

Recombination of constant and variable modules alters DNA sequence recognition by type IC restriction – modification enzymes

Marcel Gubler², Diego Braguglia³,
Jürg Meyer⁴, Andrzej Piekarowicz¹ and
Thomas A. Bickle

Department of Microbiology, Biozentrum, Basel University, CH-4056 Basel, Switzerland and ¹Institute of Microbiology, Warsaw University, 00-046 Warsaw, Poland

²Present address: Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

³Present address: Institut Suisse de Recherches Experimentales sur le Cancer, CH-1066 Epalinges s/Lausanne, Switzerland

⁴Present address: Zahnärztliches Institut, CH-4056 Basel, Switzerland

Communicated by T.A. Bickle

***EcoR124* and *EcoDXXI* are allelic type I restriction–modification (R–M) systems whose specificity genes consist of common structural elements: two variable regions are separated by a constant, homologous region containing a number of repetitive sequence elements. *In vitro* recombination of variable and constant elements has led to fully active, hybrid R–M systems exhibiting new and predictable target site specificities. Methylation of synthetic DNA sequences with purified, hybrid modification methylases was used to confirm the proposed recognition sequences. The results clearly demonstrate the correlation between protein domains and target site specificity. Our data suggest that a bacterial population may switch the recognition sequences of its type I R–M system by single recombination events and thus is able to maintain a prokaryotic analogue of the immune system of variable specificity.**

Key words: DNA restriction and modification/DNA sequence restriction/evolution of sequence specificity/type IC restriction enzymes

Introduction

The physical interaction between bacterial restriction–modification (R–M) systems and DNA is basically threefold: R–M systems recognize specific sequences on DNA; they cleave DNA which contains non-protected target sequences; and they prevent cleavage of DNA by methylation of adenosine or cytosine residues within target sequences. R–M systems have been classified as type I, II and III based on subunit composition, cofactor requirements and the enzymatic reactions that they catalyse (Bickle, 1987). Most of the R–M systems known to date are type II. They are extensively used in recombinant DNA technology primarily because restriction and modification reactions are carried out by two different enzymes and cleavage of DNA occurs precisely within, or close to, the sequences that they recognize. In addition, type II restriction enzymes have very simple reaction conditions: they require no cofactor other than Mg^{2+} to cleave DNA. Type III R–M systems are

somewhat more complex in that the restriction enzyme is composed of two different subunits, requires ATP and Mg^{2+} to cleave DNA and, in the presence of *S*-adenosyl methionine (AdoMet), can also act as a modification methylase (Hadi *et al.*, 1983). Type I R–M systems consist of three different subunits, they require Mg^{2+} , ATP and AdoMet to digest DNA, and they exhibit high ATPase activity upon incubation with non-modified substrate DNA (see Bickle, 1982 for a review). The process leading to cleavage of DNA by type I restriction enzymes at seemingly random sites up to several thousand base pairs (bp) distant from their recognition sequence has been explained by ATP-stimulated translocation of the enzymes along the DNA (Studier and Bandyopadhyay, 1988).

All the genetic loci coding for type I R–M systems identified so far are organized in two transcriptional units, one containing the genes *hsdM* and *hsdS*, the other containing *hsdR* only (Sain and Murray, 1980; Suri and Bickle, 1985; Price *et al.*, 1989). Genetic analysis showed that all three gene products are required for restriction, whereas *hsdM* and *hsdS* gene products are sufficient for modification methylation. The specificity of type I R–M systems is determined by the *hsdS* gene: mutations in *hsdS* abolish restriction and methylation activity while complementation with an allelic *hsdS* gene restores both activities and confers the specificity of the complementing allele (Boyer and Roulland-Dussoix, 1969; Hubacek and Glover, 1970; Fuller-Pace *et al.*, 1985; Skrzypek and Piekarowicz, 1989). In fact, detection of genetic complementation, and later of DNA homologies and antigenic cross-reactivity (Murray *et al.*, 1982) led to the recognition of three distinct families of type I R–M systems: A, B and C (see Bickle, 1987 for a review). In this report, we focus on members of the C family which are encoded by large conjugative plasmids: *EcoR124*, *EcoR124/3* (Firman *et al.*, 1985) and *EcoDXXI* (Skrzypek and Piekarowicz, 1989). However, localization on plasmids seems not to be a necessary criterion for type IC R–M systems. The recent finding of extensive homology between chromosomal DNA of the *Escherichia coli* *prf* locus and the *EcoR124/3* *hsd* genes (Linder *et al.*, 1990) indicates a more widespread occurrence of type IC R–M systems than originally anticipated.

