Characterization of the cloned *Bam*HI restriction modification system: its nucleotide sequence, properties of the methylase, and expression in heterologous hosts

Joan E. Brooks*, Peter D. Nathan, David Landry, Laura A. Sznyter⁺, Phyllis Waite-Rees[§], Catherine L. Ives, Laurie S. Moran, Barton E. Slatko and Jack S. Benner
New England Biolabs, Beverly, MA 01915, USA

Received October 19, 1990; Revised and Accepted January 14, 1991

EMBL accession no. X55285

ABSTRACT

The *Bam*HI restriction modification system was previously cloned into *E. coli* and maintained with an extra copy of the methylase gene on a high copy vector (Brooks et al., (1989) *Nucl. Acids Res.* 17, 979–997). The nucleotide sequence of a 3014 bp region containing the endonuclease (*R*) and methylase (*M*) genes has now been determined. The sequence predicts a methylase protein of 423 amino acids, M_r 49,527, and an endonuclease protein of 213 amino acids, M_r 24,570. Between the two genes is a small open reading frame capable of encoding a 102 amino acid protein, M_r 13,351. The M.*Bam*HI enzyme has been purified from a high expression clone, its amino terminal sequence determined, and the nature of its substrate modification studied. The *Bam*HI methylase modifies the internal C within its recognition sequence at the N4 position. Comparisons of the deduced amino acid sequence of M.*Bam*HI have been made with those available for other DNA methylases: among them, several contain five distinct regions, 12 to 22 amino acids in length, of pronounced sequence similarity. Finally, stability and expression of the *Bam*HI system in both *E. coli* and *B. subtilis* have been studied. The results suggest *R* and *M* expression are carefully regulated in a 'natural' host like *B. subtilis*.

INTRODUCTION

As restriction modification systems are cloned and characterized, more features of the systems are emerging. Recently a new type of DNA methylation, N4-methylcytosine (m⁴C), was discovered (1). Now three different types of DNA methylases have been identified and cloned from restriction systems: those producing N6-methyladenine (m⁶A), 5-methylcytosine (m⁵C) or N4-methylcytosine (m⁴C) (2).

DNA sequence determination of the cloned systems has also generated more interest in methylases since, in contrast to the

endonucleases, they can be divided into families on the basis of protein sequence similarities. There are now two such families, corresponding to the m⁶A and m⁵C methylases, respectively (3–7). The third group of methylases, generating m⁴C, closely resemble m⁶A methylases at the protein sequence level and appear to be part of that family (8, 9).

*Bam*HI was one of the first restriction systems to be discovered (10). Its endonuclease (*R.Bam*HI) and methylase (*M.Bam*HI) have been purified and characterized (11, 12). Its cloning was unusual in that it required two steps: a vector carrying the *Bam*HI methylase gene (*bam*HIM) and expressing *M.Bam*HI at high levels in the host had to be established before the endonuclease gene (*bam*HIR) could be successfully introduced (13). Here we report further characterization of the system, including nucleotide sequence of the genes, and characterization of the cloned methylase. In addition, from the nucleotide sequence we have found a third component of the *Bam*HI system. Its derived protein sequence shows similarities to DNA binding proteins; recent work has shown it to exert a regulatory effect on both *R.Bam*HI and *M.Bam*HI activity (14). We conclude by describing expression of the system in both *E. coli* and *B. subtilis* and by speculating on a cause of the differences in expression within the two hosts.

MATERIALS AND METHODS

Bacterial Strains and Media

E. coli strains K802 and RR1 have been described elsewhere in detail (15). *B. subtilis* strain 168trpC2 was from Dr. Ken Bott.

E. coli were grown in Luria broth (13). Media were supplemented with the antibiotics ampicillin (Ap), 100 µg/ml, kanamycin (Kn, 50 µg/ml); chloramphenicol (Cm, 37.5 µg/ml) as required. *B. subtilis* cultures were grown in nutrient broth (Difco) supplemented with 5 µg/ml of Cm as required.

Plasmids and Phage

The plasmids pUC19 and pBR328 (16, 17) were used in several subcloning experiments. The *E. coli* / *B. subtilis* shuttle vector

* To whom correspondence should be addressed

Present addresses: ⁺ Georgetown Medical School, Washington, DC 20007 and [§] Department of Chemistry, Florida State University, Tallahassee, FL 32306 USA

pHP13 (18, 19) was from Dr. Sierd Bron. Plasmids pBamM1.8, pBamM2.2 and pBamRM5.0 were described in (13). Plasmids carrying *Bam*HI genes were named according to a previously described system of nomenclature (20).

Plasmid DNA was prepared by the alkaline lysis method (21) or by the cleared lysate method (22) followed by CsCl/ethidium bromide ultracentrifugation (21).

Phage stocks of λ_{vir} (New England Biolabs collection) were prepared from plate lysates (21). All phage dilutions were done in SM solution (21).

Cloning Enzymes and Techniques

All restriction endonucleases were produced at New England Biolabs and used as recommended. *Hind*III-cleaved pBR328, *Bam*HI-cleaved pUC19, and *Hinc*II-cleaved pUC19 vectors were treated with calf intestine alkaline phosphatase (Boehringer Mannheim) as previously described (21).

Individual DNA fragments from restriction digests of plasmids were separated by agarose gel electrophoresis using Molecular Biology Grade Agarose (IBI), excised and then purified using a Unidirectional Electroelutor (IBI) as previously described (13). All ligations used 1000 units of T4 DNA Ligase (New England Biolabs) in a 100 μ l reaction volume under recommended conditions.

E. coli transformations were done using the CaCl_2 -heat shock method (23). *B. subtilis* transformations were done by the procedure of Bott and Wilson (24).

*Bam*HI Plasmid Constructs

The plasmids pBamM Δ 1.9 and pBamM2.1 were deletion derivatives of pBamM2.2. To construct pBamM Δ 1.9, 2 μ g of pBamM2.2 was digested to completion with a twentyfold excess of *Xba*I under recommended conditions. After a 2 hr incubation the reaction was diluted to 1 ml, ligated overnight and used to transform *E. coli* K802 cells. To construct pBamM2.1, 5 μ g of pBamM2.2 was digested with an excess of *Nsi*I, extracted with phenol, then chloroform, and precipitated in ethanol. After resuspension in restriction buffer, the plasmid was digested with an excess of *Pst*I, diluted to 1 ml in ligation buffer, religated and transformed.

The plasmids pBam1.0, pBam1.5 and pBam0.7 were all constructed using the gel purified 5.0kb *Hind*III fragment from pBamRM5.0 (Figure 1). To construct pBam1.0 and pBam1.5, 5 μ g of the *Hind*III fragment was digested with a twentyfold excess of *Bal*I and then ligated to pUC19 that had been linearized at the *Hinc*II site and treated with alkaline phosphatase. Plasmids were isolated from Ap-resistant transformants and screened for appropriate-size inserts. The plasmid pBam0.7 was made by sequentially cutting 5 μ g of the gel purified fragment with *Bal*I and *Bgl*II and ligating it to 0.1 μ g pUC19 that had been digested with *Hinc*II and *Bam*HI.

The gel purified 5.0 kb *Hind*III fragment from pBamRM5.0 was also used in making constructs for the expression experiments. To create pBamRM5.0-328, 1 μ g of the gel purified fragment was ligated to 0.1 μ g pBR328 cleaved with *Hind*III and dephosphorylated. The ligation mix was used to transform *E. coli* K802 cells containing pBamM2.2 (expressing *M. Bam*HI), and the transformants selected on Cm plates. To produce pBamRM2.7-HP13, 2 μ g of the *Hind*III fragment was digested with *Bgl*II and then ligated to pHP13 plasmid digested with *Hind*III and *Bam*HI. The ligation mix was similarly used to transform K802(pBamM2.2) and the Cm resistant transformants were selected.

Analysis of *Bam*HI Endonuclease and Methylase Activity

In vitro endonuclease activity was determined as described previously (13). Phage restriction was tested using λ_{vir} to measure *in vivo* endonuclease activity, as in (20).

