JOURNAL OF BACTERIOLOGY, Jan. 1988, p. 197–202
0021-9193/88/010197-06$02.00/0
Copyright © 1988, American Society for Microbiology
Vol. 170, No. 1

Nucleotide Sequence of the *Salmonella typhimurium mutS* Gene Required for Mismatch Repair: Homology of MutS and HexA of *Streptococcus pneumoniae*

LYNNE T. HABER, PATTY P. PANG,† DAVID I. SOBELL, JOHN A. MANKOVICH, AND GRAHAM C. WALKER*

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

Received 2 June 1987/Accepted 5 October 1987

The *mutS* gene product of *Escherichia coli* and *Salmonella typhimurium* is one of at least four proteins required for methyl-directed mismatch repair in these organisms. A functionally similar repair system in *Streptococcus pneumoniae* requires the *hex* genes. We have sequenced the *S. typhimurium mutS* gene, showing that it encodes a 96-kilodalton protein. Amino-terminal amino acid sequencing of purified *S. typhimurium* MutS protein confirmed the initial portion of the deduced amino acid sequence. The *S. typhimurium* MutS protein is homologous to the *S. pneumoniae* HexA protein, suggesting that they arose from a common ancestor before the gram-negative and gram-positive bacteria diverged. Overall, approximately 36% of the amino acids of the two proteins are identical when the sequences are optimally aligned, including regions of stronger homology which are of particular interest. One such region is close to the amino terminus. Another, located closer to the carboxy terminus, includes homology to a consensus sequence thought to be diagnostic of nucleotide-binding sites. A third one, adjacent to the second, is homologous to the consensus sequence for the helix-turn-helix motif found in many DNA-binding proteins. We found that the *S. typhimurium* MutS protein can substitute for the *E. coli* MutS protein in vitro as it can in vivo, but we have not yet been able to demonstrate a similar in vitro complementation by the *S. pneumoniae* HexA protein.

Mismatched base pairs can arise during homologous recombination of allelic genes, by chemical modification of DNA, or from errors made by DNA polymerase. Repair of mismatched DNA base pairs has been invoked to explain a variety of genetic phenomena, including gene conversion in *Neurospora* spp. and other fungi (25, 38), postmeiotic segregation in *Saccharomyces cerevisiae* (49), high negative interference and gene conversion in lambda phage crosses (28, 47, 48), and the existence of high- and low-efficiency transforming markers in *Streptococcus pneumoniae* (8, 18). Mismatch repair has been studied most intensively in *Escherichia coli*, *Salmonella typhimurium*, and *S. pneumoniae*. Several reviews of mismatch repair have been published recently or will be published in the near future (5, 26, 35; M. Meselson, *In* K. B. Low, ed., *The Recombination of Genetic Material*, in press).

Extending the suggestion of Tiraby and Fox (42) that mismatch repair reduces the mutation rate by correcting replication errors, Wagner and Meselson (45) suggested that repair might be targeted to the nascent strand by some special condition, such as undermethylation. *E. coli* and *S. typhimurium mutS*, *mutL*, *mutH*, and *uvrD* (*mutU*) mutants as well as *S. pneumoniae hexA* and *hexB* mutants, all of which are defective in mismatch repair, exhibit an elevated spontaneous mutation frequency (6, 19, 40, 42, 43). The mismatch repair systems of both *E. coli* and *S. pneumoniae* repair transition mismatches much more efficiently than transversions (7, 15, 19, 44). However, the two systems appear to differ in their method of recognizing the daughter strand. A significant component of this recognition in *E. coli* appears to be based on the state of N^6-adenine methylation at d(G-A-T-C) sites. The strongest in vivo evidence for methyl-directed repair has come from transfection experiments with hemimethylated lambda heteroduplexes, in which repair occurs preferentially on the unmethylated strand (34, 36). This repair requires the products of the *mutS*, *mutL*, *mutH*, and *uvrD* (*mutU*) genes. In contrast, methylation appears unlikely to direct strand discrimination in *S. pneumoniae*, since the prototypic Hex$^+$ strain does not methylate its d(G-A-T-C) sites, and transforming DNA is mature with respect to methylation (5). Instead, it has been suggested that, in *S. pneumoniae*, single-strand breaks direct repair to the donor strand in transformation and to the nascent strand in replication (12).

An in vitro repair system has been developed which monitors the conversion of a mismatch in a hemimethylated bacteriophage fd heteroduplex to restriction endonuclease sensitivity (22, 23). Consistent with in vivo results, the products of the *mutS*, *mutL*, *mutH*, and *uvrD* genes are required for repair in this system.

Although the mechanism of methyl-directed mismatch repair in *E. coli* is not yet fully understood, biochemical activities have been ascribed to several of the components of the system. DNA mismatches are bound by purified MutS protein (41), and single-stranded DNA is bound both by MutS in crude extracts and by purified MutS (30; Haber, Pang, and Walker, unpublished data). In vivo and in vitro experiments have suggested that the MutH protein is involved in strand discrimination (15, 26). The *uvrD* gene product has been identified as helicase II (14, 16, 17). To date, no activity has been reported for the MutL protein.

As a protein that binds to mismatched DNA base pairs, MutS would be expected to play a central role in mismatch correction. Indeed, MutS appears to participate in at least three different repair strategies. Besides the repair system directed by Dam-mediated methylation of d(G-A-T-C) sites,

* Corresponding author.
† Present address: Division of Biology, California Institute of Technology, Pasadena, CA 91125.

MutS is also required for two other, less efficient, mismatch repair processes. One of these processes acts on symmetrically methylated DNA and may serve to repair mismatches produced during recombination (9, 10). Another mismatch repair system corrects C-to-T transitions at the internal C of the Dcm methylase sequence d(C-C-A/T-G-G) or subsets thereof (20) and also requires $mutL^+$ and dcm^+ (20, 35). Thus, once a mismatch is recognized by MutS, an *E. coli* cell can repair it by one of several different methods, depending on the status of the DNA, the mismatch involved, and the surrounding sequence.

