Cloning and characterization of the *Hpa*II methylase gene

Charles O.Card, Geoffrey G.Wilson, Karin Weule[1], Joseph Hasapes[1], Antal Kiss[+] and Richard J.Roberts[1,*]

New England Biolabs, Inc., 32 Tozer Road, Beverly, MA 01915 and [1] Cold Spring Harbor Laboratory, PO Box 100, Cold Spring Harbor, NY 11724, USA

Received January 23, 1990, Accepted February 13, 1990 EMBL accession no. X51322

ABSTRACT

The *Hpa*II restriction-modification system from *Haemophilus parainfluenzae* recognizes the DNA sequence CCGG. The gene for the *Hpa*II methylase has been cloned into *E. coli* and its nucleotide sequence has been determined. The DNA of the clones is fully protected against cleavage by the *Hpa*II restriction enzyme *in vitro*, indicating that the methylase gene is active in *E. coli*. The clones were isolated in an McrA$^-$ strain of *E. coli*; attempts to isolate them in an McrA$^+$ strain were unsuccessful. The clones do not express detectable *Hpa*II restriction endonuclease activity, suggesting that either the endonuclease gene is not expressed well in *E. coli*, or that it is not present in its entirety in any of the clones that we have isolated. The derived amino acid sequence of the *Hpa*II methylase shows overall similarity to other cytosine methylases. It bears a particularly close resemblance to the sequences of the *Hha*I, *Bsu*FI and *Msp*I methylases. When compared with three other methylases that recognize CCGG, the variable region of the *Hpa*II methylase, which is believed to be reponsible for sequence specific recognition, shows some similarity to the corresponding regions of the *Bsu*FI and *Msp*I methylases, but is rather dissimilar to that of the SPR methylase.

INTRODUCTION

Haemophilus parainfluenzae possesses two Type II restriction-modification systems, designated *Hpa*I and *Hpa*II (1). The enzymes of the *Hpa*II system recognize the symmetric double-stranded DNA sequence 5'-CCGG-3' (2). The *Hpa*II endonuclease cleaves the sequence between the two cytosines (2) and the *Hpa*II methylase modifies the inner cytosine to form 5-methylcytosine (3) We report here the cloning, into *E. coli*, of the gene for the *Hpa*II methylase and the determination of its nucleotide sequence. The predicted sequence of the protein is compared with other methylases that form 5-methylcytosine within DNA (m^5C-methylases). The genes for a number of m^5C-methylases have been cloned and sequenced (cited in 4). The derived amino acid sequences of the enzymes show extensive, and often very close, similarities (5–7). It is of especial interest to compare the sequence of the *Hpa*II methylase with the sequences of three other m^5C-methylases, *Msp*I (8), *Bsu*FI (9) and SPR (10,11), that also recognize the sequence 5'-CCGG-3'.

MATERIALS AND METHODS

Bacterial strains, plasmids and phages

Wild type *Haemophilus parainfluenzae*, and *Escherichia coli* K-12 strains, K802 ($HsdR_K^-$ $HsdM_K^+$, McrA$^-$ McrB$^-$) and RR1 ($HsdR_B^-$, $HsdM_B^-$, McrA$^+$ McrB$^-$) were from the New England Biolabs strain collection. They were originally provided by Drs. Jane Setlow, Helen Revel, and Raymond Rodriguez, respectively. *E. coli* K802 and RR1 are available from the American Type Culture Collection, catalog numbers 33526 and 31343, respectively

Enzymes and chemicals

Restriction enzymes, the Klenow fragment of *E. coli* DNA polymerase I, exonuclease III, S_1 nuclease, T4 DNA ligase and cleaved, dephosphorylated pBR322 were obtained from New England Biolabs, and were used according to the manufacturer's instructions. ^{35}S-α-dATP (>1000 Ci/mmole) and ^{32}P-α-dATP (>2000 Ci/mmole) were purchased from New England Nuclear. All other chemicals were of reagent grade quality.

DNA preparation

Total cellular DNA from 10 gm of frozen *H. parainfluenzae* cells was purified using lysozyme/osmotic-shock/detergent lysis, phenol/chloroform extraction, RNase digestion and isopropanol precipitation, by the method previously described (12).

Plasmid DNAs were extracted from cells by lysozyme/Triton X-100 lysis, and were purified by CsCl/Ethidium bromide ultracentrifugation (13). Plasmid mini-preps were prepared by the alkaline-SDS procedure (14)

Transformation

E. coli strains K802 and RR1 were made competent by the $CaCl_2$/low-temperature procedure (15,16). Transformation mixtures (0.25 μg DNA in 12.5 μl; 100 μl 67 mM $CaCl_2$, 5 mM

* To whom correspondence should be addressed

+ Present address Institute of Biochemistry, Biological Research Center of the Hungarian Academy of Sciences, POB 521, 6701 Szeged, Hungary

Na_3citrate, 50 mM NaCl; 200 μl competent cells, 0°C) were warmed to 42°C for 3 min, then diluted into Luria broth (LB) and incubated at 37°C for three hr. The cultures were spread onto Luria agar (LA) plates containing 50 μg Ampicillin (Ap)/ml or 25 μg Tetracycline (Tc)/ml, and then incubated overnight at 37°C to select for transformants.

