Sequence Analysis, Expression, and Conservation of *Escherichia coli* Uracil DNA Glycosylase and Its Gene (ung)*

Umesh Varshney, Terry Hutcheon, and Johan H. van de Sande

From the Department of Medical Biochemistry, Faculty of Medicine, The University of Calgary, Calgary, Canada T2N 4N1

(Received for publication, December 14, 1987)

The complete nucleotide sequence of the *Escherichia coli* ung gene is described. Transcription initiation and termination sites were determined by S1 nuclease and RNase mapping. The common prokaryotic -35, -10, and the ribosome binding site sequences are represented by TTGTCCTGTA, TAAGCTA, and AGGAGAG at their respective locations. A putative hairpin transcription terminator structure is present at the major transcription terminator sites. The open reading frame of the ung gene codes for a protein of 229 amino acids (25,664 daltons). The molecular weight, amino acid composition, and the N-terminal amino acid sequence of the uracil DNA glycosylase purified from *E. coli* cells match with the open reading frame of the ung gene. The protein sequence analysis shows that the N-terminal methionine is cleaved off in the mature protein. The in *vitro* transcription coupled translation of the ung gene directs the synthesis of a protein which comigrates with uracil DNA glycosylase. Also, the CNBr cleavage of the protein synthesized in *vitro* confirms the positions of the methionines deduced from the DNA sequence. The levels of ung gene expression remain constant up to the early stationary phase, but decline in the late stationary phase of the *E. coli* culture. The *E. coli* gene showed a strong sequence homology to *Shigella*, a weak sequence homology to *Salmonella* and *Citrobacter*, and a very weak sequence homology to *Proteus* genes. No sequence homologies were seen for *Pseudomonas*, *Clostridium*, *Micrococcus*, and several eukaryotic genomes.

DNA glycosylases excise damaged or unconventional bases from DNA and initiate the DNA repair pathway. DNA glycosylases have been identified and purified from prokaryotic and eukaryotic sources (1, 2). These enzymes are present in all organisms examined so far with the possible exception of *Drosophila* where a search for several DNA glycosylases has failed (3, 4). Recently, however, at least two DNA glycosylases which excise oxidized thymine (6) and uracil residues have been identified in *Drosophila*.

Uracil DNA glycosylases are found to be the most abundant of all the glycosylases in the cell. Both a nuclear and an organellar uracil DNA glycosylase are found in eukaryotic cells (1, 2). Uracil DNA glycosylase excises uracil residues from the DNA which can arise as a result of either misincorporation of dUMP residues by DNA polymerase or due to deamination of cytosine. None of the uracil DNA glycosylases studied so far require metal ions for their activity and both single-stranded and double-stranded DNAs containing uracil are used as substrates. However, the enzyme from most sources has higher activity with single-stranded substrates. The tetramer (dU), is the shortest substrate for the bacterial enzymes. The uracil DNA glycosylases are product inhibited by free uracil and a limited number of its derivatives, for example, 6-aminouracil, 5-azauracil, and 5-fluouracil. The uracil DNA glycosylases have been shown to excise some of the analogues that are effective as inhibitors if incorporated into DNA. Another category of inhibitors of uracil DNA glycosylases is represented by *Bacillus subtilis* phage, PBS2, and *Escherichia coli* phage T5-induced proteins (6, 7). *E. coli* uracil DNA glycosylase has been purified to homogeneity (8). The enzyme is a single polypeptide monomer of about 25 kDa molecular mass and contains a single residue of cysteine which is not involved in the catalytic activity of the enzyme (2). The uracil DNA glycosylase (ung) gene of *E. coli* has also been cloned and overexpressed in *E. coli* (9). However, neither the amino acid sequence of the enzyme nor the nucleotide sequence of the ung gene has been reported.

In order to study the mechanism of action of uracil DNA glycosylase, the complete nucleotide sequence of the *E. coli* ung gene was determined and is presented in this paper. The open reading frame of the gene, confirmed by N-terminal protein sequence analysis, codes for a protein of 25,664 Da which contains a single cysteine residue towards the C terminus. Furthermore, the structure, expression, and some aspects of the ung gene conservation among other organisms are also reported.

**MATERIALS AND METHODS**

**Purification, Amino Acid Analysis, and N-terminal Sequence Analysis of Uracil DNA Glycosylase**

Uracil DNA glycosylase was purified to homogeneity from *E. coli* BD 438 exactly as described previously (8). The BD 438 strain of *E. coli* harbors the cloned uracil DNA glycosylase (ung) gene of *E. coli* (9) on a plasmid (pBD15) and was kindly provided to us by Dr. B. K. Duncan of Duncan Laboratories, Philadelphia, PA. Fluorescence loss assays (slightly modified from Ref. 10) were performed to follow the activity of uracil DNA glycosylase through various purification steps, whereas [1H]uracil release assays (8) were performed for the final purification analysis. Amino acid analysis of the pure protein was carried out on an automated Beckman 6300 amino acid analyzer following its hydrolysis in 6 N HCl at 110 °C for 24 h in evacuated, sealed tubes. Cysteine and tryptophan were not analyzed. Sequence analysis was carried out by an automated sequence analyzer (Applied Biosystems 477A).
Preparation of DNA

Genomic— Cultures of E. coli (wild type, HB101, JM109), Proteus vulgaris, Shigella sonnei, Citrobacter freundii, Salmonella typhimurium, and Pseudomonas aeruginosa were grown in LB broth (11). Cells from log phase cultures (30 ml) were harvested by centrifugation and treated with 4 mg of lysosome in 4 ml of 50 mM Tris-HCl (pH 7.8), 20 mM Na,EDTA, 5% sucrose, 0.5% Triton X-100 (STEB buffer) for 5 min at room temperature. The cells were lysed by the addition of 1% SDS and digested with proteinase K (50 µg/ml) at 60°C for 2 h. The DNA was precipitated with 1 volume of ice-chilled isopropanol alcohol in the presence of 0.3 M sodium acetate, spotted out, washed with 70% ethanol, resuspended in 4 ml of 50 mM Tris-HCl (pH 7.8), 5 mM Na,EDTA, and digested with RNase A (50 µg/ml) at 37°C for one to two h. The DNA was further purified by phenol/chloroform extractions, recovered by ethanol precipitations (11), and resuspended in 1 ml of 10 mM Tris-HCl, 1 mM Na,EDTA (pH 7.8).