In this paper, we report the sequence of the *S. typhimurium mutS* gene and show that significant homology exists between it and the *S. pneumoniae hexA* gene. Although these two organisms use different methods of strand discrimination, other elements of their mismatch repair processes may show mechanistic similarities. We were able to show that the *S. typhimurium* MutS protein can substitute for the *E. coli* MutS protein in vitro as it can in vivo, but we were not able to demonstrate a similar in vitro complementation by the *S. pneumoniae* HexA protein.

MATERIALS AND METHODS

Bacteriophage strains and plasmids. The wild-type *E. coli* strain used for the in vitro assay was MM294A, which has the genotype *pro-82 thi-1 endA1 hsdR17 supE44*. The *mutS201*::Tn*5* or *mutS215*::Tn*10* mutations were transduced into MM294 by P1 transduction. Bacteriophage M13mp8 was purchased from New England Biolabs, and M13mp8 containing a G-to-A transition in the unique *Pst*I site was provided by J. Essigman (21). The cloning of the *mutS* gene into pGW1811 has been described (30). pGW1825 is a *Bgl*II deletion of pGW1811 which overproduces MutS protein (unpublished data).

Nucleotide sequence determination. The 1.1-kilobase *Cla*I-*Pst*I, 2.2-kilobase *Pst*I-*Sal*I, and 2.5-kilobase *Pst*I-*Sma*I fragments of pGW1811 were isolated. Each fragment was then digested with *Sau*3AI, *Alu*I, and *Hae*III, and the resulting fragments were shotgun cloned into M13mp11. The DNA sequence was determined by using the dideoxynucleotide termination method (39). Additional clones which crossed the *Pst*I sites were isolated and sequenced (Fig. 1).

Purification and sequencing of MutS protein. MutS protein overproduced by the plasmid pGW1825 was purified by the method of Su and Modrich (41). The amino-terminal sequence was determined on 800 pmol of pure MutS protein by using an Applied Biosystems gas phase microsequencer.

In vitro mismatch repair assay. Heteroduplexes were prepared by the method of Kramer et al. (15) from linear duplex M13mp8 DNA which had been fully methylated in vitro and single-stranded circular DNA from an M13mp8 mutant containing a G-to-A transition in the unique *Pst*I site (21). The resulting hemimethylated heteroduplex contained a G/T mismatch, with a methylated wild-type strand. Cell extracts were prepared and the mismatch repair assay was performed essentially as described by Lu et al. (22).

RESULTS

Amino acid sequence of the MutS protein. The nucleotide sequence of the *S. typhimurium mutS* gene contains one continuous reading frame of 2,559 base pairs and potentially encodes a protein with a calculated molecular weight of 95,650 (Fig. 2). This value is in agreement with previous reports that the *S. typhimurium mutS* protein has an M_r of 98,000 (30, 31), and the *E. coli* protein has an M_r of 97,000 (41), both as determined by sodium dodecyl sulfate-polyacrylamide electrophoresis. Both in vivo (30, 31) and in vitro (see below) complementation experiments have shown that the *E. coli* and *S. typhimurium mutS* genes are functionally equivalent. We identified the correct initiator methionine and confirmed the initial portion of the deduced amino acid sequence by amino-terminal amino acid sequencing of purified *S. typhimurium* MutS protein (Fig. 2).

The MutS protein is relatively rare and has been estimated to be present at only about 10 to 20 molecules per cell on the basis of *mutS'-lacZ*$^+$ fusions (M. Radman, personal communication). However, MutS protein can be extensively overproduced by placing the *mutS* gene under control of the P_L promoter (41; Pang, Haber, and Walker, unpublished data). This is consistent with the absence of any strong homology to promoter consensus sequences immediately upstream of the *mutS* coding region (37).

Since we have found that the MutS protein binds to single-stranded DNA (30; Haber, Pang, and Walker, unpublished data), we compared its sequence with that of known single-strand binding proteins. No significant homology was found between the MutS protein and the *E. coli ssb* protein or the bacteriophage T4 single-strand binding protein, gp32. This failure to see homology is not surprising, since it is known that the single-strand binding proteins characterized to date do not all have related primary structures. For example, the *E. coli ssb* protein and T4 gp32 do not share any obvious sequence homology (3).

Homologous MutS and HexA gene products. Certain similarities between the *E. coli mutS* gene and the *S. pneumoniae hexA* gene had prompted speculation as to whether the two genes were related (5). Inactivation of either gene results in an elevated spontaneous mutation frequency in its respective organism. Also, the molecular weight of the *hexA* gene product had been reported as 86,000 to 94,000 (2, 24), close to the 96,000 molecular weight of the *mutS* gene product.

The nucleotide sequence of the *S. pneumoniae hexA* gene

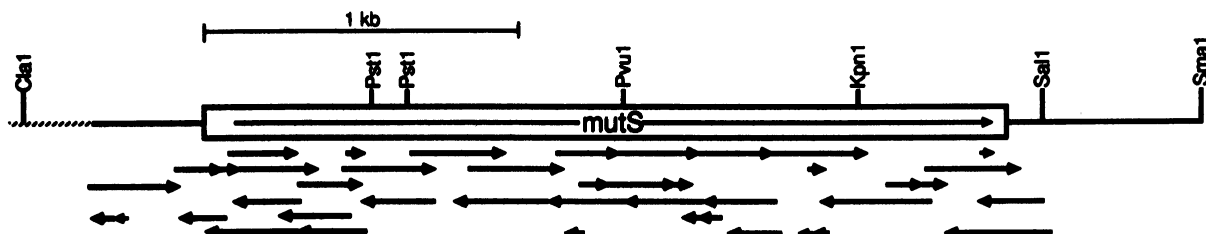

FIG. 1. Sequencing strategy of the *mutS* gene. Horizontal arrows show the direction and extent of sequences determined. The direction of transcription (30) and deduced coding region of the *mutS* gene are shown. The hatched line indicates a sequence from phage lambda. kb, Kilobase.

