et al.*, 1999). The 580 amino acid protein contains an N-terminal methyl-binding domain, MBD (residues 82 – 147) and a C-terminal glycosylase domain (residues 401 – 580) with homology to *E. coli* thymine glycol glycosylase (*Endo*III) and 8-oxoG:A specific adenine glycosylase (MutY). A CpG sequence context was preferred, but not absolutely required since mismatches in other sequences were processed at lower rates (Hendrich *et al.*, 1999). Furthermore, the methylation status of CpG·TpG mismatches or CpG·UpG mismatches did not affect the glycosylase activity of MBD4 under certain reaction conditions. Nevertheless, the enzyme binds more strongly to 5-meCpG·TpG mismatches than to unmethylated mismatches or 5-meCpG·meCpG matches. It was therefore proposed that the function of MBD4 is to counteract the mutagenic effects of deamination of 5-meC to thymine. In this function, the MBD domain may be responsible for the binding specificity and the glycosylase domain for catalysis (Hendrich and Bird, 2000).

Surprisingly, MBD4 has also been reported to have 5-meC DNA glycosylase activity, especially in hemimethylated DNA. It might therefore contribute to active changing of the methylation status in cells (Zhu *et al.*, 2000a). This, as well as the mismatch repair function, are potential functions that could be relevant to the possible role of MBD4 in cancer prevention (reviewed by Bellacosa, 2001). Furthermore, MBD4 mutations were reported in 14 cases of human cancer with microsatellite instability, but not in tumour without microsatellite instability. All of the mutations were deletions of A (one or two) in a polyA track, resulting in frame shift and generation of a stop codon. While six of the cancers also carried a mutation in *MLH1*, with which MBD4 interacts, five of the tumours did not carry *MLH1* mutations and in three the *MLH1/MSH2* status was not known (Riccio *et al.*,

1999). These findings suggest that MBD4 may be important in prevention of cancer.

A fifth family of uracil DNA-glycosylases

Recently, a new uracil-DNA glycosylase was identified in the crenarchaeon *Pyrobaculum aerophilum* by using a search algorithm targeted to the catalytic motifs I and III (Aravind and Koonin, 2000) of the UDG superfamily (Sartori *et al.*, 2002). This UDG, called *Pa*-UDGb, surprisingly contained no polar residue in the catalytic motif I (GLAPA) which is proposed to activate or orient water in the other UDGs. The enzyme furthermore had a unusual broad substrate specificity, including U, 5-HmU, 5-FU and εC. In addition, *Pa*-UDGb was able to remove hypoxanthine, and thus is the first member of the UDG superfamily able to remove both pyrimidines and purines (Sartori *et al.*, 2002). The authors thus proposed that *Pa*-UDGb represented a fifth UDG family, that evolved in organisms living at elevated temperatures to counteract the mutagenic threat of both cytosine and adenine deamination.

Evolution of uracil-DNA glycosylases

The classification of DNA glycosylases into superfamilies, can be based on characteristic sequence motifs as defined in the Pfam database (Bateman *et al.*, 2002). The two largest superfamilies are uracil DNA glycosylases (UDG) and helix – hairpin – helix – GPD glycosylases (HhH – GPD). The HhH – GPD superfamily is named after its hallmark motif (Thayer *et al.*, 1995), a helix – hairpin – helix (HhH) and Gly/Pro rich loop (GP) followed by a conserved aspartate (D). UNG, SMUG1 and TDG belong to the UDG superfamily, which is the most important in terms of uracil processing. MBD4 belongs to the HhH – GPD family, members of which remove a large variety of lesions, including uracil, oxidised bases and certain mismatches, particularly A mismatched to G or 8-oxoG.

The phylogenetic distribution of DNA repair genes has been discussed in detail by Eisen and Hanawalt (1999), while Aravind and Koonin (2000) have specifically analysed the UDG superfamily. The HhH – GPD glycosylases are widespread in both Archaea, bacteria and eukaryotes, and are believed to represent a very ancient gene family. However, the UDG superfamily shows a more non-even distribution pattern. TDG is the most widespread UDG gene family, as it is found both in Archaea, bacteria and eukaryotes. It is distantly related to the AUDG gene family, which is mainly found in Archaea and bacteria, and it seems likely that they share an ancient ancestor. Neither the UNG nor the SMUG1 gene families are found in Archaea. SMUG1 has so far been found only in some eukaryotes. Based on the conservation pattern in the minor-groove intercalation loop it was suggested

that SMUG1 may have evolved from an UNG-like enzyme by rapid divergence, possibly to meet special requirements for repair in multicellular animals (Aravind and Koonin, 2000). However, a more traditional phylogenetic analysis tends to place SMUG1 closer to TDG/MUG and AUDG, and more distant from UNG (Sartori *et al*., 2001). These studies are based on relatively few sequences with low similarity. This makes alignment and analysis difficult, and the interpretation of these data should be done with caution. The origin of SMUG1 therefore remains an open question.

