Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 4468–4472, July 1985
Genetics

Nucleotide sequence of the *Dpn* II DNA methylase gene of *Streptococcus pneumoniae* and its relationship to the *dam* gene of *Escherichia coli*

(restriction enzymes/DNA methylation/cloning in Gram-positive bacteria/bacterial evolution/heteroduplex DNA base mismatch repair)

BRUNO M. MANNARELLI, T. S. BALGANESH, BILL GREENBERG, SYLVIA S. SPRINGHORN, AND SANFORD A. LACKS

Biology Department, Brookhaven National Laboratory, Upton, NY 11973

Communicated by Hamilton O. Smith, March 12, 1985

ABSTRACT The structural gene (*dpnM*) for the *Dpn* II DNA methylase of *Streptococcus pneumoniae*, which is part of the *Dpn* II restriction system and methylates adenine in the sequence 5′-G-A-T-C-3′, was identified by subcloning fragments of a chromosomal segment from a *Dpn* II-producing strain in an *S. pneumoniae* host/vector cloning system and demonstrating function of the gene also in *Bacillus subtilis*. Determination of the nucleotide sequence of the gene and adjacent DNA indicates that it encodes a polypeptide of 32,903 daltons. A putative promoter for transcription of the gene lies within a hundred nucleotides of the polypeptide start codon. Comparison of the coding sequence to that of the *dam* gene of *Escherichia coli*, which encodes a similar methylase, revealed 30% of the amino acid residues in the two enzymes to be identical. This homology presumably reflects a common origin of the two genes prior to the divergence of Gram-positive and Gram-negative bacteria. It is suggested that the restriction function of the gene is primitive, and that the homologous restriction system in *E. coli* has evolved to play an accessory role in heteroduplex DNA base mismatch repair.

Strains of *Streptococcus pneumoniae* contain one of two complementary and incompatible restriction systems (1, 2). Some strains contain the endonuclease *Dpn* I, which is unusual in that it acts only on the methylated DNA sequence 5′-G-m^{6}A-T-C-3′ (3, 4); the DNA in these strains is not methylated at this site. Other strains contain the complementary endonuclease *Dpn* II, which acts only on unmethylated 5′-G-A-T-C-3′ sites (4, 5). The latter strains contain a DNA methylase that methylates adenine at these sites. The genes for the *Dpn* II DNA methylase and endonuclease appear to be linked because they are simultaneously transferred in bacterial transformation by chromosomal DNA (2).

A segment of chromosomal DNA that expresses the *Dpn* II DNA methylase, but not the endonuclease, was recently cloned in the *S. pneumoniae*/pMP5 host/vector system (6). The recombinant plasmid containing the methylase gene could be transferred to a *Dpn* I-containing strain only when expression of *Dpn* I was turned off by a mechanism as yet unknown (7). In the present work the gene encoding the methylase was identified, and its nucleotide sequence was determined. The DNA sequence adjacent to the structural gene was also examined to explore possible mechanisms controlling its expression. A likely promoter for its transcription was identified.

The *dam* gene (8) of *Escherichia coli* encodes a methylase with the same specificity as the *Dpn* II methylase (4). The nucleotide sequence of the *dam* gene was recently determined (9), and it was of interest to compare the amino acid sequences deduced for the two polypeptides of similar function but different origin. Despite the considerable evolutionary divergence of the source bacteria, one being Gram-positive and the other Gram-negative, significant homology was detected between the protein products of their chromosomally located methylase genes. The apparently common origin of the genes raises interesting questions relating to the evolution and function of DNA methylation. Implications of the present work for defense against viral invasion, heteroduplex DNA base mismatch repair, and both positive and negative control of gene function in prokaryotes and eukaryotes are discussed.

MATERIALS AND METHODS

Bacterial Strains and Plasmids. Strains of *S. pneumoniae* used as recipients in transformation were 762 (*malM558 end-1*) and 777 (*malDXMP581 end-1*); both strains are derivatives of Rx1, which has the null restriction phenotype (2). Strains 678 (*end-1*) and 697 (*end-1 str^r*) have the *Dpn* II phenotype. Plasmids used were pMP8 (*tet^r dpnM$^+$*) (6), pLS69 (*tet^r malM$^+$*) (10), and pLS139 (*tet^r*). Plasmids were transferred to *Bacillus subtilis* strain MB11 (*lys-3 metB10 hisH2*) as described with selection for tetracycline resistance (Tc^r) (11).

Plasmid DNA Preparation. Purified plasmids were prepared by the method of Currier and Nester (12). Crude plasmid extracts called alkaline lysates (10) and cleared lysates (11) were prepared as previously described.

Culture Growth and Transformation. Cultures were grown in a semisynthetic medium based on casein hydrolysate (13) and were supplemented with 0.2% sucrose and a 1:50 dilution of fresh yeast extract. Transformation was carried out as previously described (10). Tc^r transformants were selected with tetracycline at 1.0 μg/ml. Maltose-utilizing transformants were selected by substituting maltose for sucrose and eliminating the fresh yeast extract. Clones were isolated in pour plates containing 1% agar. To select plasmids carrying the *dpnM* gene, a mixed plasmid preparation from a bulk culture of transformants obtained with ligated DNA was treated with *Dpn* II and used to transform a fresh recipient culture, as previously described (6).