-370 -350 -330 -310 -290 -270
GATCTGACGAGGCTACTCTTCGCTGGATTACGTAGGTTTTTTCGTCAAACTGGCTTTTACCCCGAAACCAGCAGGAGCTTCTCTGTGGTCATCGAAAACAGTGTAGAGTTTTTGAGCT

-250 -230 -210 -190 -170 -150
GGGTTAGCTTCGAAAACGAATTTTTCCAGAGCGACTGAGTGCAGGCTTAACATTGATACTTAACTTGTTGATATGAAAGGCTGGCTGAGAGTTGTTTTCAAACTATCACCCTGTTGTAG

-130 -110 -90 -70 -50 -30
TCAAGATGATGAGATGTAACTGTCACATTTAGGCTACAGGTGAGCTACAAATTTATGGGTTTATGTGAGTTTATAAGGGTTTTACCTGGGGCCACATGAGCACTTGAAAGAATAAAAAA

-10 10 30 50 70
ACGAGTAAAATCAAT ATG AAT GAG TCA TTT GAT AAG GAC TTC TCC AAC CAC ACC CCG ATG ATG CAG CAG TAT CTC AAG CTG AAA GCC CAG CAT
Met Asn Gly Ser Phe Asp Lys Asp Phe Ser Asn His Thr Pro Met Met Gln Gln Tyr Leu Lys Leu Lys Ala Gln His

90 110 130 150
CCG GAG ATC CTG CTC TTT TAT CGT ATG GGA GAC TTT TAC GAG CTG TTT TAT GAC GAC GCA AAG CGT GCG TCG CAG TTG CTC GAT ATT TCG
Pro Glu Ile Leu Leu Phe Tyr Arg Met Gly Asp Phe Tyr Glu Leu Phe Tyr Asp Asp Ala Lys Arg Ala Ser Gln Leu Leu Asp Ile Ser

170 190 210 230 250
CTG ACC AAA CGC GGC GCA TCG GCT GGC GAA CCT ATC CCT ATG GCG GGT ATC CCG CAC CAC GCC GTA GAA AAC TAC CTC GCG AAA CTG GTC
Leu Thr Lys Arg Gly Ala Ser Ala Gly Glu Pro Ile Pro Met Ala Gly Ile Pro His His Ala Val Glu Asn Tyr Leu Ala Lys Leu Val

270 290 310 330
AAT CAG GGC GAA TCC GTC GCT ATT TGC GAA CAG ATT GGC GAT CCG GCC ACC AGC AAA GGT CCC GTT GAA CGT AAG GTC GTG CGT ATC GTT
Asn Gln Gly Glu Ser Val Ala Ile Cys Glu Gln Ile Gly Asp Pro Ala Thr Ser Lys Gly Pro Val Glu Arg Lys Val Val Arg Ile Val

350 370 390 410 430
ACG CCT GGC ACT ATC AGC GAT GAA GCC CTG TTA CAG GAG CGT CAG GAT AAC CTG CTG GCG GCT ATC TGG CAG GAT GGT AAA GGT TAC GGT
Thr Pro Gly Thr Ile Ser Asp Glu Ala Leu Leu Gln Glu Arg Gln Asp Asn Leu Leu Ala Ala Ile Trp Gln Asp Gly Lys Gly Tyr Gly

450 470 490 510
TAC GCC ACG CTG GAT ATC AGC TCC GGG CGC TTT CGT CTG AGC GAA CCC GCC GAC CGT GAA ACG ATG GCC GCA GAG CTG CAG CGC ACC AAT
Tyr Ala Thr Leu Asp Ile Ser Ser Gly Arg Phe Arg Leu Ser Glu Pro Ala Asp Arg Glu Thr Met Ala Ala Glu Leu Gln Arg Thr Asn

530 550 570 590 610
CCC GCC GAA CTG TTG TAT GCC GAA GAT TTT GCT GAA ATG GCG TTA ATA GAG GGA CGC CGG GGT CTG CGC CGT CGC CCG TTG TGG GAG TTT
Pro Ala Glu Leu Leu Tyr Ala Glu Asp Phe Ala Glu Met Ala Leu Ile Glu Gly Arg Arg Gly Leu Arg Arg Arg Pro Leu Trp Glu Phe

630 650 670 690
GAG ATT GAT ACC GCC CGC CAG CAG TTA AAT CTG CAG TTC GGT ACG CGG GAT CTG GTC GGT TTC GGC GTC GAA AAT GCC TCG CGT GGA TTA
Glu Ile Asp Thr Ala Arg Arg Gln Leu Asn Leu Gln Phe Gly Thr Arg Asp Leu Val Gly Phe Gly Val Glu Asn Ala Ser Arg Gly Leu

710 730 750 770 790
TGT GCG GCA GGC TGC CTG TTA CAG TAC GTA AAA GAT ACC CAG CGC ACC TCC CTG CCG CAT ATT CGT TCC ATT ACG ATG GAA CGC CAG CAG
Cys Ala Ala Gly Cys Leu Leu Gln Tyr Val Lys Asp Thr Gln Arg Thr Ser Leu Pro His Ile Arg Ser Ile Thr Met Glu Arg Gln Gln

810 830 850 870
GAC AGC ATC ATT ATG GAT GCC GCG ACC CGC CGC AAT CTG GAA ATT ACC CAG AAC CTG GCC GGT GGT GTC GAA AAT ACC CTC GCC GCG GTG
Asp Ser Ile Ile Met Asp Ala Ala Thr Arg Arg Asn Leu Glu Ile Thr Gln Asn Leu Ala Gly Gly Val Glu Asn Thr Leu Ala Ala Val

890 910 930 950 970
CTT GAC TGT ACC GTG ACG CCA ATG GGT AGC CGA ATG CTT AAA CGC TGG CTG CAT ATG CCG GTA CGA AAT ACC GAC ATA TTA CGT GAA CGC
Leu Asp Cys Thr Val Thr Pro Met Gly Ser Arg Met Leu Lys Arg Trp Leu His Met Pro Val Arg Asn Thr Asp Ile Leu Arg Glu Arg

990 1010 1030 1050
CAG CAG ACC ATC GGC GCC TTG CAG GAC ACC GTC AGC GAA CTG CAA CCG GTG CTG CGT CAG GTC GGC GAT TTG GAG CGT ACT CCT TGC GCG
Gln Gln Thr Ile Gly Ala Leu Gln Asp Thr Val Ser Glu Leu Gln Pro Val Leu Arg Gln Val Gly Asp Leu Glu Arg Thr Pro Cys Ala

