Nucleic Acids Research, Vol. 19, No. 22 ***6183–6190***

The sequence specificity domain of cytosine-C5 methylases

Saulius Klimašauskas[+], Janise L.Nelson[§] and Richard J.Roberts[*]
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

Received August 29, 1991; Revised and Accepted October 15, 1991

ABSTRACT

Prokaryotic DNA[cytosine-C5]methyltransferases (m^5C-methylases) share a common architectural arrangement of ten conserved sequence motifs. A series of eleven hybrids have been constructed between the *Hpa*II (recognition sequence: Cm^5CGG) and *Hha*I (recognition sequence: Gm^5CGC) DNA-methylases. The hybrids were over-expressed in *E.coli* and their *in vivo* methylation phenotypes investigated. Six were inactive by our assay while five of them retained partial methylation activity and full specificity. In all five cases the specificity matched that of the parent methylase which contributed the so-called variable region, located between conserved motifs VIII and IX. This was the only sequence held in common between the active hybrids and for the first time provides unequivocal evidence that the specificity determinants of the mono-specific m^5C-methylases are located within the variable region. Correlation of the hybrid methylase structure with the efficiency of methylation suggests that conserved motif IX may interact with the variable region whereas motif X most probably interacts with the N-terminal half of the molecule.

INTRODUCTION

Prokaryotic DNA-methylases are an attractive class of enzymes in which to study site specific protein-DNA interactions. They recognize short sequences on DNA with a specificity that matches that of the better-known restriction enzymes, which they usually accompany. Little is known about the manner in which they interact with DNA. DNA methylases lack sequence homology with their companion restriction enzymes. Nor do they resemble other proteins which interact specifically with DNA, suggesting that the methylases probably recognize their specific targets by some new mechanism. There are three major classes of DNA-methylases that differ in the nature of the modification introduced: N^6-methyladenine, N^4-methylcytosine or 5-methylcytosine (m^5C). The first two classes of enzymes, which methylate exocyclic amino groups of adenine or cytosine bases, share two types of conserved domains and can be grouped into several subclasses (1, 2).

Two families of m^5C-methylases are known. The first contains mono-specific methylases that recognize and modify a single DNA recognition sequence. This family occurs predominantly as the modification partner of restriction-modification systems, although a few such methylases have no counterpart restriction enzyme (for reviews see 2–4). The second family contains multi-specific methylases that each recognize and methylate several different DNA sequences. This group has few members, which so far are limited to enzymes encoded by *Bacillus* bacteriophages. Both families comprise the most structurally uniform group of DNA-methylases and show an overall common architecture (5–7). Ten conserved sequence motifs, 10–20 amino acids long, separated by diverse spans are found in all m^5C-methylases so far sequenced (7).

Of special interest is a large variable region that lies between motifs VIII and IX. In the mono-specific family of m^5C-methylases this variable region is 80–120 amino acids long. In the multi-specific family the variable region is much longer and contains between 200 and 300 amino acids. While all m^5C-methylases show overall similarity because of the conserved motifs, whenever similarities are found between the variable regions of the mono-specific family it is in pairs of methylases that recognize identical or related DNA sequences (7–9). This led to the hypothesis that this region is responsible for DNA sequence recognition. This hypothesis is bolstered by results obtained from studies of the multi-specific class of methylases. For instance, one multi-specific enzyme, SPR, recognizes the sequences GGCC, CCGG and CCWGG. Mutants have been isolated that are defective in their ability to recognize one of these three sequences, but are still able to methylate the other sequences (10). These mutations map to the variable region. Similar experiments are not possible for the mono-specific methylases. Several hybrids have been constructed between multi-specific methylases allowing the methylation specificities of the hybrid

* To whom correspondence should be addressed

+ Permanent address: Institute of Applied Enzymology FERMENTAS, 232028 Vilnius, Lithuania

§ Present address: Abbott Laboratories, Dept 93D, AP9A, Abbott Park, N. Chicago, IL 60064, USA

to be correlated with the presence of certain sections of the variable region (11, 12).

The three multi-specific enzymes, Phi3T, SPR and Rho11s, each include GGCC as one of their recognition specificities. However, their variable regions show limited similarity to the variable regions of the mono-specific enzymes *Bsp*RI, *Bsu*RI, *Hae*III or *Ngo*PII (5, 7) which also recognize GGCC. Moreover, the variable regions of the mono-specific methylases *Msp*I and *Bsu*FI (recognition sequence CCGG) as well as *Eco*RII and dcm (recognition sequence CCWGG) do not resemble any segment within the variable region of the SPR methylase (8), which possesses the capacity to modify both of these targets. The low sequence similarity between the variable regions of these two families of m^5C-methylases may reflect substantial differences in structural organization of the specificity domains as well as the sequence recognition mechanisms.

The main goal of our work is to identify the region responsible for DNA sequence recognition in the mono-specific DNA-methylases. We have devised experiments to test whether the variable region is responsible for DNA sequence specificity. We have constructed a series of hybrids between the mono-specific methylases, M.*Hpa*II (recognition sequence: Cm^5CGG) and M.*Hha*I (recognition sequence: Gm^5CGC), and tested the specificity of these hybrid enzymes. These methylases have relatively high overall amino-acid sequence similarity (13) except in their variable regions. Furthermore the position of the base to be methylated is similar in both recognition sequences. M.*Hha*I has been cloned and sequenced (14), purified to apparent homogeneity and characterized both biochemically and kinetically (15, 16). Some biochemical data are also available on the *Hpa*II methylase (17) which has recently been cloned and sequenced (13).

