Cloning and characterization of the genes encoding the *Msp*I restriction modification system

P M.Lin*, C.H.Lee+ and R.J.Roberts

Cold Spring Harbor Laboratory, PO Box 100, Cold Sprıng Harbor, NY11724, USA

Received January 27, 1989, Revised and Accepted March 16, 1989 EMBL accession no X14191

ABSTRACT

The genes encoding the *Msp*I restriction modification system, which recognizes the sequence 5′ CCGG, have been cloned into pUC9. Selection was based on expression of the cloned methylase gene which renders plasmid DNA insensitive to *Msp*I cleavage *in vitro*. Initially, an insert of 15 kb was obtained which, upon subcloning, yielded a 3 kb *Eco*RI to *Hind*III insert, carrying the genes for both the methylase and the restriction enzyme. This insert has been sequenced. Based upon the sequence, together with appropriate subclones, it is shown that the two genes are transcribed divergently with the methylase gene encoding a polypeptide of 418 amino acids, while the restriction enzyme is composed of 262 amino acids. Comparison of the sequence of the *Msp*I methylase with other cytosine methylases shows a striking degree of similarity. Especially noteworthy is the high degree of similarity with the *Hha*I and *Eco*RII methylases.

INTRODUCTION

The *Msp*I restriction enzyme was originally isolated in this laboratory from an organism present as a contaminant in a culture of *Xanthomonas malvacearum*. A local clinical pathology laboratory characterized the organism as a *Moraxella* species, and that formed the basis for the name *Msp*I. However, it should be noted that subsequent tests of this organism suggest that the original identification was incorrect and it has since been variously characterized as an *Acinetobacter* species, or a *Flavobacterium* species. The correct taxonomic designation remains in doubt. For historical reason, it is still referred to as *Moraxella* species.

Interest in the enzyme *Msp*I was greatly stimulated by the unexpected finding that it differed from the enzyme *Hpa*II that had been discovered earlier, but which recognized the same sequence. It was found that the *Msp*I endonuclease was able to cleave DNA that was modified at the internal cytosine residue of the sequence CCGG (1). This is the position that is modified by the *Hpa*II methylase, and which provides protection against the action of the *Hpa*II endonuclease (2). It is now known that the *Msp*I methylase forms 5-methylcytosine at the outer cytosine residue (3) in contrast to the *Hpa*II methylase which modifies the inner cytosine residue (2). The pair of enzymes *Hpa*II and *Msp*I have found great utility in the analysis of eukaryotic DNA since comparative digests can indicate whether the cytosine residue in the CG dinucleotide is methylated or not.

Over the last few years, many genes have been cloned encoding methylases and restriction endonucleases. Of particular relevance to the present studies has been the finding that among methylases that specifically modify cytosine residues, there is a great deal of similarity at the amino acid level (4–6). However, so far, it has not been possible to relate this similarity directly to function. In contrast to the sequence conservation among cytosine methylases, only slight similarities have been detected among adenine methylases (4,7) and no similarities have been found among restriction enzymes (4).

Previous work to characterize the *Msp*I restriction modification system had led to the cloning of a large fragment of DNA that encoded the *Msp*I methylase gene, but this clone was reported not to express the endonuclease gene (3). Using a similar approach, we have been successful in obtaining clones that carry and express both the *Msp*I methylase and restriction endonuclease genes. Similar success in cloning these genes has also been reported recently (8).

MATERIALS AND METHODS

Bacterial strains, plasmids and phage

Moraxella species was originally isolated in this laboratory. *E. coli* strains RR1 (9), MM294 and DH1 (10), JM101, JM103 and JM107 (11), were grown in either LB or YT media (12).

Plasmids pUC8 and pUC9 (11) were used as vectors in initial cloning experiments, and plasmid DNA was isolated by the alkaline lysis method (13), and further purified by the cesium chloride−ethidium bromide procedure (12). The phages λ_{vir} and M13 mp18 and mp19 were obtained from Drs. A. Bukhari and J. Messing respectively.

Enzymes and chemicals

Restriction endonucleases, T4 ligase and T4 polynucleotide kinase were obtained from New England Biolabs, calf intestinal alkaline phosphatase was obtained from Boehringer−Mannheim. The Klenow fragment of *E. coli* DNA polymerase I was obtained from Bethesda Research Laboratories. Enzymes were used according to the manufacturer's specifications. Synthetic linkers were obtained from New England Biolabs. ^{35}S-α-dATP (>1000 Ci/mmole) and ^{32}P-α-dATP (>2000 Ci/mmole) were purchased from New England Nuclear. All other chemicals were of reagent grade quality.

Cloning of the MspI restriction system

Moraxella species cell DNA was extracted and purified by the method of Marmur (14). 50 μg of *Moraxella* species DNA was partially digested with 40 units of *Eco*RI at 37° for 5−30 minutes. Aliquots of the solution were taken at intervals of 5 minutes and the reaction stopped by adding SDS to 0.4%. The aliquots were mixed, heated at 65° for 10 minutes, and subjected to electrophoresis on a 1.4% agarose gel. Fragments in the size range 5−20 kb were eluted, purified and ligated to pUC9 DNA which had previously been cleaved with *Eco*RI, and dephosphorylated. The reaction mixture was used to transform *E. coli* RR1 competent cells according to standard procedures (12). Typically, 0.2 μg pUC9 DNA and 0.4−1.6 μg of *Moraxella* species DNA were ligated with 100 units T4 DNA ligase for 30 minutes at room temperature. Following incubation the reaction mixture was added to 200 μl of competent cells. After 30 minutes incubation, the transformation mixture was diluted to 1 ml with LB broth, and the cells plated on LB agar plates containing ampicillin (100 μg/ml). Following overnight incubation, the cells were scraped from the plate and combined for the preparation of plasmid DNA. Plasmid DNA was prepared from this culture using the alkaline lysis method followed by cesium chloride centrifugation (13). Purified DNA from the mixed population of plasmids was digested with the *Msp*I restriction endonuclease (20 units/μg), and the resulting digest used to retransform competent *E. coli* RR1. Individual transformants were recovered from LB agar plates containing ampicillin (100 μg/ml). The plasmid DNA from individual colonies was tested for its sensitivity to the *Msp*I restriction endonuclease, while cell extracts from each colony were tested for the presence of the *Msp*I restriction endonuclease as described below.

