*Proc. Natl. Acad. Sci. USA*
Vol. 82, pp. 4331–4335, July 1985
Biochemistry

# *umuDC* and *mucAB* operons whose products are required for UV light- and chemical-induced mutagenesis: UmuD, MucA, and LexA proteins share homology

(SOS responses/DNA repair/pKM101/*recA*)

KAREN L. PERRY, STEPHEN J. ELLEDGE, BARBARA B. MITCHELL, LORRAINE MARSH, AND GRAHAM C. WALKER

Biology Department, Massachusetts Institute of Technology, Cambridge, MA 02139

*Communicated by Mary Lou Pardue, March 11, 1985*

ABSTRACT **The products of the *Escherichia coli umuDC* operon and its plasmid-borne analog, *mucAB*, are required for mutagenesis caused by UV light and by many chemicals. We have determined the nucleotide sequences of *umuDC* and *mucAB* and present comparisons of these sequences. The two operons are 52% homologous at the nucleotide level. Open reading frames corresponding in position and size to the *umu* and *muc* genes have been identified. The reading frames of *umuD* and *umuC* overlap by 1 base pair, and the reading frames of *mucA* and *mucB* overlap by 13 base pairs. The predicted amino acid sequences of the UmuD and MucA proteins are 41% homologous; those of the UmuC and MucB proteins are 55% homologous. Considerable homology has also been detected between UmuD, MucA, and the COOH-terminal domains of the LexA repressor and the repressors of phage λ, 434, and P22. Complementation analyses reveal that MucA protein cannot substitute for UmuD in a *umuD*$^-$*umuC*$^+$ host and that MucB protein cannot substitute for UmuC in a *umuD*$^+$*umuC*$^-$ host. Potential regulatory sequences have been identified in *umuDC* and *mucAB*.**

The *Escherichia coli umuDC* gene products function in the cellular process that is required for mutagenesis caused by ultraviolet (UV) radiation and by a variety of chemicals. This process has been referred to by such terms as "error-prone repair" and "SOS processing" (for reviews, see refs. 1 and 2). *umuDC* mutants are nonmutable by UV light and by many chemicals and are slightly sensitive to the lethal effects of these agents (3–5). They are also defective in mutagenesis and reactivation of UV-irradiated bacteriophage λ and UV-irradiated single-stranded phage (4, 6). The biochemical mechanism of repair mediated by *umuDC* has not yet been determined. *umuDC* is a member of the SOS regulatory network and as such is under the coordinate control of the RecA and LexA proteins (5).

The plasmid pKM101 suppresses the deficiencies of *umuDC* mutants (7). pKM101 was derived from its clinically isolated parent by a series of *in vivo* manipulations (8) and has played an important role in the Ames *Salmonella* strains for detecting carcinogens as mutagens (9). The *mucAB* (*m*utagenesis: *U*V and *c*hemical) operon of pK101 is required for the suppression of the phenotype of *umuDC* mutants (10). This suppression is *recA*$^+$*lexA*$^+$ dependent as *mucAB* is also under the transcriptional control of the LexA repressor (11). Homology to *mucAB* has been detected among several mutagenesis-enhancing plasmids (12) and may be widespread (13).

The *umuDC* and *mucAB* operons have been recently cloned and their gene products identified (10, 14, 15). Each operon encodes two proteins with approximate molecular weights of 16,000 and 46,000. The *mucA* and *umuD* genes, which encode the $M_r$ 16,000 proteins, are each promoter proximal. Complementation analyses reveal that, for each operon, both gene products are required for the expression of mutagenesis and repair (10, 14, 15). The previous characterizations of *mucAB* and *umuDC* have revealed structural and functional similarities that suggest the two operons share a common evolutionary origin (10, 14, 15). In this paper, we report the DNA sequence determination of *umuDC* and *mucAB* and compare those nucleotide sequences and the predicted amino acid sequences of the *muc* and *umu* proteins.