UNG is very widespread in bacteria, and has also been identified in most eukaryotes. It has been suggested that UNG was introduced into eukaryotes by horizontal gene transfer (Aravind and Koonin, 2000), possibly from the mitochondrial genome (Eisen and Hanawalt, 1999). UNG is also the only of these gene families that is found in a large number of viruses, indicating another possible mechanism for horizontal gene transfer. Furthermore, there are more, and more recently established pseudogenes from UNG than from other DNA glycosylases. A Blast search (Altschul *et al*., 1990) of the Ensembl human genome sequence (Hubbard *et al*., 2002) identifies at least four certain UNG pseudogenes, compared to 0–2 pseudogenes for other human DNA glycosylase genes (F Drabløs, unpublished).

Sequence comparison of the different gene families within the UDG superfamily identifies several conserved sequence motifs, indicating a common 3D-fold for all UDG-type proteins. This is confirmed by the known 3D structures for UNG and TDG/MUG, as they share a common 3D-fold classified as an alpha beta 3-layer (aba) sandwich in the CATH classification system (Pearl *et al*., 2000). A direct structural super-position using the DALI software (Holm and Sander, 1993) for MUG (1mug) *vs* UNG (3eug) gives a highly significant probability score (Z-score 7.3) with a residual means score (RMS) or the 3D structures of 4.0 Å over 145 residues, despite just 13% sequence identity in the corresponding sequence alignment. It therefore seems realistic to assume that all UDG-type proteins share this common fold.

There are at least five relevant sequence motifs in the UDG sequences that can be related to structural features in UNG (Parikh *et al*., 2000a). The first is the water-activating motif (GQDPYH in human UNG). It is also identified as Motif 1 by Pearl (2000) or Motif I by Aravind and Koonin (2000), and is relatively well conserved in all UDG sequences. This loop coordinates the catalytic-centre water molecule just below the C1′ deoxyribose. The proline-rich motif (PPPPSL) which is directly involved in DNA interaction in UNG is not well-conserved in other UDG sequences. However, when comparing the UNG and MUG structures the main difference is an extra loop outside the actual binding region in UNG. The essential part of the proline-rich motif that is in direct contact with the DNA backbone is structurally conserved in MUG. The uracil-recognition motif

(GVLLLN) is conserved mainly in the UNG proteins. However, the actual region seems to be well conserved in most sequences, the variation is mainly with respect to the specific residues found at each position. The motif contributes to uracil recognition by hydrogen bonding to polar atoms of the uracil ring. In the uracil binding pocket there is also a favourable stacking interaction between uracil and a well-conserved phenylalanine residue found between the water-activating and the proline-rich motifs, in addition to the tyrosine from the water-activating motif mentioned above. The uracil-recognition motif is then followed by the glycine-serine motif (LWGS), which is identified as Motif II by Aravind and Koonin. This motif is also involved in DNA interaction. In particular the glycine is well conserved, probably because a side chain at this position would interfere with the close contact between the protein and the DNA backbone. The final motif is the minor-groove intercalation loop motif (HPSPLS), identified as Motif 2 by Pearl and Motif III by Aravind and Koonin. This motif shows some variation, but the histidine and the first proline are conserved in most sequences, except in SMUG1. The histidine is one of the active site residues, and forms hydrogen bond with uracil.

The crystal structures show that UNG has a highly conserved, single domain α/β topology with a central, four-stranded parallel β-sheet surrounded by eight α-helices, classified as an alpha beta 3-layer (aba) sandwich in CATH. This architecture family is a very large one, with 70 topologies listed in CATH. The central β-sheet consists of parallel strands in the order 2-1-3-4.