During methylase purification, column fractions were assayed using ligated *Bam*HI linkers (d[pCGGATCCG]; New England Biolabs) as the substrate DNA. Reactions containing 100 mM NaCl, 50 mM Tris, pH7.5, 10 mM EDTA, 5 mM β -mercaptoethanol, 80 μ M [^3H] S-adenosyl-L-methionine (79.7 Ci/mmol; New England Nuclear), and 0.5 μ g ligated linkers, in 25 μ l volume, were incubated for 30 min at 37°C. The reactions were stopped by the addition of 1 mM unlabeled S-adenosyl-L-methionine, spotted onto EPM2001 glass fiber filters (Whatman), and processed as described in (25). Methylase units were quantified using a protection assay (25) in which 1 μ g of *M. Bam*HI modified λ DNA was challenged with 25 units of *Bam*HI endonuclease for 20 min. One unit of *Bam*HI methylase is defined as the amount of enzyme which, in a 50 μ l reaction volume after one hour at 37°C, is required to completely protect 1 μ g of λ DNA against *Bam*HI endonucleolytic cleavage.

*Bam*HI Methylase Purification and Protein Sequencing

Two liters of K802 cells containing pBamM1.8 were grown to saturation at 37°C. The cells were harvested by centrifugation and the pellet (8 g) stored at -70°C. The frozen cells were thawed on ice and resuspended in 20 ml Sonication Buffer (10 mM KPO_4 , 10 mM β -mercaptoethanol, 100 mM NaCl, 2 mM EDTA, pH 7.2).

Lysozyme was added to a final concentration of 750 μ g/ml and the suspension incubated for 30 min at 4°C and 15 min at room temperature. All subsequent steps were done at 4°C. The cells were disrupted by sonication and the cell lysate spun at 15,000 $\times$ g for 30 min. A polyethylene glycol precipitation step was used to remove nucleic acids: the supernatant was adjusted to 0.5M NaCl followed by the addition of polyethylene glycol (PEG 6000) to a final concentration of 7.5%. The solution was stirred for 2.25 hr and then clarified by centrifugation at 12,000 $\times$ g for 45 min.

The supernatant was diluted to a final NaCl concentration of 0.1 M by the addition of Buffer A (20 mM KPO_4 , 10 mM β -mercaptoethanol, 2 mM EDTA, pH 7.4). It was then applied to a P11 phosphocellulose (Whatman) column (10 $\times$ 2.5 cm) in Buffer A with 0.1 M NaCl, and the column washed with 3 vol Buffer A. The methylase was eluted with a 500 ml linear gradient of 0.1 M to 1.0 M NaCl. Five ml fractions were collected and assayed for methylase activity; the activity eluted at 0.7 M NaCl. The most active fractions (15 ml) were pooled and concentrated threefold using a Centricon (Amicon) chamber. The concentrate was dialyzed against Buffer A with 0.05 M NaCl and 50% glycerol and stored at -20°C.

Before sequencing, 10 μ g of sample was loaded onto a Vydac C4 (5 μ m, 4.6 $\times$ 300 mm) 300 Angstrom reverse phase column. The column was developed with a linear gradient of 5% to 100% acetonitrile in 0.1% trifluoroacetic acid over 45 min with a flow rate of 1 ml/min with detection at 214 nm. Four individual peaks were collected, lyophilized and 5% of each subjected to SDS-polyacrylamide electrophoresis. The protein moiety present in the major peak ran with an apparent molecular weight appropriate for the *Bam*HI methylase and this peak was used for protein sequencing.

The sequential degradation of the methylase protein was performed with an Applied Biosystems Model 470A gas phase sequencer (26). The first nineteen phenylthiohydantoins were

identified by high-performance liquid chromatography on an IBM Cyano (5 μ m, 4.5 $\times$ 250 mm) column with slight gradient modifications from those previously described (27).

Identification of the *M.BamHI* Modified Base and Determination of Its Position

Identification of the *M.BamHI* product was accomplished by TLC analysis as described in (28). The substrate, 80 ng of the nonphosphorylated *BamHI* 12 base linker (d(CGCGGAT-CCGCG); New England Biolabs) was methylated for two hr at 37°C in a 50 μ l reaction containing 50 mM Tris, pH 7.5, 10 mM EDTA, 5mM β -mercaptoethanol, 0.8 mM [3 H]S-adenosyl-L-methionine (79.7 Ci/mmol; New England Nuclear) and 40 units of *M.BamHI*. Subsequently 20 ng of the methylated product (containing 3000 cpm) were digested to deoxynucleosides and chromatographed as previously detailed (28).

The position of the methylated base was also determined. Two twenty base oligonucleotides, d(CTAGATAGGATCCTATCTAG), designated B1 and d(pCTAGATAGGATCCTATCTAG) designated pB1, synthesized on a Biosearch 8600 automated DNA synthesizer were used as substrates. The mixture of 160 μ g of pB1 was methylated under the conditions described above. It was mixed with [32 P] labeled B1 (approximately 3×10^4 cpm) and 75 μ g of unmodified pB1. The mixture was partially digested with venom phosphodiesterase and the products analyzed as described (28).

DNA Sequencing

Sanger dideoxy sequencing (29, 30) was performed by a modification of the 'collapsed plasmid' protocol (31–33). The templates used are shown in Figure 1. For the majority of sequencing reactions, premixed [35 S] DNA dNTP/ddNTP sequencing reagents (New England Biolabs), Klenow fragment of DNA Polymerase I (New England Biolabs), and α -[35 S] dATP (600–1200 Ci/mmol, New England Nuclear) were used. For sequencing the region between nucleotides 1862 and 2094 it was necessary to use templates containing a mixture of two plasmids. In these sequencing reactions, the primers were labeled with γ -[32 P]ATP (New England Nuclear, 3000 Ci/mmol) using T4 Polynucleotide Kinase (New England Biolabs). Unlabeled dATP (4 μ M), in lieu of α -[35 S] dATP was added to the sequencing reaction as recommended. When sequencing regions which were shown to be prone to 'compressions', *Thermus aquaticus* DNA polymerase (New England Biolabs) was substituted for Klenow polymerase in the sequencing reactions, and the dNTP/ddNTP ratios adjusted as recommended.

All reactions were subjected to electrophoresis on 6–8% standard, 'wedge' or buffer gradient sequencing gels (34–36). Both strands in the region containing the endonuclease gene and methylase genes were completely sequenced. When required, primers were synthesized on a Biosearch 8600 automatic synthesizer and purified by gel electrophoresis and column chromatography. Data were introduced into the computer utilizing a Grafbar Digitizer (SAC Corporation) and analyzed using computer programs from Cold Spring Harbor Laboratory (37) and the University of Wisconsin Genetics Computer Group (38).

Sequence Comparisons

The COMPARE and DOTPLOT programs combined with the BESTFIT program (38) from the University of Wisconsin Genetics Computer Group were used to compare the amino acid sequences of the *BamHI* proteins to those of other known endonuclease and methylase sequences. Sequences of the

Figure 1. Organization of the *BamHI* system and sequencing strategy. (A) Restriction map of the cloned region of the *B. amyloliquefaciens*H chromosome indicating the position and orientation of the endonuclease (R) and methylase (M) genes, and the intervening open reading frame (ORF). (B) Map positions of the plasmid subclones used as sequencing templates. Broken lines represent sequences within plasmids pBamM2.2 and pBamM2.1 not found in the *B. amyloliquefaciens*H chromosome (13). Discontinuous lines in pBam1.5 and pBamRM5.0 indicate a 1500 bp region not included on the restriction map. The black bar on pBamRM5.0 indicates the region between the *XmnI* and *BalI* sites that was sequenced using a mixture of pBamM1.8 and pBamRM5.0 as template and [32 P] labeled primers. (C) Position, orientation and extent of various sequencing reactions are indicated by the arrows. Arrows beginning with circles (○→) correspond to sequencing reactions primed with unique primers; those with vertical lines (|→) indicate those primed with universal primers. Both strands were completely sequenced for the entire 3014 bp region.

following endonucleases were used for comparison: *BsuRI*, *DdeI*, *DpnI*, *DpnII*, *EcoRI*, *EcoRV*, *HhaII*, *HinfI*, *PaeRI*, *PstI*, *SinI*, *TaqI* (primary sequence references given in (2)); *EcoRII* (39); *MspI* (40); *RsrI* (41); and *SmaI* (42).

M.BamHI was compared to the Type II methylases listed in (8) with the addition of *HaeIII* (43), *HpaII* (44), and *NgoPII* (45), m5C methylases; *DpnA* (46), *EcaI* (47) and *RsrI* (48), m6A methylases; and *SmaI* (42) and *BamH2* (49), methylases of undetermined type.

