The EMBO Journal vol.6 no.10 pp.3177-3183, 1987

Formamidopyrimidine-DNA glycosylase of *Escherichia coli*: cloning and sequencing of the *fpg* structural gene and overproduction of the protein

Serge Boiteux, Timothy R.O'Connor and Jacques Laval

Group 'Réparation des lésions chimio- et radio-induites', UA 147 CNRS and U 140 INSERM, Institut Gustave-Roussy, 94805 Villejuif Cédex, France

Communicated by T.Lindahl

An *Escherichia coli* genomic library composed of large DNA fragments (10-15 kb) was constructed using the plasmid pBR322 as vector. From it 700 clones were individually screened for increased excision of the ring-opened form of *N*7-methylguanine (2-6-diamino-4-hydroxy-5*N*-methylformamidopyrimidine) or Fapy. One clone overproduced the Fapy-DNA glycosylase activity by a factor of 10-fold as compared with the wild-type strain. The Fapy-DNA glycosylase overproducer character was associated with a 15-kb recombinant plasmid (pFPG10). After subcloning a 1.4-kb fragment which contained the Fapy-DNA glycosylase gene (*fpg*$^+$) was inserted in the plasmids pUC18 and pUC19 yielding pFPG50 and pFPG60 respectively. The cells harbouring pFPG60 displayed a 50- to 100-fold increase in glycosylase activity and overexpressed a 31-kd protein. From these cells the Fapy-DNA glycosylase was purified to apparent physical homogeneity as evidenced by a single protein band at 31 kd on SDS-polyacrylamide gels. The amino acid composition of the protein and the amino acid sequence deduced from the nucleotide sequence demonstrate that the cloned fragment contains the structural gene coding for the Fapy-DNA glycosylase. The nucleotide sequence of the *fpg* gene is composed of 809 base pairs and codes for a protein of 269 amino acids with a calculated mol. wt of 30.2 kd.

Key words: DNA alkylation/DNA repair/formamidopyrimidine-DNA glycosylase gene/cloning and sequence/*Escherichia coli*

Introduction

Alkylating agents such as *N*-methylnitrosourea or *N*-methyl-*N'*-nitro-*N*-nitrosoguanidine are potent mutagens and carcinogens. They interact with DNA yielding alkylated bases and phosphotriesters (Singer and Grunberger, 1983). The main base modification products are 7-methylguanine, 3-methyladenine, and O^6-methylguanine. Whereas in *Escherichia coli* mutagenesis is currently attributed to O^6-methylguanine (Loechler *et al.*, 1984), killing is associated with purines alkylated at the *N*3 position (Evensen and Seeberg, 1982; Boiteux *et al.*, 1984). In *E. coli* lesions introduced in DNA by methylating or ethylating agents are repaired either by a DNA alkyltransferase encoded by the *ada* gene (Demple *et al.*, 1985; Nakabeppu *et al.*, 1986) or by DNA glycosylases encoded either by the *tag*A (Sakumi *et al.*, 1986) or the *alk*A gene (Nakabeppu *et al.*, 1984).

Although *N*7-methylguanine is the major alkylation product in DNA, this lesion does not seem to be harmful to the cell as it does not interfere with *in vitro* DNA synthesis (Boiteux and Laval, 1983), and persists in DNA for generations *in vivo* (Lawley and Orr, 1970; Karran *et al.*, 1982). However *N*7 alkylation favours the cleavage of the imidazole ring yielding 2-6-diamino-4-hydroxy-5*N*-methylformamidopyrimidine (Fapy) (Robins and Townsend, 1963; Haines *et al.*, 1962). Several observations suggest that the imidazole ring-opened form of *N*7 methylguanine might play a significant role in processes leading to mutagenesis and/or cell death by alkylating agents. *In vitro* DNA synthesis experiments have shown that the Fapy residues inhibit *E. coli* DNA polymerase I synthesis, suggesting that these lesions may have a role in cell death (Boiteux and Laval, 1983). There is a specific DNA-glycosylase in *E. coli* and in mammalian cells which is responsible for the removal of this lesion (Chetsanga and Lindahl, 1979; Chetsanga *et al.*, 1981; Lagravere *et al.*, 1985; Margison and Pegg, 1981).

However, despite the circumstantial evidence for the biological importance of the Fapy-DNA glycosylase, the evaluation of the physiological role of this enzyme, and the role of Fapy residues in killing and mutagenesis, remain speculative and await the isolation of an enzyme defective mutant. Isolation of a Fapy-DNA glycosylase mutant of *E. coli* has not been achieved by screening for mutants which are sensitive to alkylating agents (Yamamoto *et al.*, 1978; Kataoka *et al.*, 1983). Therefore, we cloned the Fapy-DNA glycosylase gene of *E. coli* into a multicopy plasmid as the first step in elucidating the biological role of this enzyme. We constructed a plasmid library of *E. coli* DNA and screened it for an increase in Fapy-DNA glycosylase activity in individual crude lysates. From this library we have isolated a clone which overproduces the Fapy-DNA glycosylase activity as compared with the wild-type background. We show that this overproducer character is associated with a 15-kb recombinant plasmid. We propose that the gene coding for this protein be named *fpg* for Fapy-DNA glycosylase. Subcloning experiments localized the *fpg* gene within a 1.4-kb DNA fragment. The nucleotide sequence of the fragment and amino acid composition of the pure Fapy-DNA glycosylase show that the 1.4 kb fragment contains the *fpg*$^+$ gene which codes for the enzyme in *E. coli*.

Results

Construction of recombinant plasmids which overproduce the Fapy-DNA glycosylase activity

The search for a plasmid carrying the *fpg* gene was based on the assumption that such a plasmid would result in the overproduction of the Fapy-DNA glycosylase in the host cell. We constructed a plasmid library by inserting *Sau*3A1-generated fragments of DNA from *E. coli* strain HB1100 into the *Bam*HI site of pBR322 (see Materials and methods and Figure 1). The recombinant plasmids were used to transform *E. coli* strain HB101, which is *rec*A$^-$.