1070 1090 1110 1130 1150
TCT GGC GCT GCG CAG CCG TCC GCG CGA TCT CGC CCG AAT GCG TCA CGC CTT CCA GCA GCT GCC GAA TTG CAC GCT CAG CTT GAA ACC GTT
Ser Gly Ala Ala Gln Pro Ser Ala Arg Ser Arg Pro Asn Ala Ser Arg Leu Pro Ala Ala Ala Glu Leu His Ala Gln Leu Glu Thr Val

1170 1190 1210 1230
GAC AGC GCG CCG GTA CAG CGT TGC GTA AAA AAA TGG GCG ATT TCG CCG AGC TGC GCG ACC TCC TGG AAC GCG CCA TTA TTG ACG CGC CGC
Asp Ser Ala Pro Val Gln Arg Cys Val Lys Lys Trp Ala Ile Ser Pro Ser Cys Ala Thr Ser Trp Asn Ala Pro Leu Leu Thr Arg Arg

1250 1270 1290 1310 1330
CGG TAC TGG TCG CGA CGG CGC GTT ATT GCG CCC GGC TAC CAT GAA GAG CTG GAC GAA TGG CGC GCG CTG GCG GAC GGC GCC ACC GAT TAT
Arg Tyr Trp Ser Arg Arg Arg Val Ile Ala Pro Gly Tyr His Glu Glu Leu Asp Glu Trp Arg Ala Leu Ala Asp Gly Ala Thr Asp Tyr

1350 1370 1390 1410
CTC GAT CGT CTG GAA ATT CGC GAG CGC GAG CGT ACC GGG CTG GAT ACG CTG AAA GTC GGC TAT AAC GCG GTA CAT GGT TAT TAC ATT CAG
Leu Asp Arg Leu Glu Ile Arg Glu Arg Glu Arg Thr Gly Leu Asp Thr Leu Lys Val Gly Tyr Asn Ala Val His Gly Tyr Tyr Ile Gln

1430 1450 1470 1490 1510
ATT AGC CGT GGT CAA AGC CAT CTG GCG CCC ATC AAT TAT GTG CGT CGC CAG ACG CTG AAA AAT GCC GAA CGC TAC ATT ATT CCT GAG CTT
Ile Ser Arg Gly Gln Ser His Leu Ala Pro Ile Asn Tyr Val Arg Arg Gln Thr Leu Lys Asn Ala Glu Arg Tyr Ile Ile Pro Glu Leu

1530 1550 1570 1590
AAA GAG TAC GAA GAT AAA GTT CTG ACC TCG AAA GGC AAA GCG CTG GCG CTG GAA AAA CAG CTA TAT GAC GAA TTG TTT GAT CTG CTG TTG
Lys Glu Tyr Glu Asp Lys Val Leu Thr Ser Lys Gly Lys Ala Leu Ala Leu Glu Lys Gln Leu Tyr Asp Glu Leu Phe Asp Leu Leu Leu

1610 1630 1650 1670 1690
CCG CAT CTG GCC GAT TTA CAG CAG AGC GCC AAC GCG CTG GCG GAG CTG GAT GTG CTG GTT AAC CTG GCC GAA CGC GCC TGG ACG CTG AAT
Pro His Leu Ala Asp Leu Gln Gln Ser Ala Asn Ala Leu Ala Glu Leu Asp Val Leu Val Asn Leu Ala Glu Arg Ala Trp Thr Leu Asn

1710 1730 1750 1770
TAC ACC TGC CCG ACA TTT ACC GAT AAA CCC GGT ATC CGT ATT ACC GAA GGT CGC CAC CCG GTG GTT GAA CAG GTA CTG AAT GAG CCG TTT
Tyr Thr Cys Pro Thr Phe Thr Asp Lys Pro Gly Ile Arg Ile Thr Glu Gly Arg His Pro Val Val Glu Gln Val Leu Asn Glu Pro Phe

1790 1810 1830 1850 1870
ATC GCT AAC CCG CTC AAT TTA TCG CCG CAG CGC CGG ATG TTG ATC ATT ACC GGC CCC AAT ATG GGC GGT AAA AGT ACC TAT ATG CGA CAA
Ile Ala Asn Pro Leu Asn Leu Ser Pro Gln Arg Arg Met Leu Ile Ile Thr Gly Pro Asn Met Gly Gly Lys Ser Thr Tyr Met Arg Gln

1890 1910 1930 1950
ACG GCA TTG ATT GCC CTG CTG GCC TAT ATC GGC AGT TAC GTT CCG GCG CAA AAC GTG GAA ATC GGC CCC ATC GAC CGT ATT TTT ACC CGC
Thr Ala Leu Ile Ala Leu Leu Ala Tyr Ile Gly Ser Tyr Val Pro Ala Gln Asn Val Glu Ile Gly Pro Ile Asp Arg Ile Phe Thr Arg

1970 1990 2010 2030 2050
GTC GGC GCA GCG GAC GAC CTG GCC TCC GGG CGT TCG ACC TTT ATG GTG GAG ATG ACC GAA ACC GCG AAC ATT CTG CAT AAT GCC ACG GAA
Val Gly Ala Ala Asp Asp Leu Ala Ser Gly Arg Ser Thr Phe Met Val Glu Met Thr Glu Thr Ala Asn Ile Leu His Asn Ala Thr Glu

2070 2090 2110 2130
AAC AGT CTG GTA TTG ATG GAT GAA ATC GGG CGC GGT ACC TCC ACG TAT GAC GGG CTG TCG TTG GCC TGG GCC TGC GCG GAG AAT CTC GCG
Asn Ser Leu Val Leu Met Asp Glu Ile Gly Arg Gly Thr Ser Thr Tyr Asp Gly Leu Ser Leu Ala Trp Ala Cys Ala Glu Asn Leu Ala

2150 2170 2190 2210 2230
AAT AAG ATT AAA GCG TTA ACG CTG TTC GCT ACC CAC TAC TTC GAG CTG ACC CAA TTA CCG GAG AAA ATG GAA GGC GTG GCG AAC GTC CAC
Asn Lys Ile Lys Ala Leu Thr Leu Phe Ala Thr His Tyr Phe Glu Leu Thr Gln Leu Pro Glu Lys Met Glu Gly Val Ala Asn Val His