Preparation of genomic DNA libraries

Purified *H. parainfluenzae* DNA was diluted to approximately 100 μg/ml in 10 mM Tris.HCl pH 7.5, 10 mM $MgCl_2$, 10 mM mercaptoethanol, 100 mM NaCl. In separate reactions, 100 μl aliquots of the DNA were digested at 37°C for 1 hr with varying amounts (1 unit enzyme/μg DNA to 0.001 units/μg) of the restriction endonucleases *Bam*HI, *Bcl*I, *Bgl*II, *Eco*RI, *Hin*dIII and *Pst*I. Reactions were terminated by heating to 75°C for 15 min, then the digests were examined by 1% agarose gel electrophoresis. Tubes in which substantial, but incomplete, digestion had occurred were selected for ligation: 8 μg (80 μl) of each partially-digested DNA preparation was mixed with 4 μg of cleaved, dephosphorylated, pBR322, and then ligated with T4 DNA ligase at 17°C for 4 hr in 150 μl reaction volumes

The ligations were extracted with chloroform, then transformed into competent *E. coli* K802 or RR1 (8 μg total DNA in 100 μl, added to 1 ml 67 mM $CaCl_2$, 5 mM Na_3Citrate, 50 mM NaCl, and 2 ml competent cells). Transformed mixtures were cultured in LB, then spread onto LB plates containing Ap or Tc. Following overnight growth, the plates were each flooded with 2.5 ml of 10 mM Tris.HCl pH 7.5, 10 mM $MgCl_2$, and the transformants from each ligation were scraped together to form cell libraries. 1 ml of each cell library was inoculated into 500 ml LB containing Ap or Tc. The cultures were grown to saturation, the cells were collected by centrifugation, and the plasmid populations that they carried were purified by CsCl/ethidium bromide ultracentrifugation, to create plasmid libraries.

Selection of methylase clones

Each plasmid library was diluted to approximately 30 μg/ml in *Hpa*II digestion buffer (10 mM Tris HCl pH 7.5, 10 mM $MgCl_2$, 10 mM mercaptoethanol, 10 mM KCl). A 100 μl aliquot of each was digested to completion at 37°C for 1 hr with 30 units of *Hpa*II The digests were transformed into *E. coli* K802 or RR1, cultured in LB for several hr, then plated onto LA containing Ap or Tc. Surviving colonies were collected and the plasmids that they carried were individually purified by mini-prep procedure and analyzed.

Endonuclease assays of crude cell extracts

50 ml overnight cultures were centrifuged at 4,000 rpm for 10 min. The cell pellets were resuspended in 2.5 ml of 10 mM Tris.HCl pH 7.5, 1 mM Na_2EDTA, 10 mM mercaptoethanol, 1 mg lysozyme/ml, and left on ice for 2 hr. The suspensions were frozen at −20°C overnight, then thawed on ice, and lysed by mixing with an equal vol of 10 mM Tris.HCl pH 7.5, 1 mM Na_2EDTA, 10 mM mercaptoethanol and 0.01% Triton X-100. 1.5 ml of the lysed suspensions were micro-centrifuged for 10 min. at 4°C. The supernatants were withdrawn and then assayed by serial dilution into 100 μl aliquots of *Hpa*II digestion buffer containing 50 μg phage λ DNA/ml. Each assay series contained from 1 to 0.001 ml of supernatant/μg λ DNA. The tubes were incubated at 37°C for 1 hr, then electrophoresed through 1% agarose gels containing 0.5 μg EtdBr/ml The gels were illuminated with UV light and examined for the occurrence of the characteristic pattern of *Hpa*II digestion of λ DNA.

DNA sequence determination

The 2-Kb *Hin*dIII fragment carrying the *Hpa*II methylase gene was recloned from pCC*HpaIIM*2-1 into M13mp19 in both orientations. RF DNA, isolated from the two clones, was digested with *Xba*I or *Sac*I and unidirectional deletions were prepared by the combined action of exonuclease III, S_1-nuclease, the Klenow fragment of DNA polymerase I and DNA ligase (17). Single-stranded phage DNA was isolated from the shortened clones and sequenced using the chain termination procedure (18,19). Usually a synthetic primer (TCCCAGTCACGACGT) was used that is complementary to the M13 sequence immediately adjacent to the site of insertion. Thin sequencing gels (20) were used throughout and contained either 5, 6, or 8% polyacrylamide Both strands were sequenced.

Computer Analysis

Computer analysis was performed on a DEC VAX11/750 and a SUN Microsystems 3/60. Primary data was stored and overlaps were established using the programs ASSEMBLER (21), M13 and SEQ (22). Analysis was carried out using additional programs described elsewhere (23−28). Homology searches used the PIR database version 19, the GenBank database version 59 and the EMBL data library version 18.

RESULTS

Isolation of *Hpa*II methylase clones

Plasmid clones carrying the gene for the *Hpa*II methylase were isolated by selecting for recombinants that had protectively modified themselves against digestion by *Hpa*II. Details of this procedure have been discussed previously (29). In brief, *H. parainfluenzae* DNA was purified, digested with a number of restriction enzymes, and the resulting fragments were ligated to appropriately cleaved preparations of the plasmid vector, pBR322. The ligation mixtures were transformed into *E. coli* K802; the transformants from each ligation were pooled, and the plasmid populations that they contained were purified. The mixed plasmid populations were then digested with *Hpa*II to destroy unmethylated molecules, and the digests were transformed back into *E. coli* K802 to recover survivors. The surviving transformants were individually picked and the plasmids that they contained were separately purified by a mini-preparation procedure, and examined by restriction enzyme digestion and agarose gel electrophoresis.

