J. Mol. Biol. (in press)
Accepted 15/7/94

A Symmetrical Model for the Domain Structure of Type I DNA Methyltransferases

G. G. Kneale

Biophysics Laboratories
University of Portsmouth
Portsmouth PO1 2DT,
U.K.

SUMMARY

Type I DNA methyltransferases are complex multisubunit enzymes which methylate a specific base in each half of an asymmetric bipartite DNA recognition sequence. The specificity (S) subunit contains two corresponding DNA sequence recognition domains, plus a number of conserved regions which interact with two modification (M) subunits to form a trimeric enzyme of the form M_2S. The way in which the subunits interact with DNA in a pseudo-symmetric fashion has long been unclear. Analysis of internal sequence repeats in the S-subunit shows the occurence of significant homologies between the central conserved domain and sequences near the N- and C-termini. On the basis of this "split repeat", a "circular" organisation of the domains of this subunit is proposed that provides the required symmetry for interacting with the M-subunits and with the target DNA sequence. In the proposed model, one M-subunit interacts with the N- and C-terminal conserved regions of the S-subunit, which are thereby brought into close proximity. The second M-subunit makes equivalent contacts with repeated sequences in the central conserved domain. The model suggests a more general scheme for the imposition of pseudo-dyad symmetry on protein subunits that have internal repeats by making equivalent contacts with additional subunits.

Keywords: DNA-protein interactions, restriction-modification, methyltransferase, domain organisation

Type I DNA methyltransferases are complex multisubunit enzymes which methylate a specific base in each half of an asymmetric bipartite DNA recognition sequence [Bickle and Kruger , 1993, Wilson and Murray, 1991]. The specificity (S) subunit contains two corresponding DNA sequence recognition domains, plus a number of conserved regions which are believed to interact with two modification (M) subunits to form a trimeric enzyme of the form M_2S [Taylor et al., 1992; Dryden et al., 1993]. An additional subunit (R) is required to form the endonuclease but the precise stoichiometry of this enzyme has never been firmly established. The specificity (S) subunit of type I restriction-modification systems has been a the focus of considerable interest, since it is responsible for the complex DNA recognition properties of these enzymes. Three related families of type I systems have been identified to date and a number of sequences within the S-genes are highly conserved within each family [Gough, and Murray, 1983;Kannan et al., 1989; Tyndall et al., 1994]. It is believed that at least some of these conserved domains play a role in maintaining contacts with the M-subunit, the sequences of which are also strongly conserved within a family [Sharp et al., 1992].

The S-polypeptides of all type I systems possess a central conserved sequence separating the two (variable) DNA recognition domains, plus at least one additional conserved domain near the N- or C-terminus (Fig.1). The central conserved domain has been proposed to act as a spacer between the two variable domains, and thus to determine the length of the non-specific base sequence that separates the two half-recognition sites in the target DNA sequence [Price et al., 1987]. The precise role of the additional conserved domains is not known, but there is evidence that they also interact with the M-subunit [Cooper and Dryden, 1994].

Domain swapping experiments have clearly shown that the two variable domains of the S-subunit are responsible for determining specificity [Fuller-Pace and Murray , 1986; Gubler et al., 1992], each being responsible for the recognition of one half of the DNA target site located approximately one helical turn apart. The M-subunits of the methyltransferase must be positioned close to each of the DNA recognition sites, matching the pseudo-dyad symmetry of the DNA backbone, but cannot make base-specific contacts (except, perhaps, for the adenine that is always methylated). The precise location of the M-subunits on the DNA sequence must therefore be determined indirectly by interactions with the S-subunit.

The detailed organisation of the conserved domains has been difficult to visualise, given the

stoichiometry (M_2S) of the multisubunit methylase [Taylor et al., 1992; Dryden et al., 1993], and the likely requirements for partial dyad symmetry of the system. Furthermore, the insertion of various amino acids into the "spacer" region of the S-subunit of *Eco*R124I gives recognition properties that vary in a non-predictable way [Gubler and Bickle, 1991], suggesting that this domain cannot be a simple alpha helical linker between the DNA recognition domains as previously thought . It has been suggested that the required symmetry could result from interactions between the two M-subunits [Abadjieva et al., 1993]. However, extensive interactions between M-subunits seems unlikely since they do not dimerise in the absence of the S-subunit [Taylor, 1992]. Thus the required symmetry is most likely to be provided by the S-subunit, even though it is present in only one copy in the methyltransferase.