The resulting ampicillin-resistant transformants were purified and individually grown in liquid meidum. Approximately 700 clones from the plasmid library were screened for an increase in Fapy-DNA glycosylase activity in crude lysates and compared with the background activity of HB101 and HB101/pBR322. Only one clone C496 showed a significant increase in the activity of the enzyme. Table I shows that this clone C496 overproduces the Fapy-DNA glycosylase activity by a factor of 14 compared

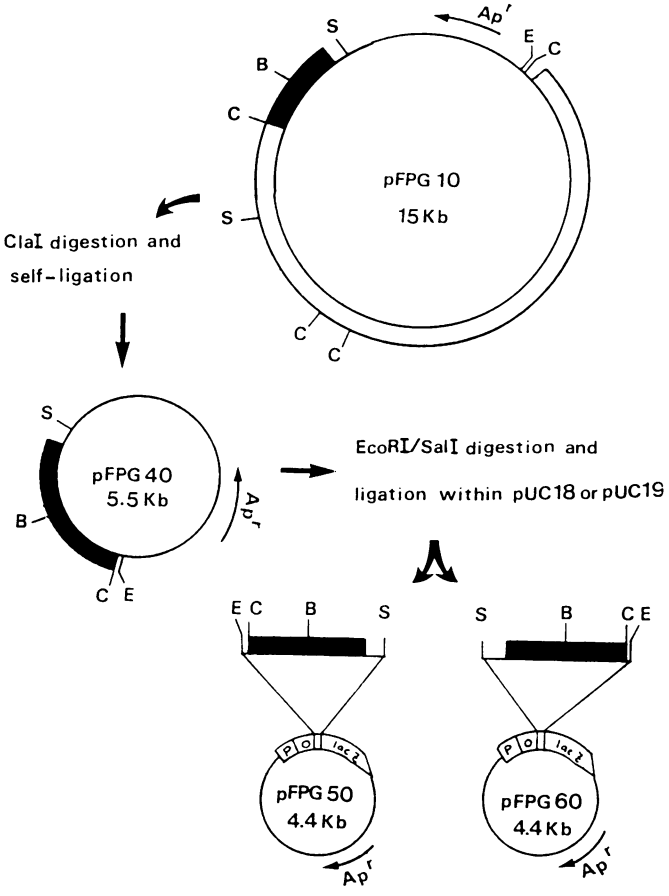

Fig. 1. Construction of plasmids that overproduce Fapy-DNA glycosylase. Plasmid pFPG10 was digested with *Cla*I and self-ligated to produce the pFPG40 plasmid (_____, plasmid DNA; ■□ *E. coli* DNA). The 1.4-kb fragment of *E. coli* DNA which confers the overproducer character is indicated in the thick black region (■). *Eco*RI and *Sal*I digestion of pFPG40 yielded a 1.7-kb fragment which was ligated into *Eco*RI–*Sal*I-digested pUC18 and pUC19 yielded plasmids pFPG50 and pFPG60. Restriction sites: *Cla*I (C), *Bam*HI (B), *Sal*I (S), *Eco*RI (E). Ap^r, ampicillin resistance gene; P and O, lac promoter and operator region; *lacZ*, β-galactosidase gene.

Table I. Isolation of a clone overproducing the Fapy-DNA glycosylase

Strain	Relevant properties				
	Amp[a]	Tet[a]	MMS[a]	Fapy-DNA glycosylase activity[b]	Relative enzyme activity[c]
HB101	S	S	S	12	1.0
HB101/pBR322	R	R	S	13	1.1
C497[d]	R	S	S	13	1.1
C496	R	S	S	173	14.4

Cells from the *Sau*3AI library in pBR322 were screened for sensitivity to ampicillin, tetracycline or methylmethanesulphonate and for Fapy-DNA glycosylase activity.
[a]Ampicillin concentration: 50 μg/ml, tetracycline concentration: 10 μg/ml, methylmethane sulphonate concentration: 1 mM.
[b]Units/mg protein, 1 unit released 1 pmol of Fapy in 5 min at 37°C. In crude lysates, >95% of ethanol-soluble material was identified by h.p.l.c. as Fapy bases.
[c]Compared with the HB101 background.
[d]C497 is an arbitrary clone of the plasmid library.
S, sensitive; R, resistant.

Table II. Fapy-DNA glycosylase activity in crude lysates of *E. coli* HB101 cells hosting various plasmids

Plasmid	Fapy-DNA glycosylase specific activity (units/mg protein)	Relative enzyme activity[a]
pBR322	10	1.0
pFPG10	120	12.0
pFPG40	116	11.6
pUC18	10	1.0
pFPG50	540	54.0
pFPG70[b]	11	1.1
pUC19	10	1.0
pFPG60	688	68.8
pFPG80[b]	10	1.0

[a]Relative Fapy-DNA glycosylase activity in plasmid-containing cells to the Fapy-DNA glycosylase activity in *E. coli* strain HB101.
[b]The plasmids pFPG70 and pFPG80 were obtained by subcloning the *Eco*RI/*Bam*HI DNA fragment from pFPG40 into pUC18 and pUC19 (see Figure 1).

with the control strains. Furthermore C496 was ampicillin resistant, tetracycline-sensitive and remained sensitive to methylmethane sulphonate (Table I). These results suggest that C496 was harbouring a recombinant plasmid which was responsible for the increase in the Fapy-DNA glycosylase activity over the HB101 background. In fact a 15-kb plasmid isolated from C496 was able to confer the Fapy-DNA glycosylase overproducer characteristics when introduced into HB101 and other *E. coli* strains (JM103, HB1100, AB1157, data not shown). This plasmid which causes a 12-fold increase in glycosylase production in HB101 was designated pFPG10 (Figure 1 and Table II). A restriction map of the pFPG10 plasmid is shown in Figure 1. Subcloning of pFPG10 by *Cla*I digestion and self-ligation generated the pFPG40 plasmid and localized the *fpg*$^+$ gene responsible for the overproduction of the Fapy-DNA glycosylase on a 1.4-kb region of *E. coli* DNA.

Overproduction of the Fapy-DNA glycosylase

To overproduce the Fapy-DNA glycosylase the gene was cloned into the pUC18 and pUC19 runaway system. A 1.7-kb *Sal*I/*Eco*RI fragment of pFPG40 containing the 1.4-kb *E. coli* DNA which carries the *fpg* gene and two flanking regions of pBR322 DNA (23 and 275 bp) was isolated (Figure 1). This 1.7-kb fragment was inserted into pUC18 and pUC19 to generate the two plasmids pFPG50 and pFPG60 in the orientations indicated in Figure 1. The 0.75-kb *Bam*HI/*Eco*RI fragment from pFPG40 was inserted into the pUC18 and pUC19 digested with *Bam*HI and *Eco*RI to create the plasmids pFPG70 and pFPG80 respectively.