Restriction Mapping and Subcloning of Chromosomal DNA. Except for *Dpn* II (14) restriction endonucleases and other enzymes came from commercial sources and were used as indicated by the supplier. Fragments were analyzed by gel electrophoresis in 1% agarose or 5% polyacrylamide. Mixtures of fragments from different plasmids were ligated as described (10) for use in transformation. Methylation of plasmid DNA at G-A-T-C sites was demonstrated by resistance to cleavage by *Dpn* II.

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.

Abbreviations: Tc^r, tetracycline resistance (resistant); kb, kilobase pair(s); bp, base pair(s).

Methylase Assay. DNA methylase activity in cell extracts was measured as described (6) with substrate DNA from strain 213, a *Dpn* I producer.

Nucleotide Sequence Determination. Plasmids were further purified by treatment with RNase and gel filtration. After cleavage with restriction enzymes, the DNA fragments were treated with phosphatase and labeled at their 5′ ends with [γ-^{32}P]ATP and polynucleotide kinase. After cutting with a second restriction enzyme, DNA sequences were determined by the method of Maxam and Gilbert (15).

RESULTS

Localization of the Methylase Gene in Cloned DNA. A 3.7-kilobase-pair (kb) *Bam*HI fragment of the chromosome of the *Dpn* II-containing strain 697 of *S. pneumoniae* was originally inserted into the vector pMP5 in both possible orientations to give the recombinant plasmids pMP8 (Fig. 1) and pMP10 (6). These plasmids contain the *dpnM* gene, which encodes the *Dpn* II DNA adenine methylase. To define the position of this gene in the cloned DNA, a series of plasmids reduced in size were prepared. An attempt was made to remove the 3.8-kb *Bst*EII fragment from pMP8. Transformation of cells with ligated *Bst*EII-cleaved pMP8 fragments gave no Tc^r transformants containing plasmids that lacked the 3.8-kb piece. However, one plasmid had apparently undergone a 3.4-kb deletion in that region to remove one *Bst*EII site and give the structure shown as pMP12 (Fig. 1). The removal of the 4.6-kb *Cla* I piece from pMP12 gave pMP7. Host cells carrying pMP7, which were otherwise devoid of the enzyme, produced the *dpnM* methylase. This was indicated both by the susceptibility of plasmid DNA from these cells to *Dpn* I but not *Dpn* II (data not shown) and by measurement of DNA methylase activity in extracts (Table 1).

Localization of *dpnM* within the truncated chromosomal insert of pMP7 was accomplished by ligating the 2.7-kb *Nco* I fragment of that plasmid to a 5.4-kb *Nco* I fragment of pLS69, which contains the *malM* gene and the replication apparatus of pMV158 that is present in all of the plasmids shown in Fig. 1. Cleavage with *Nco* I destroyed the *tet* gene of pLS69, so cells of the maltose-negative strain 777, which produces neither the methylase nor *Dpn* I, were transformed with the ligated DNA, and pLS60 was obtained by selecting, first, for plasmids that conferred the maltose-utilizing phenotype, and then, among these, for those resistant to *Dpn* II. Expression of methylase by pLS60-containing cells showed that the *dpnM* gene lies within the 2.0-kb *Nco* I/*Cla* I segment of the insert. The gene was further localized to the 1.5-kb *Hae* III fragment within this segment as shown in pMP13 (Fig. 1). The latter plasmid was obtained by ligating the mixture of fragments resulting from *Hae* III cleavage of pMP7 and pLS69. Screening of plasmids that conferred Tc^r and were resistant *in vitro* to *Dpn* II yielded pMP13. This plasmid contains a 1.2-kb direct repeat in the *tet* region that results in frequent deletion of the intervening segment containing the *dpnM* gene. Consequently, plasmid preparations from pMP13-containing cells examined by gel electrophoresis showed a deleted plasmid as well as pMP13. Only the latter plasmid was completely resistant to *Dpn* II treatment (data not shown). Plasmid pMP6 was obtained by removing the 2.4-kb *Eco*RI piece from pMP7. This plasmid failed to express the methylase, which indicated that the *Eco*RI site lies within the coding region.

The levels of methylase activity expressed by the variously configured plasmids are similar (Table 1). This suggests that the methylase fragment carried its own transcriptional promoter in all of the cases studied. The putative promoter and the entire *dpnM* gene must reside within the 1.5-kb *Hae* III segment of the cloned DNA.