2250 2270 2290 2310
CTG GAT GCG CTG GAA CAC GGC GAT ACT ATC GCG TTT ATG CAC AGC GTG CAG GAC GGC GCG GCA AGT AAG AGC TAT GGC CTG GCG GTC GCG
Leu Asp Ala Leu Glu His Gly Asp Thr Ile Ala Phe Met His Ser Val Gln Asp Gly Ala Ala Ser Lys Ser Tyr Gly Leu Ala Val Ala

2330 2350 2370 2390 2410
GCG CTT GCG GGC GTC CCT AAA GAA GTC ATC AAA CGC GCG CGC CAG AAA CTC CGC GAG CTG GAA AGC ATC TCG CCC AAT GCG GCG GCA ACG
Ala Leu Ala Gly Val Pro Lys Glu Val Ile Lys Arg Ala Arg Gln Lys Leu Arg Glu Leu Glu Ser Ile Ser Pro Asn Ala Ala Ala Thr

2430 2450 2470 2490
CAG GTG GAC GGC ACG CAA ATG TCG TTG CTT GCC GCC CCG GAG GAG ACA TCG CCT GCC GTT GAA GCG CTG GAG AAT CTC GAT CCG GAC TCC
Gln Val Asp Gly Thr Gln Met Ser Leu Leu Ala Ala Pro Glu Glu Thr Ser Pro Ala Val Glu Ala Leu Glu Asn Leu Asp Pro Asp Ser

2510 2530 2550 2570 2590
CTG ACG CCA CGC CAG GCG CTG GAG TGG ATC TAT CGG CTG AAA AGT CTG GTG TAAACGTGTCAGACCCGATAGCGCCGCGCTATCGGG
Leu Thr Pro Arg Gln Ala Leu Glu Trp Ile Tyr Arg Leu Lys Ser Leu Val END

FIG. 2. Nucleotide and deduced amino acid sequence of the *S. typhimurium mutS* gene. Underlined amino acids are those identical to the corresponding residues in the *S. pneumoniae* HexA protein when the sequences have been aligned optimally. The sequence of the boxed amino acids was confirmed by amino-terminal sequencing of purified MutS protein.

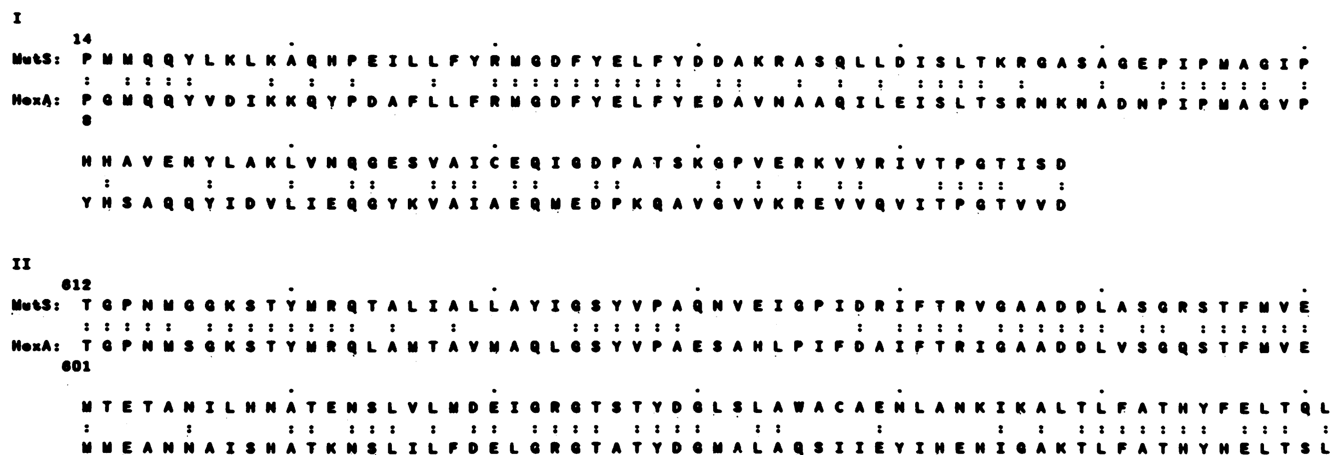

FIG. 3. Regions of strong homology between the MutS protein and the HexA protein. Identical residues are indicated by colons.

has been determined (33), and the prediction that HexA has an M_r of 95,000 brings the molecular weight of the protein even closer to that of MutS than was previously reported. We and Priebe et al. (33) compared the deduced amino acid sequences of the MutS and HexA proteins and found that they are approximately 36% homologous overall. This homology extends throughout the entire length of the two proteins and requires only a few small gaps for optimal alignment, although the greatest similarity is at the amino and carboxyl termini (Fig. 2).

Alignments of the MutS and HexA sequences in the two regions where they are most similar are presented in Fig. 3. Region I is near the amino terminus and consists of a stretch of 109 amino acids that are 54% homologous at the amino acid level and 58% homologous at the nucleotide level. A stretch of 10 amino acids is conserved exactly between the two proteins, suggesting that this region may have particular functional importance.

Region II is near the carboxyl terminus and contains 122 amino acids. The amino acid sequences within this region are 62% homologous, and the nucleotide sequences are 60% homologous. This region contains two stretches of high homology, one starting at amino acid 612 of MutS and having 13 of 14 exact amino acid matches and the other starting at amino acid 657 of MutS and having 15 of 17 exact amino acid matches. The first highly homologous stretch is also homologous with a consensus nucleotide-binding site, which is found in many ATPases (1, 11, 46) (Fig. 4A).

A number of DNA-binding proteins contain a helix-turn-helix motif that has been implicated in protein-DNA interactions (29). Both the MutS and HexA sequences contain homology to the consensus sequence for this motif, starting at amino acid residues 771 and 760 of MutS and HexA, respectively (Fig. 4B). Secondary structure analysis of the sequence by the method of Chou and Fasman (4) predicts the appropriate structures. The amino acid sequences of HexA and MutS are 40% homologous for this section, which falls 37 amino acids to the carboxyl-terminal side of region II.