## MATERIALS AND METHODS

**Bacterial Strains and Plasmids.** The *E. coli* strains used were TK612 (3) and GW1103 (5). Plasmids used were pGW270 (16); pGW1700 (10); pKM101IS46::Tn*5* (17); pKM101 *mucA421*::Tn*5* and pKM101 *mucB1044*::Tn*5* (10, 16); pSE115 and pSE116 (14). Phage M13 mp8 (18) was obtained from New England Biolabs.

**Nucleotide Sequence Determination.** The DNA sequence was determined using the dideoxynucleotide-termination method (19) as described by Messing *et al.* (20). Sequence analysis was aided by the computer programs of G. De Vos and E. Gubbins.

## RESULTS AND DISCUSSION

**Amino Acid Sequences of the *umu* and *muc* Proteins.** The nucleotide sequence of *umuDC* contains two continuous reading frames of 417 and 1266 base pairs (bp) (Fig. 1). These correspond to the positions of the *umuD* and *umuC* genes (14, 15), respectively, and potentially encode proteins with calculated molecular weights of 15,064 and 47,681. These values are in good agreement with the molecular weights for UmuD and UmuC determined by $NaDodSO_4$/polyacrylamide gel electrophoresis (14, 15). The DNA sequence of *umuDC* has been determined independently by Kitagawa *et al.* (22).

The nucleotide sequence of *mucAB* contains two continuous reading frames of 435 bp and 1260 bp. These potentially encode proteins with calculated molecular weights of 16,371 and 46,362, respectively. The molecular weights of the MucA and MucB proteins have been previously estimated by $NaDodSO_4$/polyacrylamide electrophoresis as 16,000 and 45,000 (10). These data suggest that the nucleotide and amino acid sequences shown in Fig. 2 are those of the pKM101 *mucA* and *mucB* genes and their proteins.

**Overlapping Reading Frames and Regulatory Sequences.** The *mucA* and *mucB* genes, as well as the *umuD* and *umuC* genes, are overlapping. *mucA* and *mucB* overlap by 13 bp;

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.

Abbreviation: bp, base pair(s).