Expression of the Methylase Gene in *B. subtilis*. From the above data it appears likely that *dpnM* is the structural gene for the methylase. However, inasmuch as some strains of *S. pneumoniae* make the methylase and some do not, it is conceivable that the structural gene is present in all pneumococcal strains and that *dpnM* encodes a regulatory element, which allows expression of the structural gene. Thus it was desirable to transfer the *dpnM* gene into a foreign host, such as *B. subtilis*, to check whether it alone encoded the methylase. This was accomplished by using pLS62, a plasmid similar in structure to pLS60 (Fig. 1) and obtained in the same construction experiment, but that contained in addition the 1.4-kb *Nco* I fragment of pMP7 (located between 5.0 and 6.4 on the map) inserted into the *Nco* I site of pLS60 located at 7.0 (Fig. 1) to restore a functional *tet* gene. Expression of methylase in *B. subtilis* containing pLS62 was detected by sensitivity of the cellular DNA, both plasmid and

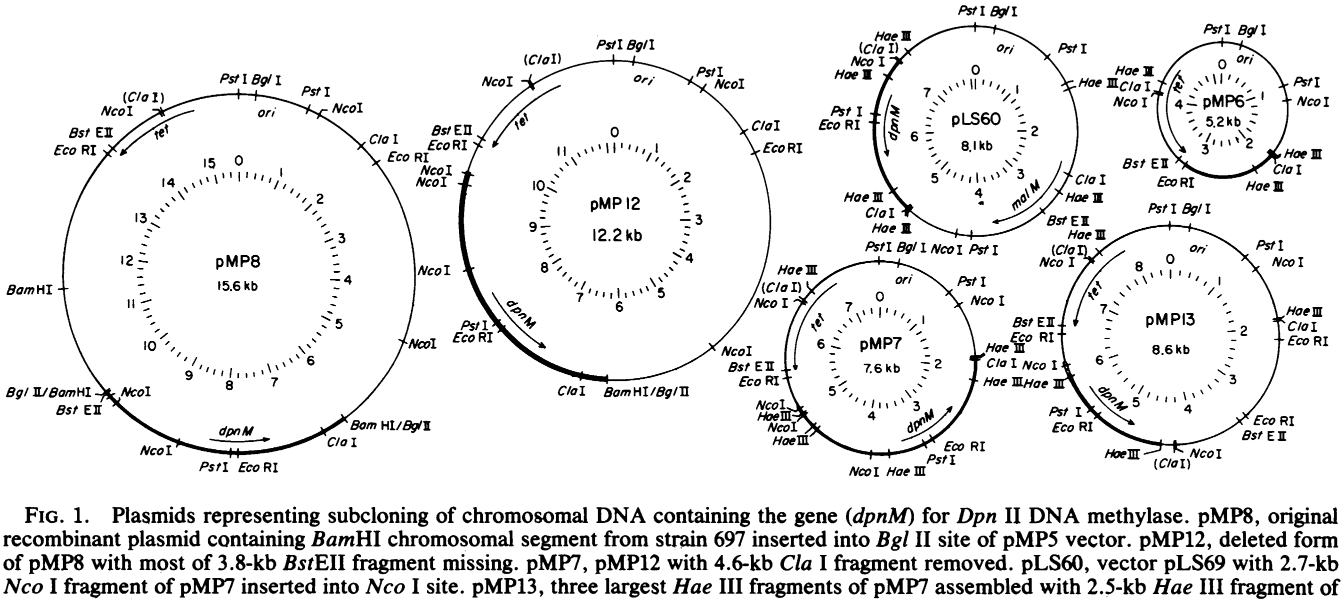

FIG. 1. Plasmids representing subcloning of chromosomal DNA containing the gene (*dpnM*) for *Dpn* II DNA methylase. pMP8, original recombinant plasmid containing *Bam*HI chromosomal segment from strain 697 inserted into *Bgl* II site of pMP5 vector. pMP12, deleted form of pMP8 with most of 3.8-kb *Bst*EII fragment missing. pMP7, pMP12 with 4.6-kb *Cla* I fragment removed. pLS60, vector pLS69 with 2.7-kb *Nco* I fragment of pMP7 inserted into *Nco* I site. pMP13, three largest *Hae* III fragments of pMP7 assembled with 2.5-kb *Hae* III fragment of pLS69; contains 1.2-kb direct repeat of distal part of *tet* gene; localizes *dpnM* in 1.5-kb *Hae* III segment. pMP6, pMP7 with 2.4-kb *Eco*RI fragment removed. Solid bar, chromosomal DNA contiguous to *dpnM*; hatched bar (pMP13), noncontiguous chromosomal DNA. Arrows indicate extent of structural gene and direction of transcription: *tet*, tetracycline resistance; *malM*, amylomaltase; *dpnM*, methylase; *ori*, putative origin of replication; (*Cla* I), *Cla* I site blocked by methylation at G-A-T-C-G-A-T.

Table 1. DNA methylase activity in strains of *S. pneumoniae* containing plasmids that carry the *dpnM* gene

Host strain	Plasmid	DNA methylase, pmol/hr per mg of protein
762	None	<10
762	pMP7	3600
762	pMP8	2740
762	pMP13	2260
777	None	<10
777	pLS60	1690
678	None	345

chromosomal, to *Dpn* I and its resistance to *Dpn* II (Fig. 2). A control strain of *B. subtilis* containing an unrelated recombinant plasmid, pLS139, showed no such methylation of its DNA. It is therefore concluded that *dpnM* is the structural gene for the methylase.

Nucleotide Sequence of the Methylase Gene. The nucleotide sequence on both DNA strands of the *dpnM* gene carried by plasmid pMP7 was determined by chemical cleavage of end-labeled DNA (15). A detailed restriction map of pMP7 was prepared; sites for 17 restriction enzymes are shown in Fig. 3. The strategy used for the DNA sequence determination of a 2025-bp segment containing the *dpnM* gene is indicated below the map.