The HhH-GPD superfamily, to which MBD4 belongs, has a very different architecture. It consists of two α-helical domains classified as mainly alpha orthogonal bundles by CATH. Some members of the HhH–GPD superfamily, including MBD4, also contain a methyl-CpG binding domain (MBD) which has an alpha beta 2-layer sandwich architecture. It has been shown that MBD4 acts as a G:T and G:U specific thymine and uracil glycosylase (Hendrich *et al*., 1999; Petronzelli *et al*., 2000). The 3D structure for MBD4 is not known. However, structures are available for several other members of this superfamily, and it can be assumed that important structural features are conserved. The structure of human OGG1 in complex with oxoG·C-containing DNA (1ebm) (Bruner *et al*., 2000) and of *Methanobacterium thermoformicicum* MIG (1kea) (Mol *et al*., 2002) can be used as examples. The HhH–GPD-fold consists of a four-helix bundle domain and a six-helix barrel domain, with the active site and the HhH motif located at the interface between these domains. Whereas most DNA-binding proteins seem to use a charged surface rich in lysine and arginine residues to bind backbone phosphates, the DNA binding surface of OGG1 is nearly charge neutral. However, a large number of α-helices have the N-terminal end oriented towards the DNA, maximizing helix-dipole interactions. Only one helix, which is part of the helix–hairpin–helix (HhH) motif,

is actually in contact with DNA. As seen in other DNA glycosylases the DNA residue is fully extruded from the DNA helix and inserted into the active-site pocket of the enzyme.

The functional roles of the different glycosylases and coupling to downstream steps in BER

UNG1 and UNG2, together with SMUG1 are the only known DNA glycosylases with preference for single-stranded DNA. UNG-proteins are highly selective for uracil, but remove 5-fluorouracil and certain oxidised pyrimidines with very low efficiency (Krokan *et al.*, 1997). While U in U:G mispairs is generally preferred over U:A, the preference is dependent on the sequence context (Eftedal *et al.*, 1993). The efficient removal of uracil from single-stranded DNA is puzzling since it leaves a non-informative lesion without the information in a complementary strand. Single-stranded DNA is probably mainly found temporarily in transcribed genes and very close to the replication fork. Abasic sites resulting from uracil-removal in single-stranded DNA at the replication fork could be handled by at least three different mechanisms; (i) regression of the replication fork and repair by short patch or long patch BER, (ii) recombination repair using the old strand at the other side of the fork, (iii) translesion DNA synthesis. Regression of a replication fork stalled at a single-strand lesion is well established in *E. coli* and is an ATP-requiring process. It may in principle apply to all types of lesions that stall the replication fork, including abasic sites (Robu *et al.*, 2001). Recombination using information from the sister chromatid at stalled replication forks (Gruss and Michel, 2001), as well as translesion synthesis across abasic sites are well established processes in bacteria. Interestingly, repair of abasic sites in chromosomal DNA in *E. coli in vivo* has been demonstrated to require both BER, recombination repair and translesion synthesis (Otterlei *et al.*, 2000). Therefore, abasic sites resulting from the action of UNG at the replication fork is possibly unlikely to be dealt with by BER alone.

After removal of uracil by a uracil-DNA glycosylase and cleavage of the resulting abasic site by AP-endonuclease (APE1/HAP1), the BER pathway splits into two branches. For more comprehensive overviews, the reader should consult other recent reviews (Dogliotti *et al.*, 2001). The presumed major track is the short patch pathway. It uses the dRPase activity of DNA polymerase β to cleave 3′ of the abasic site, thus releasing deoxyribose-5-phosphate. Then Pol β inserts C or T, depending on the template base. Finally DNA ligase 3 seals the nick, aided by the scaffold protein XRCC1. The alternative long patch pathway largely uses replication proteins (Dogliotti *et al.*, 2001) and may take place in replication foci (Otterlei *et al.*, 1999). This pathway requires Pol ε and/or δ, as well as the trimeric sliding clamp and polymerase processivity factor proliferating cell nuclear antigen (PCNA) and

the clamp loader replication factor C (RFC). Repair synthesis is stimulated by Pol β, which may be important in the first step of polymerization. The structure specific endonuclease FEN1 removes the 2–8 nucleotide displaced 'flap' of DNA and DNA ligase 1 seals the nick (Dogliotti *et al.*, 2001).