MATERIALS AND METHODS

Materials

E.coli strains ER1648, Δ(*mcrBC-hsdRMS-mrr*)2::Tn*10*, *mcr-A1272*::Tn*10*, and ER1727 which is ER1648 with F′ *lac proAB lac*I^q Δ*(lacZ)M15* (18) were kindly provided by E.Raleigh. Plasmids pCC*hpaII*M2−1 and pNW2081 were a kind gift of G.G.Wilson and the plasmid pUHE25-2 was donated by U.Deuschle. All restriction endonucleases and T4-ligase were obtained from New England BioLabs, AmpliTaq polymerase was from Perkin-Elmer Cetus, T4-DNA ligase from BRL, Sequenase 2.0 DNA sequencing kit from USB, Calf Intestinal Phosphatase and Nuclease P1 from Boehringer Mannheim, labeled chemicals from NEN and Amersham, 2′-deoxy-5′-mononucleotide standards from Sigma, PEI-F cellulose plates from J.T.Baker, QIAGEN columns from Qiagen. All primers used for PCR and DNA sequencing were synthesized in the Cold Spring Harbor Oligonucleotide Core Facility, except #6 which was purchased from New England BioLabs (mismatches with templates are underlined):

#1:GCCTACAATATAAATTCTTTC, #2:TTTCGCAATGATCTCAATATTC,
#3:AATCCCTTCACGATTTACCTCC, #4:CGGAAATTACACCCTAGAGAG,
#5:CGATGCATGCGAGATGTGTTA, #6:AGAATTCTCATGTTTGACAGCTTATCATCG,
#7:TTTTGAGCATGCTATTTCTTT, #8:TAGAGTCGACCTGCAGGC,
#9:TTGGTGTGCCACAAA, #10:CTTAGATTCAAATTGTGAGCGG,
#11:GATAGTTATAAAGTCCACCCGTCTGATGCATCAGCGTATAA,
#12:TATTTAGTAAACGGGAAGACACGTAAAATGACCCCTCGAGAA,
#13:AACACCAGTGCTTTTATGAAAACAAATCATATAGATACG,
#14:GATAGTTATGTTATTCCGGTTTCAACCAGCCAAGCATATAA,
#15:AGATTCGCATGCTTGAAATAA, #16:ATCAACAGGAGTCCAAGCTCAG,
#17:TCCTAAATTTTGCATGGCAAT, #18:GTTTCCTCTAAGGTAATATCTC,
#19:AACCATGATAAAGGTAGGAC, #20:ATCAATATACTGAGTTGATAGG,
#21:GCCAATAACTTAAAGAGGGCG, #22:ATGATTGAAATAAAAGATAAAC,
#23:GCTGAGTGCGTTTATTCTAAT, #24:GAAAACCATTCCTGATCACG,
#25:ACAGTTCGACTTGGTATTGTA, #26:AACTATCTGGGTAGC.

Construction of the hybrids

Splicing through PCR deletion. First, plasmid pJ505, which contains the *Hpa*II methylase gene closely followed by the *Hha*I methylase gene in a head to tail orientation, was constructed by subcloning the *hpaIIM* and *hhaIM* genes from pCC*hpaII*M2-1 (13) and pNW2081 (14), respectively, into the pUC19 vector. The hybrids **P1** and **P2** were the constructed by a PCR deletion scheme. Two oligonucleotides (#1, #2 or #3, #4, respectively) were then used to prime an inverted PCR reaction (see Figure 1) (19). The control plasmids encoding each wild type methylase (**H0** and **P0**) were also constructed using this approach by specifically deleting the fragment encoding the other methylase.

Transfer to the pUHE25-2 vector. To clone methylase genes onto pUHE25-2 vector an *Sph*I endonuclease site was created by PCR mutagenesis. One of the primers in each pair matched the sequence downstream of the *Hin*dIII site of the corresponding gene in pJ505 while the other was designed to create the *Sph*I site at the ATG start codon of the methylase (Figure 2). pJ505 was amplified in the presence of the corresponding pair of primers (#5, #8 or #15, #6), followed by a double *Sph*I-*Hin*dIII cut and subsequent ligation in the presence of the precut and dephosphorylated vector. Transformation of competent ER1727 cells resulted in recombinant clones, **P0** and **H0**, carrying each methylase gene in the desired vector. The constructs **P1** and **P2** were transferred into pUHE25-2 by subcloning the *Nsi*I-*Hin*dIII fragment containing part of the hybrid methylase gene to replace the corresponding fragment from pHSP0-1.

Fusion by overlap extension. Five hybrids, **P4**, **P5**, **P6**, **H2** and **H3**, were made by a PCR technique, which fuses two fragments that have an overlap of twenty or more nucleotides (20). In Reaction 1, a fragment of the desired gene was prepared by amplifying a natural template with two primers one of which contained the specific priming sequence from one gene together with 20 nucleotides at its 5′-end which matched the specific site adjacent to the fusion point in the second gene (see Table 1 for primers and templates). This gel-purified fragment was mixed

Table 1. Construction of the hybrid methylases by PCR fusion through overlap extension. For each PCR, templates and primers are shown except for the SOE step (Reaction 3) where the fragments generated in previous reactions were used as templates.

Hybrid	Reaction 1	Reaction 2	Reaction 3	Final Plasmid
P4	pJ505 #11/#8	pJSP1-2 #5/#6	#5/#8	pHSP4-7
P5	pJ505 #14/#6	pJ505 #5/#8	#5/#6	pHSP5-4
P6	pJ505 #11/#8	pHSP2-1*	#10/#8	pHSP6-1
H2	pJ505 #12/#8	pJ505 #15/#7	#15/#8	pHSH2-2
H3	pJ505 #15/#13	pJSP2-6 #9/#6	#15/#6	pHSH3-1

* no Reaction 2. the plasmid was used directly in Reaction 3.

with a DNA template, usually prepared by standard PCR (Reaction 2 – see Table 1), coding for the remaining part of the construct and subjected to PCR primed by two oligomers that aneal to the terminal regions of each DNA fragment (Reaction 3 – see Table 1). This results in a hybrid in which the fusion is defined by the intermediate primer and the termini by the two distal oligonucleotides. Finally, the PCR products were deproteinized and cut with *Sph*I and *Hin*dIII endonucleases. In some cases *Dpn*I was also included to destroy all template DNAs (the templates were all obtained from *dam*$^+$ strains of *E. coli*). After gel purification, the fragments were ligated into the precut vector plasmid.

Swapping through restriction endonuclease sites. To complete the desired set of chimeric methylases we took advantage of some restriction endonuclease sites present within the gene sequences. The hybrids **H1** and **H4** were constructed by replacing the *Pfl*MI-*Hin*dIII fragment in the construct **H3** with those from **P3** and **P5**, respectively. Similarly, the small *Bbs*I fragment of **P4** was swapped with that in **H2** to make the hybrid **P3**, and the small *Xho*I fragment in **P5** was replaced by the analogous fragment from the construct **H2** to yield the hybrid **H6**. Sequencing revealed some amplification errors, which would lead to amino acid sequence alterations, in the hybrids **P1** and **P4**. The defective regions were localised in the *Sph*I-*Nsi*I fragment of **P1** and the *Sph*I-*Mun*I fragment of **P4**. These were replaced by the analogous wild-type fragments from the hybrid **P5**.