The DNA sequence was examined for the presence of open reading frames in both directions (Fig. 4). When the sequence was read from right to left in the orientation shown in Fig. 3, no open reading frame greater than 300 bp in length was found. When read from left to right, only three open reading frames longer than 160 bp were observed. The two leftmost were in phase 3, and the third, which slightly overlapped the second, was in phase 1, as shown in Fig. 4. The second open reading frame, beginning at nucleotide 394, with an ATG start codon at 439, and extending to 1290, corresponds to the *dpnM* gene. It is the only complete coding sequence contained with the *Nco* I/*Cla* I segment of pLS60 and the *Hae* III segment of pMP13, and it is interrupted by the *Eco*RI site in the cloned methylase fragment. The *dpnM* gene would thus encode a polypeptide of 284 amino acid residues or 32,903 daltons. The open reading frame that extends past the *Cla* I site, and hence out of the chromosomal segment cloned in pMP7, has two possible start codons near its origin. The polypeptide encoded by this chromosomal gene would be at least 26,000 daltons.

The nucleotide sequence of the DNA strand corresponding to the mRNA presumably transcribed from the region for which the sequence was determined is shown in Fig. 5. The amino acid sequence encoded by the methylase gene is also shown. The start site indicated is the only ATG or GTG codon found within 375 nucleotides from the preceding TAG stop codon. However, unlike the situation found for other pneumococcal genes (16), there is no sequence corresponding to a strong ribosome binding site (17) associated with the methylase start site. Similarly, there is no ribosome binding sequence preceding the first ATG codon in the third open reading frame. In both cases, however, an identical sequence, A-A-T-T-T-C-T...4 or 5 nucleotides...T-A-T-A...9 or 10 nucleotides..., precedes the putative start codon. It is conceivable that this sequence plays a role in translation of the products of these genes. It should be pointed out, though, that (*i*) the methylase reading frame overlaps this start site in the third reading frame and (*ii*) that reading frame has a strong ribosome binding site associated with its second potential start site, which lies past the methylase gene.

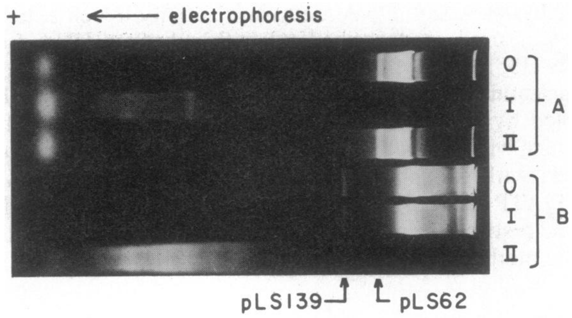

FIG. 2. Expression of the *dpnM* methylase gene in *B. subtilis*. Cleared lysates of strain MB11 containing pLS62 (*A*) or pLS139 (*B*) were subjected to electrophoresis in 1% agarose without prior treatment (0) or after treatment with *Dpn* I (I) or *Dpn* II (II). Both pLS62 and pLS139 contain the *tet* gene; only pLS62 contains *dpnM*. Expression of *dpnM* makes DNA sensitive to *Dpn* I and resistant to *Dpn* II. Arrows mark the covalently closed forms of the plasmids.

Evidence showing the similarity in amount of DNA methylase synthesized by cells containing plasmids with variously truncated chromosomal inserts in different orientations (Table 1; ref. 6) suggests that the methylase gene carries its own transcriptional promoter and that this promoter is situated within the *Hae* III fragment containing the *dpnM* gene. A putative promoter sequence beginning at nucleotide 343, T-T-G-A-T-A...17 nucleotides...T-A-A-A-A-T, corresponds to the *E. coli* consensus sequence for promoters (18) with only a single base deviation in each of the RNA polymerase binding sites.

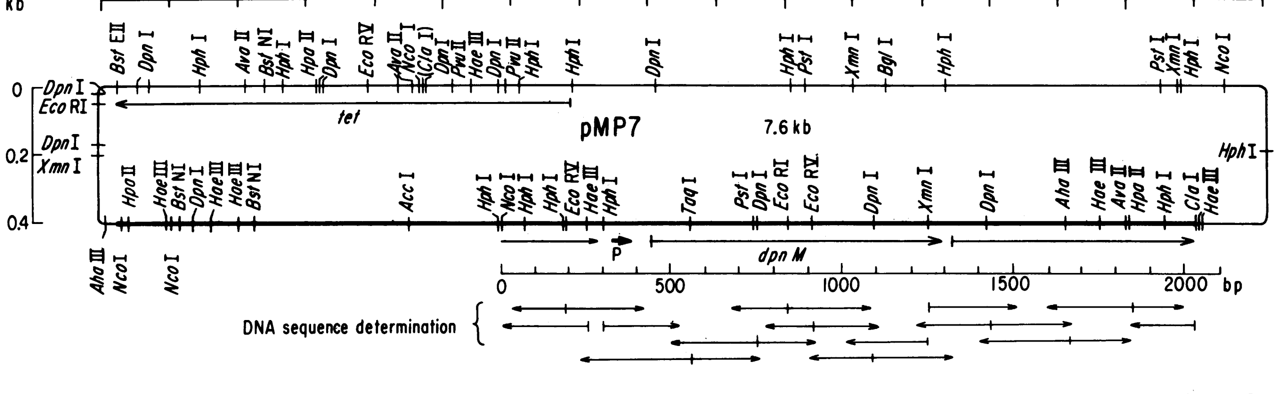

FIG. 3. Physical map of plasmid pMP7. Thin line, vector portion; solid bar, chromosomal insert. All restriction sites are shown for the enzymes indicated except *Taq* I for which only a single site used in determining DNA sequence is shown. Nucleotide sequence was determined for segments labeled at restriction sites indicated by vertical marks and extending to arrowheads. Arrows under bar indicate open reading frames and putative promoter (*P*). bp, Base pairs.