Genomic DNAs of Clostridium perfringens and Micrococcus lysodeikticus were purchased from Sigma and further purified by phenol/ chloroform extractions and ethanol precipitations. The DNA was resuspended in 10 mM Tris-HCl, 1 mM Na,EDTA (pH 7.8) at a concentration of approximately 1 mg/ml.

Plasmid DNA—Plasmid DNAs were prepared from overnight cultures of the plasmid harboring E. coli cells using the boiling method of lysis (12) and further purified by CsCl-ethidium bromide gradient centrifugation (13).

Subcloning and DNA Sequence Analysis

A 1.4-kb HpaI fragment (size revised to 1.5 kb following sequence analysis) containing the ung gene (9) was isolated and subcloned into the Smal site of a multifunctional vector pTZ19R (Pharmacia LKB Biotechnology Inc.) in clockwise and anticlockwise orientations, pTZUng2 and pTZUng4, respectively. Further constructs used in these studies were described in detail in Fig. 5. Standard recombinant DNA techniques were used throughout the whole procedure (11).

DNA sequence was obtained from both strands of DNA by a modified enzymatic chain termination method (14) and chemical degradation method (15).

Synthesis of RNA Probes

RNA probes were prepared by carrying out in vitro transcriptions in the presence of [α-32P]UTP. The vector used in these studies, pTZ19R, contains a T7 phage promoter. The transcription reactions (16) with some modifications were performed as follows. About 1 µg of the linearized plasmid DNA was incubated in a 25-µl reaction consisting of 40 mM Tris-HCl (pH 7.5), 6 mM MgCl2, 2 mM spermidine, 10 mM dithiothreitol, 20 units of RNA guard (Pharmacia), 500 µM each of ATP, GTP, and CTP, 100 µM of UTP, 50 µCi of [α-32P]UTP (AmerhamCorp., specific activity approximately 2000 Ci/mmol), and 7 units of T7 RNA polymerase (Pharmacia) at 38°C for 1 h. The DNA template was digested by incubating with 10 units of RNase-free DNase I (Pharmacia) at 37°C for 10 min. The reaction was then treated with proteinase K (50 µg/ml) in the presence of 0.5% SDS at 45°C for 30 min, extracted once with phenol, and fractionated on a G-75 column to separate the unincorporated nucleotides. The RNA probe was recovered by ethanol precipitation in the presence of carrier yeast total RNA (50 µg/ml) and resuspended in 1 ml of 10 mM Tris-HCl, 1 mM Na,EDTA (pH 7.8).

RNA was recovered by ethanol precipitations at different times during the growth cycle as total nucleic acids. Cultures were chilled in ice and centrifuged at 5000 × g for 5 min. The pellet was resuspended in 0.1 ml of the STEB buffer and quickly lysed by the addition of 0.4 ml of 7 M urea, 2% SDS, 10 mM Tris-HCl (pH 7.5), 0.15 M NaCl, and 1 mM Na,EDTA at 60°C. The total nucleic acids were then vigorously extracted three times in H2O-saturated phenol at 60°C, followed by two to three extractions with chloroform/isomyl alcohol (24:1). The nucleic acids were ethanol precipitated, washed with 70% ethanol, and resuspended in 0.5 ml of H2O.

S1 Nuclease Mapping

Transcription Initiation Site—Plasmid pTZUng4B was digested with BamHI and 5′-end labeled with T4 polynucleotide kinase (19) and [γ-32P]ATP (Du Pont-New England Nuclear, specific activity approximately 7000 Ci/mmol). Secondary digestion with Drai yielded a 0.38-kb fragment (Drai-BamHI, see Fig. 2A), which was eluted from a 5% polyacrylamide gel by the diffusion method (11). This fragment, 5′-end labeled, contained the 5′-flanking region, promoter region, and part of the coding region sequences. The labeled fragment was ethanol precipitated with carrier yeast total RNA and used as a probe for S1 nuclease mapping to determine the transcription initiation site (20). Aliquots (250,000 cpn) of this probe were lyophilized with 50 µg each of the total nucleic acids extracted from log phase cultures of E. coli HB101 (wild type) and E. coli JM109 harboring the plasmid pTZUng2. The nucleic acids were resuspended in 25 µl of hybridization buffer (80% formamide, 400 mM NaCl, 40 mM PIPES, 1 mM Na,EDTA, pH 6.5), overlaid with mineral oil, denatured at 85°C for 5 min, and hybridized at 49°C for about 16 h. At the end of hybridization, the reaction was diluted with 400 µl of the ice-chilled S1 nuclease buffer (0.1 M NaC1, 30 mM sodium acetate, pH 4.6, 1.6 mM Mn,SO4). Aliquots (200 µl) were digested with 250 or 500 units of S1 nuclease (Pharmacia) at 37°C for 45 min. The reaction was stopped by phenol/ chloroform/isomyl alcohol (25:24:1, v/v) extraction and the nucleic acids were recovered by precipitation with 2.5 volumes of ethanol in the presence of carrier yeast total RNA (50 µg/ml) and resuspended in 10 µl of sequencing dye. Aliquots (6 µl) were analyzed on 6% polyacrylamide, 8 M urea gels. Chemical sequencing reactions (A + G and C + T) of the probe were used as markers (15).

Transcription Termination Site—Plasmid pTZUng2B was digested with BamHI and 3′-end labeled by using T4 DNA polymerase (Pharmacia) and [γ-32P]dCTP (Amersham, specific activity approximately 3000 Ci/mmol). Secondary digestion with EcoRI resulted in a 0.75-kb fragment which was 3′-end labeled on the antisense strand in the presence of [γ-32P]dCTP. The labeled fragment was purified by polyacrylamide gel and used for S1 nuclease mapping as described above except that the hybridization was performed at 56°C. Markers were obtained by 3′-end labeling of the MapI-digested pBR322 using T4 DNA polymerase and [γ-32P]dCTP.