HB101 cells were transformed with the plasmids described above, and the Fapy-DNA glycosylase activity in crude lysates was assayed. Table II shows that cells which host the pFPG50 and pFPG60 plasmids produce 54- and 68-fold more enzyme respectively than the host cells. The reason(s) for the higher production of enzyme in cells hosting the pFPG60 plasmid compared with the cells hosting the pFPG50 plasmid may be a difference in plasmid copy number and/or the ability of the cloned *fpg*$^+$ gene to utilize the *lac* promoter. In fact the production of the glycosylase in JM103 cells carrying a *lac*I^q mutation and the pFPG60 plasmid is stimulated by isopropyl-β-D-thiogalactopyranoside (IPTG) (data not shown). These data suggest that the orientation of transcription of the *fpg* gene in the pFPG60 plasmid

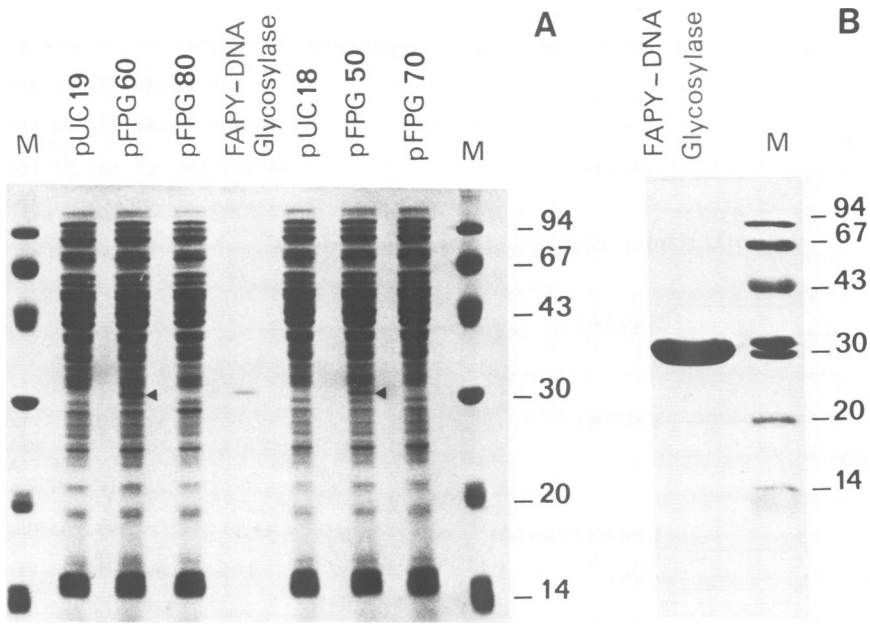

Fig. 2. Overproduction of a 31-kd protein in HB101 cells containing pFPG50 and pFGP60 and homogeneous Fapy-DNA glycosylase purified from HB101 cells hosting pFPG60. (**A**) HB101 cells containing pFPG50, pFGP60, pFPG70, pFPG80, pUC18 and pUC19 were grown in 20 ml LB broth at 37°C and were lysed as described in Materials and methods. Fifty micrograms of protein from each lysate was loaded in each lane of a 15% SDS−-polyacrylamide gel. The crude lysates were also analysed for Fapy-DNA glycosylase activity (Table II). Each lane represents the lysate from HB101 harbouring a different plasmid: pUC19, pFPG60, pFGP80, pUC18, pFPG50 and pFPG70. One lane was loaded with 1.0 μg of pure Fapy-DNA glycosylase (Fraction V from Table III). Molecular weight standards were phosphorylase b (94 kd), bovine serum albumin (67 kd), ovalbumin (43 kd), carbonic-anhydrase (30 kd) trypsin inhibitor (20.1 kd) and α-lactalbumin (14.4 kd). (**B**) Pure Fapy-DNA glycosylase (7.5 μg (Fraction V from Table III), was chromatographed as described in panel A. M, molecular weight standards as in Figure 2A.

Table III. Purification of the Fapy-DNA glycosylase[a]

Fraction	Total activity ($\times 10^3$ units)	Total protein (mg)	Specific activity (units/mg)	Purification (-fold)	Recovery (%)
Crude extract	632	551	1147	1.0	100
QMA anion exchange	571	545	1047	0.9	90
Phospho-ultrogel	298	40	7450	6.5	47
Gel filtration ACA54	279	5.6	49 821	43.4	44
F.p.l.c. Mono S HR5/5	73	1.0	73 000	63.6	12

[a]Isolation of the Fapy-DNA glycosylase from 5 g HB101 hosting pFPG60. The glycosylase is >1% of the total soluble proteins in the cell.

is the same as that from the *lac* promoter. The HB101 cells which contain the pFPG70 and pFPG80 plasmids do not show any increase in the glycosylase activity, which implies that the complete *fpg* gene was not contained in the DNA of the *Eco*RI/*Bam*HI fragment (Table II).

The proteins from the crude extracts of HB101 cells harbouring different plasmids which were assayed in Table II were also subjected to SDS−polyacrylamide gel electrophoresis as displayed in Figure 2A. The cells which contain the pFPG50 and pFPG60 plasmids show a reinforced 31-kd protein band compared with the four other crude extracts which were analysed. Additionally the quantity of the 31-kd protein produced by the cells hosting pFPG60 is greater than the quantity produced by the cells hosting pFPG50 (Figure 2A). This correlates well with the enzymatic activity assay which was detailed above. The cells harbouring the pFPG80 and pFPG70 plasmids do not overproduce the 31-kd protein (Figure 2A). These results show that the 31-kd protein produced by the pFPG50 and pFPG60 plasmids has the same electrophoretic mobility as the purified Fapy-DNA glycosylase as indicated in Figure 2A. Together the results of the enzymatic activity and gel electrophoresis suggest that the 31-kd protein overproduced in cells containing the pFPG50 and pFPG60 plasmids is the Fapy-DNA glycosylase.

Purification of the Fapy-DNA glycosylase

The Fapy-DNA glycosylase was purified to apparent physical homogeneity by taking advantage of the overproduction of the protein in HB101 cells hosting the pFPG60 plasmid. The purification is described in Materials and methods and summarized in Table III. The enzyme was purified 64-fold at the final step of the purification with a 12% recovery. SDS−polyacrylamide gel electrophoresis of the purified enzyme in Figure 2B shows that a single 31-kd protein is the final result of the steps outlined in Table III. More than 99% of the products released from [^{3}H]-Fapy-poly(dG-dC)·poly(dG-dC) by the purified enzyme were Fapy bases. The purified enzyme did not release *N7* or *N3*-methylguanine or *N*3-methyladenine from dimethylsulphate (DMS)-treated DNA (data not shown).