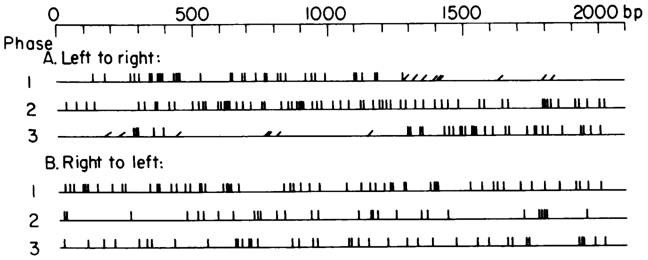

FIG. 4. Open reading frames in the vicinity of the *Dpn* II methylase gene. Vertical marks indicate positions of nonsense codons. Oblique marks indicate potential ATG start codons in the three longest frames, reading from left to right.

Comparison with the *E. coli dam* Methylase. Brooks *et al.* (9) have recently determined the nucleotide sequence of the *dam* methylase of *E. coli*, which, like the *Dpn* II methylase of *S. pneumoniae*, methylates adenine in DNA at G-A-T-C sites (4). The amino acid sequence encoded by the *dam* gene is compared to that encoded by the *dpnM* gene in Fig. 5. A considerable homology between the two proteins is evident. The *dpnM* polypeptide contains 284 amino acid residues; the *dam* polypeptide is 6 residues shorter. When deletions in the *dam* gene were postulated to maximize the correspondence, 30% of the residues in the two proteins were identical. Although the portion of the *dpnM* gene coding the amino end of the protein was considerably deleted in *dam*, and the carboxyl-terminal portions were very different, stretches of strong homology were found throughout the polypeptide chains. For example, correspondences of 75% or more are evident for amino acid stretches 14–21, 41–47, 141–151, and 191–203 (numbered in the *dpnM* chain from the amino end). The distribution and extent of similarity between the amino acid residues of the two proteins indicate that the genes that encode them share a common evolutionary origin.

Homology between the *dpnM* and *dam* genes is evident also at the nucleotide level, but only in the regions of amino acid correspondence. Whereas the first two bases in the codons for corresponding amino acids show 95% (160/168) identity, the third (or wobble) bases correspond in only 40% (34/84) of these codons. Codon bases outside the corresponding amino acid sequences show 27% (153/561) identity, which is only 2% higher than the random expectation. Overall base correspondence in the two genes is 43%. Apparently the *dpnM* and *dam* genes have evolved separately over a sufficiently long period of time that without functional constraints causing the retention of particular amino acids, the base composition was randomized by mutation.