RNase Mapping

An antisense RNA probe for the RNase protection experiments was derived by in vitro transcriptions of pTZUng4, linearized with EcoRI. About 250,000 cpn of the probe were lyophilized with 25 µg of total nucleic acids, resuspended in 15 µl of hybridization buffer, and overlaid with mineral oil as above. The nucleic acids were hybridized at 56°C for 16 h without any prior denaturation step. After hybridization, the reaction was digested with 200 µl of 30 mM Tris-HCl (pH 8.0), 375 mM NaCl, 2 mM Na,EDTA, pH 8.0, and digested with 25 µg of RNase A and 25 units of RNase T1 at 34°C for 1 h. The contents were then digested with proteinase K (50 µg/ml) for 30 min at 45°C in the presence of 0.5% SDS, extracted with phenol, and precipitated with 2.5 volumes of ethanol. Nucleic acids were resuspended in 10 µl of an 80% formamide dye mixture and 0.5 µl aliquots were loaded onto 5% polyacrylamide, 8 M urea gels. The MapI- or XhoII-digested pBR322 was 3′-end labeled using T4 DNA polymerase and [γ-32P]dCTP and used as size markers.

Southern Blot Analysis

Genomic DNAs (0.5 µg each) were cleaved with restriction endonucleases according to the suppliers recommendations and electrophoresed on a horizontal and submersed 0.9% agarose gel using TAE electrophoresis buffer (40 mM Tris-HCl, 2 mM acetic acid, 2 mM Na,EDTA, 1 µg/ml EtBr, pH 8.1). The DNA was transferred onto a nylon membrane (Zeta probe, Bio-Rad) using 0.4 M NaOH as the transfer buffer (17).

The hybridizations were performed essentially as described previously (18) except that the 10× Denhardt's solution was replaced with 1% Carnation low fat dairy milk powder during prehybridization and by 0.6% during hybridization and the first post-hybridization wash. The SDS concentration was increased to 0.5% throughout and the hybridization and the washings were carried out at 68°C in the absence of any formamide. The probe for hybridization was derived by in vitro transcription of Drai linearized pTZUng4S (Fig. 5).

Preparation of RNA

RNA was prepared from 10-ml cultures at different times during the growth cycle as total nucleic acids. Cultures were chilled in ice and centrifuged at 5000×g for 5 min. The pellet was resuspended in 0.1 ml of the STEB buffer and quickly lysed by the addition of 0.4 ml of 7 M urea, 2% SDS, 10 mM Tris-HCl (pH 7.5), 0.15 M NaCl, and 1 mM Na,EDTA at 60°C. The total nucleic acids were then vigorously extracted three times in H2O-saturated phenol at 60°C, followed by two to three extractions with chloroform/isomyl alcohol (24:1). The nucleic acids were ethanol precipitated, washed with 70% ethanol, and resuspended in 0.5 ml of H2O.

The abbreviations used are: SDS-PAGE, sodium dodecyl sulfate-polyacrylamide gel electrophoresis; kb, kilobase pair; PIPES, 1,4-piperazinediethanesulfonic acid.
**In Vitro Transcription Coupled Translations**

The S-30 extracts of *E. coli* (21, 22) were purchased from Amer- sham. Reactions (5 μl) containing 0.5 μg of circular or linearized plasmid DNA were carried out in the presence of [35S]methionine according to the supplier’s recommendations. The reaction was then diluted to 25 μl with H2O and 2-μl aliquots were analyzed on 15% SDS-polyacrylamide gels (SDS-PAGE; Ref. 23). Proteins were detected by autoradiography with or without fluorography (24).

For CNBr digestion, a 20-μl reaction, containing 2.5 μg of circular plasmid pTZUng2 was carried out in the presence of 60 pCi of [3H]Leu. At the end, the reaction was passed through a 0.5-m1 hydroxylapatite (Bio-Rad) column, using 12 mM P3, 200 mM KCl, 1 mM dithiothreitol (pH 7.4) buffer. The flow through was mixed with 50 μg of carrier bovine serum albumin and dialyzed against 10 mM Tris-HCl (pH 7.5), 60 mM NaCl, 1 mM Na2EDTA. Aliquots containing about 5 μg of bovine serum albumin were lyophilized and resuspended in 50 μl of 70% formic acid containing 20 mM CNBr (25). The reaction was incubated at room temperature for 24 h. The contents were recovered by freeze-drying and analyzed on a 17% SDS-PAGE (23). The gels were fluorographed by using Amplify (Amersham) and autoradiographed using Kodak X-AR films at -70 °C.

**RESULTS**

**Purification of Uracil DNA Glycosylase**

Uracil DNA glycosylase was purified (approximately 2500-fold) to homogeneity from an overproducing strain (BD438) of *E. coli* by the method of Lindahl et al. (8). The purity of the protein was shown by analyzing a 2-μg aliquot on a 15% SDS-PAGE (23). Coomassie Blue staining of the gel showed only a single band represented by uracil DNA glycosylase (lane 1, Fig. 1A). The purified protein comigrates (lanes 1 and 3, Fig. 1B; lanes 14 and 15, Fig. 6) with bovine chymotrypsinogen A of M, = 25,666 (26). Uracil DNA glycosylase from the same preparation was occasionally seen to migrate as a doublet (lane 1, Fig. 1B) which was also previously observed (8). The amino acid analysis of these two bands following their electroelution from the gel was found to be identical (data not shown). It is not clear at present why the purified protein occasionally runs as a doublet.