Clone selection and analysis

Following PCR amplification and cloning, typically 10 – 20 colonies were chosen for analysis of the insert length between the *Sph*I and *Hin*dIII sites. This was performed by an *in situ* PCR-amplification of unlysed cells (21) primed at the outer boundaries of the *Sph*I-*Hin*dIII region (primers #10 and #16) and subsequent gel electrophoresis of the resulting fragments. The selected colonies were grown to late logarithmic phase in LB medium containing ampicillin (160 mg/l) and the synthesis of the hybrid protein was induced by incubating the culture for 2 – 6 hours in the presence of 0.4 – 1 mM IPTG. To test the expression of the full length hybrid protein, a 0.2 ml aliquot of each culture was analysed as a crude extract by SDS-PAGE (22). The selected plasmid DNA mini-preparations were examined by a set of diagnostic restriction endonucleases and one or two good candidates were selected for sequencing. Sequencing of the coding regions was performed by dideoxy-termination method using Sequenase 2.0 DNA Sequencing kit (USB), α-[^{35}S]-deoxyadenosine-5′-triphosphothioate (300Ci/mmol) and a set of specific primers (#1 – #4, #9, #10, #16 – #26). Plasmid DNAs for sequencing were prepared from uninduced 100 ml cultures by standard alkaline procedure and purified by precipitation with PEG (22).

In vivo methylation at the *Hpa*II and *Hha*I sites (second position of the sequences CCGG and GCGC) was examined by digestion of the plasmid DNA (0.2 – 1 μg) with an excess of R.*Hpa*II, R.*Msp*I or R.*Hha*I (2 – 16h, 5 – 20u). The resulting fragments were resolved by agarose gel electrophoresis in the presence of ethidium bromide (22). Experiments were performed with at least two independent DNA preparations. Direct analysis of the methylation status at CCGG sites was performed using a modification of the scheme of Cedar *et al.* (23). To remove RNA the plasmids were purified on QIAGEN mini-columns according to the manufacturer's recomendations followed by fragmentation with an excess of R.*Msp*I and dephosphorylation with alkaline phosphatase (CIP). The resulting DNA, after deproteinization, was labeled at its 5′-ends using T4-polynucleotide kinase and γ-[^{32}P]-ATP, desalted on a Sephadex G-50 spun-column, and digested to mononucleotides with nuclease P1. Treatment with 50mM $NaIO_4$ was used to remove traces of 5′-ribonucleotide contaminants (24) and the reaction mixture, containing unlabeled 5-methylated and standard 2′-deoxycytidine-5′-monophosphates, was chromatographed on PEI-cellulose TLC plates in one or two dimensions (23). ^{32}P-labeled compounds were detected by autoradiography while unlabeled standards were localized by inspection under UV-illumination.

RESULTS

Hybrid construction by deletion

Hybrids between the *Hpa*II and *Hha*I methylases, that might retain function, were constructed so that the junctions would lie at the motifs that are conserved among all m^{5}C-methylases. To facilitate this construction and to allow flexibility in the selection of junction points, we first constructed a plasmid, pJ505, that contained the *Hpa*II methylase immediately upstream of the *Hha*I methylase in a head to tail orientation (Figure 1). We then prepared exact deletions at the appropriate point in the two methylase genes, adapting a PCR method originally devised for site-specific mutagenesis (19). In the first hybrid, **P1**, the junction was chosen so that the resulting hybrid contained the N-terminal sequence from the *Hpa*II methylase gene up to and including motif VIII, which immediately precedes the start of the variable region. This was then fused to the variable region and remaining C-terminal sequence from the *Hha*I methylase gene. A second hybrid, **P2**, had the junction positioned at motif IX, immediately

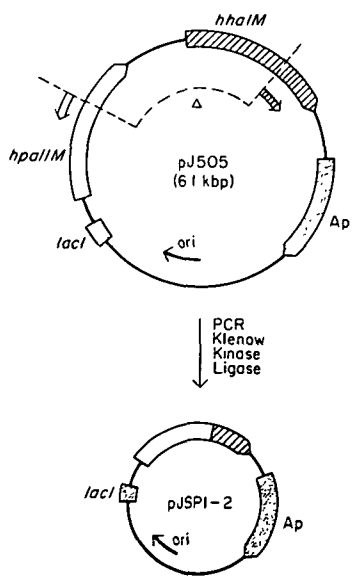

Figure 1. Construction of recombinant plasmids encoding the hybrid methylases using PCR-mediated deletion. The oligonucleotide primers, shown by the open and shaded arrows, which correspond to internal regions of each methylase were used to prime an 'inverted' PCR reaction in which synthesis proceeded outward from the methylase genes and through the vector sequences. Religation of the PCR product then generated a specific hybrid, with the junction being defined by the 5′ ends of the two primers.

downstream from the *Hpa*II variable region (see Figure 7 for structures of all hybrids).

These two constructs were analysed for their methylation capacity *in vivo* by isolating plasmid DNAs from growing cultures and testing for their sensitivity to digestion *in vitro* with the two cognate endonucleases, R.*Hpa*II and R.*Hha*I. Both restriction enzymes are unable to cleave DNA if their substrate sequences, CCGG and GCGC respectively, are methylated at the second cytosine residue in either one or both strands (25–28). Control DNAs, carrying either the *Hpa*II or *Hha*I methylase genes under the same promoter, were immune to the action of the cognate restriction enzymes, but were readily fragmented by the second endonuclease. **P1** was completely fragmented by R.*Hpa*II, but digestion by R.*Hha*I gave several very faint bands in addition to the expected digestion pattern (not shown, but see Figure 5 for the results with an over-expressed version of this hybrid). This proved reproducible but indicated an extremely low level of protection presumably because the hybrid methylase produced was a very inefficient enzyme. The plasmid DNA encoding **P2** could be completely digested by both restriction enzymes.