Assay for MspI modification activity in vivo

Individual colonies were used to inoculate small cultures (20 ml), and the cells grown to saturation at 37°. Plasmid or chromosomal DNA was prepared from the clone and tested for its sensitivity to the *Msp*I restriction endonuclease. DNA (1 μg) was digested in a reaction mixture containing 6 mM NaCl, 10 mM Tris.HCl (pH 7.4), 10 mM $MgCl_2$, 1 mM dithiothreitol and 5 units *Msp*I. The reaction mixtures were analyzed by gel electrophoresis and modification was detected by a lack of digestion. Parallel reactions containing 2 μg bacteriophage lambda DNA were used as a positive control.

Restriction endonuclease assay

Individual colonies were used to inoculate 10 ml LB broth containing ampicillin (100 μg/ml). The culture was grown to saturation, cells were harvested by centrifugation, and resuspended in 100 μl of a solution containing 10 mM Tris.HCl (pH 7.4), 1 mM Na_2EDTA. The suspension was sonicated (3×10 sec.) and the cell debris removed by centrifugation. The supernatant was used as a source of the *Msp*I restriction endonuclease. Typically 1,2,5 or 10 μl aliquots were used to digest 1 μg bacteriophage lambda DNA in a 50 μl reaction. The products were analyzed by agarose gel electrophoresis.

DNA sequence analysis

Small restriction fragments of the 3 kb fragment containing the *Msp*I genes were prepared and inserted into the vectors mp18 or mp19. These were used as templates in the chain termination procedure for DNA sequencing (15,16). Usually a synthetic primer (TCCCAGTCACGACGT) was used that is complementary to the M13 sequence immediately adjacent to the site of insertion. Thin sequencing gels (17) were used throughout and contained either 5, 6, or 8% polyacrylamide.

Computer Analysis

Computer analysis was performed on a DEC PDP11/44 and a SUN Microsystems 3/60. Primary data was stored and overlaps were established using the programs ASSEMBLER (18), M13 and SEQ (19). Analysis was carried out using the IGSUITE and PCGENE programs (IntelliGenetics) and additional programs described elsewhere (20–25). Homology searches used the PIR database version 17, the GenBank database version 57 and the EMBL data library version 16.

RESULTS

Cloning the MspI restriction system

The initial cloning of the *Msp*I restriction modification system into *E. coli* RR1 was achieved by preparing a partial *Eco*RI digest of *Moraxella* species DNA and cloning fragments in the size range 5 to 20 kb in the vector pUC9. A mixed population of recombinant plasmid DNAs were prepared, digested with the *Msp*I restriction endonuclease *in vitro*, and the mixture retransformed into *E. coli* RR1. This provides a strong selection for colonies that contain a recombinant plasmid expressing the *Msp*I methylase. A similar technique has been used previously to clone the *Msp*I (3) and other methylases (eg. 26). Several clones were isolated and tested individually for the presence of the *Msp*I methylase and/or endonuclease. Initially, small plasmid DNA preparations were made and the DNA tested for its resistance to the *Msp*I restriction endonuclease *in vitro*. Those clones that tested positive in this assay, and hence contained the methylase gene, were regrown and cell extracts from them prepared. These were tested in crude extracts for the presence of the *Msp*I restriction endonuclease by digesting bacteriophage lambda DNA. One clone designated pM748 contained both chromosomal and plasmid DNA that was fully resistant

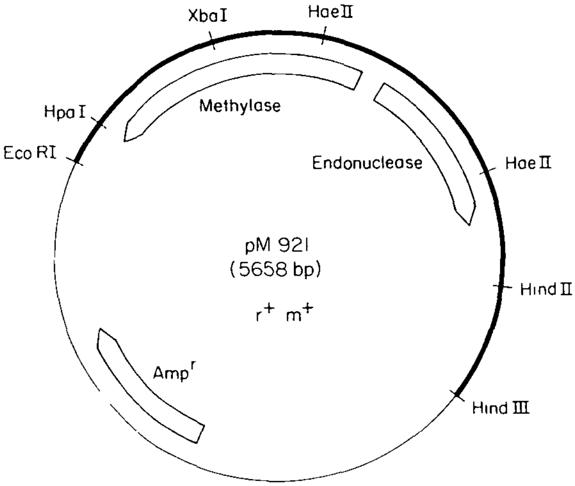

Figure 1. Schematic diagram of plasmid pM921, which contains a 3 kb insert of *Moraxella* species DNA in the vector pUC9 and which expresses both the restriction and modification genes. The thick lines indicate *Moraxella* species DNA and the thin lines indicate pUC9 vector sequences. The locations of the *Msp*I methylase and restriction endonuclease genes are indicated.

to the action of the *Msp*I endonuclease *in vitro* and crude extracts from this strain were shown to contain the *Msp*I restriction endonuclease. This clone contained five *Eco*RI fragments with a total insert size of 15 kb. It proved to be fairly unstable. Attempts to propagate this plasmid in *E. coli* strains DH1, JM103 or MM294 were unsuccessful, presumably because of the *mcr*B phenotype (36) of these strains. The clone could be transferred successfully to HB101, although the efficiency of transformation of this strain was 10 to 20-fold lower than RR1.