*Hpa*II-methylated plasmids were recognized by their resistance to digestion by *Hpa*II, and their sensitivity to digestion by *Msp*I *Msp*I also recognizes the sequence CCGG but cleaves whether or not it has been methylated by the *Hpa*II methylase (30). Several recombinant plasmids meeting these criteria were identified among the survivors from the libraries prepared with *H. parainfluenzae* DNA that had been digested individually with either *Eco*RI, *Hin*dIII or *Pst*I. They were not found among the survivors from the libraries prepared with *Bam*HI, *Bcl*I, or *Bgl*II.

The *Hpa*II methylase clones from the *Hin*dIII-library carried a single, 2-Kb *Hin*dIII fragment in common; they displayed complete resistance to *Hpa*II-digestion, indicating that they were fully methylated. The clones from the *Eco*RI-library carried a common 6-Kb *Eco*RI fragment; those from the *Pst*I-library, carried a common 10-Kb *Pst*I fragment. The *Eco*RI-, and *Pst*I-clones displayed incomplete resistance to *Hpa*II-digestion, indicating that they were only partially methylated. Restriction mapping established that the fragments were nested; the *Hin*dIII

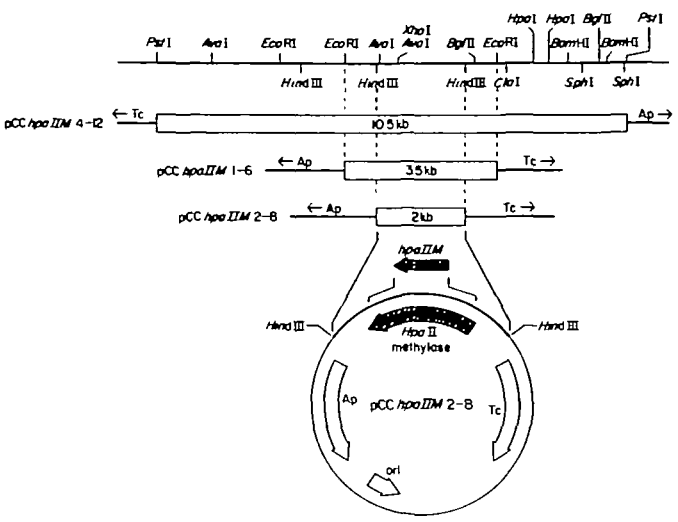

Figure 1. Schematic diagram of plasmid pCC*hpaII*M2-1, which contains a 2-Kb insert of *H parainfluenzae* DNA in the vector pBR322 and which expresses the methylase gene. Also shown are schematics for two larger plasmids containing the flanking sequences of the *H parainfluenzae* genome

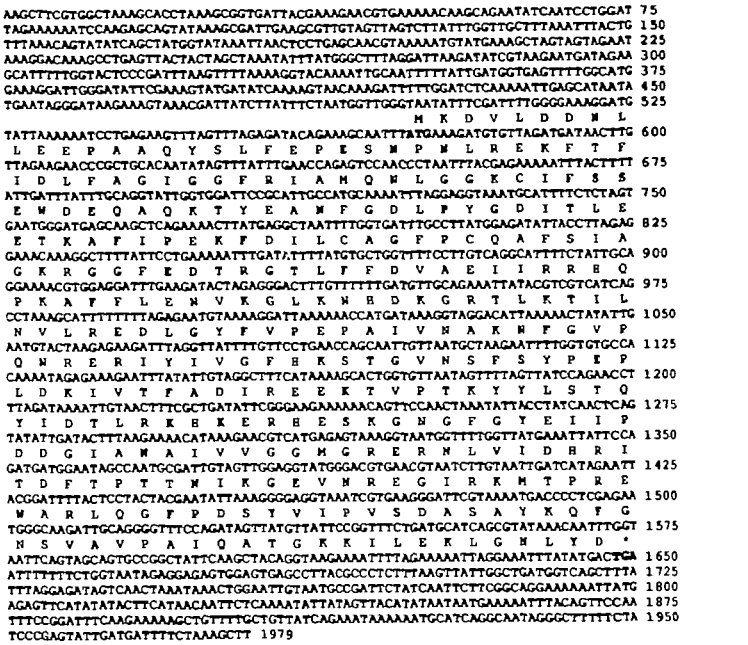

Figure 2. The sequence of the 2-Kb *Hin*dIII fragment of pGW1 containing the *Hpa*II methylase gene The translation for that gene is shown above the DNA sequence beginning at the first AUG codon within the reading frame.

fragment is located in the middle of the *Eco*RI fragment which is, itself, located in the middle of the *Pst*I fragment (Fig. 1).

Where is the *Hpa*II endonuclease gene?

The methylase selection procedure that we have used to isolate the *Hpa*II methylase gene frequently yields large fragments of DNA that encode both methylase and endonuclease genes (4). In all cases where both genes have been cloned they have been found to lie adjacent to one another (4). Among the *Hpa*II methylase clones that we isolated, the 10-Kb *Pst*I-fragment extends approximately 4-Kb beyond each end of the *Hin*dIII-fragment that contains the methylase gene. Since 4-Kb is ample to code for both a large restriction endonuclease (average length: 280 aa) and a large intergenic region we examined these clones for production of the endonuclease. Cell extracts were prepared from the *Hin*dIII-, *Eco*RI- and *Pst*I-clones, and each extract was assayed for *Hpa*II endonuclease activity. No activity was detected. This suggests that either the gene for the endonuclease is not present or its expression in *E. coli* is too low to be detected.

Another possibility is that the endonuclease gene is present in the *Pst*I-clones, but is defective. Some restriction-modification systems cannot be transferred to *E. coli* in a single step, probably because the DNA of the recipient cell becomes irreparably cleaved before the methylase is able to protect it. Such systems, which include *Dde*I (31) and *Bam*HI (32) can only be cloned in two steps, even though both genes of the system lie in close proximity and could be retrieved on a single restriction fragment. In those cases, the methylase gene must be cloned first so that the resulting clones provide a permissive host for the endonuclease gene.