All type I R–M systems whose target sites have been determined recognize bipartite sequences consisting of two half-sites, each of three to five nucleotides, separated by a non-specific spacer region of six to eight nucleotides. Recombination of the *hsdS* genes of the *Salmonella* type I R–M systems *StySBI* and *StySPI* gave rise to the hybrid systems *StySQI* and *StySJI* which recognize hybrid target sites that are combinations of the original SB and SP half-sites (Bullas *et al.*, 1976; Fuller-Pace *et al.*, 1985; Nagaraja *et al.*, 1985; Gann *et al.*, 1987). It was concluded that in the wild-type and the recombinant HsdS polypeptides, amino-terminal domains specify the 5' half and carboxy-terminal domains specify the 3' half of bipartite recognition

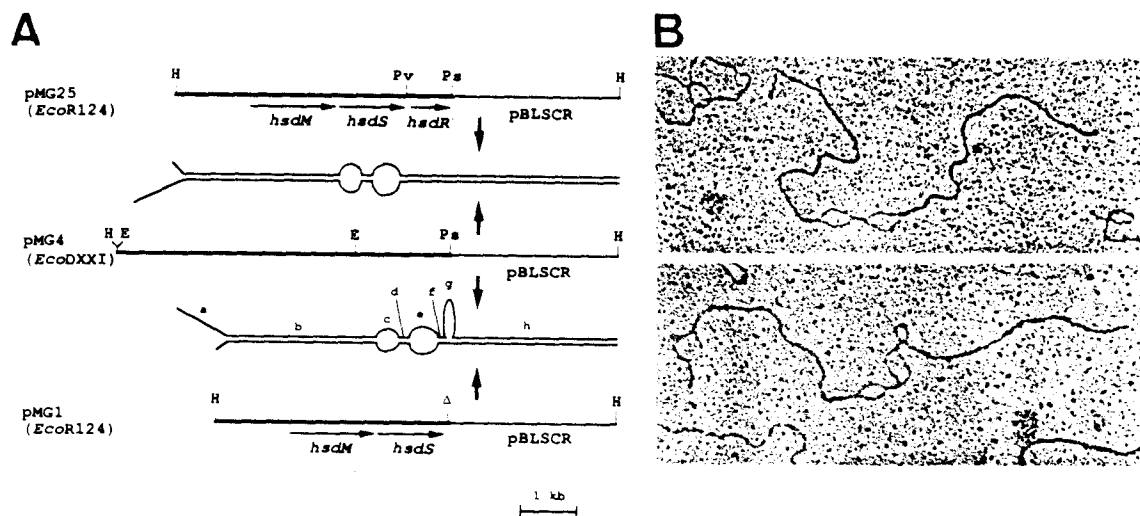


Fig. 1. Heteroduplex structures of *hsd* DNA drawn schematically (A) and as seen in the electron microscope (B). Thin lines in pMG25, pMG4 and pMG1 represent pBLSR SK(+) vector DNA. Loops of single-stranded DNA are not drawn to scale, their exact lengths are given in Table I.

sequences. A different correlation between HsdS polypeptide structure and target site specificity was observed for the type I R–M systems *EcoR124I* and *EcoR124/31* (Price *et al.*, 1989). It was found that a differing number of short repetitive elements in the centre of the otherwise completely identical *hsdS* genes determine the number of nucleotides in the non-specific spacer between target half-sites.

