Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 8860–8864, December 1988
Biochemistry

A human 200-kDa protein binds selectively to DNA fragments containing G·T mismatches

JOSEF JIRICNY*, MELYA HUGHES, NATALIE CORMAN, AND BRIAN B. RUDKIN

Friedrich Miescher-Institut, P.O. Box 2543, CH-4002 Basel, Switzerland

Communicated by Diter von Wettstein, August 16, 1988 (received for review July 11, 1988)

ABSTRACT **G·T mispairs, the sole mismatch type that can arise in "resting" mammalian DNA (through spontaneous hydrolytic deamination of 5-methylcytosine), are corrected *in vivo* with high efficiency and mostly to a G·C. We identified a protein factor, present in HeLa cell extracts, that binds selectively to DNA substrates containing this mismatch. The partially purified protein was shown by gel-filtration chromatography and UV cross-linking experiments to have an apparent molecular mass of 200 kDa. Its binding to G·T mispairs was not influenced by sequences flanking the mismatch, but methylation of guanines either within the mismatch itself or in its immediate vicinity abolished the formation of the protein–DNA complex. The protein appears to lack both endo- and exonuclease activities and requires neither magnesium nor zinc nor ATP for binding. We discuss the possible role of this protein in a repair pathway, which helps mammalian cells counter the mutagenic effect of the hydrolytic deamination of 5-methylcytosine.**

Methylation of cytosines within CpG dinucleotides in the DNA of higher eukaryotes plays an important role in the regulation of gene expression (1–3). For the differentiated state of the cell to be maintained, the tissue-specific pattern of DNA methylation must be protected from the effects of spontaneous hydrolytic deamination of 5-methylcytosine to thymine (4), which was estimated to generate about 12 G·T mispairs per human genome per cell cycle (5, 6).

We recently reported that simian cells possess a specific repair pathway, which counteracts the effects of 5-methylcytosine deamination by correcting G·T mispairs almost exclusively to G·C pairs (7). We describe here the identification, substrate specificity, and partial purification of a human DNA-binding protein, which binds solely to DNA duplexes containing a single G·T mispair. We propose that this G·T binding protein may be a component of the mammalian G·T mismatch repair pathway.

MATERIALS AND METHODS

Preparation of the Cell Extracts. The entire procedure described below was carried out at 4°C. The harvested HeLa cells [30-g aliquot obtained from 80-liter suspension culture (8)] were suspended in three packed cell volumes of buffer A (25 mM Hepes KOH, pH 8.0/1 mM EDTA/1 mM benzamidine/2 mM 2-mercaptoethanol/0.5 mM phenylmethylsulfonyl fluoride/0.5 mM spermidine/0.1 mM spermine) and homogenized in a Dounce homogenizer (pestle B, 10 strokes). One volume of glycerol was added, followed by a saturated neutralized solution of ammonium sulfate (11 ml per 100 ml of homogenate). The mixture was allowed to stand for 30 min, transferred into polycarbonate bottles, and centrifuged in a Beckmann Ti 70 rotor for 90 min at 60,000 rpm at 4°C. The use of polyamines in the buffer represents an improvement on the method described by Manley (8), as it resulted in the formation of tight pellets containing, in addition to the cell debris, the nucleic acids.

Bandshift Assays. The synthetic 34-mer oligonucleotide 5′-AGCTTGGCTGCAGGTYGACGGATCCCCGG GAATT-3′, ^{32}P-labeled at the 5′ end, was annealed (9) with its complementary oligonucleotide 5′-AATTCCCGGGGATCCGTCRACCTGCAGCCAAGCT-3′, where R and Y represent purine (A, G) and pyrimidine (T, C) nucleosides, respectively. Forty femtomoles of the resulting duplex G·C, G·T, A·T, or A·C was added to a reaction mixture containing 25 mM Hepes·KOH (pH 8), 0.5 mM EDTA, 10% (vol/vol) glycerol, 0.01 mM $ZnCl_2$, 0.5 mM dithiothreitol, and poly[d(I·C)-d(I·C)] (1 μg per 10 μg of protein) in a total vol of 20 μl. Five microliters of the cell extract or the column fraction (1–10 μg of total protein) was then added and the mixture was allowed to stand at room temperature for 30 min. Five microliters of the mixture was then loaded on a 6% nondenaturing polyacrylamide gel (10). Electrophoresis was carried out at 10 V/cm until the bromophenol blue dye, loaded in an adjacent well, migrated ≈7 cm. The dried gels were autoradiographed for 45 min at −80°C (Figs. 1*A* and 2*D*).