To test if the *Hpa*II system behaved similarly, the *Pst*I fragment was recloned *de novo*, into cells that already contained the *Hpa*II methylase gene. The 2-Kb *Hin*dIII fragment, encoding the *Hpa*II methylase gene, was recloned into pACYC184, and the new plasmid was transformed into *E. coli* K802. DNA from these cells was prepared and shown to be resistant to *Hpa*II digestion. *H. parainfluenzae* DNA was digested with *Pst*I, and a DNA fraction that included fragments in the 10-Kb size range was gel-purified. The fraction was ligated to pBR322 and transformed into the *Hpa*II-premethylated cells. Clones carrying the 10-Kb *Pst*I fragment were detected by colony hybridization to a nick-translated probe made from the unique right end of the *Pst*I fragment. Several new *Pst*I-clones were isolated and were shown to contain plasmids with the expected structures; however, when extracts of these clones were analyzed, no *Hpa*II endonuclease activity was detected.

Effects of Mcr function on *Hpa*II methylase plasmids

The *Hpa*II methylase clones were isolated from libraries prepared in *E. coli* K802. When similar libraries were prepared in *E. coli* RR1, as they were during our first attempts to clone the gene, no *Hpa*II methylase clones were recovered. The difference, it was later discovered, was due to the *E. coli* McrA function, which specifically restricts *Hpa*II methylated DNA (33,34). The McrA gene functions normally in RR1, but it is defective in K802 (35). pBR322 DNA that is methylated *in vitro* with *Hpa*II transforms K802 at normal efficiency, but it transforms RR1 at only 2% of normal efficiency. The *Hpa*II methylase clones were also found to transform K802 normally, but to transform RR1 at approximately 5% of the normal efficiency (data not shown) Transformants of RR1 that do arise and contain the *Hpa*II methylase show an unhealthy colony morphology: they form flat, translucent colonies, as opposed to the dome-shaped, opaque colonies that RR1 normally forms. Examination of the plasmids carried by fourteen independent *Hpa*II-transformants of RR1 revealed that the plasmids were all normal, as judged by restriction enzyme analysis and in the degree of *Hpa*II-modification that they displayed.

Sequence of the *Hpa*II methylase gene

The smallest *Hpa*II methylase plasmid clone displays complete resistance to *Hpa*II endonuclease digestion, suggesting that it possesses a fully functional *Hpa*II gene. The nucleotide sequence of the 2-Kb *Hin*dIII fragment inserted in this plasmid was

determined on both strands and is shown in Figure 2. The sequence contains only one internal open reading frame long enough to encode the *Hpa*II methylase. The first methionine codon in this reading frame lies at position 574 and the frame continues until a TGA terminator at position 1648. This terminator is followed by three more in-frame translational stops, TAA, TAG and TGA, five, six and eleven triplets further downstream. This open reading frame predicts a protein containing 358 amino acids with a calculated molecular weight of 40,406 daltons. Previous studies of the *Hpa*II methylase have detected proteins of approximately 38,500 and 41,500 daltons which showed methylase activity (36). The exact relationship between these protein species was not established, although it was suggested that the higher molecular weight form might represent a precursor species (36). The higher molecular weight would correlate well with it being the product of the 358 amino acid open reading frame

Sequence comparison with other modification enzymes

Sequences have been reported for thirty four methylase genes including fifteen that form N^6-methyladenine, one that forms N^4-methylcytosine and eighteen that form 5-methylcytosine (referenced in 5–7, 37 and this paper). The FASTA program (26) was used to compare the *Hpa*II methylase with each of these methylase sequences as well as the complete contents of the PIR, GenBank and EMBL databases. The major similarities detected were with other m^{5}C-methylases (Table 1) It can be seen that the *Hpa*II methylase shows the greatest similarity to the *Hha*I and *Bsu*FI methylases and an alignment between these three sequences is shown in Figure 3. Overall the sequence of the *Hpa*II methylase displays the same pattern of ten conserved sequence similarities, that have been found in all known m^{5}C-methylases (6,7) (Figure 3).

The interval between conserved regions VIII and IX in m^{5}C-methylases forms the so-called variable region. The length and sequence of this region varies considerably from enzyme to enzyme. In the *Bacillus* phage methylases, which recognize multiple target sequences, the variable region has been shown to be responsible for sequence specificity (49–51). Based upon the overall common architecture of the prokaryotic m^{5}C-methylases, it is likely that this variable region is also responsible for sequence specificity in the case of the monospecific methylases. Four m^{5}C-methylases that all recognize the sequence GGCC have been found to possess very similar variable regions (putative recognition domains); these are *Hae*III (41), *Ngo*PII (42), *Bsp*RI (44) and *Bsu*RI (47) It was, therefore, of great interest to compare in detail the variable regions of the *Hpa*II, *Msp*I, *Bsu*FI and SPR methylases, all of which recognize the sequence CCGG.