We show by *in vitro* recombination of the *hsdS* genes of *EcoR124I*, *EcoR124/31* and *EcoDXXI* [which recognize GAA(N₆)RTCG, GAA(N₇)RTCG and TCA(N₇)RTTC, respectively] that type IC R–M systems have the potential to change their target site specificities by simple means and in various ways. Although these prokaryotic enzymes are not as diverse as eukaryotic immunoglobulins, we postulate that there are parallels between type I R–M enzymes as primitive but efficient bacterial immune systems and the more sophisticated immunological repertoire of higher eukaryotes.

Results

Close relationship between *EcoR124I* and *EcoDXXI* *hsd* loci

We have used electron microscopic analysis of *EcoR124I* and *EcoDXXI* heteroduplex DNA in order to determine the degree of homology between the two allelic R–M systems and, in particular, to localize structural differences in the *hsdS* genes which may be responsible for the site specificity of each individual system. DNA fragments covering parts of the completely sequenced *EcoR124I* and the restriction mapped *EcoDXXI* *hsd* regions were hybridized and analysed by electron microscopy. Figure 1A shows the DNA probes used for hybridization and a schematic drawing of the observed heteroduplex structures of which two representative examples are depicted in Figure 1B. The result of the hybridization of pMG25 (*EcoR124I*) with pMG4 (*EcoDXXI*) DNA revealed strong conservation of both *hsd* regions over the entire length of *hsdM*, *hsdS* and the 5' portion of the *hsdR* genes with the exception of two small regions of non-homology located at the position of the *hsdS* genes. To localize the non-homologous DNA regions more precisely, the *EcoR124I* clone pMG1 lacking the 700 bp long

Table I. Lengths of segments of heteroduplexes

Segment	a	b	c	d	e	f	g	h
Length	920	3090	420	180	540	170	800	2950
±SD	120	210	70	50	80	40	–	–

The data (in nucleotides) shown as means ± SD are based on the known lengths of segments g and h. Twenty individual molecules were measured. The positions of the segments are shown in Figure 1A.

PvuI–*PstI* 5' fragment of *hsdR* (Figure 1A) was hybridized with the *EcoDXXI* clone pMG4. The position of the additional loop of *EcoDXXI* *hsdR* DNA was taken as the start point to measure the length of the double-stranded and single-stranded sections (a–h in Figure 1A) of the heteroduplex molecules. Their lengths are listed in Table I. They allow dissection of the *hsdS* specificity genes into distinct structural and possibly also functional regions; there are two regions of homology, one of 180 bp (see Table I for standard deviations) in the centre and one of < 170 bp at the 3' end of the *hsdS* genes. However, there are two large regions of non-homology, one of 420 bp at the 5' end and a larger one of 540 bp towards the 3' end of the *hsdS* genes. Based on the fact that no differences other than those in the *hsdS* genes have been detected, we postulate that the *EcoR124I* and *EcoDXXI* R–M systems exhibit a high degree of structural conservation and that the regions of non-homology in the *hsdS* genes encode protein domains which may be responsible for DNA target site specificity as was found for the A and B families of type I R–M systems (Nagaraja *et al.*, 1985; Gann *et al.*, 1987; Kannan *et al.*, 1989).

Partial DNA sequence analysis of the *EcoDXXI* *hsdS* gene revealed that the central region of homology shown in Figure 2 contains a 12 bp sequence repeated three times. Exactly the same 12 bp sequence has been found to be repeated twice in the *EcoR124I* and three times in the *EcoR124/31* *hsdS* genes (Price *et al.*, 1989). It was shown for the *EcoR124I* and *EcoR124/31* R–M systems that two and three 12 bp repeats in the centre of *hsdS* correlate with six and seven non-specific nucleotides, respectively, separating the half-sites of the bipartite recognition sites. It is therefore not surprising that *EcoDXXI*, recognizing a site with seven