It was first shown for type IA systems that two regions of the S-subunit (ca. 60 amino acids long) are related by approximate translational symmetry, one in the central conserved domain and one within the C-terminal conserved domain [Argos, 1985]. A similar repeating structure is apparent in type IB systems, but here the major repeat is near the N-terminus, although there is also a smaller repeat (ca. 20 amino acids) close to the C-terminus [Kannan et al., 1989]. In type IC systems, the main repeat is near the C-terminus [Tyndall et al., 1994]. However, further investigation of internal sequence homologies in type IC systems shows that there is an additional small repeat within the N-terminal conserved region of the sequence (Fig. 2a). This N-terminal sequence (D') corresponds to a region of the central conserved domain (D) which is lacking from the C-terminal conserved domain , thus forming a "split repeat" (Fig. 2b).

This "split repeat" suggests that if the C-terminus were linked to the N-terminus, the components of the repeating sequences would then be effectively continuous; moreover, the conceptual "circularisation" of the sequence means that the conserved domains are then related by approximate dyad symmetry (Fig. 3). What might this mean in structural terms? Fig. 4(a) shows how the conserved regions of the S-subunit can now make pseudo-symmetrical contacts with each of the M-subunits. There need be no direct interaction between the N- and C-terminal repeats of the S-subunit; they are held in proximity by specific interactions made with the M-subunit, equivalent to those made by the central conserved domains with the second M-subunit. Presumably the extra few amino acid residues that are present at the N- and C- termini of S do not contact M and thus do not interfere with this interaction.

Superficially, the situation resembles the circular organisation of secondary structure elements found in

certain proteins [Luger et al., 1989]. However, the "circular" structure proposed by Luger et al. is maintained by interactions between adjacent secondary structure elements within a single domain; in the model proposed for the S-subunit, it is the domains themselves that are arranged in a circular fashion. Moreover, the circular organisation proposed for the S-domains is created by their interaction with other components (i.e. the M-subunits), thus converting translational symmetry (direct repeats) into rotational symmetry (a pseudo-dyad).

One region of the central domain (B) shows conservation between related enzymes, but there is little homology to the equivalent region in the C-terminal domain (B'). The lack of symmetry suggests that this linker region does not make any specific contacts with M, and can be envisaged as an "elbow" joint within the two conserved "arms". Since extra amino acids (of various types) can be inserted at this site which can affect the length of the non-specific spacer in the target sequence [Gubler and Bickle, 1991] and in some cases lead to "promiscuous" recognition (i.e.variation in the length of the non-specific spacer of the DNA recognition sequence), this "elbow" joint must have a certain degree of flexibility. There must be a matching flexibility in the M-subunit that allows some tolerance in the precise positioning of the contact regions. That no equivalent insertion is required in the homologous domain (B') reflects the fact that there is additional flexibility built in here, since in this arm the C' and D' domains are not covalently linked. It has been noted that promiscuous recognition with respect to spacer length in the DNA recognition sequence is more pronounced for methylation activity than for restriction activity, and thus that the R subunit (only found in the endonuclease) may interact with the "elbow" [Gubler and Bickle, 1991].