It had been reported previously (3) that a clone containing a single *Eco*RI fragment contained the methylase gene but not the restriction endonuclease gene. Consequently, in an attempt to prepare a smaller subclone containing both genes, plasmid pM748 was partially digested with *Eco*RI and *Hin*dIII. The resulting partial digest was ligated into the vector pUC9 that had been previously digested with *Eco*RI and *Hin*dIII. The recombinants were selected for those containing the methylase gene by *in vitro* digestion with *Msp*I and surviving colonies were screened individually for the presence of the *Msp*I restriction endonuclease. One clone called pM921 was found to contain both activities and was characterized further. This plasmid turned out to contain an insert of 3 kb that was bounded by a *Hin*dIII site on one end and an *Eco*RI site at the other end (Figure 1). There were no internal sites for either enzyme.

Location of the MspI methylase gene

To determine the location of the methylase gene the 3 kb insert from pM921 was excised with *Eco*RI and *Hin*dIII and then partially digested with *Sau*3AI. The resulting fragments were ligated into pUC9 that had been cleaved with *Eco*RI and *Bam*HI or *Hin*dIII and *Bam*HI. One clone, pM121, containing a 1.55 kb *Eco*RI to *Sau*3AI insert was found to express the methylase gene. Other clones harboring longer inserts also expressed methylase activity while shorter clones such as pM300 showed no activity. In a separate experiment, the sub-clone pM924, in which the internal *Hae*II fragment had been deleted, was found to

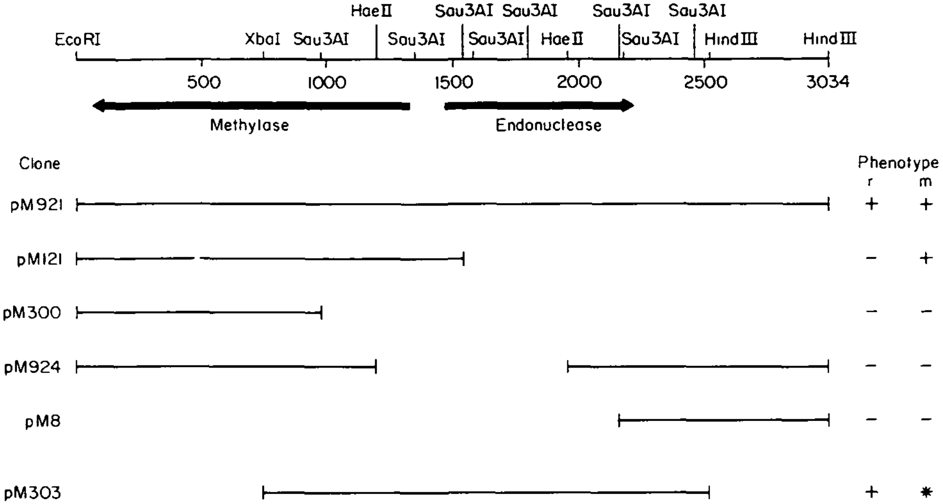

Figure 2. Deletion analysis of plasmid pM921. The thin horizontal lines indicate the DNA sequences remaining in the various deletion derivatives. The phenotypes were scored by *in vitro* analysis as described in Materials and Methods. The * indicates that the methylase gene was carried by a second plasmid present in the cell.

be stable but showed neither methylase nor restriction activities. Attempts to delete the *Eco*RI to *Xba*I fragment were unsuccessful as judged by the failure to obtain any viable clones. This result together with the properties of pM303 (see below) indicated that the methylase gene spanned the *Xba*I site (Figure 2).

Location of the restriction endonuclease gene

The 1.55 kb *Eco*RI to *Sau*3AI fragment containing the methylase gene was isolated, the cohesive ends were filled in using the Klenow fragment of DNA polymerase I, *Eco*RI linkers were added, and the construct was ligated into the *Eco*RI site of the plasmid PACYC184. Clones were isolated in which the methylase gene was cloned in either orientation and both expressed the methylase gene sufficiently well to render the plasmid DNAs fully resistant to *Msp*I *in vitro*. One of these clones was used as a permissive host