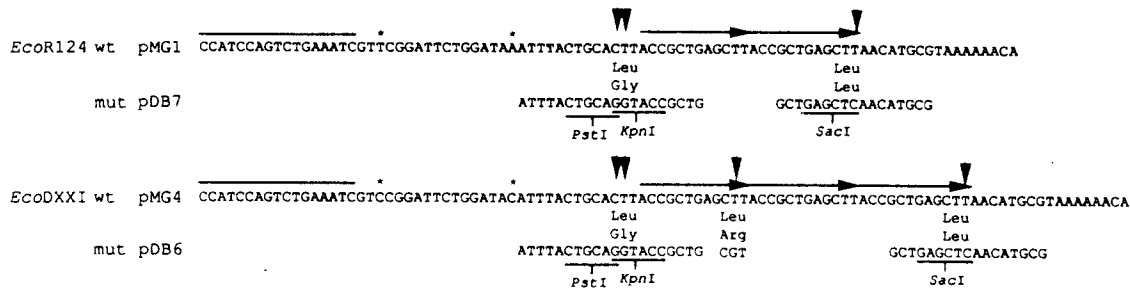


Fig. 2. DNA sequence of the region spanning the central repetitive elements in the *EcoR124I* and *EcoDXXI* *hsdS* genes. pMG1 and pMG4 represent the wild-type DNA sequences; the arrows indicate the 12 bp sequence elements that are repeated twice in *EcoR124I* and three times in *EcoDXXI*. Asterisks mark sequence differences in the central region of homology of both *hsdS* genes. Arrowheads indicate the point mutations introduced in the *EcoR124I* and *EcoDXXI* wild-type genes by the oligonucleotides shown below the pMG1 and pMG4 sequences. They created the *KpnI* and *SacI* restriction sites and incidentally a *PstI* site in pDB6 and pDB7. The overscored sequence represents the oligonucleotide used for the DNA sequence determination of the *KpnI* and *SacI* mutations in the *EcoR124I* and *EcoDXXI* *hsdS* genes and for the sequence confirmation of the *KpnI* and *SacI* junction sites in the four hybrid *hsdS* genes constructed in this work.

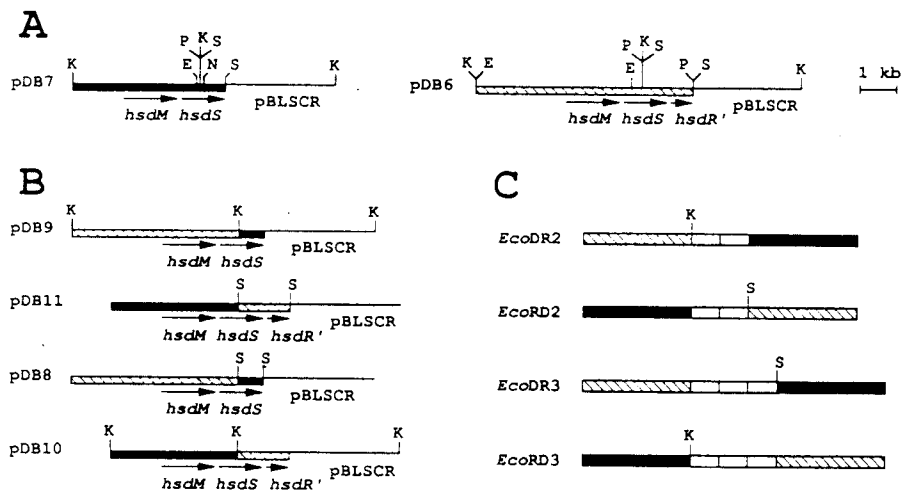


Fig. 3. Map of plasmids expressing wild-type and hybrid modification methylases and primary structure of the corresponding hybrid specificity polypeptides. (A) shows the plasmids pDB7 and pDB6 which contain the *EcoR124I* and *EcoDXXI* *hsdS/hsdM* operons, respectively, after insertion of *KpnI* and *SacI* sites. (B) shows plasmids pDB8, 9, 10 and 11 which are composed of a combination of *KpnI* or *SacI* fragments from pDB7 and pDB6. They express modification methylases with hybrid specificity polypeptides of which the structure is drawn schematically in (C) (not to scale). Black bars represent *EcoR124I* and hatched bars *EcoDXXI* sequences. Open bars indicate the repetitive elements in the hybrid HsdS polypeptides.

non-specific nucleotides, contains three 12 bp repeats in the centre of *hsdS*. We believe that this repetitive sequence codes for a structurally important constant polypeptide portion, previously assigned the function of an interdomain linker, which bridges two variable domains in Hsd specificity polypeptides (Gubler and Bickle, 1991).