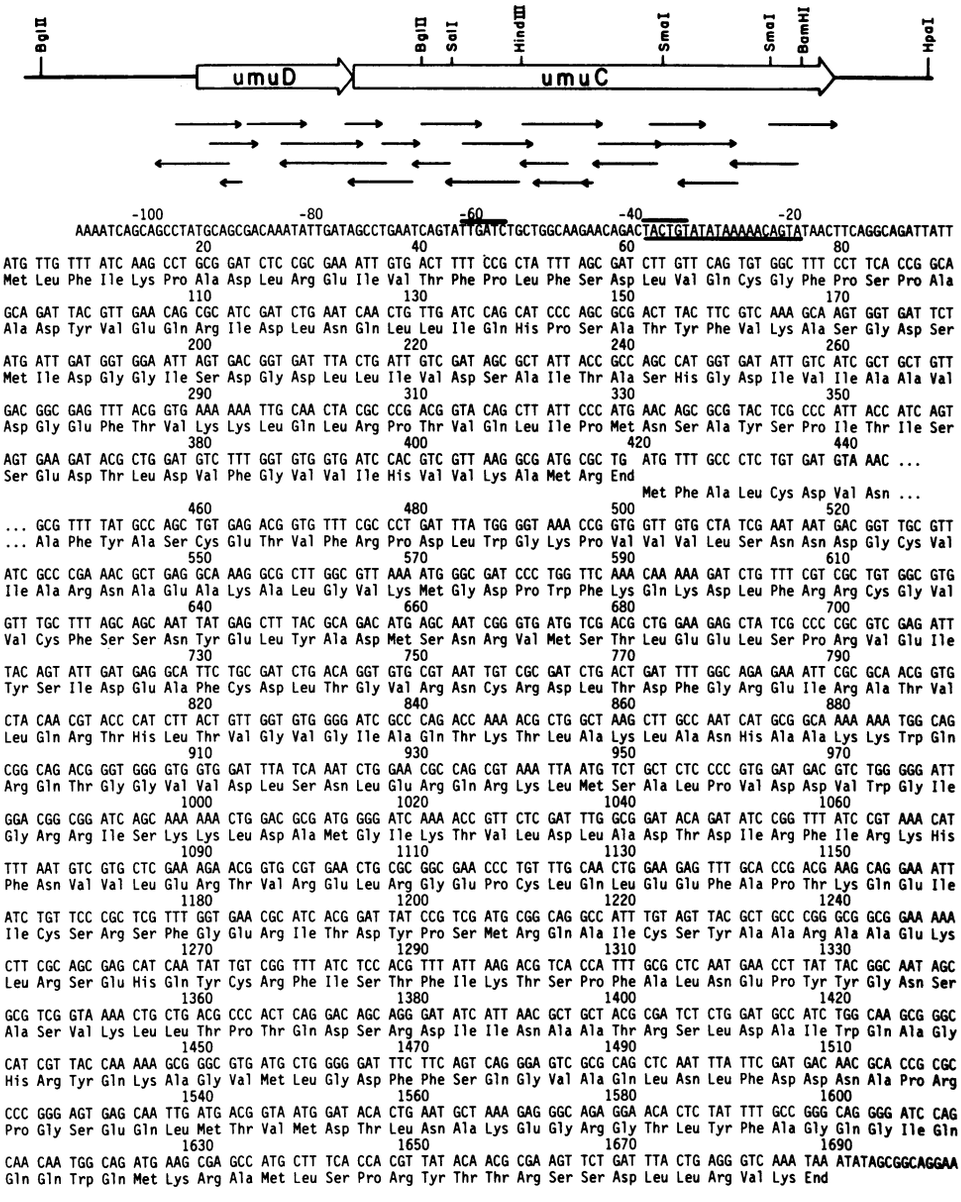


FIG. 1. Nucleotide and predicted amino acid sequences of *E. coli umuDC*. (*Upper*) A partial restriction map of *umuDC* and flanking regions of pSE115. Open arrows correspond to the coding regions of *umuD* and *umuC* and indicate the direction of transcription. Arrows below represent DNA sequence determined from specific *Bgl* II/*Bam*HI, *Bgl* II/*Hin*dIII, *Fnu*DII, *Hpa* II, *Sal* I, *Sau*3AI, *Sma* I/*Hpa* I, and *Taq* I fragments cloned into M13 bacteriophage vectors. Nucleotides are numbered from the beginning of the *umuD* initiation codon. Two open reading frames corresponding to the *umuD* and *umuC* genes extend from nucleotides +1 to +420 and from nucleotides +420 to +1686, respectively. A discontinuity has been introduced between nucleotides +443 and +444 in order to maintain the numbering frame. The sequence of a portion of the 5′ flanking region has been published (21), and we report a correction of that sequence at the −62-bp position. The underlined portion of the 5′ flanking region indicates a potential LexA recognition site; the lines above indicate a possible promoter sequence.

*umuC* overlaps *umuD* by 1 bp. Whether these overlaps play roles in translational coupling or in some other regulatory phenomenon is presently unknown. Sequences with homology to known ribosome binding sites (23) are located just upstream of the *umuC* (-TAAGGcGaTG-) and *mucB* (-TA tcGAG-) coding regions as well as just upstream of the *umuD* (-TcAGGcaG-) and *mucA* (-TttGGAaG-) coding regions.