DISCUSSION

A cloned 1.5-kb segment of chromosomal DNA from a *Dpn* II strain of *S. pneumoniae* was necessary and sufficient for production of the *Dpn* II methylase in a host strain that did not otherwise produce the enzyme. The only intact open reading frame within this segment, which codes for a poly-

```
GCC CAT GGT GGC TAT CTA TTC ACG CTA TGT GAC CAA ATC AGT GGT TTG GTG GTT ATC TCG CTG GGA CTT GAT GGG GTG ACA CTC CAA TCC   90
TCT ATC AAC TAC CTT AAG GCA GGA AAA CTC GAC GAT GTA TTG ACC ATT AAA GGA GAA TGT GTC CAT CAA GGT CGT ACA ACC TGT GTA ATG  180
GAT GTG GAT ATC ACC AAT CAA GAA GGC AGA AAT GTC TGC AAA GCA ACC TTT ACC ATG TTT GTC ACA GGC CAG CGG TCA GAA GAC AGA CAG  270
GTA AGG ATA TAA AGA GTA ATT TAA TAG AGT GAG GTG AAT TTT GGA ATT TAT AAG ATT TGC CTT ACT CAT TTT TTG ATA TTG ATA TGA TTT  360
AAT TCT AAA ATA GAA AAT TTA GTA ATG TGG TAG AAA ATA CAA GAG TTG TGT TTT AAT TTC TAT GGT ATA ATT AAA AGC ATG AAG ATA AAA  450
                                                                                                        MET LYS ILE LYS
                                                                                                        met lys  -   -
GAA ATA AAG AAA GTT ACT TTA CAA CCG TTC ACG AAA TGG ACA GGT GGT AAA AGA CAA TTA TTG CCT GTT ATT AGA GAA TTA ATA CCT AAA  540
GLU ILE LYS LYS VAL THR LEU GLN PRO PHE THR LYS TRP THR GLY GLY LYS ARG GLN LEU LEU PRO VAL ILE ARG GLU LEU ILE PRO LYS
 -   -  lys asn arg ala  -   -   -  phe leu lys trp ala gly gly lys tyr pro leu leu asp asp ile lys arg his leu pro lys
ACC TAT AAC AGG TAT TTC GAA CCT TTT GTT GGA GGT GGA GCT TTA TTT TTT GAT TTG GCT CCT AAA GAT GCA GTT ATT AAT GAT TTT AAC  630
THR TYR ASN ARG TYR PHE GLU PRO PHE VAL GLY GLY GLY ALA LEU PHE PHE ASP LEU ALA PRO LYS ASP ALA VAL ILE ASN ASP PHE ASN
gly glu cys leu val  -  glu pro phe val gly ala gly ser val phe leu asn thr asp phe ser arg tyr ile leu ala asp ile asn
GCT GAA CTA ATA AAT TGC TAT CAA CAA ATT AAG GAC AAT CCT CAA GAA TTG ATT GAA ATT TTG AAA GTT CAT CAG GAA TAT AAT TCA AAA  720
ALA GLU LEU ILE ASN CYS TYR GLN GLN ILE LYS ASP ASN PRO GLN GLU LEU ILE GLU ILE LEU LYS VAL HIS GLN GLU TYR ASN SER LYS
ser asp leu ile ser leu tyr asn ile val lys met arg thr asp glu tyr val gln ala ala arg glu leu phe val pro glu thr asn
GAA TAT TAT TTA GAT TTA CGT TCT GCA GAT CGT GAT GAA AGA ATA GAT ATG ATG TCC GAA GTA CAA AGA GCT GCA CGT ATT CTA TAT ATG  810
GLU TYR TYR LEU ASP LEU ARG SER ALA ASP ARG ASP GLU ARG ILE ASP MET MET SER GLU VAL GLN ARG ALA ALA ARG ILE LEU TYR MET
cys ala glu val tyr tyr gln phe  -   -  arg glu glu phe asn lys ser gln asp pro phe arg arg ala val leu phe leu tyr leu
TTG AGA GTG AAC TTT AAT GGT CTA TAT CGT GTG AAT TCT AAG AAT CAA TTT AAT GTT CCA TAT GGA CGT TAT AAG AAT CCT AAA ATT GTT  900
LEU ARG VAL ASN PHE ASN GLY LEU TYR ARG VAL ASN SER LYS ASN GLN PHE ASN VAL PRO TYR GLY ARG TYR LYS ASN PRO LYS ILE VAL
asn arg tyr gly tyr asn gly leu cys arg tyr asn leu arg gly glu phe asn val pro phe gly arg tyr lys lys pro tyr phe pro
GAT GAG GAA TTG ATA TCT GCT ATT TCA GTT TAT ATA AAT AAC AAT CAA CTA GAA ATT AAA GTG GGA GAT TTT GAA AAG GCA ATT GTA GAT  990
ASP GLU GLU LEU ILE SER ALA ILE SER VAL TYR ILE ASN ASN ASN GLN LEU GLU ILE LYS VAL GLY ASP PHE GLU LYS ALA ILE VAL ASP
glu ala glu leu tyr his phe ala glu lys ala gln asn ala phe phe tyr cys glu ser tyr ala asp ser  -  met ala arg ala asp
GTT CGA ACA GGA GAT TTT GTG TAT TTT GAC CCT CCA TAT ATT CCA TTG TCT GAG ACG AGT GCA TTT ACG TCT TAT ACT CAT GAG GGA TTC 1080
VAL ARG THR GLY ASP PHE VAL TYR PHE ASP PRO PRO TYR ILE PRO LEU SER GLU THR SER ALA PHE THR SER TYR THR HIS GLU GLY PHE
asp ala ser val  -   -  val tyr cys asp pro pro tyr ala pro leu ser ala thr ala asn phe thr ala tyr his thr asn ser phe
TCT TTT GCA GAT CAA GTA AGA TTA AGA GAT GCC TTT AAG AGA TTG AGT GAT ACA GGA GCT TAT GTT ATG TTA TCA AAT TCT TCT AGT GCT 1170
SER PHE ALA ASP GLN VAL ARG LEU ARG ASP ALA PHE LYS ARG LEU SER ASP THR GLY ALA TYR VAL MET LEU SER ASN SER SER SER ALA
thr leu glu gln gln ala his leu ala glu ile ala glu gly leu val glu arg his ile pro val leu ile ser asn his asp thr met
TTA GTA GAG GAG TTG TAT AAG GAT TTT AAT ATA CAT TAT GTT GAA GCT ACC CGA ACT AAT GGA GCA AAA TCT TCA AGT CGA GGA AAA ATT 1260
LEU VAL GLU GLU LEU TYR LYS ASP PHE ASN ILE HIS TYR VAL GLU ALA THR ARG THR ASN GLY ALA LYS SER SER SER ARG GLY LYS ILE
leu thr arg glu trp tyr gln arg ala lys leu his val val lys val arg arg ser ile ser ser asn gly gly thr arg lys lys val
TCT GAA ATT ATA GTC ACA AAT TAT GAA AAA TAA CGA ATA TAA GTA TGG AGG TGT TCT TAT GAC AAA ACC ATA CTA CAA TAA AAA TAA GAT 1350
SER GLU ILE ILE VAL THR ASN TYR GLU LYS ###
asp glu leu leu ala leu tyr lys pro gly val val ser pro ala lys lys
GAT TCT TGT TCA TTC AGA TAC GTT CAA GTT CTT ATC AAA AAT GAA ACC AGA AAG TAT GGA TAT GAT TTT TGC TGA TCC ACC TTA TTT TTT 1440
AAG TAA TGG TGG AAT ATC TAA TTC TGG GGG ACA AGT AGT TTC TGT TGA TAA AGG AGA TTG GGA TAA AAT TTC TTC ATT CGA AGA AAA ACA 1530
TGA GTT TAA TCG TAA ATG GAT TCG CCT AGC AAA AGA AGT TCT GAA GCC TAA TGG GAC GGT ATG GAT TTC AGG TAG TTT GCA CAA CAT ATA 1620
CTC AGT TGG AAT GGC ATT AGA ACA AGA AGG TTT TAA AAT TCT GAA TAA TAT TAC TTG GCA GAA AAC AAA CCC TGC CCC CAA TTT ATC TTG 1710
TCG TTA TTT TAC CCA TTC TAC TGA AAC CAT TTT ATG GGC CAG AAA AAA TGA TAA AAA AGC TCG TCA TTA CTA CAA TTA TGA TTT AAT GAA 1800
AGA ATT GAA TGA TGG AAA ACA AAT GAA AGA TGT CTG GAC CGG TTC TTT AAC AAA GAA AGT TGA AAA ATG GGC TGG GAA ACA TCC AAC TCA 1890
AAA ACC AGA GTA TTT GTT AGA ACG TAT TAT TTT AGC CTC TAC TAA AGA GGG TGA CTA TAT TCT AGA CCC ATT TGT TGG TAG TGG CAC TAC 1980
GGG TGT TGT TGC GAA GCG GTT AGG TAG AAG ATT TAT AGG TAT CGA 2025
```