As shown in Figure 2, uracil in DNA may be present in different positions relative to a replication fork, and in addition the sequence context may vary. It seems likely that the type of uracil-DNA glycosylase, as well as the mechanism of repair in the subsequent steps will depend on these factors. Uracil present in the fork prior to replication, e.g. in the G_1-phase, may mainly result from cytosine deamination and may in principle be removed by any of the four identified human uracil-DNA glycosylases. Most likely, TDG and MBD4 mainly function in CpG-contexts, while SMUG1 and UNG2 may operate in any sequence context, albeit with varying efficiency. Thus, UNG2/UNG1 are known to be less active in GC-rich regions as compared with AT-rich regions (Eftedal *et al.*, 1993, 1994; Slupphaug *et al.*, 1995). Yet, even in such sequences, the catalytic efficiencies of UNG2 and SMUG1 are much higher than those of TDG or MBD4, so the latter enzymes presumably have a limited significance in uracil-repair except in special sequence contexts. SMUG1 was recently suggested to be a main enzyme in repair of uracil in U:G mismatches (Nilsen *et al.*, 2001). However, our data strongly suggest that even in repair of U:G mispairs resulting from cytosine deamination, UNG2 may be a major player (Kavli *et al.*, 2002). UNG2 has at least as low K_m as SMUG1, and is present in replication foci as well as in the nucleoplasm. However, UNG2 is essentially excluded from nucleoli, whereas SMUG1 accumulates in nucleoli, suggesting that SMUG1 may have a specialized role in uracil-repair in nucleoli (Kavli *et al.*, 2002). Among the uracil-DNA glycosylases, only UNG2 specifically accumulates in the replication foci during S phase, and as discussed, all experimental evidence suggests that UNG2 has an important role in the removal of misincorporated uracil in replication foci.

One important problem is the following: How are deaminated cytosines that escape repair prior to replication repaired, if at all? If uracil escapes repair and directs incorporation of adenine, subsequent removal of uracil by UNG and replacement by thymine would result in a G:C→A:T transition, and this would be a poor strategy. However, UNG2 efficiently removes uracil from single-stranded DNA and may thus generate an abasic site that blocks replication. The stalled replication fork may recruit proteins required for fork regression and homologous recombination, which are alternative mechanisms to short patch repair and long patch repair in the downstream steps subsequent to uracil removal. Fortunately, APE1 requires double-stranded DNA and thus the danger of creating a double-strand break is small. Involvement of recombination in the repair of abasic sites has been documented for *E. coli* (Otterlei *et*
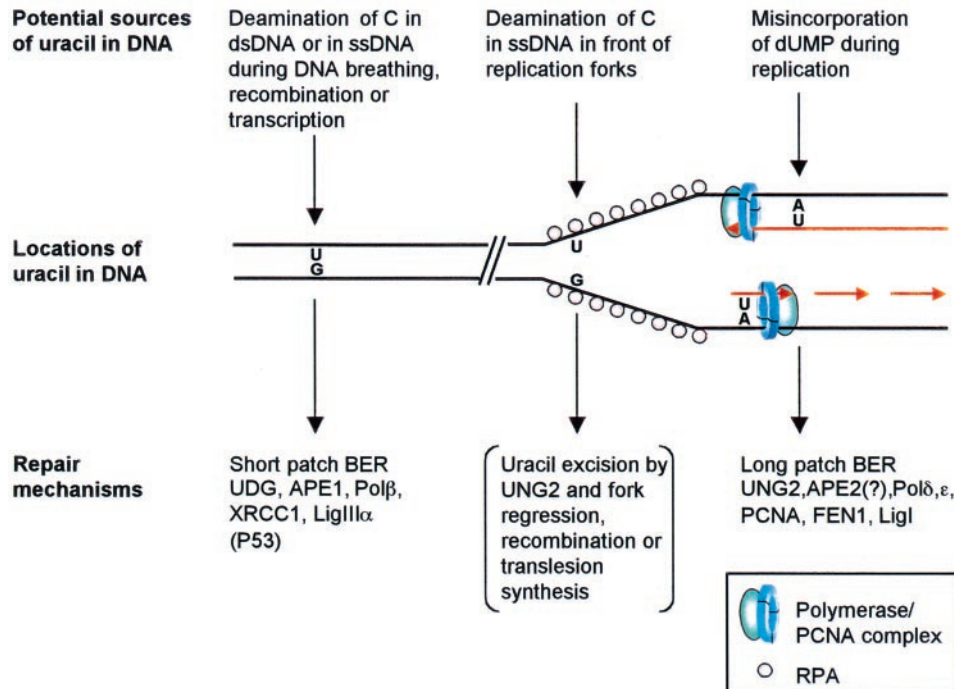
Figure 2 Uracil in different positions relative to the replication fork, and proposed mechanisms of repair in different positions

al., 2000). Furthermore, induction of deamination of cytosine by NO is strongly cytotoxic in *E. coli* cells deficient in recombination (Spek *et al.*, 2001). The finding that recombination factors are required for processing of abasic sites in bacteria suggests that this may also be the case in mammalian cells, since this basic process is highly likely to be conserved. We therefore propose that uracil in single-stranded DNA at the replication fork is incised by UNG2 and repaired by recombination or fork regression, which are both processes requiring recombination proteins.