The deduced amino acid sequences indicate that both MutS and HexA are highly charged proteins. MutS protein has an excess of 14 acidic residues (Asp and Glu minus Arg and Lys), and HexA has an excess of 34 acidic residues. Both the acidic and basic residues are distributed throughout the two proteins, and no region contains a concentration of basic residues.

In vitro mismatch repair assay. Our in vitro mismatch repair assay is conceptually similar to the one described by Lu et al. (22) but uses as the DNA substrate a hemimethylated M13mp8 heteroduplex containing a G/T mismatch. As expected, we were able to repeat their key results: (i) extracts from wild-type *E. coli* cells repaired the unmethylated strand almost 100% of the time, and (ii) extracts of *E. coli mutS*, *mutL*, and *mutH* mutants were repair incompetent but could complement each other completely. We also found that an extract of an *E. coli mutS215*::Tn*10* strain was fully complemented both by crude extracts of an *E. coli mutS* strain, which contains the cloned *S. typhimurium* $mutS^+$ gene, and by purified *S. typhimurium* MutS protein (data not shown). In contrast, to date we have been unable to detect complementation (<5%) of *E. coli mutS* extracts with extracts prepared by several different methods from either hex^+ *S. pneumoniae* or *S. pneumoniae* containing the $hexA^+$ gene on a low-copy-number plasmid. However, our ability to detect this repair is limited, due to inhibition of repair of wild-type *E. coli* extracts by our *S. pneumoniae* extracts. Repair was inhibited 30% in an assay mix containing 20% *S. pneumoniae* protein, and increasing amounts of *S. pneumoniae* extract increased the amount of inhibition.

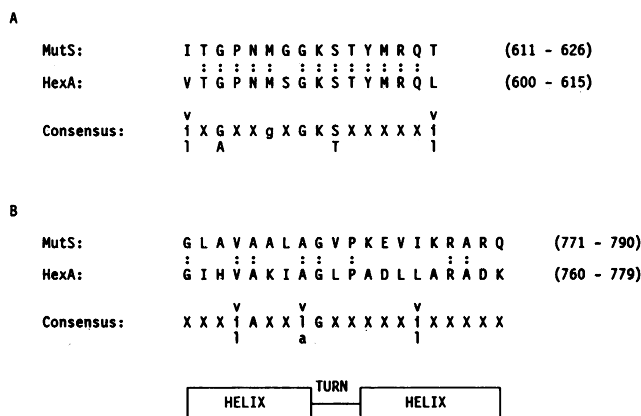

FIG. 4. MutS and HexA homology to consensus sequences. (A) Homology to the nucleotide-binding site consensus first reported by Walker et al. (46), as shown by Gill et al. (11). (B) Homology to helix-turn-helix consensus (1, 29). Uppercase letters indicate virtually invariant residues, lowercase letters indicate conserved residues (>60% of sites cited), and X's indicate variable residues.

DISCUSSION

We have determined the sequence of the *S. typhimurium* *mutS* gene and confirmed both the start site and the predicted initial protein sequence by determining the amino-terminal sequence of the purified protein. The MutS protein from the gram-negative bacterium *S. typhimurium* is approximately 36% homologous to the HexA protein from the gram-positive bacterium *S. pneumoniae*. The finding that MutS and HexA are homologous raises the possibility that, like MutS, HexA also binds mismatched DNA base pairs and single-stranded DNA. However, in spite of the similarity between the MutS and HexA proteins, the repair systems of which they are a part must differ significantly, due to their differing mechanisms of strand discrimination. A need to interact with other components of their respective repair systems might require the MutS and HexA proteins to differ in certain respects.

Our failure to observe in vitro complementation of an *E. coli mutS*::Tn*10* extract by *hexA*$^+$ *S. pneumoniae* extracts could be due to our inability to identify appropriate conditions for preparing the *S. pneumoniae* extracts, or it may indeed be due to an inability of HexA to substitute for MutS. Such a failure to substitute could be due to an inability of the *S. pneumoniae* HexA protein to interact with other components of the *E. coli* mismatch repair system; regions where the MutS and HexA proteins show little homology could be involved in such interactions. Alternatively, a failure to substitute could be due to inherent differences in the fundamental biochemical activities of the two proteins. However, given the overall homology between the two proteins, we consider this possibility less likely.

We found two regions of over 50% amino acid homology between the two proteins. The finding that the carboxy terminus contains the helix-turn-helix motif that is found in many proteins that bind double-stranded DNA (29) raises the possibility that this region plays a role in the recognition of mismatched base pairs. Most of the proteins containing this motif recognize a specific sequence. However, both the UvrB protein and the product of the *alkA*$^+$ gene, 3-methyladenine–DNA glycosylase II, also appear to contain this motif (1). Like MutS, these two proteins are involved in repair systems that can process several different DNA structures. The AlkA glycosylase recognizes several different alkylated bases, whereas the UvrABC excinuclease, of which UvrB is a component, recognizes a variety of bulky adducts. Interestingly, the carboxyl-terminal region of the MutS protein also includes homology to a consensus sequence thought to be diagnostic of nucleotide-binding sites (1, 11, 46). Whereas exogenous ATP is required for the in vitro repair reaction of Lu et al. (22, 23), it is not needed for MutS to bind mismatches or single-stranded DNA (41; Haber, Pang, and Walker, unpublished data). However, binding or hydrolysis of ATP or another nucleotide by MutS might be required for another step in the repair process. Tn*1000* insertions close to the carboxyl terminus of the MutS protein produced truncated proteins that were inactive in mismatch repair (30), suggesting that at least the carboxyl-terminal region of MutS is required for mismatch repair. One could speculate that if the carboxyl-terminal region were to be involved in the recognition of mismatches, then the amino-terminal region might be involved in binding single-stranded DNA.