Over-expression vectors for hybrids

From previous work we knew that the endogenous promoter of M.*Hpa*II was not effective in *E.coli* cells (13). The initial constructs were made in a pUC19 vector where the start codon for the methylase gene was positioned more than 0.5 kb downstream from the promoter resulting in low level production of the hybrid proteins *in vivo* (not shown). The expected reduced activity of the hybrids combined with poor expression meant that it might prove difficult to characterize any but the most active hybrids. We therefore decided to improve the expression of the modification phenotypes *in vivo* and facilitate the detection of interesting hybrids by switching to another vector that would allow a wider control of the level of protein synthesis. We also changed the strategy for constructing the hybrids so that a PCR mediated recombination method, Splicing by Overlap Extension (SOE) was employed (20).

Each methylase gene was transferred into the vector, pUHE25-2, that contains a *lac*I-repressible expression system driven by an early bacteriophage T7 promoter (H.Bujard, unpublished). The exact positioning of the start codon relative to the promoter and ribosomal binding site was achieved by cloning the desired genes on an *Sph*I to *Hin*dIII fragment such that the *Sph*I sequence GC**ATGC** overlaps with the normal start codon ATG. This resulted in a nucleotide replacement A→C at the first position of the second codon. Since an amino acid change at the second position of both methylases was unavoidable we chose to substitute the original residues with structurally similar ones: arginine for lysine (K2R) in the *Hpa*II methylase and leucine for isoleucine (I2L) in M.*Hha*I as well as in all subsequent hybrid derivatives (Figure 2). Given the great diversity in sequence and length of the N-terminal extension upstream of conserved block I in the m^5C-methylases (7) these conservative substitutions seemed unlikely to influence enzymatic activity. Indeed, we have been unable to detect differences between the wild type enzymes and these mutant derivatives in our system.

The use of the pUHE25 vector enabled us to control the *in vivo* synthesis of the desired proteins quite effectively (Figure 3). In all cases, upon induction, the hybrid was visible as a major band among proteins of the crude extract and the amounts were comparable for each hybrid (Figure 3 and data not shown). As expected, overproduction of the proteins greatly enhanced our ability to detect the methylation phenotypes *in vivo*. The **P1** hybrid gave a much stronger although still partial protection against R.*Hha*I (Figure 5). However even when over-expressed the methylase activity of **P2** was undetectable.

The hybrids

We have constructed eleven plasmids over-expressing hybrids between the *Hpa*II and *Hha*I methylases containing various combinations of domain arrangements including reciprocal ones (see Figure 7). The fusions were made at equivalent positions of the sequence motifs, with three junction points chosen to dissect the methylase sequences into four pieces (Figure 4). One junction was at the C-terminal end of motif VIII. This provided the N-terminal segment of the methylase up to the start of the variable region. A second junction was at the N-terminal end of motif IX and marked the C-terminal end of the variable region. The third junction was at the N-terminal end of motif X and enabled constructs to be made containing the intact variable region plus the cognate sequences between motifs IX and X. The structure of the plasmids coding for the hybrids was confirmed by diagnostic restriction endonuclease mapping and complete sequencing of the hybrid methylase coding region. Some of the recombinant genes acquired minor nucleotide substitutions during the construction, presumably due to PCR amplification errors.

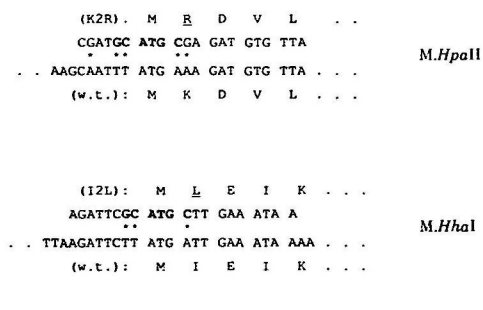

Figure 2. Mutagenesis of the N-termini of the genes encoding M.*Hpa*II and M.*Hha*I. pJ505 DNA, encoding the wild type methylases was used as a template in a PCR reaction primed by oligonucleotides # 5 or # 15. In each case, the lower sequences are for wild type protein and DNA, while the upper sequences show the mutant forms. The newly-introduced *Sph*I sites are shown in bold and the mutations, K2R in M.*Hpa*II, and I2L in M.*Hha*I, are underlined. Template-primer mismatches are marked with *.

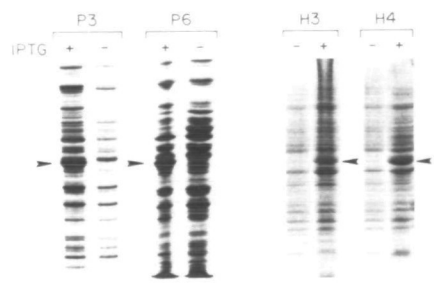

Figure 3. Expression of the hybrid methylases *in vivo*. Cell cultures were induced at late log phase with 1mM IPTG for 2h, harvested and crude extracts were analysed by SDS-PAGE (22). Uninduced controls were sampled prior to the addition of IPTG. Arrows show the bands corresponding to the over-expressed hybrid methylases.

In all but two cases the changes did not lead to protein sequence alterations. In the two mutant constructs the errors were corrected by replacing the mutant sequences with appropriate restriction enzyme fragments from the wild type gene. Again, the final constructs were checked by complete sequencing. The synthesis of the full-length proteins *in vivo* was verified by SDS-PAGE of the proteins from crude extracts (Figure 3). Experimentally determined molecular weight values were in good agreement with those deduced from the nucleotide sequences.

All plasmid DNAs over-expressing the hybrid methylases were tested for their susceptibility to cleavage, *in vitro*, with the *Hpa*II and *Hha*I restriction enzymes. This test allowed a determination of the activity, or lack thereof, for each enzyme and in the case of hybrids with detectable activity the specificity could be determined and a qualitative estimation of its efficiency of methylation. Five hybrids, **P1**, **P4**, **P5**, **P6** and **H4** showed clearly detectable activity in this assay (Figure 5). In the case of hybrid **P1** it can be seen that control of the T7 promoter is strong and activity can only be detected following induction with IPTG. Positive controls were carried out in all cases to ensure that the restriction enzymes performed as expected. Typically this included the addition of carrier bacteriophage λ DNA as shown for hybrid **P4** in Figure 5.

