Sequence, internal homology and high-level expression of the gene for a DNA-(cytosine *N*4)-methyltransferase, *M·Pvu* II

Tao Tao, Johannes Walter¹⁺, Kathleen J. Brennan[§], Melissa M. Cotterman^φ
and Robert M. Blumenthal^{*}

Department of Microbiology, Medical College of Ohio, PO Box 10008, Toledo, OH 43699 and ¹Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

Received November 18, 1988; Revised and Accepted April 4, 1989

EMBL accession no. X13778

ABSTRACT

The base sequence of the *pvuII* gene has been determined. This gene codes for a DNA-(cytosine *N*4)-methyltransferase, *M·Pvu* II. The base sequence contains a single large open reading frame that predicts a 38.3kDa polypeptide, consistent with experimental data. The *pvuII* gene contains some sequences common to DNA methyltransferases in general, but includes none of the sequences specifically conserved among DNA-(cytosine 5)-methyltransferases. The *pvuII* sequence also reveals an internal homology at the amino acid level, each half of which spans over 100 amino acids and is itself homologous to the sequences of some DNA-(adenine *N*6)-methyltransferases. A derivative of the *pvuII* plasmid was constructed to allow high-level production of *M·Pvu* II. Specifically, the composite *p_{lac}* promoter was inserted 5' to *pvuII*, intervening DNA was deleted, and the resulting construct was used to transform an *mcrB lacP* strain of *Escherichia coli*. When this transformant was induced with isopropyl-*B*-D-galactopyranoside (IPTG), growth rapidly ceased and *M·Pvu* II accumulated to the point of comprising over 10% of the total soluble protein.

INTRODUCTION

The great majority of characterized DNA methyltransferases (MTases) generate one of two bases: 5-methylcytosine (5mC) or *N*6-methyladenine (N6mA). Those two methylations have been known for about 25 years (1,2). Much more recently, a third type of DNA MTase was discovered which generates *N*4-methylcytosine (N4mC; 3–5). N4mC has been found in a broad range of bacterial species (6), and may present particular advantages to thermophilic bacteria since it is more resistant to heat-induced deamination than C or 5mC (7).

Many DNA MTases generating 5mC or N6mA have been characterized, and the base sequences of their structural genes have been determined in several cases. A comparison of these sequences has revealed two broad classes of conserved elements. One class appears to be relatively general to DNA MTase genes sequenced to date. For example, the consensus sequence DPPY (asp pro pro tyr) has been found in most (though not all) of the N6mA MTases, and in a minority of the 5mC MTases as well (*e.g.*, 8–10). The second class of conserved elements are specific to the cytosine MTases (11–14), which thus far has only included the sequences for 5mC MTases. This latter group of elements may reflect the reaction mechanism of the 5mC MTases, which appears to involve a temporary covalent linkage between the MTase and the 6 position of cytosine (15).

The *Pvu* II restriction-modification system recognizes the DNA sequence 5'-CAGCTG-3' (16). Its genes have been cloned (17). The MTase was found to act on the internal C of the substrate sequence (17,18), and to generate N4mC (18). We report here the base

1
AAGCTTTTCTTGGGATAGACCTAGCTCGCCTCGCATTTTTTTTACATTCTTTTGCAAGGGTTAA

HindIII -35 -10 +1 Potential promoter

63 MetMetVal***** Potential minigene
MetMetThr
TCGGGCTGATAAAGGATTTGTATTCTGCTCATAGACGGGATAATGATGGTATGATGACT
aGGAggt.....ATG Shine-Dalgarno

123
LeuAsnLeuGlnThrMetSerSerAsnAspMetLeuAsnPheGlyLysLysProAlaTyr
TTGAATCTACAGACTATGAGTAGCAATGATATGTTGAATTTGGAAAAAACCTGCCTAC

183
ThrThrSerAsnGlySerMetTyrIleGlyAspSerLeuGluLeuLeuGluSerPhePro
ACAACTAGTAATGGGTCTATGTATATAGGTGACTCATTGGAGCTATTAGAATCATCCCCA

Spe I

243
GluGluSerIleSerLeuValMetThrSerProProPheAlaLeuGlnArgLysLysGlu
GAAGAAAGTATTAGTCTGGTATGACTAGTCTCCCTTCGCATTACAACGAAAAAAGAA

Spe I

303
TyrGlyAsnLeuGluGlnHisGluTyrValAspTrpPheLeuSerPheAlaLysValVal
TACGGAACTTTAGAACAGCATGAGTATGTGGATTGTTTTTGTCAATTGCTAAAGTGGTT

363
AsnLysLysLeuLysProAspGlySerPheValValAspPheGlyGlyAlaTyrMetLys
AATAAGAACTAAAACCGGATGGAAGTTTTGTGTAGATTTTGGTGGGGCGTACATGAAA

423
GlyValProAlaArgSerIleTyrAsnPheArgValLeuIleArgMetIleAspGluVal
GGCGTTCAGCAGAAGTATCTATAACTTTAGAGTTTGTATCAAGAATGATAGACGAAGTG

(Bcl I)