**Structure and Expression of *E. coli ung* Gene**

**DNA Sequence Analysis**—The restriction map of the HpaI fragment containing the *E. coli ung* gene (9) is presented in Fig. 2A. The DNA sequence was determined from both strands as described under “Materials and Methods” and is shown in Fig. 2B. Only one open reading frame was found which could code for a polypeptide of the molecular weight corresponding to that of the purified enzyme. The reading frame starts with an ATG (Met) at position 533 and ends at position 1,219, which is then followed by a termination codon TAA for a total of 229 amino acids giving a protein of the expected molecular weight of 25,664. The initiation codon is preceded by a polyurine sequence 5'AGGAGAG3', starting at position 522 and ending at 528. This sequence, known as Shine-Dalgarno sequence, is a common feature of most prokaryotic genes (27) and represents the ribosome binding site in mRNA. Sequences homologous to the -10 and -35 prokaryotic promoter regions (28) are also seen in the ung gene and are shown in Fig. 2B. The -35 region represented by 5'TGTTCTGTA3' starts at position 476 and ends at position 484. The -10 region is represented by 5'TAAGCTA3', starting at position 497 and ending at 503. Towards the 3'-end, a putative stem and loop structure is present between positions 1,237 and 1,262 (Fig. 2B). The stem region of this structure consists of 8 G + C, 2 A + T, and 1 mismatch base pairs and 4 nucleotides in the loop region (estimated ΔG ∼ -29 kcal). The stem region is followed by an A + T-rich region. Out of 10 nucleotides there are 5 Thd, 3 Ado, and 1 each of Guo and Cyd. These observations strongly suggest that the putative stem and loop structure may represent the transcriptional terminator. If these analyses based on the DNA sequence are correct, S1 nuclease protection mapping must locate the transcription initiation site upstream but near to the ribosome binding site; also the transcription terminator site should be very near to the putative stem and loop structure found in the 3' region of the sequence.

**Transcription Initiation Site**—The DraI-BamHI DNA fragment (Fig. 2A), 5'-end labeled at the BamHI site, was prepared as described under “Materials and Methods.” The end labeled DNA strand represents the antisense sequence between positions 343 and 725 of Fig. 2B. This probe, containing the 5'-flanking and part of the coding region sequences was hybridized to the ung mRNA and digested with S1 nuclease. The length of the protected DNA fragment was determined by its electrophoresis on a 6% polyacrylamide, 8 m urea gel (Fig. 3A). Sequence markers G + A and C + T obtained by chemical degradation (15) of two separate aliquots of the probe were also run on the same gel. S1 nuclease protections were performed by using total nucleic acids from *E. coli* HB101 (wild type for ung) (Fig. 3A, lane 2), *E. coli* JM109 containing...
E. coli Uracil DNA Glycosylase

FIG. 3. S1 nuclease protections. A, determination of the transcription initiation site. The 0.38 kb long DraI-BamHI DNA fragment 5'-end labeled at the BamHI site of the antisense strand was hybridized to the total nucleic acids and digested with 250 units of S1 nuclease. The S1 nuclease protected material was recovered by ethanol precipitation and analyzed on sequencing gels. Lanes A and C + T represent sequencing markers from two separate aliquots of the probe. Results of S1 nuclease protection using total nucleic acids from E. coli HB101, wild type for ung (lane 2), E. coli JM109, harboring pTZUng2 (lane 3), and a heterologous source, mouse LMTK- cells (lane 4) are shown. Lane 1 is a control where the treatment was the same as in lane 2, except that it was not digested with S1 nuclease. The A marked with an asterisk represents the limit of the protected fragment. This corresponds to nucleotide T at position 517 in Fig. 2B. B, determination of transcription termination site. A BamHI to EcoRI fragment (see pTZUng2 Fig. 5) 3'-end labeled at the BamHI site of the antisense strand was used as a probe. Details and descriptions of lanes 1-4 are the same as in A. M represents markers obtained by 3'-end labeling of MspI-digested pBR322 fragments with [α-32P]dCTP and T4 DNA polymerase. Sizes of the fragments (in nucleotides) are as shown.

FIG. 2. A, restriction map and the sequence analysis strategies of the HpaII fragment containing the ung gene. Restriction sites are shown by symbols as follows: DraI (●); AcaII (O); BamHI (●); BglII (V); BstNI (●); HincII (V); HpaI (O); MstI (■); NruI (V); Sau96I (ß); and SphI (▲). Presence of BclI, MstI, and Sau96I was determined from sequence analysis. The numbers represent the size in base pairs. The arrows indicate direction and extent of sequence analysis from the site of its origin. Straight line arrows indicate sequence analysis by the chemical method, whereas wavy arrows indicate sequence analysis by the enzymatic method. B, sequence of the E. coli uracil DNA glycosylase (ung) gene. The nucleotide sequence and the amino acid sequence coded by it are numbered at the right end of each line. The -35, -10, ribosome binding site (r.b.s.), and a putative transcriptional terminator structure sequences are shown in bold face letters. The nucleotide marked < or > indicate those that can base pair and the nucleotides marked with a dot indicate mismatches in the stem of the transcriptional terminator.

pTZUng2 (Fig. 3A, lane 3), and mouse LMTK- cells (Fig. 3A, lane 4). The major protected bands in lanes 2 and 3 (Fig. 3A) comigrate and their lengths correspond to an Ado in the sequence shown in Fig. 3A. This Ado in turn corresponds to the nucleotide T at position 517 in Fig. 2B. B, determination of transcription termination site. A BamHI to EcoRI fragment (see pTZUng2 Fig. 5) 3'-end labeled at the BamHI site of the antisense strand was used as a probe. Details and descriptions of lanes 1-4 are the same as in A. M represents markers obtained by 3'-end labeling of MspI-digested pBR322 fragments with [α-32P]dCTP and T4 DNA polymerase. Sizes of the fragments (in nucleotides) are as shown.
bacteria. In another control (lane 1, Fig. 3A) where S1 nuclease was not added, no endogenous degradation of the probe is seen.

These results are in agreement with the predictions from the DNA sequence analysis and further confirm the assignment of the -10 and -35 regions of the ung gene (Fig. 2B).