Competition-Binding Assays. The competition experiments were carried out as described above, except that each reaction mixture contained, in addition to the labeled 34-mer G·T duplex, a 40-fold excess of one of the unlabeled 16-mer duplexes obtained by annealing the oligonucleotides 5′-GATCCGTCNACCTGCA-3′ and 5′-TGCAGGTNGACGGATC-3′, where N = G, A, T, or C. The relative amounts of free and protein-bound duplex were determined by densitometry of the autoradiographs of the dried gels (Fig. 1*B*).

Heparin-Sepharose Chromatography. The supernatant from the previous step was decanted, diluted 1:4 with buffer B (25 mM Hepes KOH, pH 8.0/1 mM EDTA/1 mM benzamidine/10 mM 2-mercaptoethanol/10% glycerol) and applied onto a column containing 50 ml of heparin-Sepharose (Pharmacia) equilibrated with buffer B containing 0.1 M KCl. The column was washed with 5 column vol of the same buffer and then with 2 column vol each of buffer B containing 0.3, 0.5, and 1.0 M KCl. The flow rate was maintained at 1.5 ml/min. The elution profile is shown in Fig. 2*A*. The proteins present in the 0.3 M KCl fraction were precipitated with ammonium sulfate (60% saturation) and collected by centrifugation in a Sorvall SS-34 rotor at 17,500 rpm for 20 min at 4°C.

Sephacryl S-300 Chromatography. The pellet from the previous step was resuspended in buffer B (5 ml, 0.1 M KCl) and applied onto a Sephacryl S-300 column (5 × 50 cm, Pharmacia) equilibrated with the same buffer. Elution was performed at 1.5 ml/min. The elution profile is shown in Fig. 2*B*.

Mono Q FPLC Chromatography. The S-300 fractions containing the G·T mismatch binding activity were pooled and loaded onto a Mono Q FPLC column (Pharmacia) equilibrated with buffer B (0.1 M KCl) containing kalikrein

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*To whom reprint requests should be addressed.