The proposed model fits in well with the observation that mutants of the S-subunit which are terminated shortly after the central conserved domain are able to dimerise to recognise a DNA target sequence with dyad symmetry, as recently found for both *Eco*DXXI and *Eco*R124I [Meister et al., 1993; Abadjieva et al., 1993]. The pseudo-dyad symmetry proposed for the conserved domains of S leads to a simple rationalisation of the phenomenon, without the need to invoke interactions between the two M-subunits (Fig. 4b). It would seem that the additional region of the sequence (D') at the N-terminus of S does not interfere with the interaction, and does not appear to increase the distance between the two variable domains, as the distance between the two methylated adenines remains the same in both cases. Indeed, it was shown that limited deletions could be made into the central conserved region of the HsdS gene of EcoDXXI without loss of function [Meister et al., 1993]. Deletion mapping showed that

function is retained even if the polypeptide is truncated at Trp203. Thus it can be concluded that at least the last four amino acids in the D domain are not essential in the truncated mutant. However, deletions as far as Arg190 (well into the C domain) are non-functional. For these mutants, the D domain should be redundant , since its role can be performed by the repeated D' domain at the N-terminus of the symmetry-related subunit. On the basis of the proposed model (Fig. 4b), one would predict that the entire D domain (or indeed the equivalent D' domain) would be dispensible, but that deletion of both these domains would not be tolerated.

The repeat previously identified in type IB sequences [Kannan et al., 1989] can also be regarded as a split repeat, although the major component is now found at the N-terminus. The conserved domains can therefore be arranged in an analogous way, except that the "break" in the second conserved domain would now be at a different position (shortly after the second variable domain). There is no clear N-terminal conserved domain in the S-subunits of the type IA family. Nevertheless, there is a sequence close to the N-terminus in *Eco*KI which, although not conserved within the IA family, shows some internal homology to a sequence at the C-terminal side of the central conserved region (see Fig. 2c). There is a stronger sequence homology in *Eco*BI, which is further from the N-terminus but is the correct distance away from its repeated sequence. Thus the same considerations of symmetry would suggest that the general features of the model may apply for all type I systems.

One component of the conserved regions, the A-repeat, appears to be conserved between all type I systems [Kannan et al., 1985; Tyndall et al., 1994], even though their M-subunits are very different [Sharp et al., 1992]. The function of this sequence could therefore be to interact with features common to all M-subunits, such as the S-adenosyl methionine binding site. It should be noted that the A domain is invariably positioned adjacent to each of the DNA recognition domains and would be in a suitable position to perform such a role.

Acknowledgements

I wish to thank the Wellcome Trust for financial support, and Dr I.A. Taylor , Prof. T.A. Bickle and Prof. N.E. Murray for helpful comments on the manuscript.

References

Abadjieva, A., Patel, J ., Webb, M., Zinkevich, V. and Firman, K. (1993) A deletion mutant of the Type IC restriction endonuclease *Eco*R124I expressing a novel DNA specificity. *Nucl. Acids. Res.* **21**, 4435-4443

Argos, P. (1985) Evidence for a repeating domain in type I restriction enzymes. *EMBO J.* **4**, 1351-1355

Bickle, T.A. and Kruger, D.H. (1993) Restriction endonucleases and modification methylases. *Microbiol. Rev.* **57**, 434-450

Cooper, L.P. and Dryden, D.T.F. (1994). The Domains of a Type I DNA Methyltransferase. Interactions and Role in Recognition of DNA Methylation. *J. Mol. Biol.* 236, 1011-1021

Dryden, D.T.F., Cooper, L.P. and Murray, N.E. (1993) Purification and characterisation of the methyltransferase from the type I restriction and modification system of E.coli K12. J. Biol. Chem. 268, 13228-13236.

Fuller-Pace, F.V and Murray, N.E. (1986) Two DNA recognition domains of the specificity polypeptides of a family of type I restriction enzymes. *Proc. Natl. Acad. Sci.* **83**, 9368-9372

Gough, J.A. and Murray, N.E. (1983) Sequence diversity among related genes for recognition of specific targets in DNA molecules. *J. Mol. Biol.* **166**, 1-19

Gubler, M., Braguglia, D., Meyer, J. Piekarowitz, A., and Bickle, T.A. (1992) Recombination of constant and variable modules alters DNA sequence recognition by Type IC restriction-modification enzymes. *EMBO J.* **11**, 233-240

Gubler, M. and Bickle, T.A. (1991) Increased protein flexibility leads to promiscuous protein-DNA interactions in type IC restriction-modification systems. *EMBO J.* **10**, 951-957