To define the boundaries between the conserved regions and the variable region we aligned conserved blocks VIII and IX, as defined in (7), for all eighteen m^{5}C-methylase sequences The result is shown in Figure 4. It can be seen that for the first nine amino acids beyond the wholly conserved element Q-R-R in block VIII the sequences are fairly highly conserved whereas at the tenth amino acid several non-conservative amino acid changes are seen and the sequences show signs of divergency. This formed the N-terminal boundary of the variable region for the purposes of our analysis. The C-terminal region was clearly delimited by the dramatic change that immediately preceded the fully conserved arginine residue of block IX (Figure 4). In the case of the SPR methylase the region initially chosen for detailed comparison was that part of the variable domain that had previously been implicated in recognition of the sequence CCGG (39,49,50), together with the unassigned flanking sequences. The

TABLE 1. Similarity between the *Hpa*II and other m^{5}C-methylases

Methylase	Recognition Sequence	Score	Reference
*Hha*I	G$\overset{*}{C}$GC	611	12
*Bsu*FI	$\overset{*}{C}$CGG	573	9
*Eco*RII	C$\overset{*}{C}$WGG	508	37
dcm	C$\overset{*}{C}$WGG	506	38 and A Bhagwat unpublished
*Msp*I	$\overset{*}{C}$CGG	412	8
SPR	GG$\overset{*}{C}$C, $\overset{*}{C}$CGG, C$\overset{*}{C}$WGG	389	10,11
φ11	GGCC, GAGCTC	380	39
φ3	GG$\overset{*}{C}$C, G$\overset{*}{C}$NGC	334	40
*Hae*III	GG$\overset{*}{C}$C	299	41
*Ngo*PII	GGCC	293	42
*Aqu*I	CYCGRG	281	43
*Bsp*RI	GG$\overset{*}{C}$C	245	44
*Bep*I	$\overset{*}{C}$GCG	237	45
*Sin*I	GGW$\overset{*}{C}$C	226	46
*Bsu*RI	GG$\overset{*}{C}$C	222	47
*Dde*I	$\overset{*}{C}$TNAG	203	31
*Sss*I	$\overset{*}{C}$G	89	48

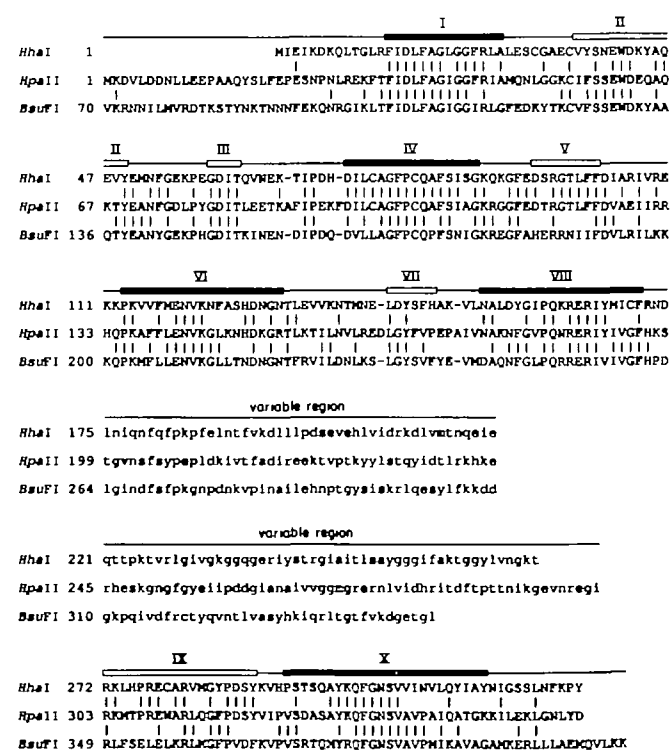

Figure 3. Alignment between the *Hpa*II, *Hha*I and *Bsu*FI methylase protein sequences The initial alignments were obtained using the program GENALIGN Further refinement was carried out manually. The first 69 residues of the *Bsu*FI sequence are omitted, since they have no counterpart in either of the other two sequences No attempt was made to align the variable region sequences, since these are shown in detail for the *Hpa*II and *Bsu*FI methylases in Figure 5.

total length of the SPR variable region is 210 amino acids of which at most 41 (residues 293–334) are potentially involved in the specific recognition of CCGG (49). The other variable regions have lengths of 103 for *Hpa*II, 84 for *Bsu*FI and 81 for the *Msp*I methylases. The SPR sequence is quite different from the sequences of the other three variable regions and no satisfactory alignment could be achieved. An alignment between the variable region sequences for the mono-specific methylases is shown in Figure 5.

It can be seen that the *Msp*I and *Bsu*FI methylases show an unusually high degree of similarity as noted previously (9). A common tripeptide DDG was found in a similar position in the three mono-specific methylase recognition domains and was used as an internal anchor point for the alignment shown in Figure 5. Some conservation of sequence is apparent between the *Hpa*II methylase and the *Msp*I and *Bsu*FI methylases in this region.

Throughout the length of the variable region the *Msp*I and *Bsu*FI show 50% identity. Within sequences lying C-terminal to the conserved DDG element, the level of identity increases to 70% (26/37) In this same C-terminal segment the *Hpa*II methylase can be aligned with 25% identity by introducing two gaps of a single amino acid. Within sequences lying N-terminal to the DDG element some similarity can be detected between the *Hpa*II, *Bsu*FI and *Msp*I methylases. Overall the *Hpa*II methylase shows greater similarity to the *Bsu*FI methylase than to the *Msp*I methylase.