RESULTS

M.BamHI Purification and Sequence Determination

M.BamHI was purified from *E. coli* K802 (pBamM1.8) cells as described in Methods and Materials. After polyethylene glycol precipitation and phosphocellulose column chromatography, one major protein band was found when aliquots from the most active fractions were analyzed by SDS-PAGE. The band, which was approximately 80% of the protein present, migrated with an apparent molecular weight of approximately 50 kDa (data not shown). This agrees well with the size previously reported for *M.BamHI* purified from *B. amyloliquefaciens*H cells (12). The methylase yield from this purification procedure is 74,000 units/g cells. The specific activity of the purified methylase is 135,000

Table 1. Identification of the *M.Bam*HI modified base^a

Nucleosides Relative R _f ^d [³ H] incorporation (cpm) ^e	Solvent D ^b		Solvent G ^c	
	m ⁵ dC	m ⁴ dC + m ⁶ dA	m ⁶ dA	m ⁴ dC + m ⁵ dC
	0.9	1.0	1.3	1.1
	21	1308	20	1197

^a Experimental procedure given in Materials and Methods and (28).^b Solvent D is 80:20 ethanol-water (v/v).^c Solvent G is 66:33:1 isobutyric acid-water-ammonium hydroxide (v/v).^d Migration distances (R_f) are relative to dT.^e [³H] incorporation has been corrected for background cpm.

units/mg when determined on λ DNA. The protein is very stable; after 18 months storage, greater than 90% of the activity remains.

Prior to sequence determination, the *Bam*HI methylase preparation was further purified by reverse phase column chromatography, as described in Materials and Methods. Aliquots from the peak fractions were analyzed by SDS-PAGE and found to contain the appropriate size protein. The protein sequence of its amino terminus was determined to be MRFFSVFDIVKKNANQLGY.

Characterization of *M.Bam*HI Modification

A series of experiments were performed with the purified *M.Bam*HI to characterize its modification product. First it was determined that under the reaction conditions used, with M13 DNA as a substrate, only *Bam*HI sites are modified (28).

The base modification made by *M.Bam*HI was identified. Ligated *Bam*HI linkers d(pCGCGGATCCGCG) served as substrate. After methylation, the substrate was enzymatically degraded to 2'-deoxyribonucleosides and then separated by TLC in two different solvent systems. The results are shown in Table 1. It is clear that *M.Bam*HI methylation produces an N4-methylcytosine residue.

The position of the modified base within the *Bam*HI recognition site was also determined. A twenty base palindromic oligonucleotide, d(pCTAGATAGGATCCTATCTAG) was used as substrate. After methylation, the substrate was mixed with the same 20mer that had been phosphorylated in the presence of γ -[³²P]ATP. The mixture was subjected to partial degradation with venom phosphodiesterase and the digestion products separated by PAGE. As shown in Table 2, the internal cytosine within the *Bam*HI recognition sequence is the one that has been modified, since [³H] is present only in oligonucleotides containing this base. These data agree with the previous report of the *M.Bam*HI methylation site, which had been determined by nearest neighbor analysis (50).

Nucleotide Sequence of the *Bam*HI Region

In previous work, we assessed the relative position and approximate boundaries of the *M* and *R* genes within the cloned *Hind*III fragment of *B. amyloliquefaciens* DNA (13). The sequence of a 3014 bp region containing the *Bam*HI system has now been completed.

The resulting nucleotide sequence is shown in Figure 2. Analysis of the sequence shows two large open reading frames corresponding in position to the endonuclease and methylase genes, respectively. The two genes are divergently oriented. Sequence of the entire region, along with the predicted sequence of the endonuclease protein is shown in Figure 2A; the reverse complement of the sequence in the region containing the *M* gene along with its predicted protein sequence is shown in Figure 2B.

Table 2. Position of the base methylated by *M.Bam*HI^a

Oligodeoxyribonucleotide ^b	Radioactivity ^c (cpm)		Ratio ^d [³ H]/[³² P]
	[³ H]	[³² P]	
d(pCTAGATAGGATCCTATCTAG)	3167	572	5.5
d(pCTAGATAGGATCCTATCTA)	4391	783	5.5
d(pCTAGATAGGATCCTATCT)	5195	1252	4.1
d(pCTAGATAGGATCCTATC)	3773	1010	3.7
d(pCTAGATAGGATCCTAT)	1263	313	4.0
d(pCTAGATAGGATCCTA)	1355	302	4.4
d(pCTAGATAGGATCCT)	667	191	3.4
d(pCTAGATAGGATCC)	1747	205	5.7
d(pCTAGATAGGATC)	1559	257	6.1
d(pCTAGATAGGAT)	252	170	1.4
d(pCTAGATAGGA)	213	183	1.4
d(pCTAGATAGG)	153	166	1.1
d(pCTAGATAG)	147	165	0.92

^a Experimental procedure given in Material and Methods and (28)^b *Bam*HI recognition site is indicated in **boldface**.^c [³H] is in m⁴C and [³²P] is in the phosphomonoester terminus of the oligodeoxyribonucleotide. Values have been corrected for background cpm and channel cross-talk.^d Increase in ratio for the oligomer ending in the methylated residue (i.e. d(pCTAGATAGGATC)) has been observed before (28, 48) and is due to the faster cleavage by venom phosphodiesterase of the unmethylated oligodeoxynucleotide.

In the *R* gene, the first ATG occurs at position 1981 (Fig. 2A). Sequence from this position had previously been shown to match that predicted by the purified endonuclease protein (13). The open reading frame ends at position 2620 with the termination codon TAG. This would generate a protein 213 amino acids in length (Mr 24,570) which agrees well with the size of the purified endonuclease as seen by SDS-PAGE (11, 13).

A comparison was made between the sequence of the methylase protein and that predicted by the nucleotide sequence of the *M* gene. This showed the translational start of the gene to be a TTG codon at position 1622 in Figure 2B. The use of TTG as an initiation code has been reported for a number of other cytosine methylase genes from Gram positive organisms (51–55). The open reading frame terminates with a TGA at position 353. This generates a protein 423 amino acids in length (Mr 49,547) which is in agreement with the size predicted by SDS-PAGE of the purified protein. Substantiation for the assignment of the termination site also comes from the construct pBam Δ 1.9 where the region between the *Xba* and *Hind*III was deleted. The deletion removes the last two amino acids from the ORF containing *bam*HIM gene. Cells containing this plasmid display only partial methylase activity.

In the intergenic region between the *R* and *M* genes lies a third small ORF aligned in the same orientation and overlapping the open reading frame of the *R* gene. This ORF is 306 nucleotides

in length and could code for a protein of 13,351 molecular weight. Although no native protein product of this ORF has yet been isolated, it has been seen in an *in vitro* transcription/translation system. As shown in Figure 2A, the first TTG in the ORF is located at position 1668 and the first ATG at position 1719. The ORF terminates with a TAA at position 1962.

The regions upstream of the three open reading frames in the system were scanned for sequences resembling both the consensus *B. subtilis* vegetative and *E. coli* major promoters. No obvious promoter-like sequences were found upstream of either the *R* gene or the small ORF. However upstream of the *bamHIM* gene are two such sequences: TTGACG N17 AATAAT (Figure 2B, position 1763) and TTGACG N19 TAAAAT (Figure 2B, position 1700). Both these sequences differ from the consensus promoter by only one base in each of the -35 and -10 regions.

The upstream regions were also screened for potential ribosome binding sites. Thirteen bp upstream of the translational start of the *M* gene (position 1635, 3B) is the sequence GAGGTG, predicted to be the ribosome binding site. It has a high degree of complementarity to the *B. subtilis* 16S RNA and good spacing from the initiation site. This conforms well with a previous

observation that genes from Gram positive bacteria initiating with a TTG codon often contain strong Shine Dalgarno complementarity (56).

In the case of the endonuclease gene, 14 bp upstream of the initiating ATG is the sequence GGGGG (position 1967, 3A), which presumably serves as the ribosome binding site. There is no good potential ribosome binding site before the first TTG (position 1701, 3A) in the intergenic ORF, but 9 bp upstream of the first ATG (position 1719, 2A) is the sequence GGAGG.

There are no *Bam*HI sites (GGATCC) found within or adjacent to the coding region for the system.