References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. (1990). *J. Mol. Biol.*, **215,** 403 – 410.
Aravind L and Koonin EV. (2000). *Genome Biol.*, **1,** RESEARCH0007.
Baker D, Liu P, Burdzy A and Sowers LC. (2002). *Chem. Res. Toxicol.*, **15,** 33 – 39.
Barrett TE, Savva R, Panayotou G, Barlow T, Brown T, Jiricny J and Pearl LH. (1998). *Cell,* **92,** 117 – 129.
Barrett TE, Scharer OD, Savva R, Brown T, Jiricny J, Verdine GL and Pearl LH. (1999). *EMBO J.*, **18,** 6599 – 6609.
Bartsch H and Nair J. (2000). *Toxicology,* **153,** 105 – 114.
Basu AK, Wood ML, Niedernhofer LJ, Ramos LA and Essigmann JM. (1993). *Biochemistry,* **32,** 12793 – 12801.
Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M and Sonnhammer ELL. (2002). *Nucleic Acids Res.*, **30,** 276 – 280.
Bellacosa A. (2001). *J. Cell Physiol.*, **187,** 137 – 144.
Bellacosa A, Cicchillitti L, Schepis F, Riccio A, Yeung AT, Matsumoto Y, Golemis EA, Genuardi M and Neri G. (1999). *Proc. Natl. Acad. Sci. USA,* **96,** 3969 – 3974.
Bellamy SR and Baldwin GS. (2001). *Nucleic Acids Res.*, **29,** 3857 – 3863.
Bennett SE, Sanderson RJ and Mosbaugh DW. (1995). *Biochemistry,* **34,** 6109 – 6119.
Bharati S, Krokan HE, Kristiansen L, Otterlei M and Slupphaug G. (1998). *Nucleic Acids Res.*, **26,** 4953 – 4959.
Boorstein RJ, Cummings Jr A, Marenstein DR, Chan MK, Ma Y, Neubert TA, Brown SM and Teebor GW. (2001). *J. Biol. Chem.*, **276,** 41991 – 41997.
Boorstein RJ, Levy DD and Teebor GW. (1987). *Mutat. Res.*, **183,** 257 – 263.
Bouhamdan M, Benichou S, Rey F, Navarro JM, Agostini I, Spire B, Camonis J, Slupphaug G, Vigne R, Benarous R and Sire J. (1996). *J. Virol.*, **70,** 697 – 704.
Brown TC and Jiricny J. (1987). *Cell,* **50,** 945 – 950.
Bruner SD, Norman DPG and Verdine GL. (2000). *Nature,* **403,** 859 – 866.
Cannon Carlson SV, Gokhale H and Teebor GW. (1989). *J. Biol. Chem.*, **264,** 13306 – 13312.
Di Noia J and Neuberger MS. (2002). *Nature,* **419,** 43 – 48.
Dinner AR, Blackburn GM and Karplus M. (2001). *Nature,* **413,** 752 – 755.

Dogliotti E, Fortini P, Pascucci B and Parlanti E. (2001). *Prog. Nucleic Acids Res. Mol. Biol.,* **68,** 3 – 27.

Drohat AC, Xiao G, Tordova M, Jagadeesh J, Pankiewicz KW, Watanabe KA, Gilliland GL and Stivers JT. (1999). *Biochemistry,* **38,** 11876 – 11886.

Eftedal I, Guddal PH, Slupphaug G, Volden G and Krokan HE. (1993). *Nucleic Acids Res.,* **21,** 2095 – 2101.

Eftedal I, Volden G and Krokan HE. (1994). *Ann. NY Acad. Sci.,* **726,** 312 – 314.

Eisen JA and Hanawalt PC. (1999). *Mutat. Res.,* **435,** 171 – 213.

Faure H, Mousseau M, Cadet J, Guimier C, Tripier M, Hida H and Favier A. (1998). *Free Radical Res.,* **28,** 377 – 382.

Friedberg EC, Walker GC and Siede W. (1995). *DNA Repair and Mutagenesis.* Washington, DC: ASM Press. pp. 698.