Hybrid *hsdS* genes encode active specificity polypeptides

In an attempt to determine the potential of related R-M systems to vary their substrate specificity by, for example, *in vivo* recombination, and in order to identify the HsdS polypeptide domains responsible for target site specificity, we decided to recombine the *hsdS* genes of *EcoR124I* and *EcoDXXI* *in vitro* at specific positions. The *EcoR124I* and *EcoDXXI* *hsdM/hsdS* operons were subjected to two rounds of site-directed mutagenesis to introduce *KpnI* and *SacI* sites just in front of and immediately after the two and three 12 bp repeats in the centre of the *EcoR124I* and *EcoDXXI* genes, respectively. Figure 2 shows the DNA sequence spanning the central repetitive elements in the wild-type plasmid clones pMG1 and pMG4 and the sequence alterations in the mutant clones pDB7 and pDB6. The mutation creating a *SacI*

restriction site is silent, whereas the mutation creating a *KpnI* site, and incidentally also a *PstI* site, changed a leucine codon to a glycine codon. In addition, a spontaneous mutation converted the leucine codon in the first repetitive element of *EcoDXXI* *hsdS* to an arginine codon. The resulting plasmids pDB7 and pDB6, whose structure is depicted in Figure 3A, were then tested in phage infection assays for their ability to express active HsdM and HsdS polypeptides. In these assays, *EcoR124I* *hsdR* was provided in *trans* on plasmid pMG3 (Gubler and Bickle, 1991) which is compatible with pBluescript SK(+) clones. Double transformants of *E. coli* cells harbouring pMG3 plus a modification methylase plasmid were infected with non-modified phage λ vir and the level of restriction was measured as efficiency of plating (e.o.p.). In all cases, the e.o.p. was between 10^{-3} and 10^{-4} , indicating that there is no significant difference in restriction activity between cells harbouring the *EcoR124I* and *EcoDXXI* wild-type plasmids pMG1 and pMG4 and those harbouring the corresponding mutant plasmids pDB7 and pDB6. Therefore, the point mutations introduced in the *hsdS* genes did not affect the activity of the HsdS specificity polypeptides, as expected from previous studies showing that the activity of these

enzymes is relatively insensitive to the exact amino acid sequence in this region (Gubler and Bickle, 1991). This finding allowed the *SacI* or *KpnI* segments of pDB7 to be recombined with those of pDB6 to construct four different plasmids, each containing either the *EcoR124I* or the *EcoDXXI hsdM* gene plus a hybrid *hsdS* specificity gene. The structure of the resulting plasmids pDB8, 9, 10 and 11 is shown in Figure 3B and a schematic structure of the hybrid *hsdS* genes is depicted in Figure 3C. For example, the hybrid specificity gene of the newly generated *EcoDR2I* system was recombined at the *KpnI* site and it consists of *EcoDXXI* sequences on the 5' side and of *EcoR124I* sequences on the 3' side of the two central 12 bp repetitive elements. On the other hand, the *hsdS* gene of *EcoDR3I* was recombined at the *SacI* site so that it consists of the same *EcoDXXI* and

EcoR124I hsdS portions at its 5' and 3' ends as *EcoDR2I* but with three central 12 bp repeats instead of two. Similarly, the hybrid systems *EcoRD2I* and *EcoRD3I* were constructed with *EcoR124I* sequences on the 5' side and *EcoDXXI* sequences on the 3' side of the repetitive elements.

In phage infection assays, hybrid plasmids in the presence of *hsdR* were shown to confer restriction of phage infection to the same level as that exerted by the wild-type *EcoR124I* system (e.o.p. between 10^{-3} and 10^{-4}). This demonstrates that all four hybrid *hsdS* genes express specificity polypeptides which possess all the features required to recognize efficiently specific DNA sequences for methylation (of the host chromosome) and for restriction (of the DNA of an infecting phage).