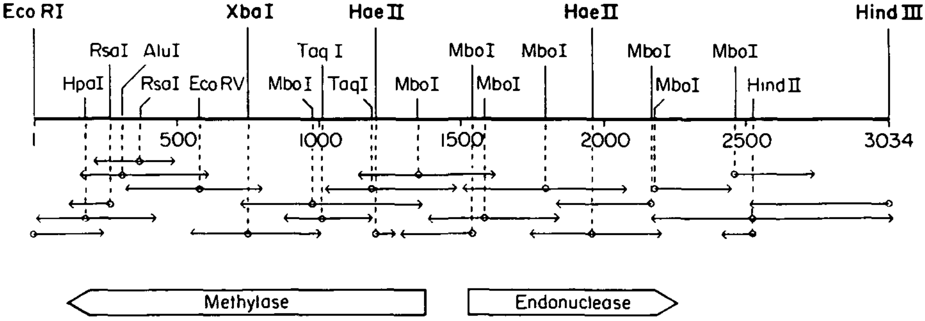

Figure 3. Sequencing strategy by which the complete sequence of the insert in pM921 was determined. The full extent of sequence obtained from individual sequencing reactions is indicated. Small circles indicate the beginning of each reading.

```
GAATTCCACTTCTTGAATATATTTAATTGATTTTATAGATGTTTTTCAATGAAAACATCA 60
TAGTCGGAAAAGTTCACTACTAACTGAATTTCGTATATGAGATTTTATGTTAAATAGCCC 120
CTCAATTAGGGGCTATATTTTTTttaAACGAGTTCTAATTCAAAGTTTTCTTGCGGGGAT 180
TGTTGGTTAACAGTTTTTAGTGCCAAACTAATCTGTTCTGCAATTTTTGTAACCACCGGA 240
ACTACGACAGAGTTACCCATTTGACGGTACATCTGAGTTCTTGATACAGGAATAACAAAA 300
TCTTTTGGAAAACCCATAATAGCTTTGCATTCATTCGTTGTTAAAAGACGGATACCTGTT 360
TCTCCATCCTTAACAAAAGTACCTGTTAATCGTTGAATTTTGTGATAGGTAGAAACTAAT 420
GTTTTAACTGCCCCAGTCGTATTTTTGTCAATTAAAGAAGGTTTACCATCATCTTTTTTA 480
AAGAGATAACTTTTTTGTAAATGCTCTGAAATGCTATATCCAGTTACATCGCTTTCTAAA 540
ACTTCACCGATATCTTTAGAAATCATTGGAGGTTTAGGAAACTCAAAGTGAATATTTTGA 600
TTTAGGAAAGCTACTAAGTAAAAACGCTTACGTTTTTGTGGGATACCAAAATGACTAGCG 660
TCTAGTACCGTATGATGAACTTTGTAGCCCATATCTTCTAGTGTTTCAATGATGACTTTT 720
AATGTATTTCCGTCATCATGATTAATGAGACCAGGAACATTTTCTAGAAATAAAACTGGG 780
GTTTTTTTTGTTTCAATGATACGAACAATTTCATGGAACATTGTTCCTTGAGTTGGATGT 840
TCAAAGCCTTCTCTTTTACCAATATGGCTAAATGGCTGACACGGAAATCCTGCACATAAA 900
ATGTCATGCTGTGGAATAGTTGTAGCCTCAACTTTTGTAATATCTCCAAAAGGGACTACA 960
CCAAAATTTGTATAATATGTAAATTTTGCAAAGGGATCTATTTCAGACGAAAAGACGCAT 1020
TTTCCGCCATTCACTTCGAATGATTGCCTTATTCCACCGATACCAGAAAATAAATCAATA 1080
AATTTGAAATCACTGCTATAAGCATCTTTTCTTTCTTCAATATGAGTAATTTCAGCTAAT 1140
TCTTCCGCATTTTTAGTAATTTGATTCTGATTGTATTTATCGAATATCTTCTTTTGGAGT 1200
GTTTTTTGAGCGCTCAGTTCCTCAAAGTTAATCTTATCTTTAAGTTTTTCTTGGAGAAAA 1260
GAGTAATAAGCAGGATGCATTTCTGTTTTACCTGATTCCCACTGTTGCCATGTTTTATCA 1320
CTAACTTCAATAATTTCAGATGCTTGCTTTTGAGTTAGATCCAATTTACTACGAATCAAT 1380
TTCAATATTTCAGGTTTcatTTGCGGTCAAAATCTATAATTTTCCTAATTATCCGTAATA 1440
AACATTATAAGTCAATCTTTCTTTATCCTACTTGTACGAAAAATATCTAAGGCATTGATA 1500
           M  R  T  E  L  L  S  K  L  Y  D  D  F  G  I  D  Q
GAGATAAAGAATGCGTACAGAACTATTAAGTAAGCTATATGATGATTTTGGGATAGATCA 1560
  L  P  H  T  Q  H  G  V  T  S  D  R  L  G  K  L  Y  E  K  Y
GTTACCTCATACCCAACATGGGGTAACTTCAGATCGACTTGGTAAGCTATATGAAAAGTA 1620
  I  L  D  I  F  K  D  I  E  S  L  K  K  Y  N  T  N  A  F  P
TATTTTGGATATTTTTAAAGATATTGAGTCTTTAAAGAAATACAACACTAATGCTTTTCC 1680
  Q  E  K  D  I  S  S  K  L  L  K  A  L  N  L  D  L  D  N  I
TCAAGAGAAAGATATATCTAGTAAGTTATTAAAAGCATTAAATCTTGATTTAGATAATAT 1740
  I  D  V  S  S  S  D  T  D  L  G  R  T  I  A  G  G  S  P  K
TATTGATGTGAGTAGTAGTGATACTGATTTAGGTCGTACCATTGCTGGCGGTAGTCCAAA 1800
  T  D  A  T  I  R  F  T  F  H  N  Q  S  S  R  L  V  P  L  N
AACTGATGCTACGATCAGGTTTACTTTTCATAATCAGTCATCAAGGCTTGTTCCTTTAAA 1860
  I  K  H  S  S  K  K  K  V  S  I  A  E  Y  D  V  E  T  I  C
TATTAAACATTCTAGTAAGAAAAAAGTATCTATTGCTGAATATGATGTGGAAACAATATG 1920
  T  G  V  G  I  S  D  G  E  L  K  E  L  I  R  K  H  Q  N  D
TACAGGTGTCGGTATTTCTGATGGTGAGTTAAAAGAGTTAATTCGAAAACATCAAAATGA 1980
  Q  S  A  K  L  F  T  P  V  Q  K  Q  R  L  T  E  L  L  E  P
CCAAAGCGCTAAGTTATTCACTCCTGTTCAAAAGCAACGCTTAACAGAATTATTGGAGCC 2040
  Y  R  E  R  F  I  R  W  C  V  T  L  R  A  E  K  S  E  G  N
ATACAGAGAGCGATTTATTCGTTGGTGCGTTACATTACGTGCTGAAAAAAGCGAAGGAAA 2100
  I  L  H  P  D  L  L  I  R  F  Q  V  I  D  R  E  Y  V  D  V
TATTTTACATCCAGACTTATTAATTCGATTTCAAGTGATTGACCGTGAGTATGTGGATGT 2160
  T  I  K  N  I  D  D  Y  V  S  D  R  I  A  E  G  S  K  A  R
AACAATCAAAAATATTGATGATTATGTAAGTGATCGGATTGCAGAAGGATCAAAAGCTAG 2220
  K  P  G  F  G  T  G  L  N  W  T  Y  A  S  G  S  K  A  K  K
AAAACCTGGATTTGGTACTGGTTTGAATTGGACTTATGCAAGTGGTAGTAAAGCGAAAAA 2280
  M  Q  F  K  G
AATGCAGTTCAAAGGCTAAAATAATGGAAAATAATAATATTTACTTGACTGAGATGGAAG 2340
TTTATCAATTAGCAGAAAATGTAGTAAACAGTATCTGTAATGATTTAAATGAAACAATAT 2400
ATAAAAAATTAGGTGGAAAACTGTCCATAGTTTGGAATACAGATGAAAGATTTAATGCAT 2460
CTGCTCAAATATTAAATAAAGCATCTGATCCACCAAATCACAGGATTACCTTATATTATT 2520
TTTTAGTTAAAGAATTATATAGAGATACAGTCAACTATCATGAGTTTGCAGAGAAGATTC 2580
ACTATCAGCCTAGCATTTTAGCTTTTCTAAATTCTATTTCAGAAATGCCTATGCTTCCTG 2640
AAATATTTATAAAAAAAGATAGTATAAACAATATGTTTATAGCTTCTTTGACTTTTATTC 2700
TCTTTCATGAGTTAGGGCATCTTATGCAGCAACATGGAAGAATTAGAGCTAGTTTGTCTG 2760
GGCAATCTATTGATGATTCAATAATTAATGAATGTAATGCTATTGATTCAATACCACTAA 2820
CGGGGAAGCAAGCTGCTATATCCCATACTACAGAATTGTCCGCAGACTCTAGTGCTATCA 2880
CTAGATGCATTTTTGAAATAATGAGACAGTTTTCATCTAAAGAGCTTATAGGAAATGATG 2940
ATAAAGGTACTTTATTTTTAGCAACAGCTCAGTTATTCGTTATTGGCGTATCGAACCTAT 3000
TTTATAGATTTAGAGGTATCAATTCAGAAAGCTT                           3034
```