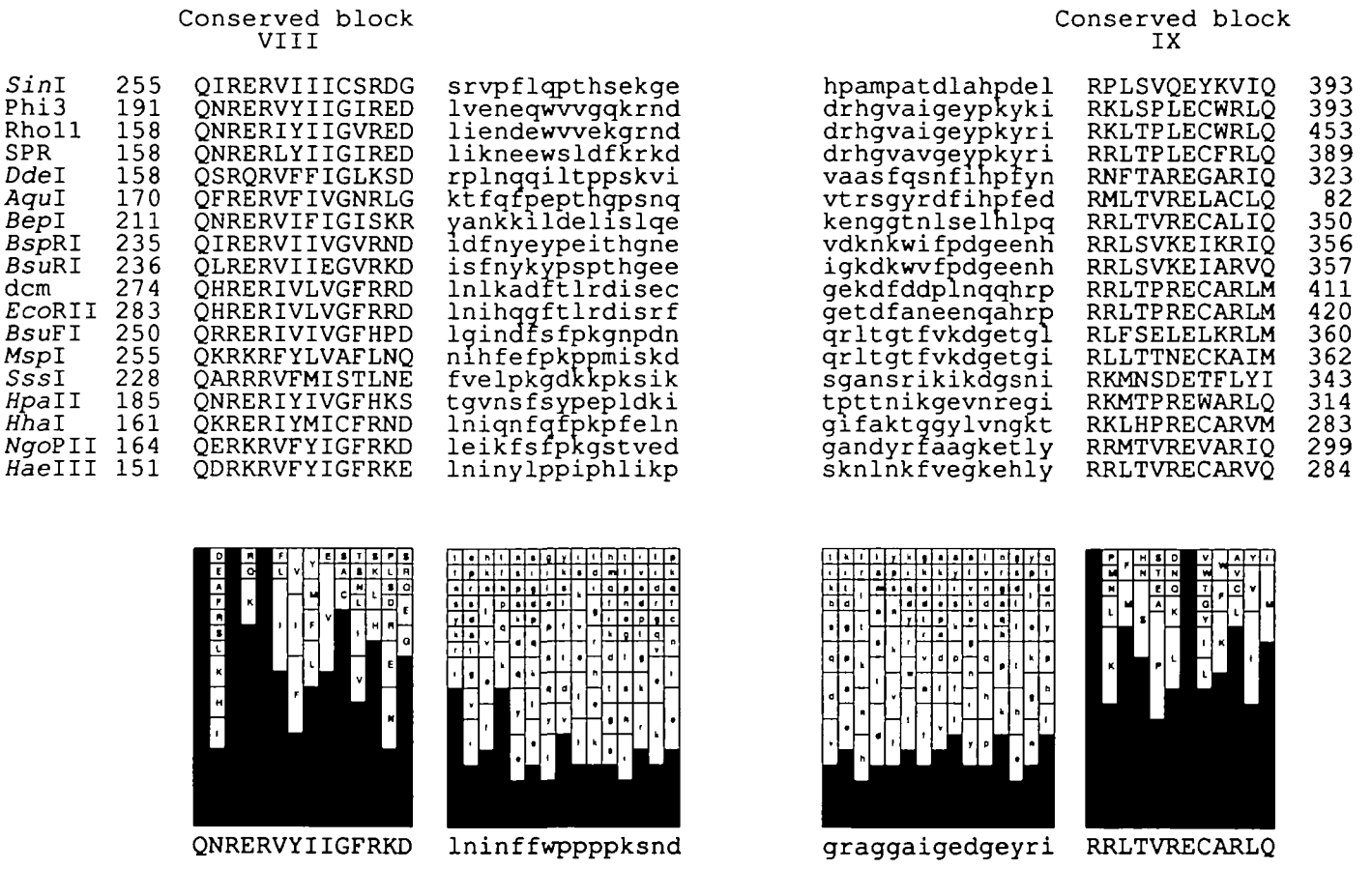

Figure 4. The boundaries of the variable region. All eighteen m^5C-methylase sequences were aligned on common blocks VIII and IX as defined in (7) The variations in amino acid occupancy at positions within and adjacent to these common blocks is indicated below the alignment The sequence written at the bottom of the figure shows the amino acids occupying the lower, black boxes in the figure

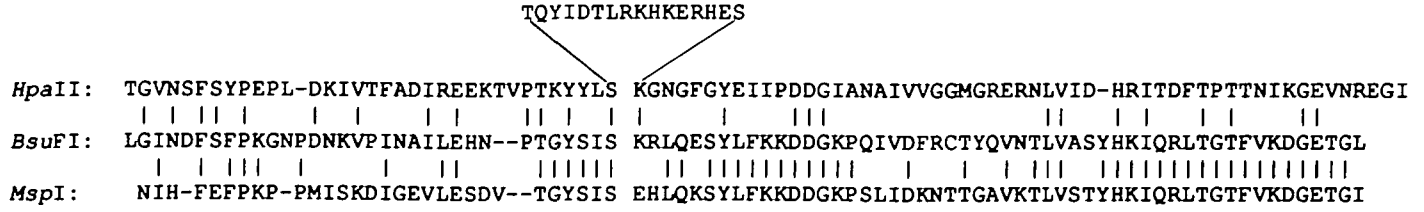

Figure 5. Alignment between the variable regions of the *Hpa*II (residues 199–302), *Bsu*FI (residues 261–348) and *Msp*I (residues 269–350) methylases. The initial alignment between the *Bsu*FI and *Msp*I methylases was produced using the program GENALIGN and the two other sequences were added manually using a program provided by Dr G Otto to identify short stretches of similar sequences that aided the subsequent alignments. The sixteen amino acids shown above the *Hpa*II sequence are considered to be an insert in that sequence, which has no counterpart in the other two sequences