Constant and variable modules of specificity polypeptides determine spacing and sequence in DNA sites

The DNA sites for *EcoR124I* and *EcoDXXI* had been determined previously: GAA(N₆)RTCG and TCA(N₇)ATTC, respectively (Price et al., 1987; Piekarowicz and Goguen, 1986). For the *EcoR124I* enzymes, the only invariant adenosine residue in the 3' half of the sequence is the site of methylation and the second A of the GAA is probably methylated because modification blocks *EcoRI* cleavage (Price et al., 1987). It is not known which adenines are methylated by *EcoDXXI*. To determine the target site specificity of the newly constructed hybrid R-M systems, we used a procedure based on the fact that modification methylases can be used to incorporate radiolabelled methyl groups into their target sites in DNA of known sequence. However, to avoid mapping the radioactive label to small regions of natural DNA substrates by restriction analysis and fluorography, we synthesized a short 45 bp double-stranded DNA fragment and cloned it into the vector pGEM-1. The resulting recombinant plasmid

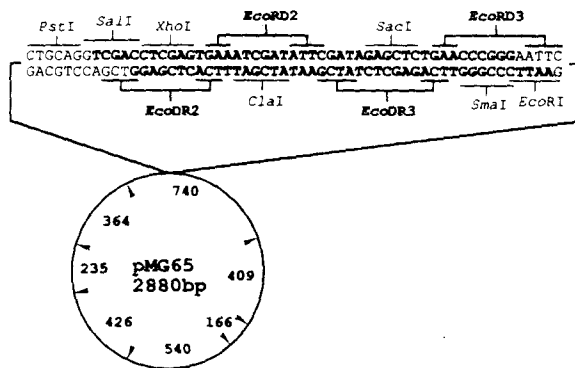


Fig. 4. Structure of plasmid pMG65 used as a substrate for *in vitro* methylation assays. The sequence of the double-stranded 45 bp oligonucleotide insert is in bold letters. The proposed bipartite recognition sites for hybrid modification methylases are over- or underscored by solid lines and sites for type II restriction sites are indicated with thin lines. *DdeI* cleavage sites in the vector pGEM-1 are shown by arrowheads and the size of the *DdeI* fragments is indicated in bp.