483
GlyPhePheLeuAlaGluAspPheTyrTrpPheAsnProSerLysLeuProSerProIle
GGTTTTTCTTGCTGAAGATTTTATTGGTTAAACCCATCAAATTACCAAGCCCTATA

543
GluTrpValAsnLysArgLysIleArgValLysAspAlaValAsnThrValTrpTrpPhe
GAATGGGTTAATAAGAGAAAAATAAGAGTTAAGATGCAGTGAATACAGTATGGTGGTTC

```

603
SerLysThrGluTrpProLysSerAspIleThrLysValLeuAlaProTyrSerAspArg
TCTAAAAACAGAGTGGCCCTAAGTCTGATATAACGAAAGTTCTAGCTCCATATAGCGATAGA
T.YAa....ttg      IHF seq

663
MetLysLysLeuIleGluAspProAspLysPheTyrThrProLysThrArgProSerGly
ATGAAAAGTTAATTGAAGACCCAGACAAATTCTATACACAAAAACTAGACCATCTGGG
AT

723
HisAspIleGlyLysSerPheSerLysAspAsnGlyGlySerIleProProAsnLeuLeu
CATGATATAGGGAAGTCGTCTCTAAAGACAATGGAGGTTCAATTCTCTCTAATTATTA
-----
Xmn I

783
GlnIleSerAsnSerGluSerAsnGlyGlnTyrLeuAlaAsnCysLysLeuMetGlyIle
CAAATATCAAACCTCAGAATCAAATGGTCAATATTTAGCCAATTGTAAGTTGATGGGTATT
-----
(sole Cys codon)
(Ssp I)

843
LysAlaHisProAlaArgPheProAlaLysLeuProGluPhePheIleArgMetLeuThr
AAAGCGCATCCCGCTAGATTCCCTGCTAAGTTACCCGAGTTTTTTATTAGATGTTAACT
-----
Hpa I, Hinc II

903
GluProAspAspLeuValValAspIlePheGlyGlySerAsnThrThrGlyLeuValAla
GAGCCAGATGATCTTGTGTGTGATATTTTTGGCGGCTCTAATACTACTGGTCTAGTCGA

963
GluArgGluSerArgLysTrpIleSerPheGluMetLysProGluTyrValAlaAlaSer
GAAAGAGAATCTCGTAAATGGATCTCGTTTGAGATGAAACCTGAATATGTTGCGGCATCG
-----
(BstY I)

1023
AlaPheArgPheLeuAspAsnAsnIleSerGluGluLysIleThrAspIleTyrAsnArg
CGGTTTCGTTTTTTAGACAATAATATAAGTGAAGAAAAATAACTGATATTTATAATCGA
T.YAA....      IHF seq

1083
IleLeuAsnGlyGluSerLeuAspLeuAsnSerIleIle***
ATCCTTAATGGTGAGTCTTTGGATTTAACTCTATTATCTAAATTTTAACTCGACCTTTT
tTgaT
-----
Dra I

1143
TGATAAAAAAGTAGTTCCCTGTAGGGCTTAGATCAGGTAGTCTTAGATATGTAAAAAAA
*****      ***      ***      ***      ***      ***

1203
AAACCGCCTCAGCAAAGGCGTTTTCTTAAAGATTGGTTCGTAGAGCAGCAACTCTTTGAA
----->      <-----
potential
rho-independent
terminator

```

Figure 2. Sequence of *pvuII*M. The numbering refers to bases, and begins at the *Hind* III site 5' to the gene. The amino acids corresponding to the single major open reading frame are also shown. Some of the restriction sites predicted by the sequence are indicated; those shown in parentheses have not yet been confirmed by experiment. Functionality of the indicated promoter, Shine-Dalgarno site, and rho-independent transcription terminator has not yet been confirmed. 'IHF seq' refers to the base sequence recognized by the *E. coli* integration host factor, and is capitalized where the *pvuII*M sequence matches the consensus.

sequence of the gene for this N4mC MTase, and find that it bears much stronger relationship to N6mA MTases than to the 5mC enzymes. We also find that this enzyme, though probably monomeric, has an element of symmetry provided by an internal homology, each half of which spans over 100 amino acids.

Figure 3. Protein standards, and extracts of an *E. coli* strain overexpressing *pvuII*M, were resolved on NaDodSO₄ acrylamide gels. Three acrylamide concentrations were used (all 29:1 acrylamide:bisacrylamide): $\circ$, 7.5%; $\bullet$, 11.25%, and $\bullet$, 15%. The closed arrows indicate the position of EF-Tu, and the open arrows show the positions of M-Pvu II. The protein standards used, with M_r in parentheses, included: phosphorylase B (97.4k), serum albumin (66k), ovalbumin (45k), glyceraldehyde-3-phosphate dehydrogenase (36k), carbonic anhydrase (29k), trypsinogen-PMSF (24k), trypsin inhibitor (20.1k), and alpha lactalbumin (14.2k).