Gruss A and Michel B. (2001). *Curr. Opin. Microbiol.,* **4,** 595 – 601.

Handa P, Roy S and Varshney U. (2001). *J. Biol. Chem.,* **276,** 17324 – 17331.

Hang B, Medina M, Fraenkel-Conrat H and Singer B. (1998). *Proc. Natl. Acad. Sci. USA,* **95,** 13561 – 13566.

Haracska L, Johnson RE, Unk I, Phillips BB, Hurwitz J, Prakash L and Prakash S. (2001). *Proc. Natl. Acad. Sci. USA,* **98,** 14256 – 14261.

Hardeland U, Bentele M, Jiricny J and Schar P. (2000). *J. Biol. Chem.,* **275,** 33449 – 33456.

Hardeland U, Bentele M, Lettieri T, Steinacher R, Jiricny J and Schar P. (2001). *Prog. Nucleic Acids Res. Mol. Biol.,* **68,** 235 – 253.

Hardeland U, Steinacher R, Jiricny J and Schar P. (2002). *EMBO J.,* **21,** 1456 – 1464.

Haug T, Skorpen F, Aas PA, Malm V, Skjelbred C and Krokan HE. (1998). *Nucleic Acids Res.,* **26,** 1449 – 1457.

Haushalter KA, Todd Stukenberg MW, Kirschner MW and Verdine GL. (1999). *Curr. Biol.,* **9,** 174 – 185.

Hendrich B and Bird A. (1998). *Mol. Cell. Biol.,* **18,** 6538 – 6547.

Hendrich B and Bird A. (2000). *DNA Methylation and Cancer,* **249,** 55 – 74.

Hendrich B, Hardeland U, Ng HH, Jiricny J and Bird A. (1999). *Nature,* **401,** 301 – 304.

Holm L and Sander C. (1993). *J. Mol. Biol.,* **233,** 123 – 138.

Horsfall MJ, Borden A and Lawrence CW. (1997). *J. Bact.,* **179,** 2835 – 2839.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I and Clamp M. (2002). *Nucleic Acids Res.,* **30,** 38 – 41.

Jiang YL and Stivers JT. (2001). *Biochemistry,* **40,** 7710 – 7719.

Kavli B, Slupphaug G, Mol CD, Arvai AS, Petersen SB, Tainer JA and Krokan HE. (1996). *EMBO J.,* **15,** 3442 – 3447.

Kavli B, Sundheim O, Akbari M, Otterlei M, Nilsen H, Skorpen F, Aas PA, Hagen L, Krokan HE and Slupphaug G. (2002). *J. Biol. Chem.* (In press). Electronic publication ahead of print, August 2, 2002.

Krokan HE, Otterlei M, Nilsen H, Kavli B, Skorpen F, Andersen S, Skjelbred C, Akbari M, Aas PA and Slupphaug G. (2001). *Prog. Nucleic Acids Res. Mol. Biol.,* **68,** 365 – 386.

Krokan HE, Standal R and Slupphaug G. (1997). *Biochem. J.,* **325,** 1 – 16.

Levy DD and Teebor GW. (1991). *Nucleic Acids Res.,* **19,** 3337 – 3343.

Lindahl T. (1974). *Proc. Natl. Acad. Sci. USA,* **71,** 3649 – 3653.

Lindahl T. (1993). *Nature,* **362,** 709 – 715.

Mellac S, Fazakerley GV and Sowers LC. (1993). *Biochemistry,* **32,** 7779 – 7786.

Missero C, Pirro MT, Simeone S, Pischetola M and Di Lauro R. (2001). *J. Biol. Chem.,* **276,** 33569 – 33575.

Mol CD, Arvai AS, Begley TJ, Cunningham RP and Tainer JA. (2002). *J. Mol. Biol.,* **315,** 373 – 384.

Mol CD, Arvai AS, Slupphaug G, Kavli B, Alseth I, Krokan HE and Tainer JA. (1995). *Cell,* **80,** 869 – 878.

Moriya M, Zhang W, Johnson F and Grollman AP. (1994). *Proc. Natl. Acad. Sci. USA,* **91,** 11899 – 11903.

Muller S, Hoege C, Pyrowolakis G and Jentsch S. (2001). *Nature Rev. Mol. Cell Biol.,* **2,** 202 – 210.

Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y and Honjo T. (2000). *Cell,* **102,** 553 – 563.