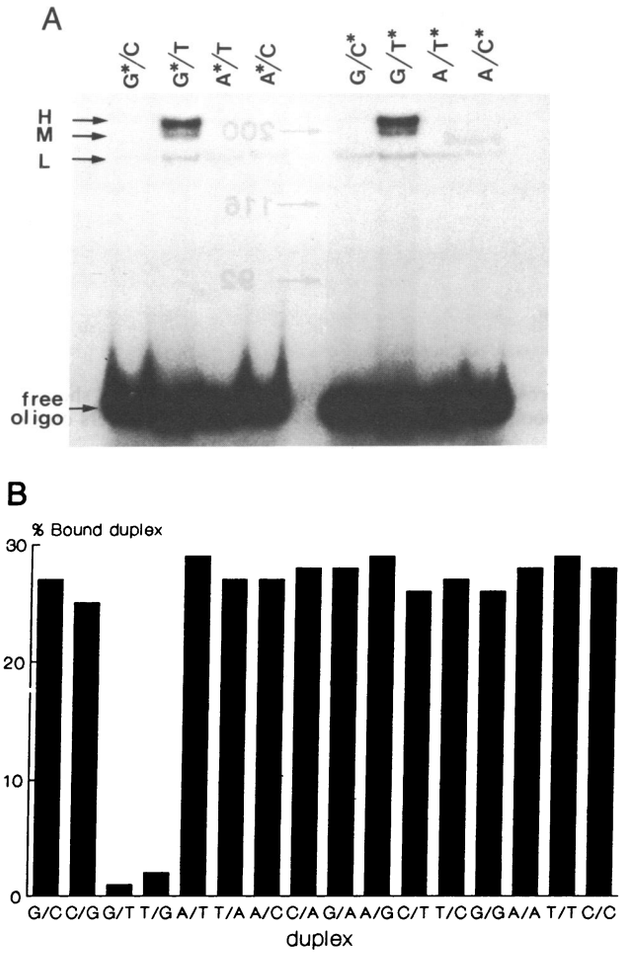

FIG. 1. Bandshift assay showing the substrate specificity of the G·T mismatch binding protein. (*A*) Labeled 34-mer oligonucleotide duplexes G·C, G·T, A·T, and A·C were incubated with a HeLa whole cell extract in the presence of excess nonspecific competitor DNA. Asterisk indicates the labeled strand. (*B*) The labeled 34-mer duplex G·T was incubated with the cell extract as in *A*, except that a 40-fold excess of unlabeled 16-mer duplexes G·C, C·G, G·T, T·G, A·T, T·A, A·C, C·A, G·G, A·A, T·T, C·C, G·A, A·G, T·C, or C·T was present in the reaction mixture. Data represent results of densitometric scans of autoradiographs of dried bandshift gels such as in *A*.

inhibitor (20 units/ml) (Calbiochem), leupeptin (2 μg/ml), and trypsin inhibitor (20 μg/ml) (Sigma). After a 10-min wash, elution was carried out at 1 ml/min with a linear gradient of 0.1–0.5 M KCl over a period of 25 min. The active fractions were pooled and flash-frozen in liquid nitrogen in 1-ml aliquots. The elution profile is shown in Fig. 2*C*, and the bandshift activity assay is shown in Fig. 2*D*.

Cross-Linking Experiments. The protein–DNA complexes were formed as described above, except that the G·C and G·T substrates were constructed by annealing the 34-mer G (see above) with the 30-mers 5′-AGCTTGGCTGCAGGTYGAC-GGATCCCCGGG-3′, where Y = C or T, and by the subsequent filling-in of the sticky end with Klenow polymerase in the presence of 5-bromo-dUTP and [α-^{32}P]dATP. The remaining steps were carried out as described (11). Denaturing sodium dodecyl sulfate (SDS)/polyacrylamide gel electrophoresis was performed using 10% gels (Fig. 3).

Methylation Interference Assays. Synthetic 90-mer oligonucleotides G_{90} and T_{90} (sequence corresponding to nucleotides 6571–6660 of M13mp9) were 5′ phosphorylated and annealed as described (9). The mismatched duplex (5 pmol), labeled either in the upper or in the lower strand, was treated with dimethyl sulfate (12). The purified methylated G·T_{90}