Figure 4. The sequence of the *Msp*I restriction and modification genes. The sequence begins at the *Eco*RI site and includes the coding strand of the restriction endonuclease gene. The translation for that gene is shown. The methylase is encoded by the complementary strand. The ınitiator codon of the methylase is indicated in lowercase letters (tac) at nucleotide 1400 and the terminator codon is indicated similarly (att) at nucleotide 146.

in which to isolate subclones from the original 3 kb insert that might contain the restriction endonuclease gene. One subclone, pM303, containing the 1.8 kb *Xba*I to *Hin*dII fragment cloned into pUC9 did indeed show restriction activity when assayed *in vitro*. Attempts

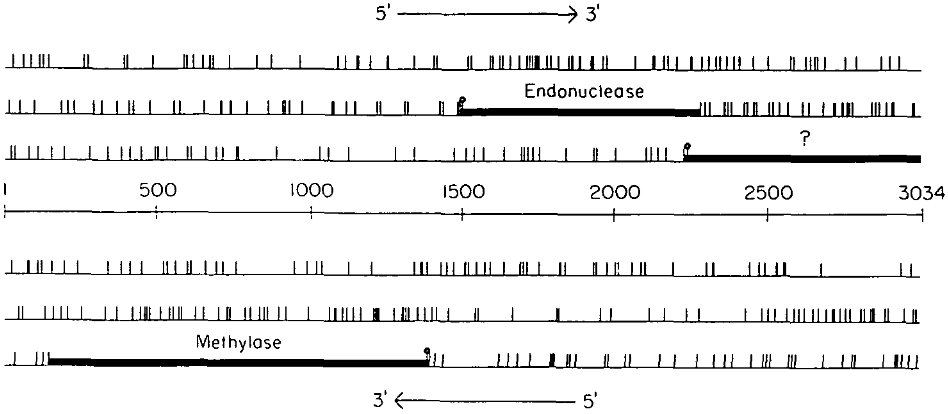

Figure 5. Reading frame analysis of the region encoding the *Msp*I restriction and modification genes. The three reading frames on both strands are presented. Terminators are indicated by small vertical lines. The initiating AUGs are indicated by vertical lines with a small circle on top. The positions of the restriction endonuclease and methylase genes are shown. The additional unidentified open reading frame, which extends uninterrupted into the adjacent vector sequences is marked with a ?

to subclone this fragment directly into *E. coli* RR1 in the absence of the methylase gene were unsuccessful. The results of the subcloning experiments are summarized in Figure 2.

Nucleotide sequence of the MspI R-M system

The nucleotide sequence of the 3,034 base pair insert in pM921 was determined by preparing subclones of the region into M13 mp18 or M13 mp19 and sequencing by the chain termination procedure. A detailed restriction map of the insert and the sequencing strategy is shown in Figure 3 and the sequence is shown in Figure 4. Three large open reading frames are present in the sequence, although one of them contains no terminator in the insert and runs into the adjacent vector sequences through the *Eco*RI site (Figure 5). Based upon the previous mapping data (Figure 2), the methylase gene and the restriction endonuclease genes can be positioned unequivocally. The methylase gene is encoded by an open reading frame that extends from base 1400 to base 147, while the restriction endonuclease gene must be encoded by the open reading frame that begins at base 1511 and extends through base 2296. Assuming that it is the first AUG in each of these reading frames at which translation begins, then the methylase will contain 418 amino acids with a predicted molecular weight of 47,664. The restriction endonuclease gene would be a 262 residue polypeptide with a molecular weight of 29,833. The two AUG codons that define the start of the genes are separated by 110 base pairs and transcription is divergent. In the case of the methylase gene, N-terminal sequence analysis of the purified methylase protein confirms that translation begins at this first AUG (B. Mollet and R.J. Roberts, unpublished).