Determination of the 3′ Termini of the ung mRNA—The 3′-end of ung mRNA was also mapped by S1 nuclease protection. The antisense strand of the BamHI-EcoRI fragment was 3′-end labeled at the BamHI site and used as a probe. This probe, containing the coding region, the 3′-flanking region sequences, and a few nucleotides of the multiple cloning site of the vector (see Fig. 5, construct pTZUng2) was prepared and hybridized as described under “Materials and Methods.” The lengths of the protected fragments represent the 3′ termini of the mRNA. Fragments in the major region of protection are about 530–535 nucleotides long (Fig. 3B), placing the 3′-end of the ung mRNA about 530 nucleotides downstream of the BamHI site. This corresponds to the region around position 1260, marked with vertical arrows in Fig. 2B.

Unlike the S1 nuclease mapping for the 5′ analysis where a fragment of fixed length represents the major transcriptional initiation site, the region of protection in 3′ analysis consists of a number of fragments varying in sizes from approximately 530–535 nucleotides. This suggests that the 3′-end of the ung mRNA is somewhat heterogeneous. Also after prolonged autoradiography, additional bands of sizes longer than 600 nucleotides were seen, suggesting some degree of run through transcription.

RNase Protection—From S1 nuclease protection experiments, transcription initiation and termination sites are placed at positions 517 and around 1260, respectively. This suggests that the total size of the ung mRNA is approximately 0.74 kb. This was confirmed by RNase protection experiments. An antisense RNA probe encompassing the whole sequence (Fig. 2B) was synthesized using T7 RNA polymerase by in vitro transcription of EcoRI-linearized pTZUng4. Following its hybridization to the ung mRNA it was digested with a mixture of RNase A and RNase T1. The size of the protected RNA probe, determined on a denaturing 6% polyacrylamide, 8 M urea gel corresponds to about 775 nucleotides, as compared to the DNA markers (Fig. 4). However, since RNA migrates slower than DNA of the same size, the size of the protected RNA probe actually corresponds to about 0.74 kb, as predicted from the S1 nuclease protection experiments. This experiment also demonstrated the relative abundance of the ung mRNA at different stages of cell culture growth (2–11 h). The amount of ung mRNA in the cells remains constant during the log phase and up to the early stationary phase at approximately 9 h (Fig. 4). A decline in the amount of ung mRNA, late in the stationary phase, is likely due to cell death. The minor amounts of ung mRNA detected in an ung− strain of E. coli during its log phase (lane ung− 5 h) correspond to the low levels of uracil DNA glycosylase activity detected in the cellular extracts of this mutant (data not shown). As expected, no protection of the RNA probe was seen when total nucleic acids from a heterogeneous source, mouse LMTK− cell line, were used (Fig. 4, lane LMTK−).

In Vitro Transcription Coupled Translations

These studies confirm the conclusions made from the sequence analysis of the ung gene and show that the product synthesized in vitro of the ung gene corresponds to the uracil DNA glycosylase purified from E. coli. Various constructs of the plasmids used for these studies are outlined in Fig. 5. The proteins synthesized in vitro were analyzed on SDS-PAGE and visualized by autoradiography. Fig. 6 shows the results of these analyses. When S-30 extracts are incubated in the absence of any DNA (lanes 7 and 9) only one major polypeptide of approximately 60 kDa is seen. This polypeptide seems to be a result of endogenous labeling (21). In another control, pTZUng4B linearized with restriction endonuclease DraI was hybridized with total nucleic acids extracted from E. coli HB101 (wild type for ung) at different times (2.5–11 h) or from an ung− strain of E. coli at 5 h. Hybrids were treated with RNase A and T1 and analyzed on sequence analysis gels. LMTK− shows a control where total nucleic acids from LMTK− cells were used. Lane C (−RNase) is another control where the treatment was the same as in the lane marked 5 h except that it was not treated with RNases. Lanes M1 and M2 show the markers obtained by 3′-end labeling of XhoII and MspI fragments of pBR322, respectively, by [α-32P]dCTP and T4 DNA polymerase. Sizes of the fragments (in nucleotides) are as shown.

**Fig. 4. RNase protections.** An antisense RNA probe, encompassing the whole HpaI fragment sequence was prepared by in vitro transcription using EcoRI-linearized pTZUng4 and hybridized to total nucleic acids extracted from E. coli HB101 (wild type for ung) at different times (2.5–11 h) or from an ung− strain of E. coli at 5 h. Hybrids were treated with RNase A and T1 and analyzed on sequence analysis gels. LMTK− shows a control where total nucleic acids from LMTK− cells were used. Lane C (−RNase) is another control where the treatment was the same as in the lane marked 5 h except that it was not treated with RNases. Lanes M1 and M2 show the markers obtained by 3′-end labeling of XhoII and MspI fragments of pBR322, respectively, by [α-32P]dCTP and T4 DNA polymerase. Sizes of the fragments (in nucleotides) are as shown.
Fig. 5. Subcloning of the ung gene and various plasmid constructs. The HpaI fragment shown on top was isolated from plasmid pRD15 (9) and subcloned into the SmaI site of pTZ19R in both orientations, clockwise pTZUng2 and anticlockwise pTZUng4. Further constructions were done as follows: pTZUng2B and pTZUng4B were obtained by a BamHI digestion, followed by religation of their parent plasmids pTZUng2 and pTZUng4, respectively. Similarly, pTZUng2S and pTZUng4S were constructed by an SphI digestion, followed by religation of the parent plasmids. Important restriction sites are shown. Blocked arrows indicate the location of different genes: e.g. uracil DNA glycosylase (ung), β-lactamase (β-lact or Amp'), and β-galactosidase Z' peptide (lac Z'). The letter P with a subscript represents the promotor of the gene, represented by a subscript. The number with brackets where shown represents base pair numbering. Sizes of different plasmids are shown in kilobase pairs.