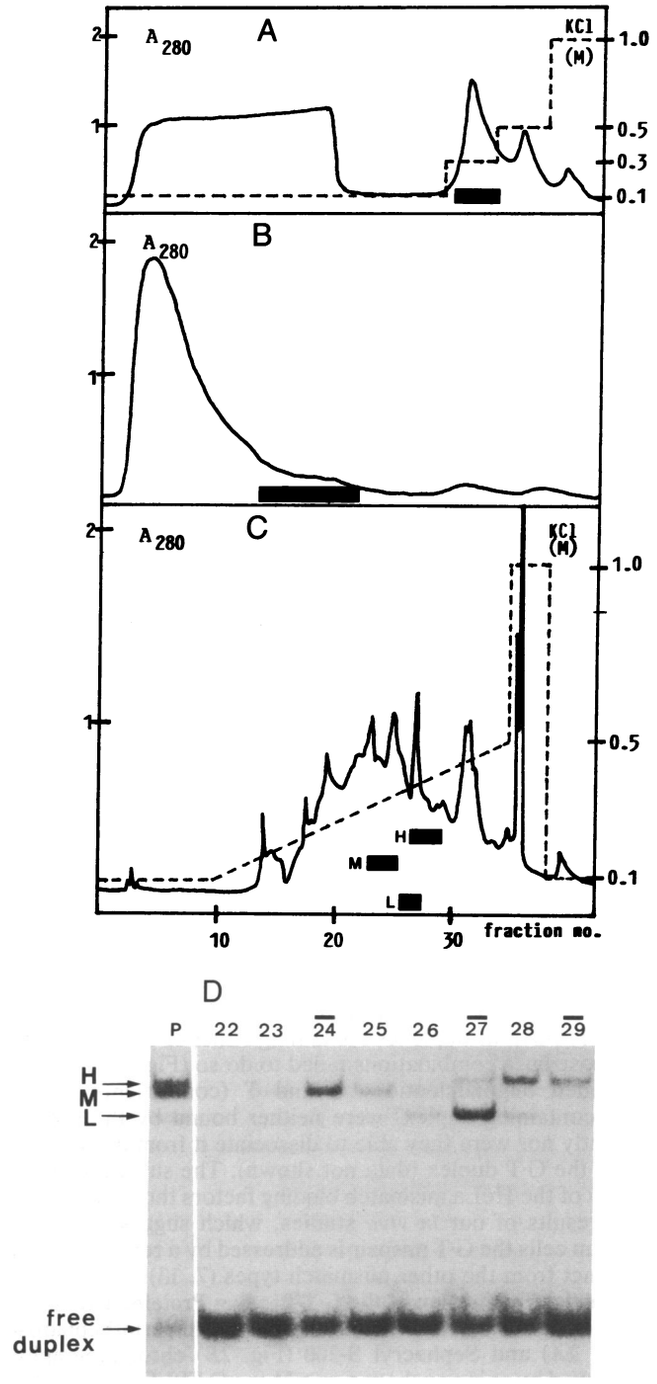

FIG. 2. Partial purification of the G·T binding protein. (*A*) Heparin-Sepharose chromatography. (*B*) Sephacryl S-300 chromatography. (*C*) Mono-Q FPLC chromatography. (*D*) Bandshift assay of activity-containing fractions, carried out with the labeled 34-mer G·T duplex. Lanes: P, S-300 pool; 22–29, Mono Q fractions. The fractions used in the cross-linking experiment (see Fig. 3) are indicated by horizontal bars.

duplexes were used in a band-shift assay, using fraction 29 from the Mono-Q FPLC experiment (Fig. 2*C*). The bands containing the free and the bound oligonucleotides were excised; the duplexes were recovered from the gel slices and alkali-treated as described (12) to generate strand cleavage at modified guanine and adenine residues (G>A reaction). Electrophoresis was carried out using an 8% polyacrylamide (29:1) sequencing gel. The autoradiograph of the dried gel is shown in Fig. 4.

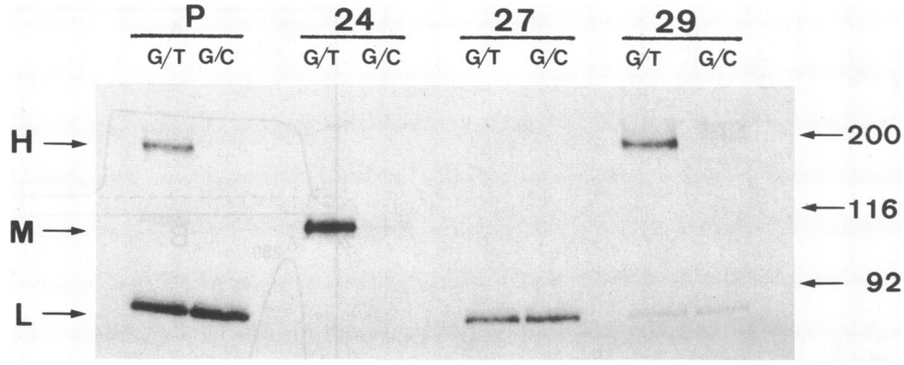

FIG. 3. Cross-linking of the G·C and G·T duplexes with the G·T binding protein-containing FPLC fractions. The figure shows an autoradiograph of the dried SDS/polyacrylamide gel. Lanes: P, Sephacryl S-300 pool; 24, 27, and 29, Mono Q fractions. Numbers on right indicate the position of the molecular mass markers (in kDa).