MATERIALS AND METHODS

Plasmids and bacterial strains

Escherichia coli strain JM107MA2 is an McrB^- derivative (17) of the *endA1 lacI^q* strain JM107 (19). The plasmid pPvuM1.9, which is a pBR322 derivative, contains the *pvuII*M gene on a 1.9kbp DNA segment inserted at the *Hind* III site (17). Plasmid pPvuM1.91 has the same DNA segment, but in the opposite orientation. IS1 insertion mutants of pPvuM1.9 were isolated following transformation of strain JM107, selecting for Ap^R survivors (17). Genetic nomenclature follows the recommendations of Szybalski, *et al.* (20).

DNA sequence determination

Double-stranded, covalently-closed, supercoiled, CsCl gradient-purified DNA was used as the template (21,22) for sequence determination by the dideoxy method (23). Primers complementary to two positions on pBR322, and to one end of IS1, were purchased from New England Biolabs, Inc. (Beverly, MA). Primers complementary to the other end of IS1 and to sequences within the MTase gene were produced in this institution on a solid phase synthesizer (Applied Biosystems, Foster City CA); these were deblocked and column-purified as recommended by the manufacturer, and used without further purification (24). Sequencing reactions were carried out with [α -³⁵S] dATP (Amersham Corp., Arlington Heights, IL), and made use of kits including either the Klenow fragment of DNA polymerase I (BRL/Life Technologies, Inc., Gaithersburg, MD) or the T7 Sequenase^R

Table 1. Codon usage in the major open reading frame of *pvuII*M.

	U	C	A	G	
U	16 UUU Phe 6 UUC 12 UUA Leu 7 UUG	9 UCU Ser 0 UCC 8 UCA 3 UCG	9 UAU Tyr 3 UAC 1 UAA Ter 0 UAG	1 UGU Cys 0 UGC 0 UGA Ter 7 UGG Trp	U C A G
C	3 CUU Leu 0 CUC 5 CUA 1 CUG	8 CCU Pro 3 CCC 9 CCA 1 CCG	3 CAU His 0 CAC 3 CAA Gln 2 CAG	2 CGU Arg 0 CGC 2 CGA 0 CGG	U C A G
A	8 AUU 5 AUC Ile 9 AUA 12 AUG Met	9 ACU 0 ACC Thr 4 ACA 1 ACG	16 AAU Asn 5 AAC 21 AAA Lys 7 AAG	8 AGU Ser 3 AGC 10 AGA Arg 0 AGG	U C A G
G	11 GUU 1 GUC Val 2 GUA 4 GUG	5 GCU 2 GCC Ala 5 GCA 4 GCG	14 GAU Asp 6 GAC 15 GAA Glu 7 GAG	8 GGU 3 GGC Gly 4 GGA 4 GGG	U C A G

DNA polymerase (USB Corp., Cleveland, OH). Most reactions were resolved by electrophoresis on buffer gradient gels (25), and other procedures were carried out as described in ref. 26. The computer programs used to manage and analyze this data have been described (27).

DNA sequence corroboration

For restriction maps, all enzymes were used under the conditions recommended by the various manufacturers. The enzymes used that had four-base specificity were *Alu* I, *Dde* I, *Hha* I, *Hpa* II, *Rsa* I and *Sau*3A I. Digests were resolved by electrophoresis on 25cm 1.4% agarose gels, or on 12cm 4% NuSieve^R agarose gels (FMC Corp., Rockland, ME), in tris-borate-EDTA buffer. For analysis of proteins, whole-cell extracts were prepared by boiling 2min in sample buffer (80mM tris-HCl pH 6.8, 80mM dithiothreitol, 2% NaDodSO₄, 10% glycerol, bromphenol blue), and resolved on NaDodSO₄ acrylamide gels (28). The gels were stained with Coomassie brilliant blue as described (29). In some cases, photographs of the gels were analysed by scanning densitometry (Shimadzu Corp., Kyoto Japan; model CS-930). Protein size standards were purchased from Sigma Chemical Co. (St. Louis, MO), and apparent protein sizes were calculated as described (30).

Construction and analysis of an overexpressing derivative

The *p_{lac}* promoter (31) was isolated as a 269bp *Bam*H I fragment from plasmid pKK223-3 (32; Pharmacia-LKB Biotechnology Inc., Piscataway, NJ). Restriction endonucleases, T4 DNA ligase (Pharmacia-LKB) and calf intestinal phosphatase (Boehringer-Mannheim GmbH, Indianapolis, IN), were used in accordance with the manufacturer's recommendations. JM107MA2 cells were transformed by the CaCl₂ procedure (33), and the boiling method was used to analyze plasmid DNA from 1ml cultures (34).