Fig. 6. SDS-PAGE of the transcription coupled translation experiments. Each lane is labeled with the plasmid used to direct protein synthesis using S-30 extracts. Plasmid names, when followed by (L) indicate that these plasmids were first linearized with restriction endonuclease DraI and then used for directing protein synthesis. Lanes 1-7, 9-13, 17, and 18 were visualized following autoradiography. The contents of lanes 8-16 were electrophoresed on the same gel and the visualization of lanes 8, 14, 15, and 16 was done by Coomassie Blue staining. Since the translation products could only be visualized by autoradiography, the autoradiograph and the Coomassie Blue-stained gel were carefully aligned with the help of markers on both sides (lanes 8 and 16) to arrive at a composite picture (lanes 8-16). Description of lanes 14 and 15 is the same as for lanes 1 and 3, respectively, of Fig. 1B. MWM represents the molecular weight markers. The numbers show the sizes of the molecular mass markers in kilodaltons. The arrows indicate the ung protein and a solid triangle represents the hybrid protein directed by pTZUng2B (see "Results").

glycosylase and bovine chymotrypsinogen A were co-electrophoresed with these reactions. The products of pAT153 (lane 10) comigrate with the putative β-lactamase gene products of pTZUng2 (lane 13) and, the ung gene product (indicated by an arrow) comigrates with the purified uracil DNA glycosylase (lane 14) and bovine chymotrypsinogen A (lane 15). Therefore, the ung gene directs the synthesis of a protein under in vitro conditions which corresponds to the uracil DNA glyco-
E. coli Uracil DNA Glycosylase

**Amino Acid Composition and N-terminal Sequence Analysis of the Uracil DNA Glycosylase**

Amino acid composition of the purified protein was determined by acid hydrolysis of the purified protein (Table 1). Also shown are the amino acid compositions of uracil DNA glycosylase determined in an earlier study (8) and the amino acid composition expected from the gene sequence. The values obtained by acid hydrolysis of the protein agree with those determined from the DNA sequence, supporting the assignment of the open reading frame of the ung gene.

The amino acid sequence of 31 N-terminal amino acid residues is shown in Fig. 7. The first amino acid residue was Ala (which is the second amino acid from the DNA sequence analysis, the first being Met). This suggests that in the mature protein Met is cleaved off. The remaining amino acid sequence matches completely with the sequence deduced from the DNA sequence, confirming the start of the open reading frame.

**Conservation of E. coli Uracil DNA Glycosylase Gene Sequence**

Uracil DNA glycosylase is an enzyme of wide occurrence among prokaryotic and eukaryotic organisms. Therefore, the sequence relatedness of the E. coli gene to various bacterial and eukaryotic genes was determined. Genomic blots of EcoRI as well as HindIII digests were hybridized to an RNA probe prepared by in vitro transcription of the DraI linearized pTZung2 (see Fig. 5). This RNA probe covers most of the coding region and part of the 5' flanking region. Results of the hybridization of the prokaryotic genomic blots are shown in Fig. 8. The Shigella gene showed a strong homology to the E. coli gene, but the hybridization signals for Citrobacter and Salmonella were weakly homologous. Proteus genome also showed a weak hybridization signal but only after prolonged autoradiographic exposures of 3-4 weeks (data not shown).

**DISCUSSION**

The complete DNA sequence of the HpaI fragment possessing the E. coli ung gene (9) is reported. The 5' and the 3' boundaries of the ung mRNA were determined by S1 nuclease mapping. The antisense DNA strands 5' or 3'-end labeled at the BamHI site were used to locate the transcriptional initiation and termination sites of the ung mRNA at positions 517 and approximately 1260, respectively. The size of the ung mRNA was thus estimated to be approximately 0.74 kb. This was further confirmed by RNase protection experiments where an antisense RNA encompassing the whole sequence of the HpaI fragment was used as a probe. RNase A and T1 digestion of the hybrids of this probe with ung mRNA resulted in protection of an approximately 0.74-kb RNA fragment. Localization of the mRNA sequences also simplified the assignment of the regulatory regions of the ung gene. The ribosome binding site, −10 region, and the −35 region sequences represented by −AGGAGA-, TAAGCTA-, −GCCGGATGCCAGATTGCCACCCGGC−, etc., under less stringent conditions of hybridization and washings (55 °C) showed no sequence homologies.

**Table 1**

| Amino Acid | Composition from gene sequence | Composition from protein hydrolysis |
|------------|--------------------------------|------------------------------------|
| Ala        | 17 (7.4)                       | 17 (7.4)                           |
| Arg        | 10 (4.4)                       | 9.4 (4.1)                          |
| Asn        | 10 (4.4)                       | 9.2 (4.9)                          |
| Asp        | 17 (7.5)*                      | 18.3 (8.0)*                        |
| Cys        | 7 (3.1)                        | 10.0 (4.6)                         |
| Gin        | 14 (6.1)                       | ND*                                |
| Glu        | 28 (12.2)                      | 31.1 (13.5)                        |
| Gly        | 18 (7.9)                       | 23.1 (10.0)                        |
| His        | 13 (5.7)                       | 10.7 (4.7)                         |
| Ile        | 10 (4.4)                       | 8.7 (3.8)                          |
| Leu        | 24 (10.5)                      | 23.1 (10.1)                        |
| Lys        | 9 (3.9)                        | 12.0 (4.7)                         |
| Met        | 3 (1.3)                        | 1.9 (0.8)                          |
| Phe        | 11 (4.8)                       | 9.1 (4.0)                          |
| Pro        | 16 (7.0)                       | 16.2 (7.1)                         |
| Ser        | 12 (5.2)                       | 11.8 (5.2)                         |
| Thr        | 12 (5.2)                       | 10.4 (4.5)                         |
| Thr        | 6 (2.6)                        | ND                                 |
| Tyr        | 5 (2.2)                        | 4.8 (2.1)                          |
| Val        | 17 (7.4)                       | 15.6 (6.8)                         |
| Total      | 229                            | 221.5                              |

* Amino acid analysis was performed as described under "Materials and Methods." Values for cysteine and tryptophan were not determined. The percent values were calculated based on a total of 229 amino acids.
* Amino acid analysis from Ref. 8. The percent values were determined based on a total of 185.6 amino acids.
* Denotes combined values for Asn and Asp or Gin and Glu.
* ND, not determined.