RESULTS

HeLa Whole Cell Extracts Contain a Factor that Binds to Oligonucleotide Duplexes Containing a G·T Mismatch. Labeled synthetic oligonucleotide duplexes, either fully complementary (G·C, A·T) or containing a single G·T or an A·C mispair (G·T, A·C, respectively), were incubated with crude HeLa whole cell extracts in the presence of a large excess of nonspecific competitor DNA and then subjected to nondenaturing polyacrylamide gel electrophoresis. These bandshift experiments (10) (Fig. 1*A*) revealed that the lysates contain factors that bind specifically to the duplex containing the G·T mispair. At least three distinct DNA–protein complexes with different electrophoretic mobilities (Fig. 1*A*, bands L, M, and H) were detected. The fastest migrating band L (low) appeared to lack substrate specificity, as it was observed with all four duplexes. In contrast, the slower migrating bands M (middle) and H (high) were observed solely with the G·T duplex. A 40-fold excess of unlabeled duplexes containing either a G·T or a T·G mispair could almost totally block the labeled G·T out of the specific protein–DNA complexes M and H, whereas duplexes containing either of the 14 remaining base·base combinations failed to do so (Fig. 1*B*). Single-stranded oligonucleotides G and T (constituents of the G·T-containing duplex) were neither bound by the protein directly nor were they able to dissociate it from its complex with the G·T duplex (data not shown). The substrate specificity of the HeLa mismatch binding factors thus agrees with the results of our *in vivo* studies, which suggested that in simian cells the G·T mispair is addressed by a repair pathway distinct from the other mismatch types (7, 13).

Partial Purification of the G·T Binding Protein. The whole cell extract was further fractionated by heparin-Sepharose (Fig. 2*A*) and Sephacryl S-300 (Fig. 2*B*) chromatography, followed by anion exchange on a Mono Q FPLC column. The latter purification step led to the separation of the H, M, and L activities (Fig. 2 *C* and *D*).

The G·T Binding Factor Is a 200-kDa Protein. As the activity-containing fractions eluted from the Mono Q FPLC column contained several protein species, we needed to obtain information regarding the size of the G·T binding factor(s). We therefore carried out a series of cross-linking experiments (11) with the Mono Q FPLC fractions 24, 27, and 29, containing the factors M, L, and H, respectively (Fig. 2*D*). In these experiments, the oligonucleotide substrate contained, in addition to the radioactive label, two 5-bromo-2′-deoxyuridine residues. The assay mixture was then irradiated with UV light and the covalently linked protein–DNA complexes were resolved by electrophoresis on a 10% denaturing SDS/polyacrylamide gel. The result of this experiment (Fig. 3) confirms the presence of three protein species in the whole cell extract, corresponding to the L, M, and H bands observed in the bandshift assays. After correcting for the change in the electrophoretic mobility due to the attached oligonucleotide duplex, the molecular masses of these species were estimated to be around 80, 120, and 200 kDa, respectively. These sizes agree with those estimated from chromatography on a Superose-12 FPLC column (data not shown), suggesting that the G·T-binding protein is present in the cell lysates in its monomeric form. The cross-linking experiment also confirmed the substrate specificities of the three species. As shown in Fig. 3, the protein present in the Mono Q fraction 24 (factor M) was cross-linked almost exclusively with the G·T duplex. In contrast, factor L present in fraction 27 and, to a small extent, also in fraction 29, reacted equally well with both G·T and G·C substrates. Protein H, present in fraction 29, showed the highest substrate specificity, as it was selectively cross-linked only with the G·T mismatch-containing substrate. The amount of protein M was found to increase, relative to H, upon prolonged storage at 4°C (data not shown), suggesting that the 200-kDa species undergoes proteolytic degradation to yield the 110-kDa fragment. However, this degradation could be prevented by the addition of trypsin and kallikrein inhibitors and leupeptin to the extracts prior to the chromatographic separation steps.

The G·T Binding Protein Possesses No Apparent Endo- or Exo-Nucleolytic Activity. To further characterize the 200-kDa protein, we carried out a series of experiments aimed at elucidating whether it possessed any enzymatic activity. G·T mismatch-containing DNA substrates (a 90-mer synthetic duplex and an M13 heteroduplex) were incubated with the partially purified protein fractions in the presence or absence of a variety of cofactors. This series of experiments revealed that the binding reaction required neither magnesium nor zinc nor ATP. Furthermore, no mismatch-dependent nicking or degradation of either strand of the DNA duplex was observed under the conditions of the assay (data not shown).