Cultures were grown in DYT medium (per liter: 16g Bacto tryptone, 16g yeast extract, 5g NaCl), and induced in mid log-phase growth through the addition of IPTG to a concentration of 1mM. After shaking at 37° for 6h, the cells were pelleted, washed, frozen and thawed, brought up in 0.03 vol of gel filtration buffer (1.M NaCl, 0.01M tris pH

7.5, 0.01M 2-mercaptoethanol), and sonically disrupted. The sonicate was centrifuged for 2h at $27,000\times g$, and a portion of the supernatant was applied to a $90\text{cm}\times 2.5\text{cm}$ (dia.) column of Biogel^R A0.5M (BioRad, Richmond, CA). To assay for methylase activity, $5\text{ }\mu\text{l}$ of a column fraction was incubated with $1\text{ }\mu\text{g}$ of DNA from bacteriophage lambda in 10mM tris-HCl pH 7.5, 50mM NaCl, 10mM EDTA, 5mM 2-mercaptoethanol, and 160mM *S*-adenosyl-L-methionine. After 30min at 37° , the reaction was terminated through the addition of NaDodSO₄ (to 0.1%), and incubated for 10min at 65° . The DNA was centrifuged through a 1 ml column of Sepharose^R CL-6B (Pharmacia), and then digested with *Pvu* II endonuclease in the presence of Mg^{++} . To examine the protein content of the fractions, $20\text{ }\mu\text{l}$ was added to $80\text{ }\mu\text{l}$ of H₂O, and $100\text{ }\mu\text{l}$ of cold acetone was added in turn. After 20min on ice, the samples were centrifuged for 15min at $12,500\times g$. The pellets were brought up in sample buffer and analyzed on acrylamide gels as described above.

RESULTS

Sequence of the pvuII gene

We determined the sequence of a 1.86kbp *Hind* III fragment from plasmid p*Pvu*M1.9, which codes for M-*Pvu* II. The cloning, boundaries, and orientation of the *pvuII* gene have been described previously (17). Figure 1 illustrates the starting positions and lengths of sequence obtained with 18 of the 22 oligonucleotide primers that were used. Each arrow represents the sum of several independent experiments in which sequencing and electrophoresis conditions were varied. We made use of a collection of IS1 insertion mutants in *pvuII*, isolated by introducing p*Pvu*M1.9 into McrB^+ strains of *E. coli* (17), by sequencing these mutants with oligonucleotides complementary to the left and right ends of IS1. The resulting DNA sequence contained a single large open reading frame (ORF), and this region is shown in fig. 2 in original and translated forms. The position of this ORF is consistent with the previously determined boundaries and orientation of the *pvuII* gene.

Corroboration of the DNA sequence data

Restriction maps. Several restriction sites predicted from the sequence are indicated in fig. 2. Most of these have been confirmed. Furthermore, six endonucleases having 4-base specificities were used to digest p*Pvu*M1.9. Fragments from within the pBR322 vector were used as internal standards, and 25/25 fragments from the insert were within the limits of electrophoretic resolution of the size predicted by the sequence (not shown).

Protein size. The translated ORF shown in fig. 2 predicts a protein with a molecular weight of 38,300. This is in good agreement with experimental measurements. The product of *in vitro* transcription and translation was reported as being 34kDa (17), and the protein induced in an overexpressing strain migrates in NaDodSO₄ gel electrophoresis as 36–38kDa depending on the acrylamide concentration (fig. 3, and see below).

Figure 4. Translational potential of the *pvuII* base sequence. The base numbering on the abscissa is the same as that used in fig. 2. A: Open reading frames of at least 15-codons' length are shown. B: Filled portions indicate regions in which reading frame 3 has the highest content of Pu-N-Py codons, of the 3 reading frames. This, and the data shown in panel C, was calculated by a computer program that examined 10-codon windows, with 5-codon overlaps between adjacent windows. C: The percentage of Pu-N-Py (RNY) codons in reading frame 3. The dashed line indicates the expected percentage of such codons in random DNA sequences. Random ORFs would have an even lower content of RNY codons, as RNY codons can specify only eight of the twenty amino acids. The two horizontal bars indicate the regions of an internal sequence homology.

Figure 5. Aligned amino acid sequences from two homologous regions within M·Pvu II and from M·EcoR I. The upper sequence from M·Pvu II (amino) runs from ser¹¹ to phe¹¹³, while the lower sequence (carboxy) runs from ser¹⁸¹ to phe²⁹³. The M·EcoR I sequence underneath is from lys¹⁰⁰ to ile¹⁶², beyond which there is no further significant homology with M·Pvu II. Homology to a portion of the M·Pst I sequence is also shown. Rectangles highlight matches between the M·Pvu II domains (and where all 3 sequences match). Italic letters indicate matches or conserved substitutions between any of the sequences. A horizontal bar underlines the consensus DPPY region of M·EcoR I (see ref. 9).

Functional sequences. The DNA sequence shown in fig. 2 includes regulatory sequences at appropriate locations, though their actual use has yet to be confirmed experimentally. These include a promoter [which is known to be near to the location proposed here (17)], a Shine—Dalgarno sequence and, following the ochre termination codon, a potential rho-independent transcription terminator. In addition, the two observed hotspots for IS1 insertion (fig. 1) correspond well to binding sites for the *E. coli* integration host factor ('IHF seq', fig. 2; 35). IS1 has been shown to transpose preferentially to IHF binding sites (36).

The coding region itself exhibits two expected features. First, the codon choice is heavily

Figure 6. Construction of an inducibly-overexpressing derivative of pPvuM1.91. As described in the Results, this was a two-step process involving first insertion of DNA containing *P_{tac}* and screening for the desired orientation, then deleting DNA between the *Sma* I and *EcoRV* sites. Parentheses indicate that the *Pvu* II sites on all three plasmids are methylated.