Kannan, P., Cowan, C.M., Daniel, A.S., Gann, A.A.F., and Murray, N.E. (1989) Conservation of organisation in the specificity polypeptides of two families of Type I restriction enzymes. *J. Mol. Biol.* **209**, 335-344

Luger, K., Hommel,U., Herold, M., Hofsteenger,J. and Kirschner,K. (1989) Correct Folding of Circularly Permuted Variants of a beta/alpha Barrel Enzyme in Vivo. Science 243, 206-210

Meister, J., McWilliams, M., Hubner, P., Jutte, H., Skrzypek,E., Piekarowicz,A. and Bickle, T.A. (1993) Macroevolution by transposition: drastic modification of DNA recognition by a type I restriction enzyme following Tn5 transposition. *EMBO J.* **12**, 4585-4591

Price,C., Lingner, J., Bickle, T.A., Firman,K. and Glover, S.W. (1987) Basis for changes in DNA recognition by the *Eco*R124 and *Eco*R124/3 type I DNA restriction and modification enzymes *J. Mol. Biol.* **205**, 115-125

Sharp, P.M., Kellerher, J.E., Daniel, A.S., Cowan, G.M. and Murray, N.E. (1992) Roles of selection and recombination in the evolution of type I restriction-modification systems in enterobacteria. *Proc. Natl. Acad. Sci.* **89**, 9836-9840

Taylor, I.A. (1992) Biochemical characterisation of the EcoR124 Type I DNA modification methylase.*Ph.D thesis*, University of Portsmouth, U.K.

Taylor , I., Patel, J., Firman, K. and Kneale, G.G. (1992) Purification and biochemical charac terisation of the *Eco*R124 type I modification methylase. *Nucl. Acids Res.* **20**, 179-186

Tyndall, C., Meister, J., and Bickle, T.A. (1994) The *E.coli prr* region encodes a functional type IC DNA restriction system closely integrated with an anticodon nuclease gene. *J. Mol. Biol.* **327**, 266-274

Wilson, G.G. and Murray, N.(1991) Restriction and modification systems. *Annu. Rev. Genet.* **25**, 585-627

Figure Captions

Fig. 1 The arrangement of conserved and variable domains in the S-subunits of type IC R-M systems. The figure shows the domains of the type IC system *Eco*R124I, together with its cognate DNA recognition sequence. Types IA and IB have a similar structure, but type IA's have no clear conserved domain at the N-terminus, whilst type IB's have a longer N-terminal domain and only a small C-terminal domain.

Fig. 2 **(a)** Dot matrix plot showing internal sequence homologies in the amino acid sequence of the S-subunit of *Eco*R124I, using an 8 residue window and a threshold of 5 identical matches. The off-diagonal regions indicating direct repeats are labelled A, C and D; (B corresponds to the sequence linking A and C domains, which is conserved between all type IC's but shows weak homology internally). **(b)** The amino acid residues making up the sequence repeats are shown for the type IC enzyme *Eco*DXXI. The repeating structure of *Eco*R124I is similar but the A repeat is shorter. **(c)** "Split repeat" sequences for type IC (*Eco*DXXI, *Eco*R124I) and type IA(*Eco*KI, *Eco*BI) enzymes, showing homology between N- terminal and central conserved sequences. Amino acid identities are denoted • and homologous residues +. The sequence R190-W203 defines a critical region of the central conserved sequence in EcoDXXI mutants, as defined by deletion mapping (Meister et al.,1993).

Fig. 3 **(a)** Organisation of the repeating sequences in the S-subunits of type IC enzymes. The domains indicated correspond to those in Fig. 2. **(b)** A "circular" model for the organisation of the conserved domains in the S-subunit, locating the N- and C-terminal domains in close proximity.

Fig. 4 **(a)** Schematic model for the methyltransferase, showing how the two identical M-subunits can impose pseudo-dyad symmetry on the S-subunit by interacting with the conserved (repeated) domains which link the two variable DNA recognition domains, V1 and V2. **(b)** Two copies of a truncated S-subunit can make equivalent contacts with two M-subunits (not shown) to form a perfectly symmetrical enzyme which will recognise a palindromic DNA sequence.