The sequence similarity between *mutS* and *hexA* suggests that they arose from a common ancestor gene that evolved before the gram-positive and gram-negative bacteria diverged. Because the sequence similarity stretches over the entire length of the corresponding proteins, and only a few small gaps are necessary for optimal alignment, it is unlikely that the similarities arose by convergent evolution. This suggests that a system for mismatch correction arose early in evolution, a conclusion not too surprising considering that mismatch repair-deficient cells, at least in the case of *E. coli*, *S. typhimurium*, and *S. pneumoniae*, have a spontaneous mutation frequency increased by 10 to 1,000 (6, 19, 30, 40, 42, 43). Our observations and those of Priebe et al. (33) raise the possibility that proteins evolutionarily related to MutS and HexA might exist in widely diverged organisms, including eucaryotes. The recent isolation of *S. cerevisiae* mutants which may be defective in heteroduplex repair (49) and the development of both a cell-free *S. cerevisiae* mismatch repair system (27) and a mammalian in vivo mismatch repair assay (13) will help address such questions. It will also be interesting to determine whether any other components of the mismatch repair systems of *E. coli* and *S. typhimurium* have counterparts in *S. pneumoniae*. For example, the reported molecular weights of the *S. typhimurium* MutL protein and the *S. pneumoniae* HexB protein are sufficiently close to raise the possibility that they are related (30, 32).

ACKNOWLEDGMENTS

We thank Sanford Lacks and his colleagues for sharing with us the nucleotide sequence of *hexA* before publication and sending us *S. pneumoniae* strains and the *hexA*$^+$ plasmid. We thank R. Fishel, P. Schimmel, and members of the Walker laboratory, particularly J. Battista, for helpful discussions. We also thank R. Fishel for his improvements of the repair assay, W. Gilbert for aid with computer work, J. Essigman for his gift of the mutant M13mp8 bacteriophage, M. Radman for communicating unpublished results, P. Modrich and M. Meselson for sending copies of their reviews before publication, and S. Blackmon for performing the N-terminal sequencing.

This work was supported by grant NP-461A from the American Cancer Society. L.H. was supported by Public Health Service training grant GM07287-12 from the National Institutes of Health, and J.M. was supported by Public Health Service grant IF32GM 10792 from the National Institutes of Health.

LITERATURE CITED

1. **Backendorf, C., H. Spaink, A. P. Barbeiro, and P. van de Putte.** 1986. Structure of the *uvrB* gene of *Escherichia coli*. Homology with other DNA repair enzymes and characterization of the *uvrB5* mutation. Nucleic Acids Res. **14**:2877–2890.
2. **Balganesh, T. S., and S. A. Lacks.** 1985. Heteroduplex DNA mismatch repair system of *Streptococcus pneumoniae*: cloning and expression of the *hexA* gene. J. Bacteriol. **162**:979–984.
3. **Chase, J. W., and K. R. Williams.** 1986. Single-stranded DNA binding proteins required for DNA replication. Annu. Rev. Biochem. **55**:103–136.
4. **Chou, P. Y., and G. D. Fasman.** 1978. Prediction of secondary structure of proteins from their amino acid sequence. Adv. Enzymol. **47**:45–148.
5. **Claverys, J.-P., and S. A. Lacks.** 1986. Heteroduplex deoxyribonucleic acid base mismatch repair in bacteria. Microbiol. Rev. **50**:133–165.
6. **Cox, E. C.** 1976. Bacterial mutator genes and the control of spontaneous mutation. Annu. Rev. Genet. **10**:135–156.
7. **Cox, E. C., G. E. Degnen, and M. L. Scheppe.** 1972. Mutator gene studies in *Escherichia coli*: the *mutS* gene. Genetics **72**:551–567.
8. **Ephrussi-Taylor, H., and T. C. Gray.** 1966. Genetic studies of recombining DNA in pneumococcal transformation. J. Gen. Physiol. **49**:211–231.
9. **Fishel, R. A., and R. Kolodner.** 1984. An *Escherichia coli* cell-free system that catalyses the repair of symmetrically methylated heteroduplex DNA. Cold Spring Harbor Symp. Quant. Biol. **49**:603–609.