DISCUSSION

Like many other attempts to clone the genes for restriction-modification systems we have been able only to clone the methylase gene for the *Hpa*II system. We have isolated several clones containing sequences flanking the methylase gene, but no *Hpa*II endonuclease activity has been detected in these clones. Among nine other *Haemophilus* systems where cloning has been successful the genes for both the methylase and the restriction endonuclease have been located adjacent to one another in eight instances (4 and GGW unpublished). Only in the cases of *Hae*III and *Hpa*II have the endonucleases not been detected within clones containing sequences flanking the methylase gene (4). In neither of these two cases nor in the present case is it known whether this is because the two genes are physically separated on the genome or because the endonuclease genes are not expressed in *E coli*. Unfortunately comparison of flanking sequences with the known sequences of restriction endonuclease genes is not helpful in trying to identify an unexpressed gene. So far each restriction endonuclease gene appears to have a unique sequence and no diagnostic similarities have been detected. Unlike the situation with the *Bsu*FI methylase gene, which appears to be allelic with other *Bacillus subtilis* methylase genes (9,54), the *Haemophilus* methylase genes share no obvious sequence similarities across species.

The sequences immediately upstream of the *Hpa*II methylase gene are quite AT rich, but carry no clear similarity to typical *E. coli* promoter sequences Since a clone carrying the 2-Kb *Hin*dIII fragment shows a higher level of expression of the methylase than clones carrying longer inserts it seems likely that the endogenous promoter is not very active in *E. coli*. Rather it is likely that the increased levels of methylase expression that we have observed in the smaller clones is being driven by a pB322 promoter. This notion is supported by the observation that recloning of the *Hpa*II methylase gene into pACYC leads to full *in vivo* modification when the gene is oriented so that its expression is away from that of the tetracycline resistance gene, but only partial modification in the opposite orientation

A striking feature of the coding region is the extreme bias in codon usage. Among 359 codons, 300 have A or T in the third position and only 18 have a C in the third position. Three *Hpa*II recognition sites are found within the sequence reported; two lie within the coding region close to its C-terminus, while the third lies 230 nucleotides beyond the end of the gene. Given their location it is unlikely that these have any regulatory significance as has been proposed in other systems (52,53).

One of the most interesting aspects of the *Hpa*II methylase sequence is its comparison with the sequences of the three other known methylases that recognize the sequence CCGG. These are the *Msp*I, *Bsu*FI and SPR methylases. It should be noted that all three of these enzymes methylate the outer cytosine of the recognition sequence whereas the *Hpa*II methylase modifies the internal cytosine. In terms of overall similarity the sequences share the typical building blocks found among all known m^5C-methylases (6,7). However it is the comparison of the so-called variable regions, that is believed to be responsible for sequence recognition, that is of most interest. Among these four enzymes the *Msp*I and *Bsu*FI methylases show the highest degree of similarity to each other in this region. The *Hpa*II methylase is most similar to the *Bsu*FI methylase, and shows the least similarity to the SPR methylase. Given the overall similarity of all m^5C methylases and the clear relatedness of this family it is likely that they have all evolved from a common precursor. The extensive changes found in the SPR methylase may reflect the fact that its recognition domain is part of a complicated region that includes recognition domains for two other sequences.

When searching for common sub-sequences within the variable region we were struck by the apparent conservation of the tripeptide DDG, which we used as an anchor point when aligning the three mono-specific methylases. This is the only tripeptide found in these variable regions and is positioned similarly, with respect to conserved block IX, in all three sequences. DGG also occurs within the SPR methylase, but within a part of the variable region that has been implicated in the recognition of the sequence GGCC. It does not occur in the variable region of any other sequenced m^5C-methylase. It is interesting to note that two of the three residues are aspartic acid, which would be capable of hydrogen bonding with the bases of the DNA recognition sequence.

ACKNOWLEDGEMENTS

The authors would like to express their thanks to Drs A. Bhagwat for many helpful discussions during the course of this work and for allowing us to use the dcm methylase sequence prior to publication. Thanks are due to Dr. T. Trautner for sharing the sequence of the *Bsu*FI methylase gene prior to publication and to Dr. G. Otto for help with some of the computational aspects. We thank J. Duffy and P Renna for the artwork and photography Part of this work was supported by a grant from the National Science Foundation to RJR (DMB-8614032).

REFERENCES

1 Sharp, P A , Sugden, B and Sambrook, J (1973) Biochemistry 12 3055–3066
2 Garfin, D E and Goodman, H M (1974) Biochem Biophys Res Comm 259 108–116
3 Mann, M B and Smith, H O (1977) Nucl Acids Res 4 4211–4221
4 Wilson, G G (1988) Gene 74 281–289
5 Chandrasegaran, S and Smith, H O (1988) in Structure and Expression Volume 1 From Proteins to Ribosomes eds R H Sarma and M H Sarma (Adenine Press) pp 149–156
6 Lauster, R , Trautner, T A. and Noyer-Weidner, M (1989) J Mol Biol 206 305–312
7 Posfai, J , Bhagwat, A S , Posfai, G and Roberts, R J (1989) Nucl Acids Res 17 2421–2435
8 Lin, P M , Lee, C H and Roberts, R J (1989) Nucl Acids Res 17 3001–3011
9 Walter, J , Noyer-Weidner, M. and Trautner, T A submitted for publication
10 Posfai, G , Baldauf, F , Erdei, S , Posfai, J , Venetianer, P and Kiss, A (1984) Nucl Acids Res 12 9039–9049
11 Buhk, H-J , Behrens, B , Tailor, R , Wilke, K , Prada, J.J., Gunthert, U., Noyer-Weidner, M , Jentsch, S and Trautner, T.A (1984) Gene 29 51–61
12 Caserta, M , Zacharias, W , Nwankwo, D , Wilson, G G. and Wells, R D (1987) J Biol. Chem 262 4770–4777
13 Clewell, D B and Helinski, D R (1969) Proc Natl Acad. Sci USA 62 1159–1166
14 Birnboim, H C and Doly, J (1979) Nucl Acids Res 7 1513–1523
15 Mandel, M and Higa, A (1970) J Mol Biol 53 159–162
16. Lederberg E M. and Cohen, S.N (1974) J. Bacteriol 119 1072–1074.
17 Henikoff, S. (1984) Gene 28· 351–358
18 Sanger, F , Nicklen, S and Coulson, A R (1977) Proc Natl. Acad Sci USA 74 5463–5467
19 Messing, J (1983) Meth Enzymol 101 20–78
20 Sanger, F and Coulson, A R (1978) FEBS Letters 87 107–110
21 Gingeras, T R , Milazzo, J.P , Sciaky, D and Roberts, R J (1979) Nucl Acids Res 7 529–545
22 Blumenthal, R M , Rice, P J and Roberts, R.J. (1982) Nucl Acids Res 10 91–101
23 Staden, R (1977) Nucl. Acids Res 4 4037–4051
24 Staden R (1978) Nucl Acids Res. 5 1013–1015