Base Composition and Codon Usage

The base composition within the sequenced region is 31% G+C which is substantially lower than the 47% G+C reported for *B. amyloliquefaciens* DNA (57). The lower G+C content is even more evident when examining codon usage in the endonuclease and methylase genes. Table 3 shows a comparison of codon usage between the *bamHIR* and *bamHIM* genes and averaged values for *E. coli* and *B. subtilis* genes, respectively. The *Bam*HI gene

A	1	AAGCTTCCTT AATTCGGAA GATACCGATT TAATGCATAC TGAGATAGTG GAGATCGGCA AGGAAATGCA AGGTCAATAT
	81	AAGATTAAAT TCCATGACA TTTATGTAGT GTTTGAAGAT TCAAAAATAA GAGAAGAATT TTTACGAATA GATCATCTTT
	161	TTCCATGAGG ATCGATCCAT ATAGACTAAC TTACAATATA AAATGGACTT CTATTAGGAA GTCCATTAC CATTTTTAT
	241	CTGGTTTTCT CAATCTTTTA ATGCAGTAAA AATCTCTCTA AAGTCCTTAT CAAAAGTATA TTTTTTTTCT AGTGAGTATT
	321	TAAATATCGA CTTTTTTTAA AAGATAATTA TCATTCTAGA CTTAATGAAA TTTGGTCATT GTTTTCGTTT TTTTATCAC
	401	CTTTAGAAAT TCCCTCATGG TCTGTCTTAA AATCATCATC AATCCATAAA CCATTTTTTT TACGTAATTC TTTAACCCTA
	481	TTAGGGAACA GCTTGTTTTT TCTTCTATAA ACTTTACCTA ACAGCTTTTT ATCTTTGTCT TTTATTTTAA GTCTCTCTTT
	561	GATAATTTCA CAATTCCTCA ACTCAAAACC AATCCATTTA CGACCTAACA TTTCACTTAC AGCAAAATGT TTACCGCTTC
	641	CTCCAAACGG GTCTAAAACA ACGTCACCTT CATTGTGACT CATAGTTATT ATACGATCAA GTAATTTAAC TGATAATTCA
	721	TTGAACCTTC TATTTTTTGA ACTACTATGT CTAACAGGGT AAATATCTGA CCAACATCA GAAAGAGTTA CACCCTTGG
	801	GTTCATTTTA TTCTTGTAAC CGCCATAGTC TTTTATTCTT CTACCACAAT GAGGGCAAGT TTGTAGAGGT ATCCTTTGAA
	881	CATTAAATGT CTTAGTTTAA TCACCTTTTA CATAGTATAA AAGGCTGTAA TTTGCTGGAT ATAATCTATT CTGAATTGGA
	961	AGTCCAAAT TCATATCTAC AGTTATCCAG TTTCTAAAAT TCAATTTCCCT ATTTAAGTAC TCAGAAAGGT AAGTGTCCA
	1041	TTTTGGAATA TTATAAATGA ATAAAGAACC GCCTGGTTTT AAAACTCTGA TACACTCGTC TATCCATTTA TAATACCAAT
	1121	CCAAATATCC GCTAAAGGAA TTTTATCTGT TTACACCCCTC ATCATACTCT TTATCAAGGT TAAATGGTGG ATCAGCAAAA
	1201	ATAGTGTCCA CGTTTTTCATC AGGAACCTGT TTAACACAGT CTAACAATCC CCCATTGTAT AATTTTCCAA GATCAGTATA
	1281	AAAATAAGGT TTGATTCTT GAATAGAATT AAATTCGTTT AGTTTAAACAC TATTTTTATT TCGAGGATC TTGGCAATTT
	1361	CTTTTGTGTT ATTTAAGAAC ACATCTCTA ACCAGCCGG TATCCTGCCT AATTTTAATT CTAATCTTAC TTTAGTCATT
	1441	CCAAGATAAC TAAGTATTC ATTCAACACT TTATCTGTTG GAATAACTCC ATTTTCTTTA TAGGCGAGTA AATCCTCTCT
	1521	ATTCACATTG TAATTTTCA ATACAGCAT AATTCAGTT TCCGTATACC CTAAGTATT CGCTTTATT TTAAACAATG
		P I V Y L I L
	1601	CAAAAACAGA AAAAAACGC AATATTATAC ACCTCGTGTA ACTTATAGTC TGAG CCT ATA GTC TAC TTG ATT TTA
		S F H L N K R Q L L S G G T M T N N I R L L
	1677	TCA TTC CAT TTA AAC AAG CGT CAA TTG TTA TCA GGA GGT ACT ATG ACT AAC AAT ATC AGA TTA TTA
		*** **
		F G Q K V R Q I R L S K S N M S Q E K L A F
	1743	TTT GGT CAA AAA GTT CGT CAA ATT AGA TTG TCC AAA AGT AAT ATG TCA CAA GAA AAA CTA GCT TTT
		E C D L H R T Y I S D I E R S G T R N V S L D
	1809	GAA TGT GAC CTA CAC AGA ACC TAC ATT TCT GAT ATT GAA CGA GGA ACA AGA AAT GTT TCC TTA GAC
		N I E K I S K A L G V Q P K D L L D F T T I
	1875	AAT ATT GAA AAG ATT TCT AAA GCA CTC GGT GTT CAA CCA AAG GAC TTA CTT GAT TTT ACT ACA ATC
		G N C E Q K K *
	1941	GGC AAC TGT GAA CAA AAG AAA TAA GTGGGGGCAA ATTAGA ATG GAA GTA GAA AAA GAG TTT ATT ACT

		D E A K E L L S K D K L I Q Q A Y N E V K T
	2008	GAT GAA GCA AAA GAA TTA CTA TCT AAA GAT AAG TTA ATT CAA CAA GCA TAC AAT GAA GTT AAA ACA
		S I C S P I W P A T S K T F T I N N T E K N
	2074	TCT ATT TGT TCA CCT ATT TGG CCA GCA ACC TCA AAA ACC TTC ACG ATT AAC AAC ACC GAA AAG AAT
		C N G V V P I K E L C Y T L L E D T Y N W Y
	2140	TGT AAC GGT GTA GTA CCA ATT AAA GAA CTA TGT TAC ACC TTA CTT GAA GAT ACA TAT AAC TGG TAT
		R E K P L D I L K L E K K K G G P I D V Y K
	2206	AGA GAA AAA CCC CTT GAT ATA CTT AAA GGT GGT CCG ATT GAT GAT GAT GAT GAT GAT GAT GAT GAT
		E F I E N S E L K R V G M E F E T G N I S S
	2272	GAG TTC ATA GAA AAC AGT GAA CTT AAA CGT GTA GGT ATG GAA TTT GAA ACA GGA AAT ATT AGT TCT
		A H R S M N K L L L G L K H G E I D L A I I
	2338	GCC CAC CGT TCA ATG AAC AAA CTT CTA TTA GGA TTA AAA CAT GGC GAA ATT GAT TTG GCT ATT ATC
		L M P I K Q L A Y Y L T D R V T N F E E L E
	2404	CTT ATG CCT ATT AAA CAA TTG GCC TAT TAT CTT ACA GAT CGT GTT ACC AAT TTC GAG GAA TTA GAA
		P Y F E L T E G Q P F I F I G F N A E A Y N
	2470	CCT TAT TTT GAA CTT ACT GAA GGA CAA CCA TTT ATT TTT ATT GGA TTT AAT GCT GAG GCT TAT AAT
		S N V P L I P K G S D G M S K R S I K K W K
	2536	TCT AAT GTC CCT TTA ATT CCC AAA GGT TCT GAC GGT ATG TCA AAA CGC TCA ATT AAG AAA TGG AAA
		D K V E N K *
	2602	GAT AAA GTT GAA AAC AAA TAG TTTTGTGTA AGAGGCTTCG ATTTTGGAG TCTTTTTGT TATAAAAAATC
	2673	GCTGTTCCTA TTATGAAAAA ATTCGCGTTT ATTTGATCAT TTCTCAATCT CTTTATTCTT ATTATTTCAT TCTTCTATTA
	2753	AAACATATAT TTCTTAAGTA GCGATGGGGA GAAATAGATC TTCTCAAAAG TAACCTATTA AGAAAAAGTT ACATTACTT
	2833	TCACTAAAAT ACAGATTAAT TTAACAACGT TGAACCTGAA ATAGTTAAAA ATATGTTGCA TTCAACAGTT GATTTTCTG
	2913	CTAGAATATC TTCAACAAGTC AATACACGA GTTTATCGTC CTCTCGGCAA TCGCGCGGAA GACTGGGTCT TTCGGGAAC
	2993	GGATGCCAG AGTATATTAA AA