The *umuDC* and *mucAB* operons are transcriptionally regulated by the LexA repressor (5, 11, 14, 15). Sequences with homology to known LexA binding sites (2) and *E. coli* promoters (24) are present upstream of *umuD* (Fig. 1) and *mucA* (Fig. 2). In an accompanying paper, Kitagawa *et al.* (22) demonstrate that the sequences upstream of *umuD* are in fact the LexA binding site and promoter for *umuDC*. The −33 sequence upstream of *mucA* is the site of a mutation carried by the pKM101 mutant, pGW16 (unpublished observations). The phenotype of this mutant suggests that the *muc* genes are expressed at a higher level than normal because the mutation alters the *muc* operator, or the *muc* promoter, or both.

**The *umu* and *muc* Gene Products are Homologous.** The amino acid sequences of UmuD and MucA and of UmuC and MucB are aligned and are presented in Fig. 3. In this alignment, 41% homology exists between the UmuD and MucA amino acid sequences. The MucA and UmuD homologies are clustered, and the greatest variability is at the amino and carboxyl termini of the proteins. The homologies between UmuC and MucB are fairly evenly distributed and include 55% of the amino acid residues. Two extended blocks of conserved amino acids between position 198 and position 232 (10 and 13 amino acids, respectively) may represent functionally important sites within the proteins. The sequences predict that UmuC and MucB are highly charged and