Fig. 1

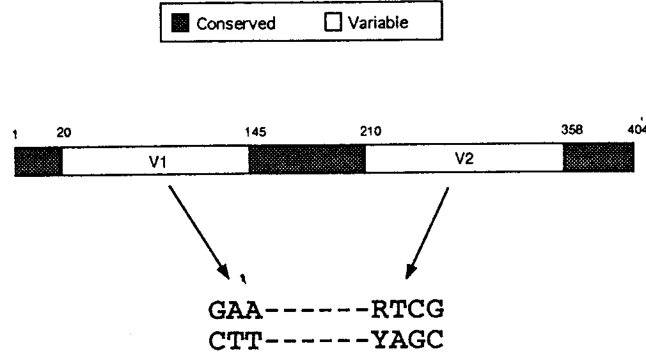
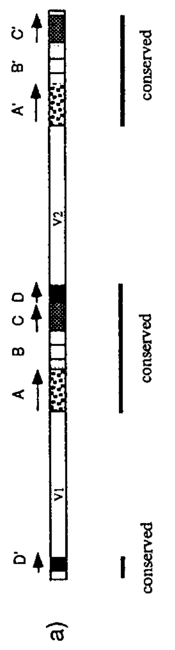


Fig. 3



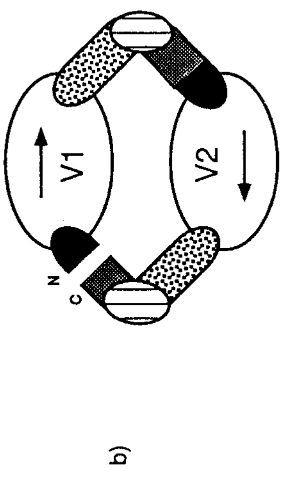
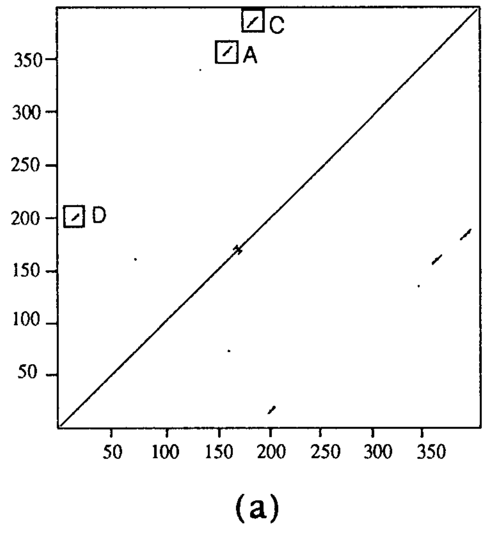


Fig. 2



(a)

```
EcoDXXI     12-22    DGVEVEWVTLG
                     +• •••• •••
            198-207  EG-EVEWKTLG

EcoR124I    11-23    DGVEVEWLPLGEI
                     +• ••••  ••••
            197-208  EG-EVEWKTLGEI

EcoKI       4-10     GKLPEGW
                     •  • • •
            207-213  GTLGEKW

EcoBI       52-64    KDGVPLIRIRDV
                     •••+ •••+ •+
            262-274  KDGIELIRVCDI
```

(c)

```
                 A                      B                      C                D
(b)  CENTRAL  PIPSPDNPEKSLAIQSEIVRILDTFTALTAELTAELTAELNMRKKQYNYYRDQLL-SF-KEG-EVEWKTLG
              ••• •  • •••  • •• •••  •  •  •  •  •  •  • ••• •••• •• •  •••• •••
     C-TERM   PIPYPSHPQKSLDEQGRIVDILDKFDAIAASITEGLPREIELRQKQYEYYRD-LLFSFPK.GVEVEWVTLG  N-TERM
              342-366                                  384-400            12-22
```

Fig. 4