Muramatsu M, Sankaranand VS, Anant S, Sugai M, Kinoshita K, Davidson NO and Honjo T. (1999). *J. Biol. Chem.,* **274,** 18470 – 18476.

Neddermann P, Gallinari P, Lettieri T, Schmid D, Truong O, Hsuan JJ, Wiebauer K and Jiricny J. (1996). *J. Biol. Chem.,* **271,** 12767 – 12774.

Neddermann P and Jiricny J. (1993). *J. Biol. Chem.,* **268,** 21218 – 21224.

Nilsen H, Haushalter KA, Robins P, Barnes DE, Verdine GL and Lindahl T. (2001). *EMBO J.,* **20,** 4278 – 4286.

Nilsen H, Otterlei M, Haug T, Solum K, Nagelhus TA, Skorpen F and Krokan HE. (1997). *Nucleic Acids Res.,* **25,** 750 – 755.

Nilsen H, Rosewell I, Robins P, Skjelbred CF, Andersen S, Slupphaug G, Daly G, Krokan HE, Lindahl T and Barnes DE. (2000). *Mol. Cell,* **5,** 1059 – 1065.

Nilsen H, Yazdankhah SP, Eftedal I and Krokan HE. (1995). *FEBS Lett.,* **362,** 205 – 209.

Olsen LC, Aasland R, Wittwer CU, Krokan HE and Helland DE. (1989). *EMBO J.,* **8,** 3121 – 3125.

Otterlei M, Haug T, Nagelhus TA, Slupphaug G, Lindmo T and Krokan HE. (1998). *Nucleic Acids Res.,* **26,** 4611 – 4617.

Otterlei M, Kavli B, Standal R, Skjelbred C, Bharati S and Krokan HE. (2000). *EMBO J.,* **19,** 5542 – 5551.

Otterlei M, Warbrick E, Nagelhus TA, Haug T, Slupphaug G, Akbari M, Aas PA, Steinsbekk K, Bakke O and Krokan HE. (1999). *EMBO J.,* **18,** 3834 – 3844.

Parikh SS, Mol CD, Slupphaug G, Bharati S, Krokan HE and Tainer JA. (1998). *EMBO J.,* **17,** 5214 – 5226.

Parikh SS, Putnam CD and Tainer JA. (2000a). *Mutat. Res.,* **460,** 183 – 199.

Parikh SS, Walcher G, Jones GD, Slupphaug G, Krokan HE, Blackburn GM and Tainer JA. (2000b). *Proc. Natl. Acad. Sci. USA,* **97,** 5083 – 5088.

Pasternack LB, Bramham J, Mayol L, Galeone A, Jia X and Kearns DR. (1996). *Nucleic Acids Res.,* **24,** 2740 – 2745.

Pearl FMG, Lee D, Bray JE, Sillitoe I, Todd AE, Harrison AP, Thornton JM and Orengo CA. (2000). *Nucleic Acids Res.,* **28,** 277 – 282.

Pearl LH. (2000). *Mutat. Res.,* **460,** 165 – 181.

Petersen-Mahrt SK, Harris RS and Neuberger MS. (2002). *Nature,* **418,** 99 – 103.

Petronzelli F, Riccio A, Markham GD, Seeholzer SH, Stoerker J, Genuardi M, Yeung AT, Matsumoto Y and Bellacosa A. (2000). *J. Biol. Chem.,* **275,** 32422 – 32429.

8948

Radany EH, Dornfeld KJ, Sanderson RJ, Savage MK, Majumdar A, Seidman MM and Mosbaugh DW. (2000). *Mutat. Res.,* **461,** 41 – 58.

Riccio A, Aaltonen LA, Godwin AK, Loukola A, Percesepe A, Salovaara R, Masciullo V, Genuardi M, Paravatou-Petsotas M, Bassi DE, Ruggeri BA, Klein-Szanto AJP, Testa JR, Neri G and Bellacosa A. (1999). *Nature Genetics,* **23,** 266 – 268.

Robu ME, Inman RB and Cox MM. (2001). *Proc. Natl. Acad. Sci. USA,* **98,** 8211 – 8218.

Rossi R, Villa A, Negri C, Scovassi I, Ciarrocchi G, Biamonti G and Montecucco A. (1999). *EMBO J.,* **18,** 5745 – 5754.

Rusmintratip V and Sowers LC. (2000). *Proc. Natl. Acad. Sci. USA,* **97,** 14183 – 14187.