B	1764	<u>TTTGACGAAC</u> TTTTGGACCA <u>AATAATAATC</u> TGATATTGTT AGTCATAGTA CCTCCTGATA <u>ACAATTGACG</u> CTGTGTTAAA	
			M R E E S Y
	1684	TGGAATGATA <u>AAATCAAGTA</u> GACTATAGGC TACAGACTAT AAGTTACACG AGGTGTATAA TA TTG CGT TTT TTT TCT GTT	
			* *****
		<u>E D I V K N K A N Q L G Y T E T E M Y A V L K</u>	
	1604	TTT GAC ATT GTT AAA AAT AAA GCG AAT CAG TTA GGG TAT ACG GAA ACT GAA ATG TAT GCT GTA TTG AAA	
		N Y N V N K K D L L A Y K E N G V I P T D K V	
	1535	AAT TAC AAT GTG AAT AAG AAG GAT TTA CTC GCC TAT AAA GAA AAT GGA GTT ATT CCA ACA GAT AAA GTG	
		L N G I L S Y L G M T K V E L E L K L G R I P	
	1466	TTG AAT GGA ATA CTT AGT TAT CTT GGA ATG ACT AAA GTA GAA TTA GAA TTA AAA TTA GGC AGG ATA CCG	
		A G L E D V F L N N T K E I A K I L E N K N S	
	1397	GCT GGG TTA GAG GAT GTG TTC TTA AAT AAC ACA AAA GAA ATT GCC AAG ATC CTC GAA AAT AAA AAT AGT	
		V K L N E F N S I Q E I K P Y F Y T D L G K L	
	1328	GTT AAA CTA AAC GAA TTT AAT TCT ATT CAA GAA ATC AAA CCT TAT TTT TAT ACT GAT CTT GGA AAA TTA	
		Y N G D C L E L F K Q V P D E N V D T I F A D	
	1259	TAC AAT GGG GAT TGT TTA GAA CTG TTT AAA CAA GTT CCT GAT GAA AAC GTG GAC ACT ATT TTT GCT GAT	
		P P F N L D K E Y D E G V T D K N S F S G Y L	
	1190	CCA CCA TTT AAC CTT GAT AAA GAG TAT GAT GAG GGT GTA ACA GAT AAA AAT TCC TTT AGC GGA TAT TTG	
		D W Y Y K W I D E C I R V L K P G G S L F I Y	
	1121	GAT TGG TAT TAT AAA TGG ATA GAC GAG TGT ATC AGA GTT TTA AAA CCA GGC GGT TCT TTA TTC ATT TAT	
		N I P K W N T Y L S E Y L N R K L N F R N W I	
	1052	AAT ATT CCA AAA TGG AAC ACT TAC CTT TCT GAG TAC TTA AAT AGG AAA TTG AAT TTT AGA AAC TGG ATA	
		T V D M K F G L P I Q N R L Y P A N Y S L L Y	
	983	ACT GTA GAT ATG AAA TTT GGA CTT CCA ATT CAG AAT AGA TTA TAT CCA GCA AAT TAC AGC CTT TTA TAC	
		Y V K G D K P K T F N V Q R I P L Q T C P H C	
	914	TAT GTA AAA GGT GAT AAA CCT AAG ACA TTT AAT GTT CAA AGG ATA CCT CTA CAA ACT TGC CCT CAT TGT	
		G R E I K D Y G G Y K N K M N P K G V T L S D	
	845	GGT AGA GAA ATA AAA GAC TAT GGC GGT TAC AAG AAT AAA ATG AAC CCA AAG GGT GTA ACT CTT TCT GAT	
		V W S D I Y P V R H S S S K N R K F N E L S V	
	776	GTT TGG TCA GAT ATT TAC CCT GTT AGA CAT AGT AGT TCA AAA AAT AGA AAG TTC AAT GAA TTA TCA GTT	
		K L L D R I I T M S T N E G D V V L D P F G G	
	707	AAA TTA CTT GAT CGT ATA ATA ACT ATG AGT ACA AAT GAA GGT GAC GTT GTT TTA GAC CCG TTT GGA GGA	
		S G T T F A V S E M L G R K W I G F E L G N C	
	638	AGC GGT ACA ACA TTT GCT GTA AGT GAA ATG TTA GGT CGT AAA TGG ATT GGT TTT GAG TTG GGG AAT TGT	
		E I I K E R L K N K D K D K K L L G K V Y E E	
	569	GAA ATT ATC AAA GAG AGA CTT AAA AAT AAA GAC AAA GAT AAA AAG CTG TTA GGT AAA GTT TAT GAA GAA	
		K N K L F P N R V K E L R K K N G L W I D D D	
	500	AAA AAC AAG CTG TTC CCT AAT AGG GTT AAA GAA TTA CGT AAA AAA AAT GGT TTA TGG ATT GAT GAT GAT	
		F R Q D H E G N S K G D K K N E N N D Q I S L	
	431	TTT AGA CAA GAC CAT GAG GGA AAT TCT AAA GGT GAT AAA AAA AAC GAA AAC AAT GAC CAA ATT TCA TTA	
		S L E * *	
	362	AGT CTA GAA TGA TAA TTATCTTTA AAAAAAGTCG ATATTAAAT ACTCACTGAA AAAAAATAT ACTTTTGATA	
	287	AGGACTTTAG AGAGATTTT ACT	

Figure 2. The nucleotide and amino acid sequence of the 3014 bp region containing the *Bam*HI endonuclease and methylase genes, and intervening ORF. (A) The complete nucleotide sequence, along with the predicted amino acid sequence of the small ORF protein (position 1656 to 1964) and endonuclease (position 1981 to 2622). The putative ribosome binding site is indicated by small asterisks (*). The 12 amino acids from the amino terminus of the endonuclease confirmed by protein sequencing are underlined. The stop codons are indicated by a large asterisk (S). The initiation (position 1620) and putative termination (position 351) codons of the methylase gene encoded on the other strand, are shown in **boldface**. (B) The nucleotide sequence and predicted amino acid sequence from the complementary strand of the region encoding the methylase gene (position 1764 to 265 in (A)). The -10 and -35 regions of two potential promoters are indicated by a double underline. The putative ribosome binding site is indicated by small asterisks (*). The nineteen amino acids from the methylase confirmed by protein sequencing are underlined. The stop codons are indicated by a large asterisk (S). The first codon of the ORF, which is encoded on the opposite strand, is shown in **boldface**.

values are very skewed in their preference for A or U over G or C in the third positions. In the *M* gene, 75% of the codons end in either A or U; in the *R* gene, 78%. As can be seen in the Table, the *Bam*HI genes' codon preference does not resemble that of strongly expressed genes of *E. coli* or *B. subtilis* (a species which is closely related to *B. amyloliquefaciens*). However the discrepancy in codon usage does not seem to seriously impair *Bam*HI expression in either *E. coli* or *B. subtilis*.

Many of the restriction modification gene sequences reported to date contain unusually low G+C content (compared to that of the host genome) and/or third position A/U codon preferences. This is true of *bspRIM*, *bsuIRM*, *ddeIRM*, *ecoRIRM*, *ecoRVRM*, *haeIIIM*, *hhaIM*, *pstIRM*, *taqIRM* (primary references given in (2)); *fokIRM* (58); *hpaIIIM* (44); *mspIRM* (40); *pvuIIIM* (9); *rsrIRM* (41, 48). The significance of this finding remains obscure, but has been suggested to reflect a relatively recent evolutionary origin of restriction systems from a progenitor organism with low G+C content (59).

Protein Sequence Comparisons

The derived amino acid sequences of the *Bam*HI proteins were compared to that of other endonucleases and methylases having

known sequences. As in all other systems studied to date, comparison of *R.Bam*HI and *M.Bam*HI reveals no significant similarities. Neither does *R.Bam*HI have great similarity to other sequenced endonucleases. This is usually the case when endonucleases are compared. However, recently there has been one instance involving the isoschizomeric pair *R.Eco*RI and *R.Rsr*I, where a high degree of similarity has been found (48).