FIG. 5. Nucleotide sequence of the *dpnM* gene encoding *Dpn* II DNA methylase and adjacent regions. Only the DNA strand reading from left to right in the orientation shown in Fig. 3 and corresponding in sequence to the presumptive messenger RNA is depicted (5′ terminus at top left). Stop codons at 280, 391, and 1291, start codons at 439, 1283, and 1319, promoter consensus sequence at 343, and a possible regulatory sequence at 415 and 1257 are indicated in boldface. The predicted amino acid sequence for the *Dpn* II methylase is shown in uppercase lettering (amino-terminal methionine at nucleotide 439). The predicted amino acid sequence for the *E. coli dam* methylase (9) is shown in lowercase; dashes indicate codons absent from the *dam* gene. Underlined amino acid residues are identical in the *dpnM* and *dam* gene products.

peptide of 33,000 daltons, corresponds to the structural gene for the methylase. It appears to be the first gene in an operon transcribed from a promoter sequence to its left as drawn in Fig. 4. The second gene, truncated in the recombinant plasmid, could conceivably encode the *Dpn* II endonuclease, which is not expressed in the recombinant clone (6). Both genes show an identical sequence preceding the first possible start codon, which otherwise lacks a ribosomal binding site.

The DNA analysis revealed a promoter sequence, within 96 nucleotides from the start site of the methylase gene, from which transcription presumably begins. The amount of enzyme made in the recombinant plasmid clones gave 5–10 times the activity present in cells of strains containing only a chromosomal gene (Table 1; ref. 6). This could be accounted for by the gene dosage in cells with the multicopy plasmids; estimates of the copy number of similar recombinant plasmids (19) ranged from 15 to 30.

The *dam* gene of *E. coli*, a Gram-negative bacterium, and the *dpnM* gene of *S. pneumoniae*, a Gram-positive bacterium, appear to be homologous. The methylase proteins that they specify are almost the same size; they contain 278 and 284 amino acid residues, respectively, of which 30% are identical. Because the identical amino acids fall into four clusters spaced throughout the polypeptide chain, it is unlikely that the similarities arose in parallel by convergent evolution. The limited degree of homology, 30%, is consistent with the presumed divergence of Gram-positive and Gram-negative bacteria over 10^9 years ago (20), inasmuch as amino acid changes at a given site in a protein evolve at a frequency of the order of 10^{-9} per year (21). We are aware of no other comparison at the level of DNA sequence between homologous chromosomal genes in Gram-positive and Gram-negative bacterial species.

In *E. coli* the *dam* methylase does not act as part of a restriction system. It has been proposed, rather, that it serves a function in the repair of base mismatches in newly replicated DNA (22). Methylation of G-A-T-C sites in the parental strand and not in the nascent strand would direct repair of the misreplicated nascent strand. Evidence for such a role of DNA methylation in the repair of heteroduplex λ phage DNA has been reported (23). However, DNA methylation may play only a minor role in mismatch repair by the *mutHLS* system of *E. coli in vivo*, being implicated in approximately 10% of the repair events. This is shown by the 10-fold greater mutator effect of mutations in the mismatch repair genes *mutH* and *mutL*, as compared to the *dam* mutation (22). The greater effect of mutations in the *mut* genes cannot be attributed to leakiness of the *dam* mutations examined because at least one of them, *dam-3*, was shown to totally eliminate G-A-T-C methylation (4). In *S. pneumoniae* the heteroduplex DNA base mismatch repair system appears to be directed by single-strand breaks in the strand to be repaired, and it was suggested that this may be the fundamental mode of strand discrimination in the *E. coli* system as well (16). According to the proposed model, mismatch repair is for the most part directed toward the nascent strand by the presence of breaks between Okazaki fragments (produced either by the replication process or by removal of incorporated deoxyuridylate residues). Hemimethylation at G-A-T-C sites would play an accessory role by enabling additional breaks in the nascent strand, in which the G-A-T-C site is not methylated.