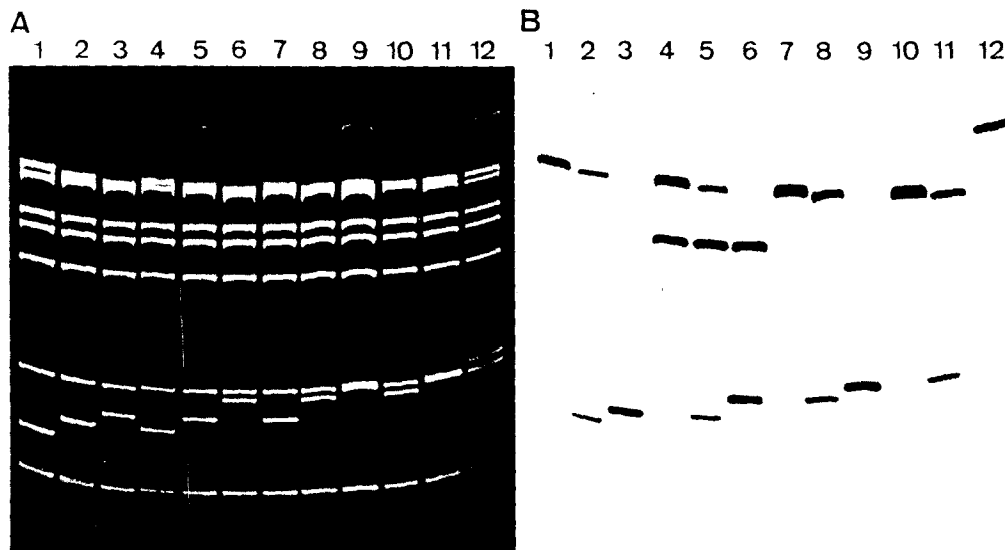


Fig. 5. *In vitro* methylation of pMG65 DNA with purified hybrid modification methylases. The DNA was labelled with [³H-methyl]AdoMet in the presence of the methylases *EcoDR2I* (lanes 1–3), *EcoRD2I* (lanes 4–6), *EcoDR3I* (lanes 7–9) and *EcoRD3I* (lanes 10–12). (A) shows an ethidium bromide stained 5% polyacrylamide gel. (B) shows the fluorogram of the same gel. The samples in every lane had been cleaved with *DdeI* plus one of the restriction enzymes listed below cutting the DNA into discrete radioactive fragments whose size is given in (bp): lane 1. *PstI* (544); lane 2. *XhoI* (536 and 204); lane 3. *ClaI* (215); lane 4. *XhoI* (536 and 409); lane 5. *ClaI* (525, 409 and 215); lane 6. *SalI* (409 and 232); lane 7. *ClaI* (525); lane 8. *SacI* (508 and 232); lane 9. *SmaI* (240); lane 10. *SacI* (508); lane 11. *SmaI* (500 and 240); lane 12. *EcoRI* (740).

pMG65 is shown in Figure 4. The synthetic DNA fragment flanked by *SalI* and *EcoRI* restriction sites was designed such that it contains four bipartite recognition sequences which represent all possible combinations of *EcoRI* 24I 5' and 3' half-sites with *EcoDXXI* 3' and 5' half-sites, respectively, separated by either six or seven non-specific nucleotides. Furthermore, to the left and right of the bipartite recognition sequences and in the middle within the non-specific spacer nucleotides, there are unique sites for type II restriction enzymes.

pMG65 DNA was methylated *in vitro* with the four purified hybrid modification methylases using [³H-methyl] AdoMet as methyl donor as described in Materials and methods. Aliquots of the modified plasmid DNA samples were then digested with the complete set of restriction enzymes for which there is a unique site in the synthetic 45 bp insert of pMG65 (Figure 4). In addition, all the samples were digested with *DdeI* to produce fragments small enough for analysis by PAGE. Figure 5 shows the polyacrylamide gel and the fluorogram of a set of selected digestions of modified pMG65 DNA. The fragment pattern and the intensity of the signals on the fluorogram unmistakably allow the determination of the site of methylation in the insert of pMG65. For example, the pMG65 DNA in lanes 1, 2 and 3 had been modified by the hybrid modification methylase *EcoDR2I*. Cleavage of this DNA with *PstI* and *DdeI* produced eight fragments of which only the largest, the 544 bp fragment, was labelled by [³H]methyl groups (lane 1 in Figure 5A and B). However, cleavage of the same DNA with *XhoI* and *DdeI* gave rise to two labelled fragments of 536 bp and 204 bp (lane 2),

each showing ~50% of the signal intensity in lane 1. Cleavage of the DNA with *Clal* and *DdeI* (lane 3) produced again only one single radioactively labelled band of 215 bp of about the same intensity as the signal in lane 1. Thus, we concluded that the *EcoDR2I* restriction site is flanked by *PstI* and *Clal* sites and that the *XhoI* site lies between the target half-sites and separates the two adenosine residues which can be modified with [³H]methyl groups. Based on this observation, we propose that the *EcoDR2I* R-M system recognizes the sequence TCA(N₆)GTCC which occurs once in the pMG65 sequence between the *PstI* and *Clal* sites and is split into two by *XhoI* (Figure 4). In a similar way, the restriction fragment and methylation pattern in lanes 7, 8 and 9 of Figure 5A and B allowed the mapping of the *EcoDR3I* recognition sequence. In pMG65 this sequence was found to be flanked by *Clal* and *SmaI* sites and split into two by *SacI*; we propose that it reads TCA(N₇)ATCC (Figure 4). These results make it clear that the amino-terminal domain of the *EcoDXXI* wild-type enzyme must recognize the 5' half of its recognition site TCA(N₇)ATTC whereas the carboxy-terminal domain of *EcoRI*24I must recognize the 3' half of its recognition site GAA(N₆)RTCC. Furthermore, the hybrid specificity polypeptides *EcoDR2I* and *EcoDR3I*, which differ only in one additional repetitive element, obey the rule that the number of repetitive elements (two or three) determines the number of non-specific nucleotides (six or seven) between the halves of split recognition sequences.