Figure 7. Induction of *E. coli* strain JM107MA2 carrying pPvuM1.91-pTac2. At time zero, isopropyl- β -D-galactopyranoside (IPTG) was added to a final concentration of 2mM (culture indicated with closed symbols). A parallel uninduced culture (open symbols) was also examined. A: Percentage of total staining material (protein; measured as A_{545} of a gel photograph; see Materials and Methods). $\square$, $\blacksquare$: EF-Tu. $\circ$, $\bullet$: M·Pvu II. B: Culture density, in units from a Klett-Summerson photometer with a 500–570nm filter. The data shown have been corrected for nonlinear absorbance at the higher culture densities using data provided by E.J. Hansen (pers. commun.).

biased in favor of A or T in the third position (240/337, or 71%; see Tab. 1). This is consistent with the overall base composition of *Proteus vulgaris* DNA, which is 40% GC/60% AT (37). Second, of the three reading frames the proposed coding frame has the highest consistency with the apparent primordial Pu-N-Py (RNY) pattern for coding sequences (fig. 4B; see ref. 38).

Internal homology in the pvuIIM gene

The pvuIIM sequence reveals a substantial internal homology, as shown in fig. 5 (and indicated in fig. 4). The amino-proximal homologous region is comprised of 103 aa, from ser¹¹ to phe¹¹³; while the carboxy-proximal region includes 113 aa, from ser¹⁸¹ through phe²⁹³. Within these regions, 39 aa are perfectly matched, including a run of 8/12 from pro^{89,264} to gly^{99,274} (fig. 5, bottom; '12' includes gaps in the homology). In addition to the 39 perfect matches, there are 11 substitutions that are conservative by the criteria of the PAM250 database (39).

High-level expression of pvuIIM

An overexpressing derivative of pPvuM1.91 was constructed in order to facilitate future structural and functional studies of M·Pvu II. Plasmid pPvuM1.91 is itself a derivative

Figure 8. An induced culture extract, containing the overexpressed product of *pvuIIM*, was resolved by gel filtration and analyzed as described in Materials and Methods. Every second fraction was assayed, and fraction numbers are indicated at the top ('C' is a control lane). **A:** 1% agarose gel of lambda DNA, incubated with fractions under methylating conditions and then challenged with *PvuII* endonuclease. The control lane shows lambda DNA that was methylated with *M·Alu I* as a positive control (*Alu I* methylation protects against *Pvu II* digestion). The bars to the right mark the positions of the uncut DNA (upper), and of the largest band from digested DNA (lower). **B:** 12.5% acrylamide NaDodSO₄ gel of the same fractions. The control lane shows the unfractionated extract. The bars to the right mark the positions of EF-Tu (upper) and of the induced band (lower).

of *pPvuM1.9*, in which *pvuIIM* has been inverted so as to be codirectional with the *amp* gene. This construction is depicted in fig. 6. The strong composite promoter p_{lac} (31) was inserted 5' to *pvuIIM*, and about 200bp of intervening DNA was subsequently deleted. Only after this deletion could *pvuIIM* be overexpressed. The deleted DNA, from the *tet* gene of pBR322, has no obvious transcription terminator sequences. It does include 3 potential stem-and-loop structures having 5bp stems (centered on pBR322 base numbers 216, 280, and 368). The construct resulting from this deletion retains intact the native *pvuIIM* promoter, and 156bp of DNA from pBR322 remains between p_{lac} and the *Hind III* site that precedes *pvuIIM*.

When strain JM107MA2(*pPvuM1.91-p_{lac2}*) [abbreviated MA2(*p_{lac2}*)] was induced with IPTG, growth rapidly ceased and a protein of the expected size accumulated to the point of comprising over 10% of total soluble protein (over 30% of the staining material on a gel from which the smaller ribosomal proteins have been run off) (fig. 7). Furthermore, when an extract of induced cells was resolved via gel filtration, the methylase activity appeared in the same fractions as the induced protein (fig. 8).

To confirm the identity of the IPTG-induced band, and to add further support to the sequence data, a deletion of *p_{lac2}* was generated by cleavage with endonuclease *Spe I*. This should create an 81bp (in-frame) deletion within the MTase gene (see fig. 2), and lead to the IPTG induction of a protein approximately 3kDa smaller. When strain MA2(*p_{lac2}-delSpe*) is induced with IPTG, its growth ceases with kinetics similar to that of MA2(*p_{lac2}*) (not shown). However, the induced band does, as expected, correspond to a smaller protein (fig. 9). This deletion also eliminates any *M·Pvu II* activity, whether or not IPTG is present (not shown).

DISCUSSION

We have presented data regarding the base sequence and overexpression of the gene for an N4mC-generating DNA MTase, *pvuIIM*. The sequence of this gene has three features of particular interest.