Results of the present work suggest that the *dam* gene of *E. coli* evolved from a methylase gene, such as *dpnM*, that was part of a restriction system. Conceivably, in the case of *E. coli*, the methylase retained its full function, but the endonuclease degenerated to give an enzyme that would produce a single-strand break at an unmethylated G-A-T-C site but could not make a double-strand break. The system therefore lost its restriction function but could now serve as an accessory for mismatch repair. Single-strand breaks have been observed in the DNA of *dam* mutants (24), and they could result from the action of such a degenerate restriction endonuclease. If these speculations are correct, the *E. coli* situation provides an interesting case study of the evolution of DNA methylation and restriction systems.

Although the *Dpn* I and *Dpn* II restriction systems of *S. pneumoniae* serve mainly as a defense against viral infection (2), and most bacterial DNA methylation may be related to this function, the present work suggests that cells can evolve other functions for DNA adenine methylation. Another evolutionary possibility might be to control gene expression. Methylation due to the *dam* function has been shown to affect transcription in both a positive manner, as in the phage Mu *mom* gene (25), and in a negative manner, as in the transposase gene of Tn*10* (26). It is conceivable that the DNA cytosine methylation thought to control transcription in eukaryotes (27) also depends on DNA cytosine methylases evolved from restriction system methylases in the prokaryotic ancestors of the eukaryotes.

This work was done under the auspices of the U.S. Department of Energy and was supported by Grants AI14885 and GMAI29721 from the National Institutes of Health.

1. Bernheimer, H. P. (1979) *J. Bacteriol.* **138**, 618–624.
2. Muckerman, C. C., Springhorn, S. S., Greenberg, B. & Lacks, S. A. (1982) *J. Bacteriol.* **152**, 183–190.
3. Lacks, S. & Greenberg, B. (1975) *J. Biol. Chem.* **250**, 4060–4066.
4. Lacks, S. & Greenberg, B. (1977) *J. Mol. Biol.* **114**, 153–168.
5. Vovis, G. F. & Lacks, S. (1977) *J. Mol. Biol.* **115**, 525–538.
6. Lacks, S. A. & Springhorn, S. S. (1984) *J. Bacteriol.* **157**, 934–936.
7. Lacks, S. A. & Springhorn, S. S. (1984) *J. Bacteriol.* **158**, 905–909.
8. Marinus, M. G. & Morris, N. R. (1973) *J. Bacteriol.* **114**, 1143–1150.
9. Brooks, J. E., Blumenthal, R. M. & Gingeras, T. R. (1983) *Nucleic Acids Res.* **11**, 837–851.
10. Stassi, D. L., Lopez, P., Espinosa, M. & Lacks, S. A. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 7028–7032.
11. Espinosa, M., Lopez, P., Perez-Urena, M. T. & Lacks, S. A. (1982) *Mol. Gen. Genet.* **188**, 195–201.
12. Currier, T. C. & Nester, E. W. (1976) *Anal. Biochem.* **76**, 431–441.
13. Lacks, S. (1966) *Genetics* **53**, 207–235.
14. Lacks, S. (1980) *Methods Enzymol.* **65**, 138–146.
15. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
16. Lacks, S. A., Dunn, J. J. & Greenberg, B. (1982) *Cell* **31**, 327–336.
17. Shine, J. & Dalgarno, L. (1975) *Nature (London)* **254**, 34–38.
18. Rosenberg, M. & Court, D. (1979) *Annu. Rev. Genet.* **13**, 319–353.
19. Lopez, P., Espinosa, M. & Lacks, S. A. (1984) *Mol. Gen. Genet.* **195**, 402–410.
20. Schwartz, R. M. & Dayhoff, M. O. (1978) *Science* **199**, 395–403.
21. Kimura, M. (1983) in *Evolution of Genes and Proteins*, eds. Nei, M. & Koehn, R. K. (Sinauer Associates, Sunderland, MA), pp. 208–233.
22. Glickman, B. W. & Radman, M. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 1063–1067.
23. Pukkila, P. J., Peterson, J., Herman, G., Modrich, P. & Meselson, M. (1983) *Genetics* **104**, 571–582.
24. Marinus, M. G. & Morris, N. P. (1974) *J. Mol. Biol.* **85**, 309–322.
25. Hattman, S. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 5518–5521.
26. Kleckner, N., Morisato, D., Roberts, D. & Bender, J. (1984) *Cold Spring Harbor Symp. Quant. Biol.* **49**, 235–244.
27. Razin, A. & Riggs, A. D. (1980) *Science* **210**, 604–610.