Comparison with other restriction and modification enzymes

Sequences have been reported for fifteen restriction endonuclease genes and twenty seven methylase genes (37 and references therein). The FASTA program (24) was used to compare the *Msp*I restriction endonuclease gene with other restriction endonuclease sequences as well as the complete contents of the PIR, GenBank and EMBL databases. No significant homologies could be detected either at the DNA or protein level. Similarly the *Msp*I

Table 1. Similarity between the *Msp*I methylase and other cytosine methylases

Methylase	Recognition Sequence	Score	Reference
HhaI	G$\overset{*}{\mathrm{C}}$GC	129.64	27
EcoRII	C$\overset{*}{\mathrm{C}}$WGG	130.97	6
BspRI	GGCC	198.22	28
NgoPII	GGCC	206.86	29
SinI	GGW$\overset{*}{\mathrm{C}}$C	208.83	30
BsuRI	GG$\overset{*}{\mathrm{C}}$C	213.16	31
Rho11	GGCC, GAGCTC	241.49	32
SPR	GG$\overset{*}{\mathrm{C}}$C, $\overset{*}{\mathrm{C}}\overset{*}{\mathrm{C}}GG, C\overset{*}{\mathrm{C}}$WGG	280.98	33,34
Phi3	GG$\overset{*}{\mathrm{C}}$C, G$\overset{*}{\mathrm{C}}$NGC	282.35	35
DdeI	$\overset{*}{\mathrm{C}}$TAG	368.98	5

When known, the sites of methylation are indicated (*)

methylase gene sequence was compared with the other methylase sequences and the complete database contents. Significant and extensive homologies were detected with other methylases known to catalyse the formation of 5-methylcytosine. To quantitate the extent of homology with the ten cytosine methylases for which sequence information has been reported, the method of Feng and Doolittle was used (25) and the results are shown in Table 1. The greatest degree of similarity is between the *Msp*I methylase and the *Hha*I and *Eco*RII methylases. A dot matrix visualisation of these similarities is presented in Figure 6 and an alignment of the three sequences produced by manual adjustment of a Needleman–Wunsch alignment is shown in Figure 7. In this alignment 120 residues are identical between *Hha*I (327 residues) and *Msp*I (418 residues), while 136 residues are identical between *Eco*RII (477 residues) and *Msp*I.

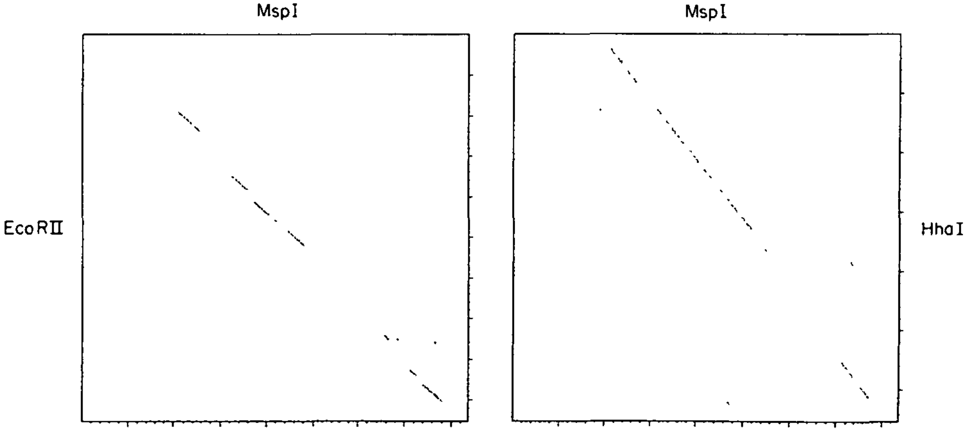

Figure 6. Dot matrix comparisons between the *Msp*I and the *Eco*RII and *Hha*I methylase amino acid sequences. The DIAGON program (22) was used with a span length of 11.

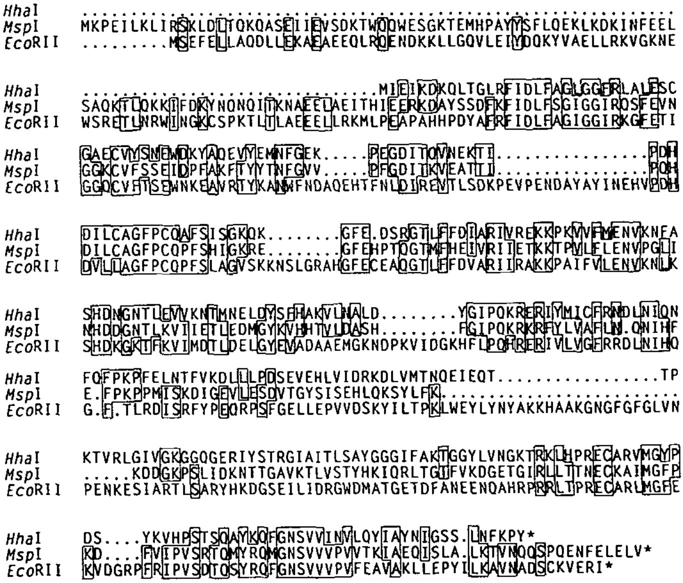
HhaI ....
MspI MKPEILKLIRSKLDLTQKQASEIIEVSDKTWQQWESGKTEMHPAYYSFLQEKLKDKINFEEL
EcoRII .........MSEFELLAQDLLEKAEAEEQLRQENDKKLLGQVLEIYDQKYVAELLRKVGKNE
HhaI ...............................MIEIKDKQLTGLRFIDLFAGLGGFRLALESC
MspI SAQKTLQKKIFDKYNQNQITKNAEELAEITHIEERKDAYSSDFKFIDLFSGIGGIRQSFEVN
EcoRII WSRETLNRWINGKCSPKTLTLAEEELLRKMLPEAPAHHPDYAFRFIDLFAGIGGIRKGFETI
HhaI GAECVYSNEWDKYAQEVYEMNFGEK.....PEGDITQVNEKTI................PDH
MspI GGKCVFSSEIDPFAKFTYYTNFGVV.....PFGDITKVEATTI................PQH
EcoRII GGQCVFTSEWNKEAVRTYKANWFNDAQEHTFNLDIREVTLSDKPEVPENDAYAYINEHVPDH
HhaI DILCAGFPCQAFSISGKQK........GFE.DSRGTLFFDIARIVREKKPKVVFMENVKNFA
MspI DILCAGFPCQPFSHIGKRE........GFEHPTQGTMFHEIVRIIETKKTPVLFLENVPGLI
EcoRII DVLLAGFPCQPFSLAGVSKKNSLGRAHGFECEAQGTLFFDVARIIRAKKPAIFVLENVKNLK
HhaI SHDNGNTLEVVKNTMNELDYSFHAKVLNALD........YGIPQKRERIYMICFRNDLNIQN
MspI NHDDGNTLKVIIETLEDMGYKVHHTVLDASH........FGIPQKRKRFYLVAFLN.QNIHF
EcoRII SHDKGKTFKVIMDTLDELGYEVADAAEMGKNDPKVIDGKHFLPQHRERIVLVGFRRDLNIHQ
HhaI FQFPKPFELNTFVKDLLLPDSEVEHLVIDRKDLVMTNQEIEQT.................TP
MspI E.FPKPPMISKDIGEVLESDVTGYSISEHLQKSYLFK.........................
EcoRII G.F.TLRDISRFYPEQRPSFGELLEPVVDSKYILTPKLWEYLYNYAKKHAAKGNGFGFGLVN
HhaI KTVRLGIVGKGGQGERIYSTRGIAITLSAYGGGIFAKTGGYLVNGKTRKLHPRECARVMGYP
MspI .....KDDGKPSLIDKNTTGAVKTLVSTYHKIQRLTGTFVKDGETGIRLLTTNECKAIMGFP
EcoRII PENKESIARTLSARYHKDGSEILIDRGWDMATGETDFANEENQAHRPRRLTPRECARLMGFE
HhaI DS....YKVHPSTSQAYKQFGNSVVINVLQYIAYNIGSS.LNFKPY*
MspI KD....FVIPVSRTQMYRQMGNSVVVPVVTKIAEQISLA.LKTVNQQSPQENFELELV*
EcoRII KVDGRPFRIPVSDTQSYRQFGNSVVVPVFEAVAKLLEPYILKAVNADSCKVERI*