When we used the enzymes *SacI*, *SmaI* and *EcoRI* to map the *EcoRD3I* recognition sequence in pMG65 (lanes 10, 11 and 12 in Figure 5A and B), we observed in lane 12 that *EcoRI* did not completely cleave the labelled 730 bp *DdeI* fragment. This indicated that the 3' half of the *EcoRD3I* recognition site overlapped with the *EcoRI* site so that DNA methylated by *EcoRD3I* was not cleavable by *EcoRI*. However, only a small portion of the total DNA had been modified by *EcoRD3I* since a substantial amount of the 740 bp fragment was cleaved by *EcoRI*, resulting in two non-radioactive fragments of 497 bp and 243 bp. Despite the low efficiency of the *in vitro* methylation reaction, the fragment and signal pattern in Figure 5 supports the assumption that *EcoRD3I* recognizes the sequence GAA(N₇)ATTC.

When pMG65 DNA modified by *EcoRD2I* and cleaved by *DdeI* was digested with either *XhoI*, *Clal* or *SacI* (lanes 4, 5 and 6 in Figure 5A and B) the following fragments were radioactively labelled: a fragment of 536 bp (lane 4), fragments of 525 bp and 215 bp with half the intensity (lane 5), and a fragment of 232 bp (lane 6). This confirmed that the recognition sequence of *EcoRD2I* reads GAA(N₆)ATTC which is located between the *XhoI* and *SacI* sites in pMG65 DNA and can be split into two halves by *Clal*. Surprisingly, an additional signal was present in lanes 4, 5 and 6 on the fluorogram on Figure 5B, indicating the *EcoRD2I* recognized and methylated at least one more site in the 409 bp *DdeI* fragment of pMG65. This was in contradiction to what we had expected because pGEM-1, the vector of pMG65, does not contain the proposed *EcoRD2I* target site, according to the DNA sequence information provided by the supplier.

Except for this discrepancy, the experimental data from the *in vitro* methylation of pMG65 with the hybrid modification enzymes clearly demonstrate the modular

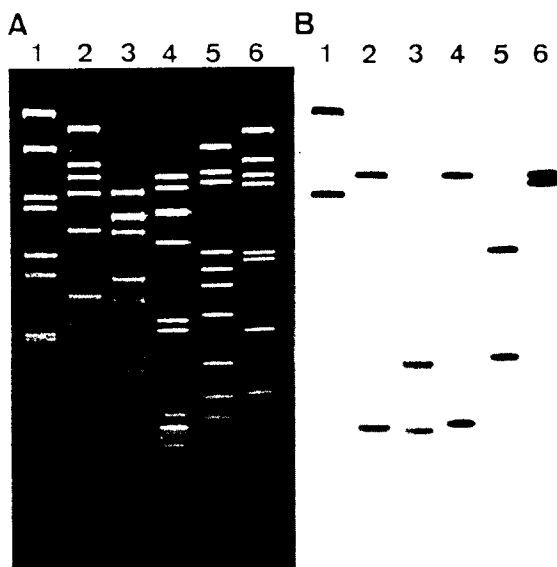


Fig. 6. Mapping of the second *EcoRD2I* target site in plasmid pMG65. (A) shows a 5% polyacrylamide gel of pMG65 DNA after *in vitro* labelling with [³H-methyl]AdoMet in the presence of the *EcoRD2I* modification methylase. Prior to gel electrophoresis, the DNA samples had been cleaved with one or two of the enzymes listed below giving rise to the radioactive fragments shown on the fluorogram in (B). The extensions of the radioactive fragments in the 2880 bp sequence of pMG65 DNA are given in parentheses for every single digest: lane 1. *ApaI* + *DdeI* (2704-564 and 603-937); lane 2. *BbvI* (7-105 and 204-623); lane 3. *Fnu4HI* (16-114 and 489-632); lane 4. *HhaI* (2596-138 and 536-636); lane 5. *MspI* (2676-63 and 496-643); lane 6. *HaeIII* (2609-130 and 333-767).