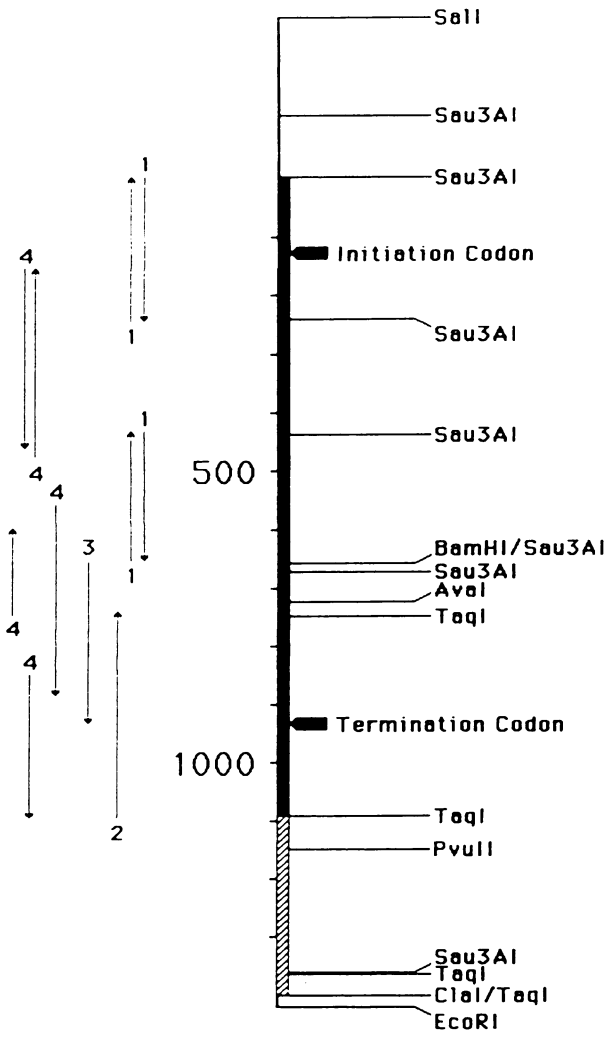

Fig. 3. Organization of the 1.4-kb fragment conferring the Fapy-DNA glycosylase overproducer character and strategy for DNA sequencing. The sites of several restriction enzymes and the length of the fragment in base pairs are indicated. The region in solid black was sequenced. The site of the methionine start and the termination codon are indicated. The numbers of the horizontal arrows refer to different clones used for sequencing: 1, *Sau*3AI fragments; 2, *Taq*I fragment; 3, *Bam*HI–*Eco*RI fragment; 4, oligonucleotide primers.

Nucleotide sequence of the fpg gene

Additional subcloning experiments showed that full glycosylase activity and overexpression of the 31-kd protein is obtained in the 1090-bp fragment between the *Sau*3AI site (position 1 on the map) and the *Taq*I site (position 1090) (Figure 3 and data not shown). Both strands have been sequenced in this region using the dideoxy method. The strategy for sequence analysis of the fragment is shown in Figure 3. The nucleotide sequence of the 1090-bp region is presented in Figure 4. There are several features of the sequence which are consistent with the fact that the Fapy-DNA glycosylase sequence is coded for in this region. There is only a single open reading frame which starts at position 126 and ends at position 932 (Figure 4). This open reading frame codes for a protein which has 269 amino acids, having a calculated mol. wt of 30.2 kd. In addition there are two sequences in the region 5′ to the open reading frame which correspond to the promoter consensus sequences (−35 and −10 sequences at positions 73 and 95 respectively) (Pribnow, 1975) and the Shine–Dalgarno-like sequence (at position 117) (Shine and Dalgarno, 1975) which indicate the start of a protein-coding region in *E. coli*. The termination codon TAA at position 933 is located before a potential stem and loop structure, suggesting termination. These sequences are indicated in Figure 4.

```
   1                GATCTACAAAGAAGCGAAAATCAAATAA

  29 TTCTCGCTTTGATGTAACAAAAAAACCTCGCTCCGGCGGGG

  70 TTTTTGTTATCTGCTTGCCCCCATATTGACTGCATCTGTT
        -35                      -10
 110 CATTCCTGGAGATGCT ATG CCT GAA TTA CCC GAA
           S.D.      Met Pro Glu Leu Pro Glu
 144 GTT GAA ACC AGC CGC CGC GGC ATA GAA CCG
     Val Glu Thr Ser Arg Arg Gly Ile Glu Pro
 175 CAT CTC GTT GGT GCA ACC ATT CTT CAT GCA
     His Leu Val Gly Ala Thr Ile Leu His Ala
 205 GTG GTG CGC AAC GGA CGC TTG CGC TGG CCG
     Val Val Arg Asn Gly Arg Leu Arg Trp Pro
 235 GTT TCA GAA GAG ATC TAC CGT TTA AGC GAC
     Val Ser Glu Glu Ile Tyr Arg Leu Ser Asp
 265 CAA CCA GTG CTT AGC GTG CAG CGG CGG GCT
     Gln Pro Val Leu Ser Val Gln Arg Arg Ala
 295 AAA TAT CTG CTG CTG GAG CTG CCT GAG GGC
     Lys Tyr Leu Leu Leu Glu Leu Pro Glu Gly
 325 TGG ATT ATC ATT CAT TTA GGG ATG TCT GGC
     Trp Ile Ile Ile His Leu Gly Met Ser Gly
 355 AGC CTG CGC ATC CTT CCA GAA GAA CTT CCC
     Ser Leu Arg Ile Leu Pro Glu Glu Leu Pro
 385 CCT GAA AAG CAT GAC CAT GTG GAT TTG GTG
     Pro Glu Lys His Asp His Val Asp Leu Val
 415 ATG AGC AAC GGC AAA GTG CTG CGC TAC ACC
     Met Ser Asn Gly Lys Val Leu Arg Tyr Thr
 445 GAT CCG CGC CGC TTT GGT GCC TGG CTG TGG
     Asp Pro Arg Arg Phe Gly Ala Trp Leu Trp
 475 ACC AAA GAG CTG GAA GGG CAT AAT GTG CTG
     Thr Lys Glu Leu Glu Gly His Asn Val Leu
 505 ACC CAT CTT GGA CCG GAG CCG CTT AGC GAC
     Thr His Leu Gly Pro Glu Pro Leu Ser Asp
 535 GAT TTC AAT GGT GAG TAT CTG CAT CAG AAG
     Asp Phe Asn Gly Glu Tyr Leu His Gln Lys
 565 TGC GCG AAG AAA AAA ACG GCG ATT AAA CCG
     Cys Ala Lys Lys Lys Thr Ala Ile Lys Pro
 595 TGG CTG ATG GAT AAC AAG CTG GTG GTA GGG
     Trp Leu Met Asp Asn Lys Leu Val Val Gly
 625 GTA GGG AAT ATC TAT GCC AGC GAA TCA CTG
     Val Gly Asn Ile Tyr Ala Ser Glu Ser Leu
 655 TTT GCG GCG GGG ATC CAT CCG GAT CGG CTG
     Phe Ala Ala Gly Ile His Pro Asp Arg Leu
 685 GCG TCA TCA CTG TCG CTG GCA GAG TGT GAA
     Ala Ser Ser Leu Ser Leu Ala Glu Cys Glu
 715 TTG TTA GCT CGG GTG ATT AAA GCG GTG TTG
     Leu Leu Ala Arg Val Ile Lys Ala Val Leu
 745 CTG CGT TCG ATT GAG CAG GGT GGT ACA ACG
     Leu Arg Ser Ile Glu Gln Gly Gly Thr Thr
 775 CTG AAA GAT TTT CTG CAA AGT GAT GGT AAA
     Leu Lys Asp Phe Leu Gln Ser Asp Gly Lys
 805 CCG GGC TAT TTC GCT CAG GAA TTG CAG GTT
     Pro Gly Tyr Phe Ala Gln Glu Leu Gln Val
 835 TAC GGG CGA AAA GGT GAG CCG TGT CGG GTG
     Tyr Gly Arg Lys Gly Glu Pro Cys Arg Val
 865 TGC GGT ACG CCG ATT GTG GCG ACT AAA CAT
     Cys Gly Thr Pro Ile Val Ala Thr Lys His
 895 GCG CAG CGG GCA ACG TTT TAT TGT CGG CAG
     Ala Gln Arg Ala Thr Phe Tyr Cys Arg Gln
 925 TGC CAG AAG TAATTCATGCGCGCCGGATGGCATACCAT
     Cys Gln Lys
 963 CCGGCATAAACGCTACGCTAACTTCGCCATCAGCGCCTGAT

1003 GGACATTCTCCGGCAGGAAATGGGTGACATCGCCCTGATGG

1045 CGCGCCACCTCTTTCACCAACGATGAAGAGATAAACGACC

1085 ACTCTTTCGA
```