10. **Fishel, R. A., E. C. Siegel, and R. Kolodner.** 1986. Gene conversion in *Escherichia coli*: resolution of heteroallelic mismatched nucleotides by co-repair. J. Mol. Biol. **188**:147–157.
11. **Gill, D. R., G. F. Hatfull, and G. P. C. Salmond.** 1986. A new cell division operon in *Escherichia coli*. Mol. Gen. Genet. **205**:134–145.
12. **Guild, W. R., and N. B. Shoemaker.** 1976. Mismatch correction in pneumococcal transformation: donor length and *hex*-dependent marker efficiency. J. Bacteriol. **125**:125–135.
13. **Hare, J. T., and J. H. Taylor.** 1985. One role for DNA methylation in vertebrate cells is strand discrimination in mismatch repair. Proc. Natl. Acad. Sci. USA **82**:7350–7354.
14. **Hickson, I. D., H. M. Arthur, D. Bramhill, and P. T. Emmerson.** 1983. The *E. coli uvrD* gene product is DNA helicase II. Mol. Gen. Genet. **190**:265–270.
15. **Kramer, B., W. Kramer, and H.-J. Fritz.** 1984. Different base/base mismatches are corrected with different efficiencies by the methyl-directed DNA mismatch-repair system of *E. coli*. Cell **38**:879–887.
16. **Kumura, K., and M. Sekiguchi.** 1984. Identification of the *uvrD* gene product of *Escherichia coli* as DNA helicase II and its induction by DNA-damaging agents. J. Biol. Chem. **259**:1560–1565.
17. **Kushner, S. R., V. F. Maples, A. Easton, I. Farrance, and P. Peramachi.** 1983. Physical, biochemical, and genetic characterization of the *uvrD* gene product. ICN-UCLA Symp. Mol. Cell. Biol. **11**:153–159.
18. **Lacks, S.** 1966. Integration efficiency and genetic recombination in pneumococcal transformation. Genetics **53**:207–235.
19. **Lacks, S. A., J. J. Dunn, and B. Greenberg.** 1982. Identification of base mismatches recognized by the heteroduplex-DNA-repair system of *Streptococcus pneumoniae*. Cell **31**:327–336.
20. **Lieb, M., E. Allen, and D. Read.** 1986. Very short patch mismatch repair in phage lambda: repair sites and length of repair tracts. Genetics **114**:1041–1060.
21. **Loechler, E. L., C. L. Green, and J. M. Essigman.** 1984. *In vivo* mutagenesis by O^6-methylguanine built into a unique site in a viral genome. Proc. Natl. Acad. Sci. USA **81**:6271–6275.
22. **Lu, A.-L., S. Clark, and P. Modrich.** 1983. Methyl-directed repair of DNA base-pair mismatches *in vitro*. Proc. Natl. Acad. Sci. USA **80**:4639–4643.
23. **Lu, A.-L., K. Welsh, S. Clark, S.-S. Su, and P. Modrich.** 1984. Repair of DNA base-pair mismatches in extracts of *Escherichia coli*. Cold Spring Harbor Symp. Quant. Biol. **49**:589–596.
24. **Martin, B., H. Prats, and J.-P. Claverys.** 1985. Cloning of the *hexA* mismatch repair gene of *Streptococcus pneumoniae* and identification of the product. Gene **34**:293–303.
25. **Mitchell, M. B.** 1955. Aberrant recombination of pyridoxine mutants of Neurospora. Proc. Natl. Acad. Sci. USA **41**:215–220.
26. **Modrich, P.** 1987. DNA mismatch correction. Annu. Rev. Biochem. **56**:435–466.
27. **Muster-Nassal, C., and R. Kolodner.** 1986. Mismatch correction catalyzed by cell-free extracts of *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA **83**:7618–7622.
28. **Nevers, P., and H.-C. Spatz.** 1975. *Escherichia coli* mutants *uvrD* and *uvrE* deficient in gene conversion of lambda heteroduplexes. Mol. Gen. Genet. **139**:233–243.
29. **Pabo, C. O., and R. T. Sauer.** 1984. Protein-DNA recognition. Annu. Rev. Biochem. **53**:293–321.
30. **Pang, P. P., A. S. Lundberg, and G. C. Walker.** 1985. Identification and characterization of the *mutL* and *mutS* gene products of *Salmonella typhimurium* LT2. J. Bacteriol. **163**:1007–1015.
31. **Pang, P. P., S. D. Tsen, A. S. Lundberg, and G. C. Walter.** 1984. The *mutH*, *mutL*, *mutS*, and *uvrD* genes of *Salmonella typhimurium* LT2. Cold Spring Harbor Symp. Quant. Biol. **49**:597–602.
32. **Prats, H., B. Martin, and J.-P. Claverys.** 1985. The *hexB* mismatch repair gene of *Streptococcus pneumoniae*: characterization, cloning, and identification of the product. Mol. Gen. Genet. **200**:482–489.
33. **Priebe, S. D., S. M. Hadi, B. Greenberg, and S. A. Lacks.** 1988. Nucleotide sequence of the *hexA* gene for DNA mismatch repair in *Streptococcus pneumoniae* and homology of *hexA* to *mutS* of *Escherichia coli* and *Salmonella typhimurium*. J. Bacteriol. **170**:190–196.
34. **Pukkila, P. J., J. Peterson, G. Herman, P. Modrich, and M. Meselson.** 1983. Effects of high levels of DNA adenine methylation on methyl-directed mismatch repair in *Escherichia coli*. Genetics **104**:571–582.
35. **Radman, M., and R. Wagner.** 1986. Mismatch repair in *Escherichia coli*. Annu. Rev. Genet. **20**:523–538.
36. **Radman, M., R. E. Wagner, B. W. Glickman, and M. Meselson.** 1980. DNA methylation, mismatch correction and genetic stability, p. 121–130. *In* M. Alacevic (ed.), Progress in environmental mutagenesis. Elsevier/North-Holland, Amsterdam.
37. **Rosenberg, M., and D. Court.** 1979. Regulatory sequences involved in the promotion and termination of RNA transcription. Annu. Rev. Genet. **13**:319–353.
38. **Rossignol, J.-L.** 1969. Existence of homogeneous categories of mutants exhibiting various conversion patterns in Gene 75 of *Ascobolus immersus*. Genetics **63**:795–805.
39. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
40. **Shanabruch, W. G., I. Behlau, and G. C. Walker.** 1981. Spontaneous mutators of *Salmonella typhimurium* LT2 generated by insertion of transposable elements. J. Bacteriol. **147**:827–835.
41. **Su, S.-S., and P. Modrich.** 1986. *Escherichia coli mutS*-encoded protein binds to mismatched DNA base pairs. Proc. Natl. Acad. Sci. USA **83**:5057–5061.
42. **Tiraby, J.-G., and M. S. Fox.** 1973. Marker discrimination in transformation and mutation of pneumococcus. Proc. Natl. Acad. Sci. USA **70**:3541–3545.
43. **Tiraby, G., and M. A. Sicard.** 1973. Integration efficiencies of spontaneous mutant alleles of *amiA* locus in pneumococcal transformation. J. Bacteriol. **116**:1130–1135.
44. **Wagner, R., C. Dohet, M. Jones, M.-P. Doutriaux, F. Hutchinson, and M. Radman.** 1984. Involvement of *Escherichia coli* mismatch repair in DNA replication and recombination. Cold Spring Harbor Symp. Quant. Biol. **49**:611–615.
45. **Wagner, R., Jr., and M. Meselson.** 1976. Repair tracts in mismatched DNA heteroduplexes. Proc. Natl. Acad. Sci. USA **73**:4135–4139.
46. **Walker, J. E., M. Saraste, M. J. Runswick, and N. J. Gay.** 1982. Distantly related sequences in the α- and β-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J. **1**:945–951.
47. **White, R. L., and M. S. Fox.** 1974. On the molecular basis of high negative interference. Proc. Natl. Acad. Sci. USA **71**:1544–1548.
48. **Wildenberg, J., and M. Meselson.** 1975. Mismatch repair in heteroduplex DNA. Proc. Natl. Acad. Sci. USA **72**:2202–2206.
49. **Williamson, M. S., J. C. Game, and S. Fogel.** 1985. Meiotic gene conversion in *Saccharomyces cerevisiae*. I. Isolation and characterization of *pms1-1* and *pms1-2*. Genetics **110**:609–646.