For hybrids providing detectable protection against R.*Hpa*II an additional direct measure of activity was performed. This analysis took advantage of R.*Msp*I, an isoschizomer of R.*Hpa*II, which cleaves both methylated and unmethylated *Hpa*II sites (29, 26). Because R.*Msp*I cleaves DNA between the two cytosine residues in the recognition sequence, CCGG, we were also able to determine the modification status of the second cytosine residue directly. Thus R.*Msp*I fragments were ^{32}P-labeled at their 5′-ends, digested to mononucleotides and the labeled mononucleotides analysed (23). The major products were identified by thin layer chromatography (TLC) as illustrated in Figure 6. One dimensional TLC (Figure 6A) provided good resolution of the major products and the assignments were confirmed by two dimensional TLC as illustrated in Figure 6B for the hybrid **H4**. In addition to the expected products, m^{5}C and C, all samples contained minor quantities of the other deoxynucleotides, presumably due to labeling of any free 5′-termini arising from nicks in the relaxed form of the substrate plasmid DNAs. All four mononucleotides were detected, even when the plasmid DNA was labeled without cleavage with R.*Msp*I (Figure 6a, lane 6). The ratio of methylated and unmethylated cytosines at *Hpa*II sites, as estimated by the one-dimensional method, correlates well with the extent of protection of each hybrid against the action of R.*Hpa*II. Thus, the hybrids have the following relative efficiencies **P0 > > P5 > H4 > > P6**. Unfortunately, no analogous isoschizomer of R.*Hha*I is available to allow a similar analysis at *Hha*I sites (GCGC).

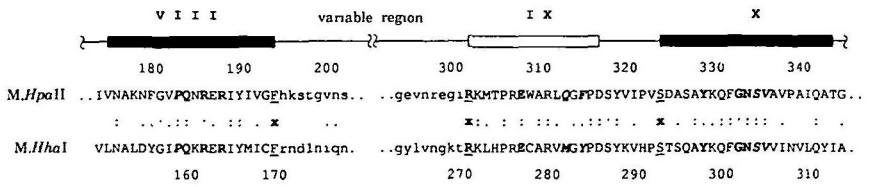

Figure 4. Partial alignment between the M.*Hpa*II and M.*Hha*I protein sequences (based on (13)). The exact positions used to form the hybrid junctions are marked by x. Identical residues are marked (:) and conservative substitutions are indicated (.). The upper solid boxes, on the schematic, indicate the highly conserved motifs, VIII and X, and the open box indicates a less-well conserved motif present in all m^{5}C-methylases (7). The variable regions (N- and C-termini shown in lower case) lie between motifs VIII and IX. Invariant residues of the motifs are shown in bold letters, highly conserved ones in bold italic and the junction residues are underlined.

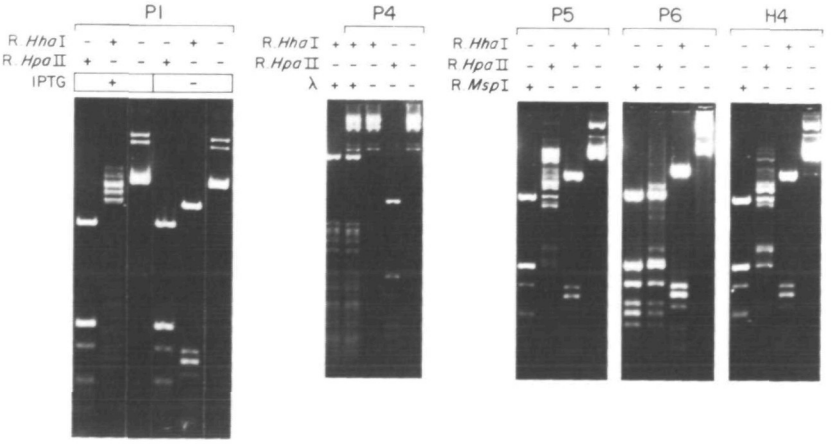

Figure 5. Restriction endonuclease analysis of the methylation potential of some hybrid methylases *in vivo*. For hybrid, **P1**, the plasmid DNA was prepared either with or without IPTG induction and challenged with the restriction endonucleases indicated. In all other cases the plasmid DNAs of the indicated hybrids were isolated after IPTG induction and challenged with an excess of R.*Hha*I, R.*Hpa*II or R.*Msp*I. In the case of hybrid, **P4**, digestions were carried out either in the presence or absence of bacteriophage λ DNA to provide a positive control. In all cases samples were analysed by 1.3% agarose gel electrophoresis. Schematic structures for the hybrids can be found in Figure 7.

In all but two cases the changes did not lead to protein sequence alterations. In the two mutant constructs the errors were corrected by replacing the mutant sequences with appropriate restriction enzyme fragments from the wild type gene. Again, the final constructs were checked by complete sequencing. The synthesis of the full-length proteins *in vivo* was verified by SDS-PAGE of the proteins from crude extracts (Figure 3). Experimentally determined molecular weight values were in good agreement with those deduced from the nucleotide sequences.

All plasmid DNAs over-expressing the hybrid methylases were tested for their susceptibility to cleavage, *in vitro*, with the *Hpa*II and *Hha*I restriction enzymes. This test allowed a determination of the activity, or lack thereof, for each enzyme and in the case of hybrids with detectable activity the specificity could be determined and a qualitative estimation of its efficiency of methylation. Five hybrids, **P1**, **P4**, **P5**, **P6** and **H4** showed clearly detectable activity in this assay (Figure 5). In the case of hybrid **P1** it can be seen that control of the T7 promoter is strong and activity can only be detected following induction with IPTG. Positive controls were carried out in all cases to ensure that the restriction enzymes performed as expected. Typically this included the addition of carrier bacteriophage λ DNA as shown for hybrid **P4** in Figure 5.

For hybrids providing detectable protection against R.*Hpa*II an additional direct measure of activity was performed. This analysis took advantage of R.*Msp*I, an isoschizomer of R.*Hpa*II, which cleaves both methylated and unmethylated *Hpa*II sites (29, 26). Because R.*Msp*I cleaves DNA between the two cytosine residues in the recognition sequence, CCGG, we were also able to determine the modification status of the second cytosine residue directly. Thus R.*Msp*I fragments were ^{32}P-labeled at their 5′-ends, digested to mononucleotides and the labeled mononucleotides analysed (23). The major products were identified by thin layer chromatography (TLC) as illustrated in Figure 6. One dimensional TLC (Figure 6A) provided good resolution of the major products and the assignments were confirmed by two dimensional TLC as illustrated in Figure 6B for the hybrid **H4**. In addition to the expected products, m^{5}C and C, all samples contained minor quantities of the other deoxynucleotides, presumably due to labeling of any free 5′-termini arising from nicks in the relaxed form of the substrate plasmid DNAs. All four mononucleotides were detected, even when the plasmid DNA was labeled without cleavage with R.*Msp*I (Figure 6a, lane 6). The ratio of methylated and unmethylated cytosines at *Hpa*II sites, as estimated by the one-dimensional method, correlates well with the extent of protection of each hybrid against the action of R.*Hpa*II. Thus, the hybrids have the following relative efficiencies **P0 > > P5 > H4 > > P6**. Unfortunately, no analogous isoschizomer of R.*Hha*I is available to allow a similar analysis at *Hha*I sites (GCGC).