Fig. 4. Nucleotide sequence of the *fpg* gene and the amino acid sequence of the Fapy-DNA glycosylase. The nucleotides are numbered from the *Sau*3AI cloning site. The predicted amino acid sequence is shown. The possible promoter sequences and the Shine–Dalgarno sequence are underlined. Palindromic sequences are shown by arrows.

Amino acid analysis

To provide further support for the protein sequence as determined by the DNA sequence, we analysed the amino acid composition

Table IV. Amino acid composition of the Fapy-DNA glycosylase[a]

Amino acid	Analysis	Predicted from nucleotide sequence
Alanine	18.5	17
Arginine	18.2	19
Aspartic acid		10
Asparagine	17.7	6
Cysteine[b]	–	6
Glutamic acid		21
Glutamine	32.2	10
Glycine	23.5	21
Histidine	10.1	10
Isoleucine	11.2	13
Leucine	33.8	35
Lysine	17.0	16
Methionine	2.6	4
Phenylalanine	6.5	6
Proline	16.2	17
Serine	15.5	15
Threonine	9.2	11
Tyrosine	8.3	8
Tryptophan[b]	–	5
Valine	16.4	19
Total		269

[a]The experimental details are described in Lederer *et al.* (1983).
[b]Cysteine and tryptophan are not determined.

of the Fapy-DNA glycosylase. The composition was determined after hydrolysis of the purified Fraction V of the enzyme (Table III). The amino acid composition is presented in Table IV. The overall amino acid composition is in good agreement with the composition deduced from the nucleic acid sequence. Furthermore the amino-terminal sequence of the protein as manually determined was Pro-Glu-Leu. This sequence exactly matched the translated nucleotide sequence (Figure 4).Since the amino terminal residue is not a methionine, it suggests that the protein is further processed by aminopeptidases. These results confirm the cloning of the *E. coli* Fapy-DNA glycosylase structural gene.

Discussion

This work reports the isolation and the nucleotide sequence of the structural gene of the Fapy-DNA glycosylase of *E. coli*. This gene consists of 809 bp which code for a protein of 269 amino acids with a calculated mol. wt of 30.2 kd. Cloning of the *fpg* gene on a plasmid vector pUC19 allows the overexpression of the *fpg* gene product, which represents over 1.0% of the total soluble proteins. The Fapy-DNA glycosylase was further purified to apparent physical homogeneity, and the amino acid composition was determined. It is in agreement with the amino acid composition deduced from the DNA sequence.

We have analysed the DNA sequence for characteristics which either conform to consensus rules for DNA sequences of the regulatory regions of *E. coli* or for unusual structural features. A recent compilation of *E. coli* promoters has proposed consensus sequences for the −35 and −10 regions of genes in *E. coli* (Harley and Reynolds, 1987). Based on this compilation we propose that the −35 region of the *fpg* gene which begins at position 73 is composed of the sequence TTGTTA compared with the consensus sequence of TTGACA for the −35 region of *E. coli*. We suggest that the −10 region indicated in Figure 4 starting at position 95 (TTGACT compared with the consensus of TATAAT) allows a 16-bp gap between the −35 and −10

```
        Pro Arg Val Phe Gly Ala
denV    CCG CGT GTT TTT GGT GCA     88
        ||| ||     ||| ||| ||
fpg     CCG CGC CGC TTT GGT GCC    338
        Pro Arg Arg Phe Gly Ala

        Lys Ile Ala Gln Arg Pro Thr Trp Tyr
denV    AAA ATT GCA CAA CGT CCT ACT TGG TAC    480
        |||  |||  ||  ||  |  ||  |  ||
fpg     AAA CAT GCG CAG CGG GCA ACG TTT TAT    914
        Lys His Ala Gln Arg Ala Thr Phe Tyr
```

Fig. 5. Homology of the amino acid sequences of the *fpg* and *den*V gene products. The homology of the nucleotide sequences is indicated by vertical lines.

regions, which conforms to the consensus distance of 16−18 bp. There are other sequences which are associated with −10 regions at positions 93 and 104, but these sequences would not be expected to be the actual −10 region based on the separation between the −35 and −10 regions (13 bp and 24 bp respectively).