Saparbaev M and Laval J. (1998). *Proc. Natl. Acad. Sci. USA,* **95,** 8508 – 8513.

Sartori AA, Fitz-Gibbon S, Hanjing Y, Miller JH and Jiricny J. (2002). *EMBO J.,* **21,** 3182 – 3191.

Sartori AA, Schar P, Fitz-Gibbon S, Miller JH and Jiricny J. (2001). *J. Biol. Chem.,* **276,** 29979 – 29986.

Savva R, McAuley Hecht K, Brown T and Pearl L. (1995). *Nature,* **373,** 487 – 493.

Sekiguchi M, Hayakawa M, Makino F, Tanaka K and Okada Y. (1976). *Biochem. Biophys. Res. Commun.,* **73,** 293 – 299.

Shapiro R. (1980). *Chromosome damage and repair.* Seeberg E, Kleppe K (ed). New York: Plenum Press, pp. 3 – 18.

Sheehy AM, Gaddis NC, Choi JD and Malim MH. (2002). *Nature,* **418,** 646 – 650.

Shroyer MJ, Bennett SE, Putnam CD, Tainer JA and Mosbaugh DW. (1999). *Biochemistry,* **38,** 4834 – 4845.

Slupphaug G, Eftedal I, Kavli B, Bharati S, Helle NM, Haug T, Levine DW and Krokan HE. (1995). *Biochemistry,* **34,** 128 – 138.

Slupphaug G, Mol CD, Kavli B, Arvai AS, Krokan HE and Tainer JA. (1996). *Nature,* **384,** 87 – 92.

Spek EJ, Wright TL, Stitt MS, Taghizadeh NR, Tannenbaum SR, Marinus MG and Engelward BP. (2001). *J. Bact.,* **183,** 131 – 138.

Stivers JT, Pankiewicz KW and Watanabe KA. (1999). *Biochemistry,* **38,** 952 – 963.

Thayer MM, Ahern H, Xing D, Cunningham RP and Tainer JA. (1995). *EMBO J.,* **14,** 4108 – 4120.

Tu YQ, Dammann R and Pfeifer GP. (1998). *J. Mol. Biol.,* **284,** 297 – 311.

Tye BK, Nyman PO, Lehman IR, Hochhauser S and Weiss B. (1977). *Proc. Natl. Acad. Sci. USA,* **74,** 154 – 157.

Um S, Harbers M, Benecke A, Pieratt B, Losson R and Chambon P. (1998). *J. Biol. Chem.,* **273,** 20728 – 20736.

Vidal AE, Hickson ID, Boiteux S and Radicella JP. (2001). *Nucleic Acids Res.,* **29,** 1285 – 1292.

Vilpo JA and Vilpo LM. (1995). *Mutat. Res.,* **316,** 123 – 131.

Waters TR, Gallinari P, Jiricny J and Swann PF. (1999). *J. Biol. Chem.,* **274,** 67 – 74.

Waters TR and Swann PF. (2000). *Mutat. Res.,* **462,** 137 – 147.

Werner RM, Jiang YL, Gordley RG, Jagadeesh GJ, Ladner JE, Xiao G, Tordova M, Gilliland GL and Stivers JT. (2000). *Biochemistry,* **39,** 12585 – 12594.

Werner RM and Stivers JT. (2000). *Biochemistry,* **39,** 14054 – 14064.

Willetts KE, Rey F, Agostini I, Navarro JM, Baudat Y, Vigne R and Sire J. (1999). *J. Virol.,* **73,** 1682 – 1688.

Wist E, Unhjem O and Krokan HE. (1978). *Biochim. Biophys. Acta,* **250,** 253 – 270.

Wittwer CU, Bauw G and Krokan HE. (1989). *Biochemistry,* **28,** 780 – 784.

Wong I, Lundquist AJ, Bernards AS and Mosbaugh DW. (2002). *J. Biol. Chem.,* **277,** 19424 – 19432.

Xiao G, Tordova M, Jagadeesh J, Drohat AC, Stivers JT and Gilliland GL. (1999). *Proteins,* **35,** 13 – 24.

Zhu B, Zheng Y, Angliker H, Schwarz S, Thiry S, Siegmann M and Jost JP. (2000a). *Nucleic Acids Res.,* **28,** 4157 – 4165.

Zhu B, Zheng Y, Hess D, Angliker H, Schwarz S, Siegmann M, Thiry S and Jost JP. (2000b). *Proc. Natl. Acad. Sci. USA,* **97,** 5135 – 5139.