Figure 7. Alignment between the *Msp*I and the *Eco*RII and *Hha*I methylase protein sequences. Initial pairwise alignments were obtained using the program GENALIGN, which uses the Needleman–Wunsch procedure. Further refinement was carried out manually.

DISCUSSION

Like other restriction and modification systems whose cloning has been reported to date the genes for the *Msp*I system lie in close proximity to one another. They can be cloned in a single step procedure in which both the methylase gene and the restriction endonuclease are introduced into an unprotected cell at the same time. While this has been observed in many systems it requires that protection of the host DNA, via modification, must precede or outcompete restriction by the endonuclease. One possibility is that expression of the restriction endonuclease gene is temporally regulated upon its initial introduction into a cell so that modification of the host cell DNA can be completed before expression of the restriction endonuclease. Alternatively the level of expression of the endonuclease may be quite low or the protein may be held in an inactive state such that it is ineffective against the intracellular DNA. As we and others (8) have noted, *E. coli* containing the cloned *Msp*I system does not show a restriction–modification phenotype *in vivo*. This suggests a low level or lack of activity *in vivo* even though the endonuclease can be detected easily *in vitro*. Production of the endonuclease in *E. coli* RR1 containing the recombinant plasmid pM921 is comparable to that in the original *Moraxella* species. While the level of expression may vary from strain to strain of *E. coli*, it seems unlikely that this can explain the previous failure to detect expression of the restriction endonuclease when the system was initially cloned (3). In that case the *E. coli* host strain used was HB101, a rec$^-$ version of the RR1 host that was used in the present study.

The predicted sizes of the restriction endonuclease and methylase polypeptides are very similar to those reported for the corresponding genes in other systems. The methylase is much larger than the restriction endonuclease as observed in all other systems studied.

This is usually attributed to the fact that most known restriction enzymes act as dimers, while most methylases act as monomers (38). There is no detectable homology between the restriction enzyme gene and the methylase gene either at the protein level or at the DNA level suggesting that they evolved independently of one another. This also applies to all other cloned restriction – modification systems and highlights an unresolved problem. How were the two components of a restriction – modification system, which must carry the same specificity, ever able to find one another?

In addition to the genes for the restriction endonuclease and the methylase the 3 kb fragment whose sequence is presented in Figure 4 contains an additional opening reading frame downstream of the restriction endonuclease gene. This reading frame continues into the surrounding vector sequences precluding an assessment of its full length. Comparison of the available sequence with all other sequences present in the PIR, GenBank and EMBL databases failed to reveal any significant homology. There is no reason at the present time to believe that the product of this open reading frame is connected to the expression of the *Msp*I restriction – modification system.

Within the 110 nucleotides separating the initiation codons of the methylase and restriction endonuclease genes there must lie the regulatory signals that direct transcription initiation. At least one of these signals appear to function adequately in *E. coli* since expression of the methylase gene has been observed when it was cloned in either orientation in pACYC184. While there are no clear sequences that would match typical *E. coli* promoter sequences, the entire intergenic region is quite A+T rich (73%). Sequences with some similarity to both −10 and −35 regions can be found, although their significance is unclear. The base composition of the intergenic region can be compared with the coding regions which are 66% A+T. The most striking feature of the coding region is the extreme codon bias. Codons with A or T in the third position account for 78% of the methylase codons and 79% of the restriction enzyme codons. There is no obvious sequence immediately beyond the coding sequences that shows any similarity to the usual *E. coli* transcriptional stop points. The only likely hairpins that can form lie 300 nucleotides downstream (bases 2615 to 2645 and 2663 to 2781 in Figure 4) of the restriction enzyme gene, but are not followed by runs of T-residues. A single *Msp*I site is found in the sequence at base 235. This lies within the coding region of the methylase close to the C-terminus and so it is unlikely that is has any regulatory significance.