A computer search for potential inverted repeat structures yielded two potential cruciform structures just before (positions 48−74) and just after (positions 945−966) the coding region of the gene (Figure 4). Both structures may be implicated in regulation of the transcription of the *fpg* gene. If the first region adopts a cruciform structure *in vivo* the formation of an RNA polymerase transcription initiation complex could be inhibited. The second region occurs 9 bp after the termination codon and has the consensus sequence of an intercistronic palindromic unit (PU* sequence) (Gilson *et al.*, 1986). Moreover this particular palindrome may signal the termination of genes transcribed in the opposite sense with termination just before the palindrome (Hudson and Davidson, 1984; Gilson *et al.*, 1986). Indeed the sequence shown in Figure 4 has an open reading frame from position 1090 (*Taq*I) to a termination codon at position 977, which is within 10 bp of the palindrome at positions 945−966.

The codon usage of the *fpg* gene showed that the glycosylase is coded for by a large number of rare codons (25% according to Konigsberg and Godson, 1983). This high percentage of rare codon usage is at the same level as proteins which are considered regulatory. Other genes responsible for DNA alkylation repair in *E. coli* the *ada*, *alk*A and *tag*A genes also are biased for the use of rare codons (26, 34 and 21% respectively) (Steinum and Seeberg, 1986). In addition there appears to be a selection of histidine codons by the *fpg* gene as only the CAT codon is observed. Other genes responsible for DNA alkylation repair in *E. coli* the *ada*, *alk*A and *tag*A genes also are biased for the use of the CAT codon (75, 75 and 100% respectively) compared with the normal usage of the CAT histidine codon in other *E. coli* genes. This low sample of genes involved in alkylation repair and the large number of possible reasons for the conservation of the CAT codon for His do not allow us to speculate further on this observation.

An homology search between some other repair-associated proteins that included the gene products of *ada*, *alk*A and *tag*A from *E. coli* and *den*V from phage T4 and the Fapy-DNA glycosylase showed that there is only a limited degree of homology between the Fapy-DNA glycosylase and the *den*V gene product. Previously the *den*V gene product was postulated to have homology with the *tag*A gene product of *E. coli* (Steinum and Seeberg, 1986). The sequences of the *den*V (Valerie *et al.*, 1984) and *fpg* proteins have two homologous regions (Figure 5). The

regions have five and six homologous amino acids, and the codons are identical in 3/5 in the first and only 1/6 in the second region.

The enzyme isolated in this study excises the secondary alkylation product of 7-methylguanine: Fapy. This enzyme is unable to remove other alkylated bases such as *N3*-methyladenine, *N3*-methylguanine and *N7*-methylguanine. However, the ring-opened form of *N7*-methylguanine (Fapy) has not been observed in bacteria treated with alkylating agents, and its occurrence in the DNA of mammalian cells has been subject to debate (Beranek *et al.*, 1983; Den Engelse *et al.*, 1986). These observations could be due to either the fact that Fapy lesions never form *in vivo*, or that the lesion is formed and removed rapidly by the enzyme. In the first case the biological substrate of the enzyme would be different from the assay for its isolation, whereas the second possibility suggests that the wild-type strains release Fapy rapidly, as was found previously for 3-methyladenine lesions (Karran *et al.*, 1982; Boiteux *et al.*, 1984). However *in vitro* the homogeneous Fapy-DNA glycosylase from *E. coli* also removes a ring-opened form of adenine after X-ray irradiation of DNA polynucleotides (Breimer, 1984). Other results show that other ring-opened forms of guanines substituted at the *N7* position with a bulky adduct (phosphoramide mustards or aflatoxin-B_1) are also excised by the purified enzyme (Chetsanga *et al.*, 1982; Chetsanga and Frenette, 1983). These results suggest that the Fapy-DNA glycosylase may have a broad substrate specificity which includes DNA residues which have been altered by a variety of agents such as small and bulky alkylating agents, X-ray irradiation and oxidizing agents. Finally, the evaluation of the physiological role for the enzyme relies on the ability to isolate an enzyme-deficient mutant in *E. coli*. The isolation of the gene is a critical step in the construction of such a mutant.

Materials and methods

Bacterial strains, plasmids and phages

We used the following strains derived from the *E. coli* K12 strain: HB101 (*hsd*R *hsd*M, *rec*A13) (Maniatis *et al.*, 1982), JM103 (Δ*lac*−*pro*) *hsd*R17/F′*lac*I^q ZΔ*M*15) (Messing *et al.*, 1981) and HB1100 (*end*A) (Durwald and Hoffmann-Berling, 1968). The vector plasmids pBR322, pUC18 and pUC19 and the M13 mp18 and mp19 phages were derived from our laboratory stocks.

Chemicals

Enzymes used for DNA manipulations, IPTG (isopropyl-β-D-thiogalactopyranoside), X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside), nucleotides and poly(dG-dC)·poly(dG-dC) were purchased from Boehringer-Mannheim and used as recommended by the manufacturer. Fapy (2-6-diamino-4-hydroxy-5*N*-methylformamidopyrimidine) for use as an internal standard in h.p.l.c. was prepared as previously described (Boiteux *et al.*, 1984). The [^{3}H]dimethylsulphate was purchased from New England Nuclear. The α-dATP labelled with ^{32}P or ^{35}S was purchased from Amersham.

General methods for DNA manipulations

The following procedures were performed using standard methods as described by Maniatis *et al.*, 1982 and by the manufacturers: preparation of plasmids, restriction enzyme digestion, agarose gel electrophoresis, DNA ligation and bacterial transformation.

Preparation of E. coli genomic DNA library

High molecular weight genomic DNA purified from *E. coli* HB1100 was partially digested with the restriction endonuclease *Sau*3AI. DNA fragments of 10−15 kb were isolated following centrifugation through a neutral sucrose gradient and ligated into dephosphorylated, *Bam*HI-digested pBR322. Following transformation of *E. coli* strain HB101 ~800 ampicillin-resistant, tetracycline-sensitive colonies were isolated. Individual clones were grown in LB broth medium containing 50 μg/ml ampicillin overnight at 37°C and stored in 15% glycerol at −70°C.

Preparation of crude lysates

Two different procedures were used to prepare the crude lysates. The first method was used to screen the library and required only 2 ml of culture while the second method was used to quantify glycosylase activity and required 20 ml of culture.

(i) Screening of the library: individual clones were grown overnight at 37°C in 1 ml of LB broth medium containing 50 μg/ml ampicillin. Two millilitres of fresh medium was inoculated with 0.1 ml of the overnight cultures and grown at 37°C until the $OD^{1\,cm}_{600\,nm}$ = 2.0. The culture was centrifuged in an Eppendorf centrifuge for 2 min at 4°C. The pellet was then resuspended in 0.2 ml of lysis buffer (300 mM Tris−HCl, 5 mM EDTA, pH 8.0). The bacterial suspension was stored for 24 h at −70°C. The frozen suspension was thawed and 20 μl of a lysozyme solution (10 mg/ml) was added. This mixture was incubated at 37°C for 5 min and frozen at −70°C for 5 min. This procedure was repeated and the lysate centrifuged for 15 min at 4°C in an Eppendorf microfuge. The pellet was then removed using a toothpick, and the supernatant was stored on ice. This crude lysate was used to assay the enzyme activity in each of the recombinant colonies. (ii) Quantification of enzyme activity. Twenty millilitres of culture was grown under the same conditions as above and resuspended in 0.5 ml of lysis buffer. The lysis procedure is identical to that described above. The lysate was centrifuged at 30 000 r.p.m. in a Beckman Ti50 rotor for 30 min at 2°C. The enzyme activity in those extracts was then evaluated.