Comparison of *M.Bam*HI to other methylases proved to be more interesting. We have shown *M.Bam*HI to be a DNA (m^4C) methylase. It shows no significant similarity to the compiled group of known or assumed DNA (m^5C) methylases. It does share regional similarities with some m^6A and m^4C methylases. Dot matrix comparisons between *M.Bam*HI and five of the methylases most similar to it are shown in Figure 3. Two of the methylases shown, *M.Hin*fl and *Dpn*A, are m^6A methylases; two, *M.Cfr*9I and *M.Pvu*II are m^4C methylases. The fifth, *M.Bam*H2, a methylase encoded by the H2 prophage of *B. amyloliquefaciens*, is known to be a cytosine methylase, but the type of modification (m^4C or m^5C) it produces has not been determined (49). Three other methylases, *Sma*I, *Rsr*I and *Eca*I also show significant sequence similarities within the same regions (data not shown).

Table 3. Codon Usage in *bamHIM* and *bamHIR* genes.

	Codon	<i>E. coli</i> ^b	<i>B. subtilis</i> ^c	<i>BamHI</i> ^d methylase gene	<i>BamHI</i> ^d endonuclease gene
Leu	UUA	6	22	48	33
	UUG	8	14	13	9
	CUA	2	6	6	14
	CUU	9	26	22	43
	CUG	69	22	6	0
Arg	CUC	7	10	4	0
	AGA	1	28	50	20
	AGG	0	9	25	0
	CGA	2	9	0	0
	CGU	58	25	25	20
Pro	CGG	3	11	0	0
	CGC	35	18	0	60
	CCA	20	19	47	30
	CCU	9	34	41	40
	CCG	65	38	12	10
Gln	CCC	6	10	0	20
	CAA	27	54	75	100
Lys	CAG	73	46	25	0
	AAA	77	75	82	84
Ala	AAG	23	25	18	16
	GCA	23	27	12	37
Val	GCU	28	28	50	37
	GCG	31	25	12	0
	GCC	19	20	25	25
	GUA	23	25	28	44
Gly	GUU	38	31	56	44
	GUG	27	19	16	0
	GUC	13	25	0	11
	GGA	5	32	31	36
Ser	GGU	49	25	45	54
	GGG	7	14	14	0
	GGC	41	30	10	9
	AGU	7	11	33	17
	AGC	22	24	14	0
	UCA	8	19	19	42
Thr	UCU	27	24	28	42
	UCG	11	10	0	0
	UCC	26	12	5	0
	ACA	6	43	41	33
Ile	ACU	24	15	53	17
	ACG	20	28	6	8
	ACC	51	14	0	42
	AUA	1	11	32	11
Asn	AUU	37	50	52	83
	AUC	62	39	16	5
	AAU	24	53	74	50
Phe	AAC	76	47	26	50
	UUU	44	64	80	67
Tyr	UUC	57	36	20	33
	UAU	41	62	65	78
	UAC	59	38	35	22
Glu	GAA	73	70	72	80
	GAG	27	30	28	20
Cys	UGU	42	48	80	100
	UGC	58	54	20	0
His	CAU	39	69	100	50
	CAC	61	31	0	50
Asp	GAU	51	64	70	89
	GAC	49	36	30	11
Met	AUG	100	100	100	100
Trp	UGG	100	100	100	100

^a Numbers given represent % use of one codon relative to all others within one amino acid group. The single codon groups Met and Trp therefore have a value of 100.

^b Values used for *E. coli* come from analysis of 25 strongly expressed nonregulatory protein genes found in reference (65).

^c Values used for *B. subtilis* come from analysis of 21 chromosomal genes found in reference (66).

^d Values for codons in the *BamHI* methylase and endonuclease genes ending in A or U appear in **boldface**.

Five regions with varying degrees of relatedness have been identified in all the methylases shown in Figure 3A and 3B. Three of the regions (A–C) are in the amino terminal half of the methylase. D and E form a long contiguous region near the carboxy terminus. Regions D and E are separated here because D shows sequence motifs present in a wide variety of other methylases that share no sequence similarity with region E (see below). There is a space of variable length between regions C and D, reminiscent of the arrangement of related regions found in m⁵C methylases (6, 7). In several of the comparisons shown (i.e. *M.BamHI* vs *M.BamH2*, *M.Cf9I*, *M.HinfI*) the space between C and D contains additional short regions of similarity. Regions A–E contain sequence motifs that are becoming increasingly familiar. The longest segment of related amino acids is found in Region D which contains the FXGXG motif present in all DNA methylases (6, 8, 47). Region B contains the (SD)PP(YF) motif that has been found in all m⁶A and m⁴C methylases (3–5, 8, 9). It has been suggested that the m⁶A and m⁴C methylases could be differentiated in this region on the bases of whether they contained (DN) PPY (m⁶A) or TSPPY (m⁴C) (8). However, since *M.BamHI* contains DPPF but produces m⁴C, this sequence does not allow discrimination between the two types of methylases.

Figure 3B shows a dot matrix comparison between *M.BamHI* and *M.BamH2*. Although the two methylases both come from *B. amyloliquefaciens*, are encoded by chromosomal genes and modify the sequence GGATCC, they show less similarity to each other than those methylases shown in 3A. The similarity between *M.BamHI* and *M.BamH2* is mainly confined to the DE region. There is little apparent relatedness on the nucleotide level; the two genes are not closely linked and will not cross hybridize under stringent conditions (13). The fact that *M.BamH2* contains a TSPPY motif not found in m⁵C methylases suggests that it is an m⁴C methylase, like *M.BamHI*.

The first one hundred residues of *M.BamHI* are apparently unique to that methylase. Other methylases have been characterized with long unique sequences on either their amino or carboxy terminal ends that could be deleted without complete loss of methylase activity (60, 61). This needs to be tested for *M.BamHI*.

In addition to comparisons of the endonuclease and methylase, the small ORF of the *BamHI* system was analyzed for similarities to other protein sequences found in GenBank. None were found. However, when the protein was compared to a small ORF from the *PvuII* system, the two ORFs show great similarity. The *PvuII* ORF, like that of *BamHI*, also appears to regulate the system (R. Blumenthal, pers. comm.). Furthermore, using the criteria outlined in (62) both ORFs satisfy all the rules for formation of helix-turn-helix structure. This suggests the ORFs are actually DNA binding proteins and may eventually help to explain their mode of action.

Expression of the Cloned *BamHI* System

It was initially necessary to clone the *BamHI* system in two steps and subsequently was found that a plasmid containing the complete system (pBamRM5.0) could not be introduced into unprotected cells. It was not clear what role the second *M* gene played: specifically, if it was necessary only for the initial establishment of the system in *E. coli*, or if it was also required for continued maintenance. The plasmid pBamRM5.0 has pA-CYC184 as its vector, a plasmid with relatively low copy number in *E. coli* (63). It seemed possible that moving the system to a

Figure 3. Comparison of amino acid sequences between *M. BamHI* and other DNA methylases. A) Dot matrix comparisons of *M. BamHI* with *M. HinfI*, *M. PvuII*, *DpnA* and *M. Cfr9I*. (B) Dot matrix evaluation of *M. BamHI* and *M. BamH2*. In (A) and (B) evaluations are done by COMPARE/DOTPLOT (38). Diagonal lines indicate regions of strong similarity. Five consistent regions of strong similarity (A–E) are labeled and the corresponding sequences are given in (C) below. (C) Amino acid sequences of regions of similarity are shown. Amino acids identical to *M. BamHI* in each region are boxed. Numbers at the end of the line indicate residue number of the first amino acid in the region. Numbering for *M. PvuII* is based on an amino terminal sequence determination of the purified protein done at New England Biolabs (J. Benner and L. Dorner, unpublished observation).

vector with higher copy number might increase *M. BamHI* expression. This, in turn, may obviate the need for a second *M.*-containing plasmid during the initial transformation step and/or later maintenance. Moving the system to a vector with higher copy number might also have the benefit of increasing *R. BamHI* expression, a second goal of our work.

We also wanted to determine how the cloned system would act in *B. subtilis*, a species closely related to *B. amyloliquefaciens*, and therefore a more 'natural' host. In particular, would the whole system be able to transform unprotected cells in one step and would the genes express better than in *E. coli*?