Figure 9. NaDodSO₄ gel analysis of four culture extracts. A '+' refers to a culture that had grown for 240min in the presence of 2mM IPTG. A '-' indicates a parallel, uninduced culture. All cultures used *E. coli* strain JM107MA2 carrying one of 2 plasmids: p*Pvu*M1.91-p*lac*2, or its derivative from which an *Spe* I fragment had been deleted. The whole-cell extracts were loaded onto parallel lanes for electrophoresis with molecular weight markers. The stained gel was photographed, and the negative subjected to scanning densitometry. The top and bottom scans (both uninduced cultures) show essentially identical profiles. The vertical lines indicate the positions of EF-Tu, M-*Pvu* II, and the protein induced from the *Spe* I deletion of *pvu*II. Some additional, minor differences are apparent in the scans, and are presumed to represent adjustments by the cell to such a disproportionate synthesis of one protein.

Comparison to the genes of 5mC-generating DNA MTases

It might be supposed that proteins catalyzing the *S*-adenosylmethionine-dependent methylation of cytosine in dsDNA, though doing so at different positions, would be related to one another fairly closely. The *pvuII*M sequence provides no support for that supposition. Several groups have reported the occurrence of conserved sequences in 5mC MTases (*e.g.*, 11–14). None of these conserved sequences are present in *pvuII*M. For example, all 5mC MTases sequenced to date, even that from mouse cells (40), contain a pro-cys pair. The cys of this pair is thought to form a covalent bond to the 6 carbon of cytosine during DNA methylation (15). The *pvuII*M sequence includes a single cys, and it is not adjacent to pro (fig. 2).

Comparison to the genes of N6mA-generating DNA MTases

In contrast to the lack of homology between *pvuII*M and the genes for 5mC MTases, this N4mC MTase gene has substantial homology to N6mA MTase genes. This has been illustrated in the case of *ecoRIM* (fig. 5). The homology is between asn¹⁰² and ile¹⁶² of *ecoRIM* (41,42), and roughly the 5' half of the internal homology of *pvuII*M. Of the 61 aa this includes, 26 perfectly match counterparts in at least one of the two *pvuII*M sequences, and another 16 represent conservative substitutions. The putative DPPY elements of all 3 sequences are in alignment with one another.

N4mC- and N6mA-generating DNA MTases may be functionally related, and perhaps evolutionarily related as well. This is suggested by more than the lack of homology between *pvuII*M and the 5mC MTase genes, together with the significant homology between *pvuII*M and N6mA MTase genes. Unlike the generation of 5mC, which proceeds via a covalent enzyme-DNA intermediate (15), N6mA is generated by a direct methylation of the N6 amino group on adenine (43). The N4 amino group of cytosine is chemically very similar to the adenine N6 amino group (*e.g.*, see fig. 3 of ref. 44), and could well be methylated by the same mechanism. Furthermore, in duplex DNA they occupy very similar positions in the major groove (see fig. 1 of ref. 45). One prediction of this work is that it may be possible for an N4mC MTase to generate N6mA (and *vice versa*) following mutation or even just under solvent and salt conditions that reduce sequence specificity ('star' conditions). It is worth noting that under star conditions, *M·EcoR* I and *M·BamH* I were seen to extensively methylate DNA outside the cognate sites (46,57), but no determination of the methylated base(s) generated was made.

Sequences of more N4mC MTase genes are needed, as are three-dimensional structures of N4mC and N6mA MTases. Nevertheless, the sequence of *pvuII*M strongly suggests that N4mC MTases are much more closely related to N6mA MTases than to 5mC MTases. This conclusion is supported by the sequence of the N4mC MTase *BamH* I (J. Brooks and D. Landry, personal communication). It is interesting that the 8/12 sequence, which is the most highly-homologous portion of the *M·Pvu* II internal homology, is represented in one other DNA MTase. That is the N6mA MTase *M·Pst* I, which recognizes a sequence similar to the *Pvu* II substrate (5' CTGCAG 3'; 38,39; see fig. 5), and is found in a related bacterial species to that producing *Pvu* II.

*Internal homology in the pvuII*M gene

The degree of internal homology found in *pvuII*M is striking. The homologous regions together comprise 206 of the 336 amino acids specified by the ORF (over 60%). Of the 206, 100 are matches or conservative substitutions. Significant internal homology does not appear to be widespread among MTases, but has been found in two other MTase genes (both of which generate N6mA): *M·PaeR7* I, and *M·Fok* I (50,51).

The internal homology may provide these MTases with a degree of structural symmetry, allowing symmetrical interaction with their DNA substrates. Where it has been examined, many (though not all) DNA MTases have been found to bind and modify DNA as monomers, with *M·EcoR* I being the best-studied example (51,52). In contrast many DNA-binding proteins, including repressors and restriction endonucleases, bind and act as symmetrical multimers (dimers or tetramers; 53–56). This symmetry is a reflection of the symmetry in the DNA sequences they act on. *M·Pvu* II recognizes a symmetrical DNA sequence, though it has not yet been shown to act as a monomer.

It is also possible that either or both occurrences of the internal homology represent catalytically-active domains. It is not yet known whether *M·Pvu* II modifies both strands in a single binding event like *M·BamH* I (57) or, like *M·EcoR* I, modifies just one strand per binding event (51). As *M·BamH* I lacks significant internal homology (J. Brooks, personal communication), it is also not clear how single- and double-methylating MTases differ structurally. We have derivatives of *pvuIIM* independently altered in each homologous domain, though these derivatives involve substantial deletions (one is the *Spe* I deletion referred to above, the others are IS1 insertions into the other of the two domains). All of these derivatives are completely inactive.