Pseudomonas, Clostridium, and Micrococcus did not show any homology under these conditions of hybridizations. A genomic blot analysis of a number of eukaryotic genomic DNAs (Neurospora, yeast, Bombyx mori, sea urchin, mouse, human, cabbage, etc.) under less stringent conditions of hybridization and washings (55 °C) showed no sequence homologies.
FIG. 7. N-terminal amino acid sequence of the uracil DNA glycosylase. PTH, phenylthiohydantoin.

FIG. 8. Study of sequence relatedness of the E. coli ung gene to those of other bacterial genomes. Genomic DNAs (0.5 μg) of each bacteria were digested with EcoRI or HindIII (as shown), electrophoresed on agarose gels, and transferred to a nylon membrane (Zeta probe, Bio-Rad). The genomic blot was hybridized to a radioactive RNA probe derived from DraI linearized pTZUng4S and washed as described under "Materials and Methods." Results obtained by autoradiography are as shown. Each lane is labeled by the bacteria from which the DNA was used. The molecular weight markers (in kb) are from λ DNA digested with HindIII.

FIG. 9. Hydrophobicity plots for ung, tag, and alkA coded DNA glycosylases. Sequences of tag and alkA were from Refs. 33 and 34. The plots were generated by a computer program based on the calculations of Kyte and Doolittle (35) using a window of 7 amino acids. (N) shows the numbers of the amino acid residues in the sequence.

in the loop. This putative transcriptional terminator is followed by an A + T-rich region. The ends of the mRNA transcripts also map near this structure.

Expression of the ung gene was studied both under in vivo and in vitro conditions. The relative abundance of the ung mRNA, determined by RNase mapping experiments at various stages of the growth suggest that there are no major changes in the level of ung gene expression up to late log or early stationary phases. Later in the stationary phase, ung mRNA levels decline presumably because of cell death.

When the in vitro synthesis of proteins in an S-50 extract was directed by plasmids pTZUng2 and pTZUng4, which contain the ung gene in opposite orientations (Fig. 5), synthesis of the same ung gene product is seen suggesting that its synthesis was directed by its own promoter. The ung gene product was also synthesized when these plasmids were linearized with restriction endonuclease DraI prior to their incubation in the extract. Since a DraI site is present at position 342 of the sequence shown in Fig. 2, the promoter of the ung gene must be located downstream of this restriction site, which agrees with the results of sequence analysis showing the presence of the -35 and -10 regions between positions 476 and 503 (Fig. 2B). In vitro transcription coupled translation studies were also carried out to identify its gene product as well as the relative strength of the ung gene promoter. As expected for a protein of minor abundance, the promoter of the ung gene is not a strong promoter. This is evident from the relative levels of β-lactamase and ung proteins (Fig. 6), where an intact plasmid has been used to direct protein synthesis. In most cases, synthesis of the ung protein is less than that of β-lactamase, however, in the case of pBD15, synthesis of the ung gene product is more than that of the β-lactamase. This is because the plasmid pBD15 is an overproducing plasmid which contains the ung gene downstream of a λ left promotor (9).
The open reading frame of the ung gene and thus the amino acid sequence of the ung protein has been confirmed through several lines of evidence. The molecular weight (25,664) and the amino acid composition of the protein coded by the open reading frame correspond exactly to those determined directly from the uracil DNA glycosylase purified from *E. coli* (Fig. 1 and Table I). The ung gene product synthesized in *vitro* comigrates with the purified uracil DNA glycosylase and has the same elution characteristics from hydroxylapatite. In addition, the CNBr digestion of the ung gene product gives rise to two major polypeptides of 14 and 10 kDa which corresponds to cleavage at the positions of methionine deduced from the gene sequence (Fig. 6). Finally, the N-terminal amino acid sequence of the first 31 residues of the uracil DNA glycosylase (Fig. 7) completely matches with the open reading frame of the gene. The N-terminal sequence analysis also revealed that the N-terminal methionine is cleaved off in the mature uracil DNA glycosylase, a phenomenon common to many prokaryotic proteins (31).

Relatedness of the *E. coli* ung gene sequences to other bacterial and eukaryotic DNAs was also determined. Among the bacteria, four Gram (−ve) members of the family Enterobacteriaceae, viz. *E. coli*, *C. freundii*, *P. vulgaris*, and *S. typhimurium*, one Gram (−ve) member of the family Pseudomonadaceae, viz. *P. aeruginosa*, and two belonging to Gram (+ve), *C. perfringens* and *M. lysodeikticus*, were studied. None of the Gram (+ve) show any sequence relatedness; the genomes of *C. perfringens* and *M. lysodeikticus* having G + C contents of approximately 26.5 and 72%, respectively, are very different from that of *E. coli* (50% G + C), and it may be that the sequences have diverged significantly. The results obtained on the homology of the *E. coli* ung gene with that of the other members of Enterobacteriaceae are similar to those obtained for many other genes (32). In the present study, however, *Salmonella* shows somewhat less homology to the *E. coli* ung gene than generally expected; it is quite possible that the *Salmonella* ung gene has diverged somewhat more than average. The *Pseudomonas* gene also did not show any detectable homology to *E. coli*. Judging from the relatedness of the *E. coli* ung gene to the members of Enterobacteriaceae, especially the *Proteus* it is not surprising that the *Pseudomonas* ung gene has diverged significantly to show no relatedness to the *E. coli* gene. Also, no homologies to eukaryotic genomes were detected even under lower hybridization stringencies. These results suggest that even though uracil DNA glycosylase is of wide occurrence among living organisms, it is not conserved in unrelated bacteria or eukaryotes.