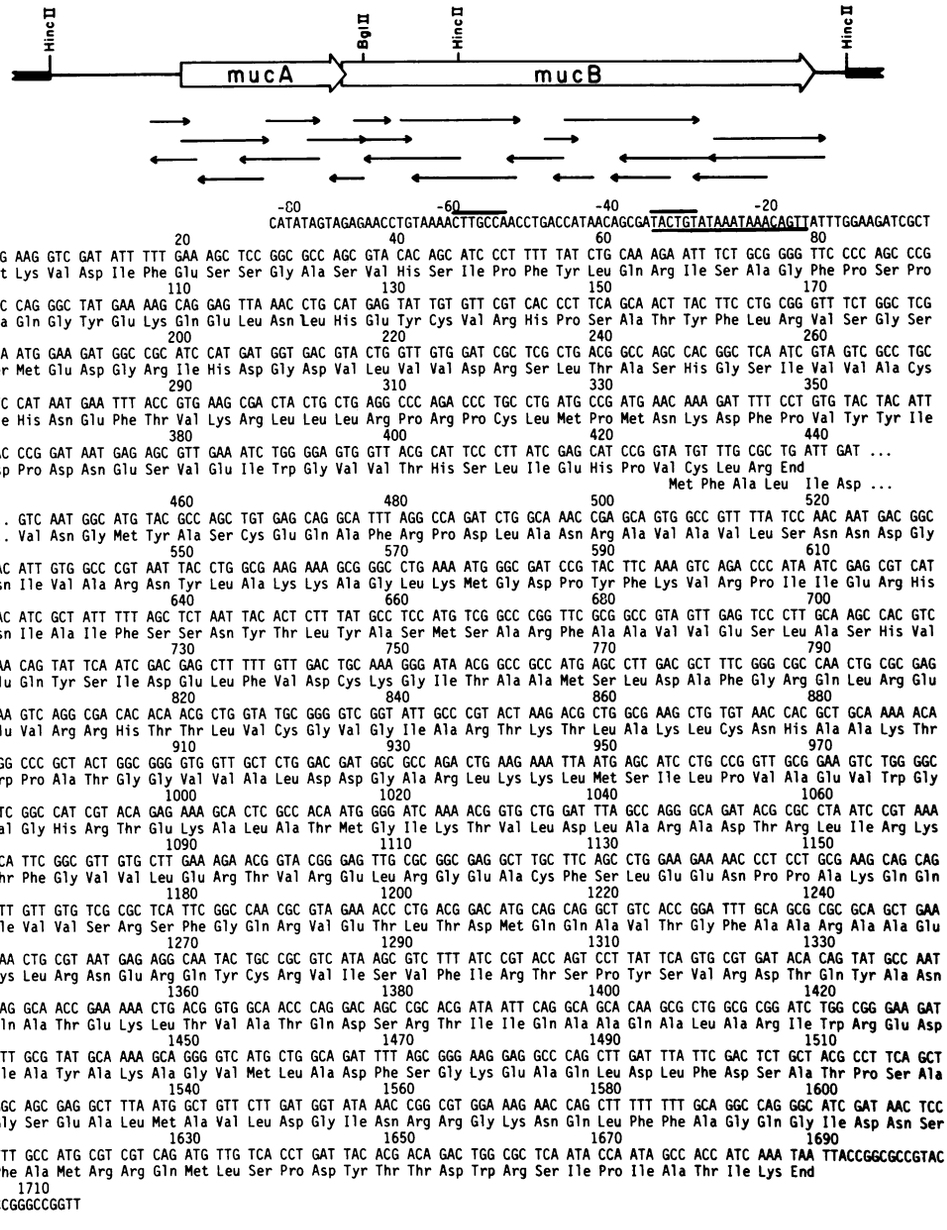


FIG. 2. Nucleotide and predicted amino acid sequences of pKM101-encoded *mucAB*. (*Upper*) A partial restriction map of *mucAB* and flanking regions of pGW1700. Heavy lines represent pKB354 vector DNA and light lines represent inserted DNA. Open arrows correspond to the coding regions of *mucA* and *mucB* and indicate the direction of transcription. Arrows below represent DNA sequence determined from *Alu* I, *Hpa* I, *Rsa* I, *Sau*3AI, and *Taq* I fragments cloned into M13 bacteriophage vectors. Nucleotides are numbered from the beginning of the *mucA* initiation codon. Two open reading frames corresponding to the *mucA* and *mucB* genes extend from nucleotides +1 to +438 and from nucleotides +426 to +1688, respectively. A discontinuity has been introduced between nucleotides +443 and +444 in order to maintain the numbering frame. The underlined portion of the 5′ flanking region indicates a potential LexA recognition site; lines above indicate a possible promoter sequence.

that UmuC has an excess of 11 basic residues (Arg+Lys minus Asp+Glu) and MucB has an excess of 12 basic residues. The fact that the *umuDC* and *mucAB* gene products play roles in UV radiation- and chemical-induced mutagenesis suggests that they may interact with nucleic acids. The strongly basic nature of the UmuC and MucB proteins (as predicted by the nucleotide sequences) raises the possibility that these proteins bind DNA.

Using the alignments shown in Fig. 3, a similarity of 52% between *umuDC* and *mucAB* exists at the nucleotide level. The homology of the nucleotide sequences is disrupted at several positions by small additions or deletions. Each addition or deletion consists of 3 bp and does not alter the reading frame. Because of the many dissimilarities between the sequences, we cannot determine whether the operons have diverged primarily by base-pair substitutions or whether compensating (+) and (−) frameshift mutations have also accumulated. The diversity between the two operons is reflected by the frequency of nucleotide changes among synonomous codons. Of the 234 synonomous codons in *umuC* and *mucB*, 58% have undergone base substitutions. There is no significant bias in the direction of these base substitutions. Codon usage is similar in *umuDC* and in *mucAB* with the following exception: 7 of the 24 isoleucine codons in *mucB* are AUA codons, which are rarely utilized by prokaryotes (25). There is only one AUA among the 21 isoleucine codons in *umuC*. Whether there is any significance to this difference is unknown.