A third, overlapping possibility is that the internal homology allows the binding of two molecules of *S*-adenosylmethionine, conferring some kinetic or regulatory advantages on the MTase. It is particularly interesting that both internally-homologous regions of *M·Pvu* II include versions of the DPPY consensus sequence, as has also been found for *M·PaeR7* I and *M·Fok* I (49,50). This consensus sequence has been suggested to be a pocket for the binding of *S*-adenosylmethionine (8–10). In *M·Pvu* II, the amino-proximal occurrence of this sequence has phe in place of tyr, as is seen in *M·EcoR* I (41,42). Both occurrences in *M·Pvu* II have ser in place of asp, which has not been found previously. The role of these sequences in *M·Pvu* II, as in other MTases, has yet to be defined.

ACKNOWLEDGEMENTS

We would like to thank J. Anderson for advice, encouragement, and hospitality during construction of the overexpressing derivative. We also thank, for advice, critiques, and shared data, J. Benner, A. Bhagwat, J. Brooks, S. Cheatham, S. Hattman, J. Heitman, J. Pflugrath, J. Posfai, R. Roberts, B. Slatko, and G. Wilson.

This work was supported by the National Science Foundation under grant #DMB-8409652, and by a Biomedical Research Support grant from the Medical College of Ohio. J.W. was supported by funds from the Burroughs-Wellcome Fund, through the undergraduate research program at Cold Spring Harbor Laboratory. K.J.B. was supported in part by a Summer Medical Research Fellowship from the Medical College of Ohio.

*To whom correspondence should be addressed

Present addresses: [†]Department of Chemistry, University of California, Berkeley, CA 94720, [§]Department of Medicine, Thomas Jefferson School of Medicine, Philadelphia, PA 19107 and [¶]Bureau of Biologics, Food and Drug Administration, Bethesda, MD 20892, USA

REFERENCES

1. Dunn, D.B., and Smith, J.D. (1958) *Biochem J.* **68**, 627–636.
2. Doskocil, J., and Sormova, Z. (1965) *Biochim. Biophys. Acta* **95**, 513–515.
3. Janulaitis, A., Klimasauskas, S., Petrusyte, M., and Butkus, V. (1983) *FEBS Lett.* **161**, 131–134.