In order to possibly determine the DNA binding/active site of the uracil DNA glycosylase, the amino acid sequences of the *E. coli* tag and alkA gene products (33, 34) were compared. Products of both of these genes catalyze excision of 3-methyl adenine residues from DNA. The product of the tag gene, like the ung gene product, is an enzyme with limited substrate specificity, however, the gene product of alkA is less specific, since it also excises other methylated bases (2). No significant sequence homologies between the uracil DNA glycosylase and the tag and alkA gene products were detected, even though there seems to be similarity in the amino acid compositions of tag and ung gene products (33). To analyze if there were some other features of the sequence in common among these three enzymes, a computer program based on Kyte and Doolittle's calculations (35) was used, to derive hydrophobicity plots by using a window of 7 amino acids. The results of this analysis are shown in Fig. 9. It is interesting to note that tag and ung gene products display larger variations between their hydrophobic and hydrophilic regions than the alkA where they are mostly scattered around the base line. It is not clear if larger changes in the hydrophobic and hydrophilic characteristics are important in conferring substrate specificities to these two enzymes.

Studies reported on substrate specificities of the uracil DNA glycosylase by other workers (1, 2) indicate that the enzyme is very specific in its recognition of the uracil residue in its substrate. Such recognition is possible only through a very specific binding site in the protein. Currently we are attempting to localize the binding/active site(s) of the enzyme by mutational analysis of the gene. Expression of these mutated genes may provide information about the active site of this enzyme.

Acknowledgments—We thank Dr. B. K. Duncan for providing us with the *E. coli* strain B443S containing the plasmid pBD15, and Drs. Sandy Kieland and R. Olafson (University of Victoria, Canada) for performing the N-terminal amino acid sequence analysis. The services provided by the protein sequencing facilities at the Medical Biochemistry Department at the University of Calgary are also thankfully acknowledged.

REFERENCES

1. Duncan, B. K. (1981) in *The Enzymes* (Boyer, P., ed) pp. 565-586, Academic Press, Orlando, FL
2. Lindahl, T. (1982) *Annu. Rev. Biochem.* 51, 61-84
3. Friedberg, E. C., Benura, T., Cone, R., Simons, R., and Anderson, C. (1978) in *DNA Repair Mechanisms* (Hanawalt, P. C., Friedberg, E. C., and Fox, C., ed) pp. 103, Academic Press, Orlando, FL
4. Green, D. A., and Deutsch, W. A. (1983) *Mol. Gen. Genet.* 192, 322-325
5. Breimer, L. H. (1986) *Biochem., Biophys. Res. Commun.* 14, 201-204
6. Cone, R., Benura, T., and Friedberg, E. C. (1980) *J. Biol. Chem.* 255, 10534-10538
7. Warner, H. R., Joosse, I. K., and Smoutad, D. P. (1980) *J. Biol. Chem.* 255, 535-538
8. Lindahl, T., Ljungqvist, S., Siegert, W., Nybert, B., and Spersøn, B. (1977) *J. Biol. Chem.* 252, 3295-3298
9. Duncan, B. K., and Chambers, J. A. (1984) *Gnat. (Amat)* 25, 211-219
10. Pullbyblank, D. E., and Morgan, A. R. (1975) *Biochemistry* 14, 5205-5209
11. Maniatis, T., Fritsch, E. F., and Sambrook, J. (eds) (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratoty, Cold Spring Harbor, NY
12. Hames, D. B., and Gimsley, M. L. (1981) *Anal. Biochem.* 113, 193-197
13. Stoughard, P., and Molin, S. (1981) *Anal. Biochem.* 116, 191-193
14. Vardaxen, Y., and van de Sande, J. H. (1987) *Biotechniques* 6, 401-411
15. Maxam, A. M., and Gilbert, W. (1980) *Methods Enzymol.* 65, 498-568
16. Milton, A. D., Kreig, P. A., Rebagliati, M. R., Maniatis, T., Zinn, K., and Green, M. R. (1984) *Nature* 311, 701-706
17. Reed, K. C. and Mann, D. A. (1985) *Nature* 313, 7027-7021
18. Vardaxen, Y., Jambrou, N., Foster, R., and Gedamu, L. (1986) *Mol. Cell. Biol.* 6, 26-37
19. Caron, G., and van de Sande, J. H. (1986) *Methods Enzymol.* 65, 76-85
20. Sharp, P. A., Berk, A. J., and Berget, S. M. (1980) *Methods Enzymol.* 65, 750-768
21. Pratt, J. M., Boulnois, J. G., Darby, V., Orr, W., Wahl, E., and Holland, I. B. (1981) *Nucleic Acids Res.* 9, 4459-4474
22. Zubay, G. (1973) *Annu. Rev. Genet.* 7, 267-287
23. Lasenfitt, U. K. (1970) *Nature* 227, 666-668
24. Bonner, W. M., and Laskay, R. A. (1974) *Eur. J. Biochem.* 46, 83-88
25. Gross, E. (1975) *Methods Enzymol.* 41, 239-255
26. Hartley, B. S., and Kaufman, D. L. (1966) *Biochem. J.* 101, 229-231
27. Shine, J., and Dalgarno, L. (1970) *Nature* 226, 34-36
28. Rosenberg, M., and Courr, D. (1970) *Annu. Rev. Genet.* 13, 319-353
29. Prinjew, D. (1975) *Proc. Natl. Acad. Sci. U.S.A.* 72, 784-788
30. Harley, C. B., and Reynolds, R. F. (1987) *Nucleic Acids Res.* 15, 2343-2360
31. Lucas-Leard, J., and Ljerman, F. (1971) *Annu. Rev. Biochem.* 40, 409-448
32. Bellin, M., and Antonius, A. J. (1980) *J. Bacteriol.* 143, 306-376
33. Steinum, A., and Seeberg, E. (1986) *Nucleic Acids Res.* 14, 3765-3773
34. Nakagawa, Y., Kondo, H., and Sekiguchi, M. (1984) *J. Biol. Chem.* 259, 13725-13729
35. Wyte, J., and Doolittle, R. F. (1982) *J. Mol. Biol.* 157, 105-132