25 Keller, C , Corcoran, M and Roberts, R J (1984) Nucl Acids Res 12 379–386
26 Lipman, D J and Pearson, W R (1988) Proc Natl Acad Sci USA 85: 2444–2448
27 Feng, D-F and Doolittle, R F (1987) J. Mol Evol 25: 351–360.
28. Devereux, J , Haeberli, P and Smithies, O (1984) Nucl Acids Res 12 387–395
29 Lunnen, K D , Barsomian, J M , Camp, R R , Card, C O , Chen, S-Z , Croft, R , Looney, M C , Meda, M M , Moran, L S , Nwankwo, D.O , Slatko, B E , Van Cott, E M and Wilson, G G (1988) Gene 74 25–32
30 Brooks, J E and Roberts, R J (1982) Nucl Acids Res 10. 913–934
31 Howard, K A , Card, C , Benner, J S , Callahan, H L , Maunus, R , Silber, K , Wilson, G and Brooks, J E (1986) Nucl Acids Res 14 7939–7951
32 Brooks, J E , Benner, J S , Heiter, D , Silber, K R , Sznyter, L , Jager-Quinton, T , Moran, L , Slatko, B E , Wilson, G G and Nwankwo, D O (1989) Nucl Acids Res 17 979–997
33 Noyer-Weidner, M , Diaz, R and Reiners, L (1986) Mol Gen Genet 205. 469–475
34 Raleigh, E A and Wilson, G (1986) Proc Natl Acad Sci USA 83 9070–9074
35 Raleigh, E A , Murray, N E , Revel, H , Blumenthal, R M , Westaway, D , Reith, A D , Rigby, P W J , Elhai, J and Hanahan, D (1988) Nucl Acids Res 16 1563–1575
36 Yoo, O J and Agarwal, K L (1980) J Biol Chem 255 6445–6449
37 Som, S , Bhagwat, A S and Friedman, S (1987) Nucl Acids Res 15 313–332
38 Hanck, T , Gerwin, N , Fritz, H J (1989) Nucl Acids Res 17 5844
39 Behrens, B , Noyer-Weidner, M , Pawlek, B , Lauster, R , Balganesh, T S and Trautner, T A (1987) EMBO J 6 1137–1142
40 Tran-Betcke, A , Behrens, B , Noyer-Weidner, M and Trautner, T A (1986) Gene 42 89–96
41 Slatko, B E , Croft, R , Moran, L and Wilson, G G (1988) Gene 74 45–50
42 Sullivan, K M and Saunders, J R (1988) Nucl Acids Res 16 4369–4387
43 Karreman, C and De Waard, A (1990) J Bacteriol 172 266–272
44 Posfai, G , Kiss, A , Erdei, S , Posfai, J and Venetianer, P (1983) J Mol Biol 170 597–610
45 Kupper, D , Zhou, J G , Venetianer, P and Kiss, A (1989) Nucl Acids Res 17 1077–1088
46 Karreman, C and de Waard, A (1988) J Bacteriol 170 2527–2532
47 Kiss, A , Posfai, G , Keller, C C , Venetianer, P and Roberts, R J. (1985) Nucl Acids Res 13 6403–6421
48 Renbaum, P , Abrahamove, D , Fainsod, A , Wilson, G G , Rottem, S and Razin, A submitted for publication
49 Wilke, K , Rauhut, E , Noyer-Weidner, M , Lauster, R , Pawlek, B , Behrens, B , and Trautner, T A (1988) EMBO J 7 2601–2609
50 Balganesh, T S , Reiners, L , Lauster, R , Noyer-Weidner, M , Wilke, K and Trautner, T A (1987) EMBO J 6 3542–3549
51 Trautner, T A , Balganesh, T S and Pawlek, B (1988) Nucl Acids Res 16 6649–6658
52 Slatko, B E , Benner, J S , Jager-Quinton, T , Moran, L S , Simcox, T G , Van Cott, E M and Wilson, G G (1987) Nucl Acids Res 15 9781–9796
53 Gingeras, T R , Theriault, G and Brooks, J E (1984) in Proceedings of the 5th International Symposium on Metabolism and Enzymology of Nucleic Acids (eds J Zelinka and J Balan) Slovak Academy of Sciences, Bratislava pp 267–275
54 Ikawa, S , Shibata, T , Matsumoto, K , Iijima, T , Saito, H and Ando, T (1981) Mol Gen Genet 183 1–6