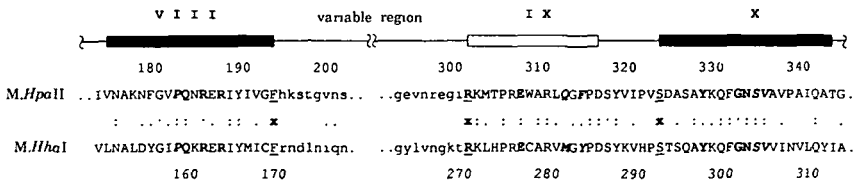

Figure 4. Partial alignment between the M.*Hpa*II and M.*Hha*I protein sequences (based on (13)). The exact positions used to form the hybrid junctions are marked by x. Identical residues are marked (:) and conservative substitutions are indicated (.). The upper solid boxes, on the schematic, indicate the highly conserved motifs, VIII and X, and the open box indicates a less-well conserved motif present in all m^{5}C-methylases (7). The variable regions (N- and C-termini shown in lower case) lie between motifs VIII and IX. Invariant residues of the motifs are shown in bold letters, highly conserved ones in bold italic and the junction residues are underlined.

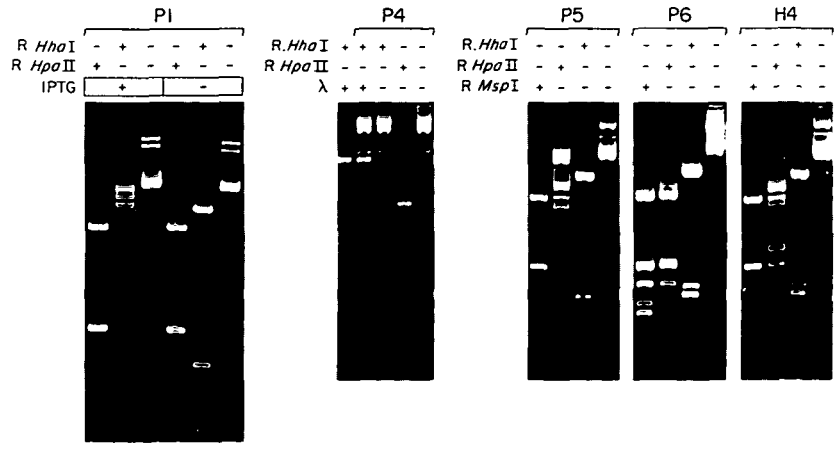

Figure 5. Restriction endonuclease analysis of the methylation potential of some hybrid methylases *in vivo*. For hybrid, **P1**, the plasmid DNA was prepared either with or without IPTG induction and challenged with the restriction endonucleases indicated. In all other cases the plasmid DNAs of the indicated hybrids were isolated after IPTG induction and challenged with an excess of R.*Hha*I, R.*Hpa*II or R.*Msp*I. In the case of hybrid, **P4**, digestions were carried out either in the presence or absence of bacteriophage λ DNA to provide a positive control. In all cases samples were analysed by 1.3% agarose gel electrophoresis. Schematic structures for the hybrids can be found in Figure 7.

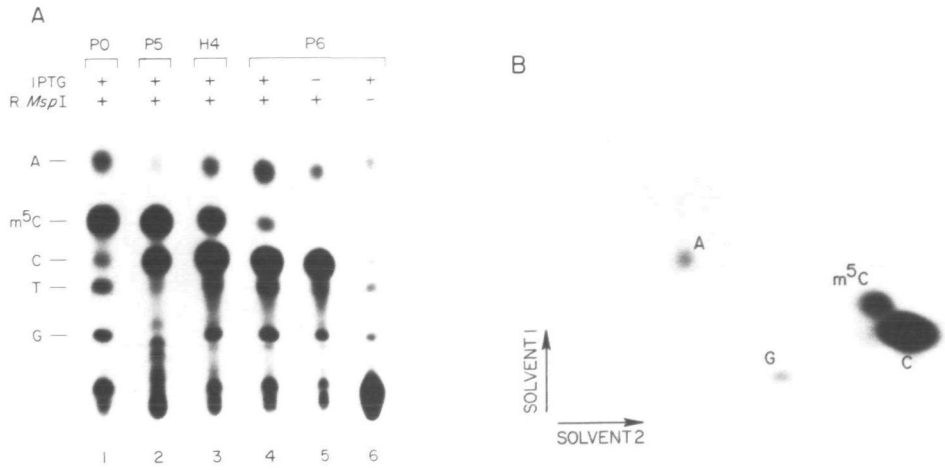

Figure 6. Methylation status at the second position of the sequence CCGG. Plasmid DNAs showing detectable protection against R.*Hpa*II action were digested to completion with R.*Msp*I, the products labeled with ^{32}P at their 5′-termini and degraded to mononucleotides. A. One dimensional TLC analysis of labeled mononucleotides from the plasmid DNAs methylated *in vivo* with the indicated hybrid methylases. Radioactive spots were identified by matching their mobility with that of synthetic 2′-deoxy-5′-mononucleotides. Lane 1—**P0** induced; lane 2—**P5** induced; lane 3—**H4** induced; lane 4—**P6** induced; lane 5—**P6** uninduced; lane 6—**P6** induced and uncut. B. Two dimensional TLC analysis of the products from lane 3 from the hybrid **H4**. The positions of the two major radioactive spots coincided with those of standard m^5dCMP and dCMP included in the applied sample.

All of the hybrids proved to be considerably weaker methylases than their wild type parents. In four cases we detected partial methylation of the plasmid during full induction of the methylase and only in one case, **P4**, was complete protection provided against the cognate restriction endonuclease (Figure 5). In contrast, the parent methylases were active enough to protect plasmid DNAs even without induction (not shown). Despite the poorer enzymatic properties the hybrids retained sequence specificity since the protection, when detectable, was always specific toward one of the cognate endonucleases rather than both.