To answer these questions, the *BamHI* system was transferred

to *E. coli* plasmid vectors with higher copy number and also a bifunctional *E. coli/B. subtilis* shuttle vector. The new constructs were tested for their ability to transform the host in the presence or absence of pBamM2.2 plasmid containing *bamHI*. Transformants were tested for endonuclease expression by both an *in vivo* assay (restriction test of λ_{vir}) and an *in vitro* assay (endonuclease activity in crude extracts). The results are shown in Table 4.

As seen in Table 4, the construct with the complete system on pACYC184 cannot transform naive *E. coli* K802 cells. It does transform K802 with the methylase plasmid efficiently and the second plasmid persists without selection. We take this as

Table 4. Characterization of Clones Carrying the *Bam*HI Restriction Modification System

Plasmid	Host	Transformation Studies		Phage Restriction (EOP) ^c	Endonuclease Expression (units/l) ^d
		Extra <i>M</i> gene needed for introduction ^a	Extra <i>M</i> gene needed for maintenance ^b		
pBamRm5.0	<i>E. coli</i>	+	+	2.8×10^{-3}	1.9×10^4
pBamRM5.0-328	<i>E. coli</i>	+	+	3.2×10^{-2}	1.5×10^4
pBamRM2.7-HP13	<i>E. coli</i>	+	—	2.4×10^{-4}	7.5×10^4
pBamRM2.7-HP13	<i>B. subtilis</i>	—	—	ND ^e	1.5×10^5

^a *E. coli* K802 and K802(pBamM2.2) were transformed in parallel with the plasmids listed and plated for Cm^R colonies. '+' indicates Cm^R colonies at high efficiency ($> 10^5/\mu\text{g}$) on K802 (pBamM2.2) but not K802 cells. *B. subtilis* 168trpC2 was transformed with the plasmid indicated and plated for Cm^R colonies. '—' indicates high transformation efficiency ($> 10^5$ colonies/ μg) in the absence of a second plasmid carrying *bamHIM*.

^b Individual colonies of *E. coli* K802 (pBamM2.2) containing the *bamHIRM* plasmid listed were streaked and grown with selection only for Cm^R. Single colony isolates were tested by Ap resistance and plasmid isolation for the presence (+) or absence (—) of pBamM2.2. *B. subtilis* 168trpC2 containing pBamRM2.7-HP13 isolates were grown and tested by endonuclease assays of crude extracts for the stable maintenance of the plasmid in the absence (—) of another copy of *bamHIM*.

^c EOP (Efficiency of Plating) is defined as the titer of λ_{vir} grown on K802 containing the test plasmid divided by the titer of λ_{vir} grown on K802 (pBamM2.2) i.e., the nonrestricting host.

^d Cultures (50 ml) containing the test plasmids were grown to mid log phase and harvested. Endonuclease activity was assayed from crude extracts as described in (13).

^e Not determined.

evidence that pBamM2.2 is required for maintenance of *bamHIR*. The cells are expressing endonuclease since λ_{vir} is restricted and endonuclease activity is found in crude extracts. (The activity found is approximately 20–30% that of *B. amyloliquefaciens* extracts.)

When the system is transferred to pBR328, a vector with higher copy number (17, 64, 65), the construct is still unable to transform unprotected cells. It does transform K802 (pBamM2.2). The unselected methylase plasmid is maintained in the cell even though it and pBamRM5.0-328 have incompatible origins of replication. Although the pBR-derived plasmid is present at higher copy number than the pACYC184 derivative in the cell, the level of endonuclease expression is comparable to that found in clones containing the system on pACYC184.

The *Bam*HI system was then cloned onto pHP13, a bifunctional vector containing the pUC *ori* as well as a *Bacillus* *ori* from pTA1060 (18, 19). As seen in the Table, the shuttle vector carrying the *Bam*HI system still cannot transform unprotected *E. coli* K802 cells; it does transfer to *E. coli* K802 (pBamM2.2) efficiently. However, after introduction of *Bam*HI on the shuttle vector, the methylase plasmid will be lost if unselected. We take this as evidence that in the case of pBamRM2.7-HP13, pBamM2.2 is required for the establishment but not the maintenance of *bamHIR*. Endonuclease activity in these cells is higher than in previous constructs as judged by both phage restriction and activity in crude extracts. Both of these properties are consistent with the higher copy number of pUC relative to pBR plasmids in *E. coli* (64, 65).

When the *Bam*HI system on the shuttle vector was transformed into *B. subtilis* cells, it was able to transfer without the need for an additional copy of the methylase gene premodifying the host. The *Bacillus* clone also showed a higher level of endonuclease activity, about twice its level in *E. coli*. This is particularly interesting since pHP13 plasmid copy number in *B. subtilis* is approximately one tenth that in *E. coli* (18, 19).

Better performance of the system in *B. subtilis* than *E. coli* may be due to better expression and/or stability of the methylase in the former host. But in addition the regulatory mechanism of the *Bam*HI system operating more efficiently in the more 'natural host' is probably playing a role (see Discussion).

DISCUSSION

*M.Bam*HI is now the fourth m⁴C methylase to be cloned, characterized and sequenced (8, 9). All four show strong regions of similarity to m⁶A methylases, but not to m⁵C methylases. Protein sequences of three of the four (i.e. *M.Bam*HI, *M.Cfr*91, and *M.Pvu*II) share have large regions of similarity (Figure 3). The fourth, *M.Mva*I, shows similarity to the other three only in one short region. That region contains the (DS)PP(FY) motif common to both m⁶A and m⁴C methylases. From the four m⁴C methylases, we now know there are no unique diagnostic sequence motifs. Therefore, there will be no easy way to distinguish them from m⁶A methylases.

The *Bam*HI system has been expressed in both *E. coli* and *B. subtilis*. In *E. coli*, it is always necessary to clone the system in two steps so as to introduce *bamHIR* into cells already expressing *bamHIM*. In *B. subtilis*, however, the system can transfer in one step and be stably maintained. Why does the *Bam*HI system behave differently in the two hosts?

*Bam*HI regulation by the ORF is probably one factor involved. We think the ORF normally acts to modulate *M* and *R* expression in *B. amyloliquefaciens* as the cells undergo physiological changes, like sporulation or transformation. The similarity of the *Bam*HI ORF to known DNA binding proteins suggests it acts as a transcriptional regulator. Experiments are now in progress to characterize its activity using both *E. coli* and *B. subtilis* clones.

Note added in proof

While this paper was being reviewed, the sequence of the *Bam*HI methylase gene was published (Vanek, P.G., Connaughton, J.F., Kaloss, W.D. and Chirikjian, J.G. (1990) Nucl. Acids Res. **18**, 6145). The nucleotide sequence it contains is identical to that reported by us in Figure 2B except for their addition of a T between the G and C (positions 1656 and 1655 in Figure 2B; position 229 in Vanek et. al.) in the intergenic region between the ORF and the *M* gene. We are currently investigating the discrepancy. They have also mistakenly identified the start of the methylase protein at the first ATG rather than the preceding TTG as determined by our sequence of the purified methylase protein.

ACKNOWLEDGMENTS

The authors wish to thank Carla Vaccaro for excellent technical assistance; Drs. S. Bron, B. Friedman, R. Yasbin and K. Bott for provision of materials and advice for the *Bacillus* work; and Drs. R. Blumenthal and E. Raleigh for helpful criticism of the manuscript and sharing unpublished data. Special thanks to Dr. Donald Comb for support of the project.