**Noninterchangeability of *umu* and *muc* Gene Products.** We have previously shown that the phenotypes of *umuDC* mutants are suppressed by plasmids encoding the wild-type *muc* operon (7, 10). We were therefore interested in examining whether, more specifically, MucA may be able to substitute for UmuD in a $umuD^{-}umuC^{+}$ host and MucB may be

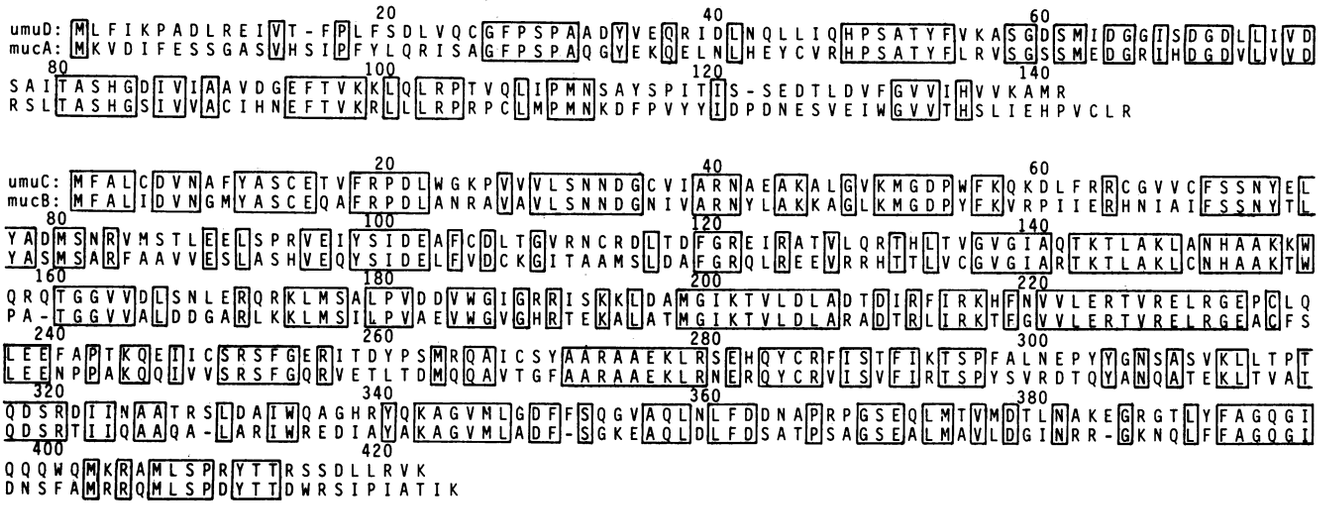


FIG. 3. Amino acid alignment of the *umu* and *muc* proteins. Amino acid sequences of the two proteins are arranged to give maximal matching. Placement of the gap at position 15 of UmuD and positions 159, 329, 351, and 384 of MucB is arbitrary within the region bounded by amino acid homologies. Placement of the gap at position 122 of UmuD is based on nucleotide homologies in the region between positions 120 and 131. Amino acids are identified by the single-letter code.

able to substitute for UmuC in a $umuD^+umuC^-$ host. To determine whether these specific substitutions are permitted, complementation between *umu* and *muc* alleles was assayed by measuring methyl methanesulfonate-induced reversion of *his-4*. TK612 ($umuD^-umuC^+$) and GW1103 ($umuD^+umuC^-$) were transformed with either pKM101 $mucA^-mucB^+$, pKM101 $mucA^+mucB^-$, of a $Muc^+$ derivative of the plasmid, pKM101IS46::Tn*5*. We chose the *umuD44* and *mucA421*::Tn-*5* alleles because they have previously been shown to be relatively nonpolar (10, 14). The nonmutability of either *umu* mutant is suppressed by the $mucAB^+$ plasmid. The nonmutability is not suppressed by the introduction of plasmids that supply only *mucA* protein or only *mucB* protein (data not shown). These results suggest that there may be a physical interaction between the two gene products of each operon and that the pairs of proteins have diverged sufficiently that interactions between UmuD and MucB and between MucA and UmuC are excluded.