Preparation of [^{3}H]Fapy-poly(dG-dC)·poly(dG-dC)

The polynucleotide was alkylated as previously described (Laval, 1977) with 5 mCi of [^{3}H]dimethylsulphate (1.9 Ci/mmol). Under these conditions 0.5% of the guanine residues are methylated. The specific activity was 2×10^4 c.p.m./μg of polynucleotide. Imidazole ring-opening of the *N7*-methylguanine residues was obtained by incubating alkylated poly(dG-dC)·poly(dG-dC) at pH 11.4 for 48 h at 25°C (Boiteux *et al.*, 1984). The [^{3}H]Fapy-poly(dG-dC)·poly(dG-dC) product was subjected to formic acid hydrolysis (75% v/v formic acid, 17 h, 37°C). The analysis of the hydrolysed material by h.p.l.c. showed that 97% of the radioactivity eluted as Fapy.

Fapy-DNA glycosylase assay

The standard incubation mixture (total volume 50 μl) contained 70 mM Hepes−KOH, pH 7.6, 100 mM KCl, 2 mM Na_2EDTA, 10% glycerol, 2500 c.p.m. of [^{3}H]Fapy-poly(dg-dc)·poly(dg-dc) and 5 μl of crude lysate. The assay reaction mxiture was incubated for 10 min at 37°C. Following incubation 50 μl of BSA−calf thymus DNA solution and 300 μl of cold ethanol (−20°C) were added to the assay reaction mixture and centrifuged in an Eppendorf microfuge at 4°C for 10 min. The supernatant was transferred to a scintillation vial, 3 ml of Beckman Ready-Solv fluid was added and the radioactivity was quantitated by scintillation counting (Boiteux *et al.*, 1984). Definition of enzyme units: 1 unit = 1 pmol Fapy released in 5 min at 37°C under the assay conditions described above.

Identification of the products released by the extracts using h.p.l.c.

The ethanol-soluble radioactive material released by the cellular extract was dried under nitrogen and resuspended in 50 μl of water with marker molecules. This fraction was further analysed by h.p.l.c. chromatography using a C_{18} μBondapak column (Waters). The mobile phase was 25 mM $NH_4H_2PO_4$, pH 4.5, containing 5% methanol (v/v). The column was isocratically developed at 1.5 ml/min. The two rotamers of Fapy internal standards eluted as two peaks at 4 and 5 min (Boiteux *et al.*, 1984). Products were detected by monitoring u.v. absorption at 254 nm and by scintillation counting of fractions.

Overproduction of the fpg gene product

HB101 cells carrying various plasmids were grown in 20-ml cultures, and crude extracts were prepared as described above. The proteins in cell extracts were analysed by SDS−polyacrylamide gel electrophoresis (Laemmli, 1970).

Purification of the Fapy-DNA glycosylase

LB broth medium (1 l) containing 50 μg/ml of ampicillin was inoculated with 20 ml of an overnight culture of the *E. coli* K12 strain HB101 carrying the pFPG60 (fpg$^+$) plasmid. The bacteria were grown at 30°C for 2 h, then supplemented with 0.5 mM IPTG and the culture incubated for 17 h at 40°C under vigorous agitation. Cells were harvested (5 g wet weight), washed and resuspended in 30 ml of buffer containing 250 mM NaCl, 50 mM Hepes−KOH, pH 8.0, 2 mM Na_2EDTA, 5 mM dithiothreitol and 5% glycerol. The cell suspension was supplemented to 1 mg/ml with lysozyme, incubated at 37°C for 15 min and frozen at −70°C for 15 min; the freezing−thawing cycle was then repeated. The lysate was centrifuged at 35 000 r.p.m. in a TFT6538 Kontron rotor for 45 min at 4°C. The supernatant was taken as crude extract (Fraction I, 30 ml). Fraction I was loaded on a QMA Anion Exchange column (Waters Accell: 54 ml bed volume), washed and equilibrated with buffer A (10 mM Hepes−KOH, pH 7.4, 2 mM Na_2EDTA, 5 mM β-mercaptoethanol and 5% glycerol) containing 250 mM NaCl. The active fractions were not retained on the column but eluted with the bulk of proteins (Fraction II, 90 ml). Fraction II was dialysed against 2 l of buffer A and applied to a Phospho-Ultrogel column (IBF-LKB, 17 ml bed volume) equilibrated with buffer A containing 50 mM NaCl. After washing with buffer A containing 100 mM NaCl the column was eluted with a linear NaCl gradient (100−800 mM). The active fractions which eluted at 500−600 mM NaCl were pooled (Fraction III, 21 ml). Fraction III was precipitated by ammonium sulphate (Schwartz−Mann) to 75% saturation. The precipitate was collected by centrifugation. The pellet was resuspended in buffer A containing 1 M NaCl to yield a

final volume of 2 ml, loaded on a gel filtration column (ACA-54 IBF-LKB, 1.6 cm diameter × 150 cm) and eluted with the same buffer. The active fractions were pooled and dialysed against buffer A (Fraction IV, 22 ml). Fraction IV was loaded onto a cationic Mono S HR5/5 column (Pharmacia) equilibrated with buffer A containing 50 mM NaCl. The column was eluted with a linear NaCl gradient (100–800 mM). The active fractions eluted at 500 mM NaCl (Fraction V, 2 ml). Protein was quantitated as described by Bradford (1976).

Determination of DNA sequence

Plasmid pFPG60 has a 1.4-kb fragment of *E. coli* DNA which contains the *fpg* gene. The sequence was determined by subcloning the 1.4-kb fragment into the replicative form of phages M13 mp18 or mp19. In addition to subcloning the 1.4-kb fragment, a series of five primers were constructed. The 20-mer oligonucleotide primers were synthesized using an Applied Biosystems oligonucleotide synthesizer. The sequence was determined using the dideoxy method (Sanger *et al.*, 1980).