Figure 7 provides a summary of the structures of the hybrid methylases and their observed properties. Two of the hybrids, **P3** and **H3**, contained exact swaps of the variable regions, but in neither case were we able to detect methylation activity *in vivo*. Thus we were not able to determine the specificity potential of this region directly. However, other hybrids allow us to define the boundaries of the specificity domain by excluding regions non-essential for this function. Thus, the methylation specificy of the hybrids **P1**, **P4** and **H4** indicates that sequences from the N-terminus up to and including motif VIII cannot be involved in target recognition. Furthermore, motif X together with the downstream C-terminal region are excluded since the methylation specificity of the hybrids **P4**, **P5** and **H4** differs from that of the parent methylases which contributed these fragments (Figure 7). Hybrid **P6** demonstrates that motif IX plays no direct role in specificity determination. Finally, a comparison of the hybrids **P4** and **P6**, which differ only in the variable region, provides firm evidence that the specificity determinant resides within the variable region.

The relative enzymatic efficiencies of the hybrid methylases could be estimated from analysis of the extent of methylation because they were all produced in comparable amounts *in vivo*. This gives some insight into the flexibility with which methylase hybrids can be constructed. First, all hybrids containing the N-terminus of M.*Hpa*II up to motif VIII are more active than their *Hha*I counterparts (compare **P1-H1**, **P4-H4**, **P6-H6**). The reason for this inequality is unclear since both methylases align well

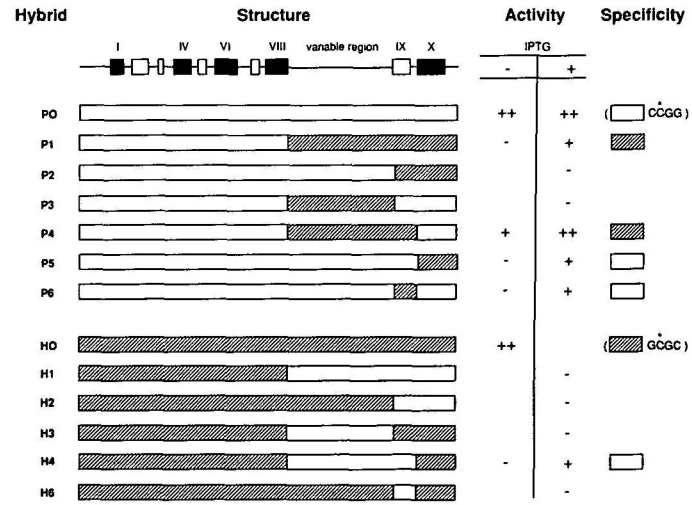

Figure 7. The structure and methylation properties of the hybrid methylases. The upper schematic diagram shows the ten conserved motifs found in m^5C-methylases (6, 7): five very highly conserved motifs (filled boxes) and five less-conserved motifs (open boxes) are connected by diverse regions (line). The parent methylases and their hybrid derivatives are shown schematically with M.*Hpa*II sequences shown as open boxes and M.*Hha*I sequences shown as hatched boxes. For each hybrid the methylation activity observed *in vivo* from either induced (IPTG +) or uninduced (IPTG −) cultures is shown under Activity. − indicates no protection, + indicates partial protection and + + indicates complete protection against the appropriate restriction endonuclease. The specificity is indicated in the right hand column. Note that 'hybrids' labeled **P0** and **H0** are the parent methylases, M.*Hpa*II and M.*Hha*I, respectively.

throughout motifs I to VIII and differ appreciably only in the extreme N-terminal extension lying upstream of motif I (13). These extensions are completely dissimilar and are 32 amino-acids long in M.*Hpa*II and 12 amino-acids long in M.*Hha*I. Second, in our system, hybrids are more active if both the variable region and the adjacent motif IX come from the same parent

methylase (compare **P1-P2**, **P5-P2**, **P4-P3**, **P4-P6**, **H4-H3**, **H4-H6**). Third, the hybrid in a pair of analogues is more efficient if the N-terminal region through motif VIII and the C-terminal region from motif X have both originated from the same parent methylase (**P4-P1**, **P6-P2**, **H4-H1**, **P0-P5**). Our most active hybrid is **P4** that satisfies all three conditions above. Similarly, the reciprocal derivative **H4** is the only one in the *Hha*I series that renders detectable methylation *in vivo*.

DISCUSSION

Previous studies of multi-specific DNA methyltransferases have shown that DNA recognition is accomplished by one of several sub-domains of the variable region. Our experiments show for the first time that the comparable variable region from the mono-specific methylases also contains the information necessary for sequence specific recognition. Although the hybrids we have made are rather poor methylases, they do show complete specificity. This retention of the wild type specificity in several of our hybrids suggests that the variable region serves as an independent domain designed for DNA recognition. We do not know whether the complete variable region is necessary for DNA recognition in the mono-specific methylases. However, in the multi-specific methylases the variable region is itself composed of several non-overlapping segments that each recognize and interact with a different specific DNA sequence. These segments can be swapped leading to predicted changes in specificity and essentially full retention of methylase activity in most cases (11).

While our results support the hypothesis about the modular structure of m^5C-methylases (6) they also highlight its limitations. The model proposes that m^5C-methylases contain a core of highly-conserved regions that are responsible for catalysis and a separate domain, the variable region, which serves to bring the catalytic machinery to the target DNA sequence. The variation in specificity toward distinct DNA sequences is thus achieved by combining different variable regions with essentially the same core structures. While the independent character of the modules is evident in the multi-specific methylases, the mono-specific methylases appear much more sensitive to perturbations introduced during domain swaps.

Among the eleven hybrids between the *Hpa*II and *Hha*I methylases shown in Figure 7, six were inactive in our assays, while five retained detectable activity. Although even in the best case, **P4**, full induction was necessary before complete protection could be detected *in vivo*. This suggests that the interchangeability of equivalent blocks among the mono-specific methylases is quite limited and methylation efficiency is critically dependent upon overall structure. The apparent flexibility of the phage methylases may be illusory, since the conserved motifs of these methylases show much closer sequence conservation than is found among the mono-specific enzymes. The multi-specific enzymes may be of relatively recent origin and have not yet diverged significantly. Alternatively, there may be a strong evolutionary pressure to preserve the original core structure which can accomodate several different specificity domains while retaining function. In contrast, the mono-specific enzymes have no inherent need for flexibility and probably evolve by accumulating compatible mutations to optimize catalysis. The conserved motif IX could be one such region that has been involved in evolutionary adjustment to accomodate its cognate variable region. It should be noted that a fully-functional hybrid has been produced between two very closely related mono-specific methylases, *Bsp*RI and *Bsu*RI, that target the same sequence, GGCC (30).