Guanine Methylation in the Immediate Vicinity of the G·T Mispair Adversely Affects the Protein Binding. Methylation interference studies, using G·T-containing synthetic 90-mer fragments, yielded invaluable information regarding the mode of binding of the 200-kDa protein to the substrate. As shown in Fig. 4, only methylation of the guanine and to some extent adenine bases situated either in the mismatch itself or immediately adjacent to it decreased the affinity of the protein for the mispair. Modification of guanine residues lying 2 nucleotides in either direction on the T-containing strand, or 2–5 nucleotides away on the G-containing strand, had no influence on the binding of the protein. This suggests that the mode of recognition of the substrate by the G·T binding protein differs from that of other sequence-specific DNA binding proteins, such as transcription factors. The

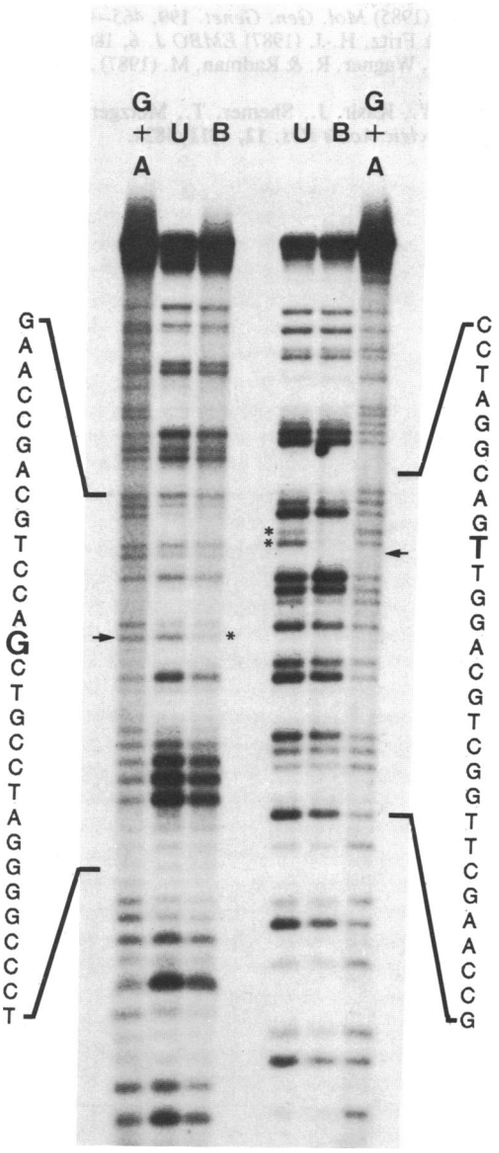

FIG. 4. Methylation interference footprinting. The ^{32}P-labeled G·T$_{90}$ duplex, partially methylated with dimethyl sulfate, was incubated with the Mono Q FPLC fraction 29. Parts of the sequences of the labeled 90-mer strands are shown along the gel, with the mismatched nucleotides shown in larger letters. The position of the mismatch is denoted by arrows. Asterisks mark the missing G- or A-specific bands. Lanes: G+A, purine-specific reference cleavage pattern of the unmodified 90-mer with formic acid (12); U, G>A-specific cleavage pattern of the unbound modified 90-mer; B, G>A-specific cleavage pattern of the bound modified 90-mer, excised from the retarded band in the bandshift assay. The three left lanes were labeled in the G strand; the three right lanes were labeled in the T strand.

latter proteins bind to their respective recognition sites by forming a series of intermolecular hydrogen bonds between the protein and the target DNA sequence. It is therefore common that guanine methylation at several individual sites within the recognition sequence interferes with protein binding. By contrast, our protein must, by definition, bind to G·T mispairs in any sequence context. It must thus recognize a specific structure rather than a specific sequence. Its binding to mismatched DNA should therefore be influenced only as the result of a modification of the recognized structure or its immediate vicinity, which is precisely what we found. The results of the methylation interference experiments support the findings of the substrate specificity bandshift studies and suggest that the G·T mispair is the true substrate for this 200-kDa protein.