**The UmuD and MucA Proteins are Homologous to LexA.** We have found that the amino acid sequences of the UmuD and MucA proteins are homologous to the COOH-terminal domains of the LexA repressor and the repressors of phage λ, 434, and P22 (Fig. 4) (26). The homologies between MucA and LexA are ≈28% and include the Ala-Gly bond corresponding to the site of the RecA-mediated cleavage of LexA (27). UmuD is 31% homologous to LexA and has a Cys-Gly bond at the cleavage position. Of the amino acids conserved between LexA and the repressors of λ, P22, and 434, 78% of these amino acids are also conserved in *umuD* and 89% are conserved in *mucA*.

The homology between the LexA and UmuD proteins raises an intriguing possibility that may explain the function of RecA in SOS mutagenesis. Activated RecA (RecA*), which promotes the proteolytic cleavage of LexA and phage repressors, is required for the induction of all SOS loci (2). Genetic evidence indicates that a function of RecA*, in addition to its capacity to facilitate repressor cleavage, is required for the expression of *umuDC*-dependent mutagenesis (2). Possibilities for this additional function of RecA* may include a mechanistic role in mutagenesis or proteolytic processing of one or more proteins that participate in SOS mutagenesis. The homology between UmuD and LexA suggests that the RecA* and UmuD proteins may interact. This interaction could yield a proteolytic cleavage that activates or unmasks the function of UmuD and MucA. The lack of conservation between the $NH_2$-terminal amino acids of UmuD and MucA may indicate that they constitute a nonfunctional, perhaps "expendable" domain. If the (only) role of RecA* in mutagenesis is to promote the removal of this domain, then the production of a UmuD or MucA peptide lacking the $NH_2$-terminal amino acids should overcome the $recA^+$ dependence of UV and chemical mutagenesis. Alternatively, an interaction between UmuD and RecA might serve some other role such as helping to direct UmuD to its site of action.

Recent evidence has suggested that mutagenesis mediated by the *umuDC* proteins differs from that mediated by the *mucAB* proteins in its requirement for "activated" RecA. It appears that *umuDC*-dependent mutagenesis, but not *mucAB*-dependent mutagenesis, is abolished by the *recA430* mutation (ref. 28; unpublished results). This mutation alters the ability of RecA to promote the proteolysis of particular repressors. RecA430 protein is able to promote the *in vitro* cleavage of the LexA repressor, but not the λ repressor (29). Perhaps the UmuD and MucA proteins interact with RecA430 protein in a manner similar to the λ and LexA repressors, respectively. The difference between *umuDC* activity and *mucAB* activity may reflect differences in the amino acid sequences of the UmuD and MucA proteins, particularly at those positions that may be required for recognition by RecA.

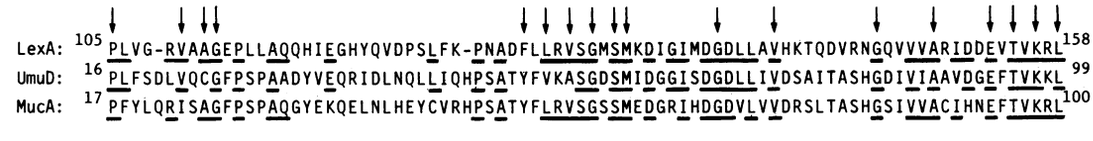


FIG. 4. Homology of the UmuD and MucA proteins to the LexA protein. Amino acid sequences of the proteins are arranged to give maximal matching. Underlined amino acids represent identities between the LexA repressor and either the UmuD or MucA proteins. The Ala-Gly cleavage site of LexA is between amino acids 112 and 113. Arrows indicate residues that are conserved among LexA protein and the repressors of phage λ, P22, and 434. Amino acids are identified by the single-letter code.

We are grateful to G. De Vos, L. Dodson, J. Krueger, P. Pang, and L. Pickering for their many helpful discussions and to D. Sobell for his generous assistance. This work was supported in part by Public Health Service Grants CA21615 from the National Cancer Institute and GM28988 from the National Institute of General Medical Sciences. S.J.E. was supported in part by Training Grant ES07020 from the National Institutes of Health and by W. R. Grace. K.L.P. was supported in part by W. R. Grace and was a Swanson Fellow. L.M. was supported by Training Grant ES07020 from the National Institutes of Health.