Our results show that foreign motifs lead to decreased functional capacity. This is clearly seen by comparing the wild type enzymes with hybrids containing just a single motif replacement (**P0** versus **P3**, **P5** and **P6**; **H0** versus **H3** and **H6**). However, certain motif combinations appear to give hybrid enzymes with less reduction in methylation efficiency. Thus, transfer of specificity from M.*Hpa*II to M.*Hha*I and vice versa occurred most efficiently when both the variable region and the adjacent conserved motif IX were transferred in concert. This would suggest that interactions occur between these regions in in the parent methylases. A similar conclusion can be drawn for the potential interaction of motif X and the N-terminal region. Presumably interactions between motifs take place to determine the overall structure of the final methylase and hybrids containing mixed motifs would be expected to be less efficient methylases than their parents. Mutagenesis of a weak hybrid methylase might yield compensatory mutations that could lead to more efficient enzymes and the analysis of these mutants might give significant insights into the function of the motifs and their interactions. Such mutants, if they exist, should be easily selected using standard procedures for cloning DNA methylases (reviewed in 2). These experiments are currently underway.

ACKNOWLEDGEMENTS

The authors thank Stacey Klein for most of the DNA sequencing work, E.Raleigh for providing the *mcr*$^-$ *E.coli* strains, G.G.Wilson and U.Deuschle for their kind gift of plasmids and A.Bhagwat for useful discussions during the early stages of this work. This work was supported by grants from the NSF (DMB-8917650) and the NIH (GM46127).

REFERENCES

1. Klimasauskas, S., Timinskas, A., Menkevicius, S., Butkiene, D., Butkus, V., Janulaitis, A. (1989) *Nucl. Acids Res.* **17**, 9823–9832.
2. Wilson, G.G. (1991) *Nucl. Acids Res.* **19**, 2539–2566.
3. Kessler, C., Manta, V. (1990) *Gene* **92**, 1–248.
4. Roberts, R.J. and Macelis, D. (1991) *Nucl. Acids Res.* **19**, 2077–2109.
5. Slatko, B.E., Croft, R., Moran, L.S., Wilson, G.G. (1988) *Gene* **74**, 45–50.
6. Lauster, R., Trautner, T.A., Noyer-Weidner, M. (1989) *J. Mol. Biol.* **206**, 305–312.
7. Pòsfai, J., Bhagwat, A.S., Pòsfai, G., Roberts, R.J. (1989) *Nucl. Acids Res.* **17**, 2421–2435.
8. Walter, J., Noyer-Weidner, M., Trautner, T.A. (1990) *EMBO J.* **9**, 1007–1013.
9. Szilak, L., Venetianer, P., Kiss, A. (1990) *Nucl. Acids Res.* **18**, 4659–4664.
10. Wilke, K., Rauhut, E., Noyer-Weidner, M., Lauster, R., Pawlek, B., Behrens, B., Trautner, T.A. (1988) *EMBO J.* **7**, 2601–2609.
11. Balganesh, T.S., Reiners, L., Lauster, R., Noyer-Weidner, M., Wilke, K., Trautner, T.A. (1987) *EMBO J.* **6**, 3543–3549.
12. Trautner, T.A., Balganesh, T.S., Pawlek, B. (1988) *Nucl. Acids Res.* **16**, 6649–6658.
13. Card, C.O., Wilson, G.G., Weule, K., Hasapes, J., Kiss, A., R.J. Roberts (1990) *Nucl. Acids Res.* **18**, 1377–1383.
14. Caserta, M., Zacharias, W., Nwankwo, D., Wilson, G.G., Wells, R.D. (1987) *J. Biol. Chem.* **262**, 4770–4777.
15. Wu, J.C., Santi, D.V. (1987) *J. Biol. Chem.* **262**, 4778–4786.
16. Wu, J.C., Santi, D.V. (1988) *Nucl. Acids Res.* **16**, 703–717.
17. Yoo, O.J., Agarwal, K.L. (1980) *J. Biol. Chem.* **255**, 6445–6449.
18. Dila, D., Sutherland, E., Moran L., Slatko, B., Raleigh, E.A. (1990) *J. Bacteriol.* **172**, 4888–4900.
19. Hemsley, A., Arnheim, N., Toney, M.D., Cortopassi, G., Galas, D.J. (1989) *Nucl. Acids Res.* **17**, 6545–6551.

20. Ho, S.N., Pullen, J.K., Horton, R.M., Hunt H.D., Pease L.R. (1990) *DNA and Protein Engineering Techniques* **2**, 50–55.
21. Joshi, A.K., Baichwal, V., Ames, G., F.-L. (1991) *BioTechniques* **10**, 42–45.
22. Sambrook, J., Fritsch, E.F., Maniatis, T (1989) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
23. Cedar, H., Solage, A., Glaser, G., Razin, A. (1979) *Nucl. Acids Res.* **6**, 2125–2132.
24. Stephenson, M.L., Zamecnik, P.C. (1967) In Grossman, L. and Moldave, K. (ed.). Methods in Enzymology, Academic Press, New York, Vol. XIIB, pp. 670–678.
25. Ehrlich, M. and Wang, R.Y.-H. (1981) *Science* **212**, 1350–1357.
26. Butkus, V., Petrauskiene, L., Maneliene, Z., Klimasauskas, S., Laucys, V., Janulaitis, A. (1987) *Nucl. Acids Res.* **15**, 7091–7102.
27. Gruenbaum, Y., Cedar, H., Razin, A. (1981) *Nucl. Acids Res.* **9**, 2509–2515.
28. Korch, C., Hagblom, P. (1986) *Eur. J. Biochem.* **161**, 519–524.
29. Waalwijk, C. and Flavell, R.A. (1978) *Nucl. Acids Res.* **5**, 3231–3236.
30. Kim, S.C., Pòsfai, G., Szybalski, W. (1991) *Gene* **100**, 45–50.