One of the most interesting features of the present sequence is the finding that the *Msp*I methylase gene shows considerable homology with the *Hha*I and *Eco*RII methylase genes. From the alignments shown in Figure 7 it can be seen that there are extensive regions of similarity between these proteins. The conserved regions that have been noted (37) among other cytosine methylases are all present. This includes the so-called PC motif that is postulated to be involved in the catalytic mechanism (39).

ACKNOWLEDGEMENTS

The authors would like to express their thanks to Drs. G. Freyer, A. Bhagwat and A. Kiss for many helpful discussions during the course of this work. Thanks are also due to C. Keller and J. Posfai for help with the computational aspects of this work. The excellent artwork of J. Duffy and fine photography of D. Greene are greatly appreciated. This work was supported by a grant from the National Science Foundation (DMB-8614032).

Present addresses: *Biology Department, Northwestern University, Xian, China and $^+$Department of Pathology, Indiana University School of Medicine, Indianapolis, IN 46223, USA

REFERENCES

1. Waalwijk, C. and Flavell, R.A. (1978) Nucl. Acids Res 5: 3231–3236.
2. Mann, M.B. and Smith, H.O. (1977) Nucl. Acids Res. 4: 4211–4221.
3. Walder, R Y , Langtimm, C.J., Chatterjee, R. and Walder, J A. (1983) J. Biol. Chem. 258: 1235–1241.
4. Chandrasegaran, S. and Smith, H.O. (1988) in Structure and Expression Volume 1: From Proteins to Ribosomes. eds. R.H. Sarma and M.H. Sarma. (Adenine Press) pp 149–156.
5. Sznyter, L.A., Slatko, B., Moran, L., O'Donnell, K.H. and Brooks, J.E. (1987) Nucl. Acids Res. 15: 8249–8266.
6. Som, S., Bhagwat, A.S. and Friedman, S. (1987) Nucl. Acids Res. 15: 313–332.
7. Lauster, R., Kriebardis, A. and Guschlbauer, W (1987) FEBS Letters 220: 167–176.
8. Nwankwo, D.O. and Wilson, G.G. (1988) Gene 64: 1–8.
9. Bolivar, F., Rodriguez, R.L., Greene, P.J., Betlach, M.C., Heynecker, H.L , Boyer, H W., Crosa, J.H. and Falskow, S. (1977) Gene 2: 95–113.
10. Bachmann, B.J. in Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology, eds. Neidhardt, F.C. et al. (Amer. Soc. Microbiol. Washington, DC) pp 1190–1219.
11. Yanisch-Perron, C., Vieira, J. and Messing, J. (1985) Gene 33: 103–119
12. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.)
13. Birnboim, H.C. and Doly, J. (1979) Nucl. Acids Res. 7: 1513–1523.
14. Marmur, J. (1961) J. Mol. Biol 3: 208–218.
15. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad Sci. USA 74: 5463–5467.
16. Messing, J. (1983) Methods in Enzymology 101: 20–78.
17. Sanger, F. and Coulson, A.R. (1978) FEBS Letters 87: 107–110
18. Gingeras, T.R , Milazzo, J.P., Sciaky, D. and Roberts, R.J. (1979) Nucl. Acids Res. 7. 529–545.
19. Blumenthal, R.M., Rice, P.J. and Roberts, R.J. (1982) Nucl. Acids Res. 10: 91–101.
20. Staden, R. (1977) Nucl. Acids Res. 4: 4037–4051.
21. Staden, R. (1978) Nucl. Acids Res. 5: 1013–1015.
22. Staden, R. (1982) Nucl. Acids Res. 10: 2951–2961.
23. Keller, C., Corcoran, M. and Roberts, R.J. (1984) Nucl Acids Res. 12: 379–386.
24. Lipman, D.J. and Pearson, W.R. (1988) Proc. Natl. Acad. Sci. USA 85: 2444–2448
25. Feng, D-F and Doolittle, R.F. (1987) J. Mol. Evol. 25: 351–360.
26. Szomolanyi, E., Kiss, A. and Venetianer, P. (1980) Gene 10 219–225.
27. Caserta, M., Zacharias, W., Nwankwo, D., Wilson, G.G. and Wells, R.D. (1987) J. Biol. Chem. 262: 4770–4777.
28. Posfai, G., Kiss, A., Erdei, S., Posfai, J. and Venetianer, P. (1983) J. Mol. Biol. 170: 597–610.
29. Sullivan, K.M. and Saunders, J.R. (1988) Nucl. Acids Res. 16: 4369–4387.
30. Karreman, C. and de Waard A. (1988) J. Bacteriol. 170: 2527–2532.
31. Kiss, A., Posfai, G., Keller, C.C., Venetianer, P. and Roberts, R.J. (1985) Nucl. Acids Res. 13: 6403–6421.
32. Behrens, B., Noyer-Weidner, M., Pawlek, B., Lauster, R., Balganesh, T.S and Trautner, T.A. (1987) EMBO J 6: 1137–1142.
33. Buhk, H.J., Behrens, B., Tailor, R., Wilke, K., Prada, J.J , Gunthert, U., Noyer-Weidner, M., Jentsch, S. and Trautner, T.A. (1984) Gene 29: 51–61.
34. Posfai, G., Baldauf, F., Erdei, S., Posfai, J., Venetianer, P. and Kiss, A. (1984) Nucl. Acids Res. 12: 9039–9049.
35. Tran-Betcke, A., Behrens, B., Noyer-Weidner, M. and Trautner, T.A. (1986) Gene 42: 89–96
36. Raleigh, E.A. and Wilson, G.G. (1986) Proc. Natl. Acad. Sci. USA 83: 9070–9074.
37. Posfai, J., Bhagwat, A.S., Posfai, G. and Roberts, R.J. submitted for publication.
38. Modrich, P. and Roberts, R.J. (1982) in 'Nucleases', eds. S.A. Linn and R.J. Roberts, Cold Spring Harbor Laboratory Press, pp311–382.
39. Santi, D.V., Garrett, C.E. and Barr, P.J. (1983) Cell 33: 9–10.

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.