REFERENCES

- Janulaitis, A., Klimasauskas, S., Petrusyte, M. and Butkus, V. (1983) *FEBS Lett.* **161**, 131–134.
- Wilson, G.G. (1988) *Gene* **74**, 281–289.
- Lauster, R., Kriebardis, A. and Guschlbauer, W. (1987) *FEBS Lett.* **220**, 167–176.
- Narva, K.E., Van Etten, J., Slatko, B.E. and Benner, J.S. (1988) *Gene* **74**, 253–259.
- Guschlbauer, W. (1988) *Gene* **74**, 211–214.
- Posfai, J., Bhagwat, A.S., Posfai, G. and Roberts, R.J. (1989) *Nucl. Acids Res.* **17**, 2421–2435.
- Lauster, R., Trautner, T.A. and Noyer-Weidner, M. (1989) *J. Mol. Biol.* **206**, 305–312.
- Klimasauskas, S., Timinskas, A., Menkevicius, S., Butkiene, D., Butkus, V. and Janulaitis, A. (1989) *Nucl. Acids Res.* **17**, 9823–9832.
- Tao, T., Walter, J., Brennan, K.J., Cotterman, M.M. and Blumenthal, R.M. (1989) *Nucl. Acids Res.* **17**, 4161–4175.
- Wilson, G.A. and Young, F.E. (1975) *J. Mol. Biol.* **97**, 123–125.
- Smith, L.A. and Chirikjian, J.G. (1979) *J. Biol. Chem.* **254**, 1003–1006.
- Nardone, G., George, S.A. and Chirikjian, J.G. (1984) *J. Biol. Chem.* **259**, 10357–10362.
- Brooks, J.E., Benner, J.S., Heiter, D.F., Silber, K.R., Szynter, L.A., Jager-Quinton, T., Moran, L.S., Slatko, B.E., Wilson, G.G. and Nwankwo, D.O. (1989) *Nucl. Acids Res.* **17**, 979–997.
- Nathan, P.D. and Brooks, J.E. (1988) *Gene* **74**, 35–36.
- Raleigh, E.A., Murray, N.E., Revel, H., Blumenthal, R.M., Westaway, D., Reith, A.D., Rigby, P.W., Elhai, J. and Hanahan, D. (1988) *Nucl. Acids Res.* **16**, 1563–1575.
- Viera, J. and Messing, J. (1982) *Gene* **19**, 259–268.
- Soberon, H., Covarrubias, L. and Bolivar, F. (1980) *Gene* **9**, 287–305.
- Haima, P., Bron, S. and Venema, G. (1987) *Mol. Gen. Genet.* **209**, 335–342.
- Haima, P., Bron, S. and Venema, G. (1988) *Mol. Gen. Genet.* **213**, 364–369.
- Gingeras, T.R. and Brooks, J.E. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 402–406.
- Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
- Clewell, D.B. and Helinski, D.R. (1969) *Proc. Natl. Acad. Sci. USA* **62**, 1159–1166.
- Cohen, S.N., Chang, A.C.Y. and Hsu, L. (1972) *Proc. Natl. Acad. Sci. USA* **69**, 2110–2114.
- Bott, K.R. and Wilson, G.A. (1968) *Bacteriol. Rev.* **32**, 370–378.
- Theriault, G., Roy, P.H., Howard, K.A., Benner, J.S., Brooks, J.E., Waters, A.F. and Gingeras, T.R. (1985) *Nucl. Acids Res.* **13**, 8441–8461.
- Strickler, J.E., Hunkapiller, M.W. and Wilson, K.J. (1984) *Anal. Biochem.* **140**, 553–566.
- Hunkapiller, M.W. and Hood, L.E. (1983) In Hirs, C.H.W. and Timasheff, S.N., (ed.), *Methods in Enzymology*, Academic Press, New York, Vol. 91, pp. 486–493.
- Landry, D., Looney, M.C., Feehery, G.R., Slatko, B.E., Jack, W.E., Schildkraut, I. and Wilson, G.G. (1989) *Gene* **77**, 1–10.
- Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
- Sanger, F. and Coulson, A.R. (1978) *FEBS Letters* **87**, 107–110.
- Chen, E.Y. and Seeburg, P.H. (1985) *DNA* **4**, 165–170.
- Hattori, M. and Sakaki, Y. (1986) *Anal. Biochem.* **152**, 232–238.
- Hattori, M., Hidaka, S. and Sakaki, Y. (1985) *Nucl. Acids Res.* **13**, 7813–7827.
- Williams, S.A., Slatko, B.E., Moran, L.S. and DeSimone, S.J. (1986) *BioTechniques* **4**, 138–147.
- Ansorge, W. and Labeit, S. (1984) *J. Biochem. Biophys. Methods* **10**, 237–243.
- Olsson, A., Moks, T., Muhlen, M. and Gaal, A.B. (1984) *J. Biochem. Biophys. Methods* **10**, 83–90.
- Keller, C., Corcoran, N. and Roberts, R.J. (1984) *Nucl. Acids Res.* **12**, 379–386.
- Devereux, J., Haeblerli, P. and Smithies, O. (1984) *Nucl. Acids Res.* **12**, 387–395.
- Bhagwat, A.S., Johnson, B., Weule, K. and Roberts, R.J. (1990) *J. Biol. Chem.* **265**, 767–773.
- Lin, P.M., Lee, C.H. and Roberts, R.J. (1989) *Nucl. Acids Res.* **17**, 3001–3011.
- Stephenson, F.H., Ballard, B.T., Boyer, H.W., Rosenberg, J.M. and Greene, P.J. (1989) *Gene* **85**, 1–13.
- Heidmann, S., Seifert, W., Kessler, C. and Domdey, H. (1989) *Nucl. Acids Res.* **17**, 9783–9796.
- Slatko, B.E., Croft, R., Moran, L.S. and Wilson, G.G. (1988) *Gene* **74**, 45–50.
- Card, C.O., Wilson, G.G., Weule, K., Hasapes, J., Kiss, A. and Roberts, R.J. (1990) *Nucl. Acids Res.* **18**, 1377–1383.
- Sullivan, K.M. and Saunders, J.R. (1988) *Nucl. Acids Res.* **16**, 4369–4387.
- Lacks, S.A., Manarelli, B.M., Springhorn, S.S. and Greenberg, B. (1986) *Cell* **46**, 993–1000.
- Brenner, B., Venetianer, P. and Kiss, A. (1989) *Nucl. Acids Res.* **18**, 355–359.
- Kaszubska, W., Aiken, C., O'Connor, C.D. and Gumpert, R.I. (1989) *Nucl. Acids Res.* **17**, 10403–10425.
- Weiner, M.P. (1986) Ph.D. Thesis, Cornell University.
- Hallman, S., Keister, T. and Gotteher, A. (1978) *J. Mol. Biol.* **124**, 701–711.
- Buhk, H.-J., Behrens, B., Tailor, R., Wilke, K., Prada, J.J., Günther, U., Noyer-Weidner, M., Jentsch, S. and Trautner, T.A. (1984) *Gene* **29**, 51–61.
- Posfai, G., Kiss, A., Erdei, S., Posfai, J. and Venetianer, P. (1983) *J. Mol. Biol.* **170**, 597–610.
- Tran-Betcke, A., Behrens, B., Noyer-Weidner, M. and Trautner, T.A. (1986) *Gene* **42**, 89–96.
- Behrens, B., Noyer-Weidner, M., Pawlek, B., Lauster, R., Balganes, T.S. and Trautner, T.A. (1987) *EMBO J.* **6**, 1137–1142.
- Kupper, D., Zhou, J.G., Venetianer, P. and Kiss, A. (1989) *Nucl. Acids Res.* **17**, 1077–1088.
- McLaughlin, J.R., Murray, C.L. and Rabinowitz, J.C. (1981) *J. Biol. Chem.* **256**, 11283–11291.
- Ehrlich, M., Wilson, G.G., Kuo, K.C. and Gehrke, C.W. (1987) *J. Bacteriol.* **169**, 939–943.
- Looney, M.C., Moran, L.S., Jack, W.E., Feehery, G.R., Benner, J.S., Slatko, B.E. and Wilson, G.G. (1989) *Gene* **80**, 193–208.
- Dila, D., Sutherland, E., Moran, L., Slatko, B. and Raleigh, E.A. (1990) *J. Bacteriol.* **172**, 4888–4900.
- Szynter, L.A., Slatko, B., Moran, L., O'Donnell, K.H. and Brooks, J.E. (1987) *Nucl. Acids Res.* **15**, 8249–8266.
- Som, S., Bhagwat, A.S. and Friedman, S. (1987) *Nucl. Acids Res.* **15**, 313–332.
- Brennan, R.G. and Matthews, B.W. (1989) *J. Biol. Chem.* **264**, 1903–1906.
- Chang, A.C.Y. and Cohen, S.N. (1978) *J. Bacteriol.* **134**, 1141–1156.
- Lupski, J.R., Projan, S.J., Ozaki, L.S. and Godson, G.N. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 7381–7385.
- Miki, T., Yasukochi, T., Nagatani, H., Furuno, M., Orita, T., Yamada, H., Imoto, T. and Horiuchi, T. (1987) *Prot. Eng.* **1**, 327–332.
- Konigsberg, W. and Godson, N. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 687–691.
- Ogasawara, N. (1985) *Gene* **40**, 145–150.