DISCUSSION

What is the possible biological role of the G·T mismatch binding protein? Mismatches arise in DNA as the result of replication errors, recombination events, and 5-methylcytosine deamination. As with any other DNA repair event, efficient mismatch correction relies on the recognition of the mispair by the cellular repair machinery. In *Escherichia coli*, for example, the role of mismatch recognition is attributed to the 97-kDa product of the *mutS* gene (14). Possessing no apparent endonuclease activity, the MutS protein was shown to bind with the highest affinity to duplexes containing mispairs most frequently associated with replication errors; i.e., G·T and A·C (15, 16). Interestingly, this protein is also involved in the very short patch (VSP) repair pathway of *E. coli*, which addresses G·T mispairs resulting from the deamination of 5-methylcytosines within the recognition sequence of DNA cytosine methylase Dcm (17). VSP repair leads almost invariably to correction of the G·T mispair to restore a G·C pair (17–19). As this finding closely parallels our results with the simian system (7), it seems conceivable that the 200-kDa protein could be one of the components of a VSP-like G·T mismatch repair pathway that may play a role in determining the efficiency and, more importantly, the directionality of correction of this lesion in mammalian cells.

The evidence pertaining to the role of the G·T binding protein in mammalian mismatch repair should come from the results of *in vitro* correction experiments with cell-free extracts. However, bandshift experiments with *Drosophila* and *Saccharomyces cerevisiae* lysates failed to indicate the presence of a G·T binding activity (data not shown). As the DNA of these organisms contains no or only negligible amounts of 5-methylcytosine (20), it seems conceivable that the G·T binding protein, identified in HeLa and, more recently, in CV-1 cell extracts, is a component of a specific repair pathway, which helps mammalian cells counteract the detrimental effects of the loss of 5-methylcytosine through spontaneous hydrolytic deamination (7).

We thank Dr. J.-P. Jost for invaluable discussions; Silvia Reiber for excellent technical assistance; Dr. L. Pick for the *Drosophila* cell extracts; Dr. M. Kuenzi and E. Haberthuer for the HeLa cells; and Drs. J.-P. Jost, Elliott Shaw, and Karin Wiebauer for critical reading of the manuscript.

1. Bird, A. P. (1986) *Nature (London)* **321,** 209–213.
2. Doerfler, W., Langner, K.-D., Knebel, D., Weyer, U., Dobrzanski, P. & Knust-Kron, B. (1985) *Prog. Clin. Biol. Res.* **198,** 133–155.
3. Gardiner-Garden, M. & Frommer, M. (1987) *J. Mol. Biol.* **196,** 261–282.
4. Riggs, A. D. & Jones, P. A. (1983) *Adv. Cancer Res.* **40,** 1–30.
5. Lindahl, T. (1982) *Annu. Rev. Biochem.* **51,** 61–87.
6. Wang, R. Y.-H., Kuo, K. C., Gehrke, C. W., Huang, L.-H. & Ehrlich, M. (1982) *Biochim. Biophys. Acta* **697,** 371–377.
7. Brown, T. C. & Jiricny, J. (1987) *Cell* **50,** 945–950.
8. Manley, J. L. (1984) in *Transcription and Translation, a Practical Approach*, eds. Hames, B. D. & Higgins, S. J. (IRL, Oxford, Washington, DC), pp. 71–88.
9. Jiricny, J., Wood, S. G., Martin, D. & Ubasawa, A. (1986) *Nucleic Acids Res.* **14,** 6579–6590.
10. Fried, M. G. & Crothers, D. M. (1983) *Nucleic Acids Res.* **11,** 141–158.
11. Chodosh, L. A., Carthew, R. W. & Sharp, P. A. (1986) *Mol. Cell. Biol.* **6,** 4723–4733.
12. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
13. Brown, T. C. & Jiricny, J. (1988) *Cell* **54,** 705–711.

14. Lu, A.-L., Welsh, K., Clark, S., Su, S.-S. & Modrich, P. (1984) *Cold Spring Harbor Symp. Quant. Biol.* **49**, 589–596.
15. Su, S.-S. & Modrich, P. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 5057–5061.
16. Su, S.-S., Lahue, R. S., Au, K. G. & Modrich, P. (1988) *J. Biol. Chem.* **263**, 6829–6835.
17. Lieb, M. (1985) *Mol. Gen. Genet.* **199**, 465–470.
18. Zell, R. & Fritz, H.-J. (1987) *EMBO J.* **6**, 1809–1815.
19. Jones, M., Wagner, R. & Radman, M. (1987) *J. Mol. Biol.* 194, 155–159.
20. Pollack, Y., Kasir, J., Shemer, T., Metzger, S. & Szyf, M. (1984) *Nucleic Acids Res.* **12**, 4811–4824.