1. Witkin, E. M. (1976) *Bacteriol. Rev.* **40,** 869–907.
2. Walker, G. C. (1984) *Microbiol. Rev.* **48,** 60–93.
3. Kato, T. & Shinoura, Y. (1977) *Mol. Gen. Genet.* **156,** 121–131.
4. Steinborn, G. (1978) *Mol. Gen. Genet.* **165,** 87–93.
5. Bagg, A., Kenyon, C. J. & Walker, G. C. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 5749–5753.
6. Defais, M. (1983) *Mol. Gen. Genet.* **192,** 509–511.
7. Walker, G. C. & Dobson, P. P. (1979) *Mol. Gen. Genet.* **172,** 17–24.
8. Mortelmans, K. E. & Stocker, B. A. D. (1979) *Mol. Gen. Genet.* **167,** 317–328.
9. McCann, J., Spingarn, N. E., Kobori, J. & Ames, B. N. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 979–983.
10. Perry, K. L. & Walker, G. C. (1982) *Nature (London)* **300,** 278–281.
11. Elledge, S. J. & Walker, G. C. (1983) *J. Bacteriol.* **155,** 1306–1315.
12. Perry, K. L., Elledge, S. J., Lichtman, M. R. & Walker, G. C. (1982) in *Proceedings of the Third International Conference on Environmental Mutagens*, eds. Sugimura, T., Kondo, S. & Takebe, H. (Univ. of Tokyo Press; Liss, New York), pp. 113–120.
13. Upton, C. & Pinney, R. J. (1983) *Mutat. Res.* **112,** 261–273.
14. Elledge, S. J. & Walker, G. C. (1983) *J. Mol. Biol.* **164,** 175–192.
15. Shinagawa, H., Kato, T., Ise, T., Makino, K. & Nakata, A. (1983) *Gene* **23,** 167–174.
16. Langer, P. J., Shanabruch, W. G. & Walker, G. C. (1981) *J. Bacteriol.* **145,** 1310–1316.
17. Winans, S. C. & Walker, G. C. (1985) *J. Bacteriol.* **161,** 402–410.
18. Messing, J. & Vieira, J. (1982) *Gene* **19,** 269–276.
19. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
20. Messing, J., Crea, R. & Seeburg, P. H. (1981) *Nucleic Acids Res.* **9,** 309–321.
21. Elledge, S. J., Perry, K. L., Krueger, J. H., Mitchell, B. B. & Walker, G. C. (1983) in *Cellular Responses to DNA Damage*, eds. Friedberg, E. C. & Bridges, B. A. (Liss, New York), pp. 353–359.
22. Kitagawa, Y., Akaboshi, E., Shinagawa, H., Horii, T., Ogawa, H. & Kato, T. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 4336–4340.
23. Shine, J. & Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA* **71,** 1342–1346.
24. Hawley, D. K. & McClure, W. R. (1983) *Nucleic Acids Res.* **11,** 2237–2255.
25. Elton, R. A., Russell, G. J. & Subak-Sharpe, J. H. (1976) *J. Mol. Evol.* **8,** 117–135.
26. Sauer, R. T., Yocum, R. R., Doolittle, R. F., Lewis, M. & Pabo, C. O. (1982) *Nature (London)* **298,** 447–451.
27. Horii, T., Ogawa, T., Nakatani, T., Hase, T., Matsubara, H. & Ogawa, H. (1981) *Cell* **27,** 515–522.
28. Blanco, M. & Rebollo, J. E. (1981) *Mutat. Res.* **81,** 265–275.
29. Roberts, J. & Devoret, R. (1983) in *Lambda II* eds. Hendrix, R. W., Roberts, J. W., Stahl, F. W. & Weisberg, R. A. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 123–144.