4. Janulaitis, A.A., Stakenas, P.S., Pyatrushite, M.P., Bitinaite, Y.B., Klimasauskas, S.I., and Butkus, V.V. (1984) *Mol. Biol. (Engl. Transl. Mol. Biol. [Moscow])* **18**, 93–105.
5. Ehrlich, M., Gama-Sosa, M., Carreira, L., Ljungdahl, L., Kuo, K., and Gehrke, C. (1985) *Nucl. Acids Res.* **13**, 1399–1412.
6. Ehrlich, M., Wilson, G.G., Kuo, K.C., and Gehrke, C.W. (1987) *J. Bacteriol.* **169**, 939–943.
7. Ehrlich, M., Norris, K.F., Wang, R.Y.-H., Kuo, K.C., and Gehrke, C.W. (1986) *Bioscience Reports* **6**, 387–392.
8. Hattman, S., Wilkinson, J., Swinton, D., Schlagman, S., MacDonald, P.M., and Mosig, G. (1985) *J. Bacteriol.* **164**, 932–937.
9. Lauster, R., Kriebardis, A., and Guschlbauer, W. (1987) *FEBS Lett.* **220**, 167–176.
10. Chandrasegaran, S., and Smith, H.O. (1987) in *Structure and expression 1: from proteins to ribosomes*, Sarma, R.H., and Sarma, M.H., ed., pp. 149–156, Adenine Press, Schenectady, NY.
11. Som, S., Bhagwat, A.S., and Friedman, S. (1987) *Nucl. Acids Res.* **15**, 313–332.
12. Sznyter, L.A., Slatko, B., Moran, L., O'Donnell, K., and Brooks, J.E. (1987) *Nucl. Acids Res.* **15**, 8249–8266.
13. Wilke, K., Rauhut, E., Noyer-Weidner, M., Lauster, R., Pawlek, B., Behrens, B., and Trautner, T.A. (1988) *EMBO J.* **7**, 2601–2609.
14. Posfai, J., Bhagwat, A.S., and Roberts, R.J. (1988) *Gene* **74**, 261–265.
15. Wu, J.C., and Santi, D.V. (1987) *J. Biol. Chem.* **262**, 4778–4786.
16. Gingeras, T.R., Greenough, L., Schildkraut, I., and Roberts, R.J. (1981) *Nucl. Acids Res.* **9**, 4525–4536.
17. Blumenthal, R.M., Gregory, S.A., and Cooperider, J.S. (1985) *J. Bacteriol.* **164**, 501–509.
18. Butkus, V., Klimasauskas, S., Petrauskienė, L., Maneliene, Z., Lebionka, A., and Janulaitis, A. (1987) *Biochim. Biophys. Acta* **909**, 201–207.
19. Yanisch-Perron, C., Vieira, J., and Messing, J. (1985) *Gene* **33**, 103–119.
20. Szybalski, W., Blumenthal, R.M., Brooks, J.E., Hattman, S., and Raleigh, E.A. (1988) *Gene* **74**, 279–280.
21. Hattori, M., and Sakaki, Y. (1986) *Anal. Biochem.* **152**, 232–238.
22. Bartlett, J.A., Gaillard, R.K., Jr., and Joklik, W.K. (1986) *BioTechniques* **4**, 208–210.
23. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
24. Sanchez-Pescador, R., and Urdea, M.S. (1984) *DNA* **3**, 339–343.
25. Biggin, M.D., Gibson, T.J., and Hong, G.F. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3963–3965.
26. Williams, S.A., Latson, B.E., Moran, L.S., and DeSimone, S.M. (1986) *BioTechniques* **4**, 138–147.
27. Blumenthal, R.M., Rice, P., and Roberts, R.J. (1982) *Nucl. Acids Res.* **10**, 91–101.
28. Laemmli, U.K. (1970) *Nature* **227**, 680–685.
29. Blumenthal, R.M., Reeh, S., and Pedersen, S. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 2285–2288.
30. Weber, K., and Osborn, M. (1969) *J. Biol. Chem.* **244**, 4406–4412.
31. DeBoer, H.A., Comstock, L.J., and Vasser, M. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 21–25.
32. Amann, E., and Brosius, J. (1985) *Gene* **40**, 183–190.
33. Mandel, M., and Higa, A. (1970) *J. Mol. Biol.* **53**, 159–162.
34. Holmes, D.S., and Quigley, M. (1981) *Anal. Biochem.* **114**, 193–197.
35. Miller, H.I., and Nash, H.A. (1981) *Nature* **290**, 523–526.
36. Gamas, P., Chandler, M., Prentki, P., and Galas, D. (1987) *J. Mol. Biol.* **195**, 261–273.
37. Sueoka, N. (1964) in *The Bacteria*, Gunsalus, I.C., and Stanier, R.Y., ed., vol. **5**, pp. 422–443, Academic Press, New York.
38. Shepherd, J.C.W. (1982) *Cold Spring Harbor Symp. Quant. Biol.* **47**, 1099–1108.
39. Dayhoff, M., Schwartz, R.M., and Orcutt, B.C. (1978) in *Atlas of Protein Sequence and Structure*, Dayhoff, M., ed., vol. **5**, pp. 345–352, Natl. Biomed. Res. Found., Silver Spring, MD.
40. Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. (1988) *J. Mol. Biol.* **203**, 971–983.
41. Greene, P.J., Gupta, M., Boyer, H.W., Brown, W.E., and Rosenberg, J.M. (1981) *J. Biol. Chem.* **256**, 2143–2153.
42. Newman, A.K., Rubin, R.A., Kim, S., and Modrich, P. (1981) *J. Biol. Chem.* **256**, 2131–2139.
43. Pogliotti, A.L., Jr., Ono, A., Subramaniam, R., and Santi, D.V. (1988) *J. Biol. Chem.* **263**, 7461–7464.
44. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S., Jr., and Weiner, P. (1984) *J. Amer. Chem. Soc.* **106**, 765–784.
45. Seeman, N.C., Rosenberg, J.M., and Rich, A. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 804–808.
46. Woodbury, C.P., Jr., and von Hippel, P.H. (1981) in *Gene Amplification and Analysis*, Chirikjian, J.G., ed., pp. 181–207, Elsevier-North Holland, New York.
47. Smith, D.I., Blattner, F.R., and Davies, J. (1976) *Nucl. Acids Res.* **3**, 343–353.

48. Walder, R.Y., Walder, J.A., and Donelson, J.E. (1984) *J. Biol. Chem.* **259**, 8015–8026.
49. 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**, 8441–8460.
50. Looney, M.C., Moran, L.S., Jack, W.E., Feehery, G.R., Benner, J.S., Slatko, B.E., and Wilson, G.G. (1989) *Gene*, *in press*.
51. Rubin, R.A., and Modrich, P. (1977) *J. Biol. Chem.* **252**, 7265–7272.
52. Smith, H.O., and Kelly, S.V. (1984) in *DNA Methylation: Biochemistry and Biological Significance*, Razin, A., Cedar, H., and Riggs, A.D., ed., pp. 39–71, Springer-Verlag, New York.
53. Modrich, P., and Roberts, R.J. (1982) in *Nucleases*, Linn, S.M., and Roberts, R.J., ed., pp. 109–154, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
54. Beyreuther, K. (1980) in *The Operon*, Miller, J.H., and Reznikoff, W.S., ed., pp. 123–154, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
55. Gussin, G.N., Johnson, A.D., Pabo, C.O., and Sauer, R.T. (1983) in *Lambda II*, Hendrix, R.W., Roberts, J.W., Stahl, F.W., and Weisberg, R.A., ed., pp. 93–121, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
56. Tjian, R. (1981) *Cell* **26**, 1–2.
57. Nardone, G., George, J., and Chirikjian, J.G. (1984) *J. Biol. Chem.* **259**, 10357–10362.

<p>This article, submitted on disc, has been automatically converted into this typeset format by the publisher.</p>
--