Amino acid composition and amino acid sequence

For the determination of amino acid composition the samples were hydrolysed at 110°C in 5.7 N HCl, 0.1% phenol in evacuated sealed tubes. Analyses were performed using an LKB 4400 amino acid analyser, operated at 0.1 OD full scale with a normal ninhydrin system as previously described (Lederer *et al.*, 1983). Manual sequencing was carried out using the DABITC (4-*N*·*N* dimethylamine benzene-4′-thiocyanate)/PITC (phenylisothiocyanate) method as described by Morel *et al.* (1984).

Computer search

Searches for open reading frames, protein homology and calculation of protein hydrophilicity were performed using BISANCE on the CITI2 system.

Acknowledgements

We are very grateful to Patricia Auffret Van der Kemp and Claudine Lagravère for their excellent technical assistance. We thank Dr Jacques Armier for the synthesis of the oligonucleotide primers, Dr Florence Lederer U25 INSERM for the amino acid composition and Dr Alain Morel from Groupe Neurobiochimique de l'Université Pierre et Marie Curie for the amino-terminal sequence. This work was supported by CNRS LA 147, the Institut National de la Santé et de la Recherche Médicale (INSERM) U 140 and by the Association pour la Recherche sur le Cancer (ARC). T.R.O'C. was a recipient of International fellowships from INSERM and ARC.

References

Beranek,D.T., Weiss,C.C., Evans,F.E., Chetsanga,C.J. and Kadlubar,F.F. (1983) *Biochem. Biophys. Res. Commun.*, **110**, 625–631.
Boiteux,S. and Laval,J. (1983) *Biochem. Biophys. Res. Commun.*, **110**, 552–558.
Boiteux,S., Huisman,O. and Laval,J. (1984) *EMBO J.*, **3**, 2569–2573.
Boiteux,S., Belleney,J., Roques,B.P. and Laval,J. (1984) *Nucleic Acids Res.*, **13**, 5429–5439.
Bradford,M.M. (1976) *Anal. Biochem.*, **72**, 248–254.
Breimer,L.H. (1984) *Nucleic Acids Res.*, **12**, 6359–6367.
Chetsanga,C.J. and Lindahl,T. (1979) *Nucleic Acids Res.*, **6**, 3673–3684.
Chetsanga,C.J., Lozon,C., Makaroff,C. and Savage,L. (1981) *Biochemistry*, **20**, 5201–5207.
Chetsanga,C.J., Polidori,G. and Mainwaring,M. (1982) *Cancer Res.*, **42**, 2616–2621.
Chetsanga,C.J. and Frenette,G. (1983) *Carcinogenesis*, **4**, 997–1000.
Demple,B., Sedgwick,B., Robins,P., Totty,N., Waterfield,M.D. and Lindahl,T. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 2688–2692.
Den Engelse,L., Jenner,A.A.J., Menkveld,G.J. and Scherer,I. (1986) In Myrnes,B. and Krokan,H. (eds), *Repair of DNA Lesions Introduced by N-Nitroso Compounds*. Norwegian University Press, pp. 257–283.
Dürwald,H. and Hoffmann-Berling,H. (1968) *J. Mol. Biol.*, **34**, 331–346.
Evensen,G. and Seeberg,E. (1982) *Nature*, **290**, 797–799.
Gilson,E., Rousset,J.P., Clement,J.M. and Hofnung,M. (1986) *Ann. Inst. Pasteur/Microbiol.*, **137B**, 259–270.
Haines,J.A., Reese,C.B. and Lord Todd (1962) *J. Chem. Soc.*, 5281–5288.
Harley,C.B. and Reynolds,R.P. (1987) *Nucleic Acids Res.*, **15**, 2343–2360.
Hudson,G.S. and Davidson,B.E. (1984) *J. Mol. Biol.*, **180**, 1023–1051.
Karran,P., Hjelmgren,T. and Lindahl,T. (1982) *Nature*, **296**, 770–773.
Kataoka,H., Yamamoto,Y. and Sekiguchi,M. (1983) *J. Bacteriol.*, **153**, 1301–1307.
Konigsberg,T. and Godson,N. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 687–691.
Laemmli,U.K. (1970) *Nature*, **227**, 680–685.
Lagravere,C., Malfroy,B., Leng,M. and Laval,J. (1985) *Nature*, **310**, 798–800.
Laval,J. (1977) *Nature*, **269**, 829–832.
Lawley,P.D. and Orr,D.J. (1970) *Chem.-Biol. Interactions*, **2**, 154–157.
Lederer,F., Ghrir,R., Guiard,B., Cortial,S. and Ito,A. (1983) *Eur. J. Biochem.*, **132**, 95–102.
Loechler,E.L., Green,C.L. and Essigman,J.M. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 6271–6275.
Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning: a Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Margison,G.P. and Pegg,A.E. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 861–865.
Messing,J.R., Crea,R. and Seeburg,P.H. (1981) *Nucleic Acids Res.*, **9**, 309–319.
Morel,A., Chang,J.Y., Cohen,P. (1984) *FEBS Lett.*, **75**, 21–24.
Nakabeppu,Y., Kondo,H. and Sekiguchi,M. (1984) *J. Biol. Chem.*, **259**, 13723–13729.
Nakabeppu,Y., Kondo,H., Kawabata,S.I., Iwanaga,S. and Sekiguchi,M. (1985) *J. Biol. Chem.*, **260**, 7281–7288.
Pribnow,D. (1975) *Proc. Natl. Acad. Sci. USA*, **72**, 784–786.
Robins,R.K. and Townsend,L.B. (1963) *J. Am. Chem. Soc.*, **85**, 252.
Sakumi,K., Nakabeppu,Y., Yamamoto,Y., Kawabata,S., Iwanaga,S. and Sekiguchi,M. (1986) *J. Biol. Chem.*, **261**, 15761–15766.
Sanger,F., Coulson,A.R., Barrell,B.E., Smith,A.J.H. and Roe,B.P. (1980) *J. Mol. Biol.*, **143**, 161–178.
Shine,J. and Dalgarno,L. (1975) *Nature*, **254**, 34–38.
Singer,B. and Grundberger,D. (1983) *Molecular Biology of Mutagens and Carcinogens*. Plenum, New York, pp. 55–78.
Steinum,A.L. and Seeberg,E. (1986) *Nucleic Acids Res.*, **14**, 3763–3772.
Valerie,K., Henderson,E.E. and de Riel,J.K. (1984) *Nucleic Acids Res.*, **12**, 8085–8096.
Yamamoto,Y., Katsuki,A., Sekiguchi,M. and Otsuji,N. (1978) *J. Bacteriol.*, **135**, 144–152.

Received on June 16, 1987; revised on July 14, 1987