Volume 17 Number 23 1989 Nucleic Acids Research

# Sequence motifs characteristic of DNA[cytosine-N4]methyltransferases: similarity to adenine and cytosine-C5 DNA-methylases

Saulius Klimašauskas, Albertas Timinskas, Saulius Menkevičius, Danguolė Butkienė, Viktoras Butkus and Arvydas Janulaitis*

Institute of Applied Enzymology, Fermentu 8, 232028 Vilnius, Lithuania, USSR

Received August 2, 1989; Revised and Accepted November 7, 1989 EMBL accession nos X16985, X17022

## ABSTRACT

The sequences coding for DNA[cytosine-N4]methyltransferases *Mva*I (from *Micrococcus varians* RFL19) and *Cfr*9I (from *Citrobacter freundii* RFL9) have been determined. The predicted methylases are proteins of 454 and 300 amino acids, respectively. Primary structure comparison of M.*Cfr*9I and another m4C-forming methylase, M.*Pvu* II, revealed extended regions of homology. The sequence comparison of the three DNA[cytosine-N4]-methylases using originally developed software revealed two conserved patterns, DPF-GSGT and TSPPY, which were found similar also to those of adenine and DNA[cytosine-C5]-methylases. These data provided a basis for global alignment and classification of DNA-methylase sequences. Structural considerations led us to suggest that the first region could be the binding site of AdoMet, while the second is thought to be directly involved in the modification of the exocyclic amino group.

## INTRODUCTION

DNA-methylases (MTases) are enzymes that transfer methyl groups from the donor S-adenosylmethionine (AdoMet) onto the adenine or cytosine residues within the sequence they recognize in DNA. Among the roles they play in cells of prokaryotes are protection against the restriction enzymes and DNA mismatch repair [1,2].

Site-specific DNA modification in bacteria usually leads to the formation of three kinds of products: N6-methyladenine (m6A), 5-methylcytosine (m5C) [2] and N4-methylcytosine (m4C). The latter type of modification has only recently been discovered [3,4]. Therefore, m4C-forming MTases despite of their wide occurence [5,6] have not yet received an exhaustive structural characterization. Among MTases for which primary structures have been published only M.*Pvu*II is known to catalyze the formation of m4C [7,8].

Prokaryotic MTases as well as restriction endonucleases due to their relatively simple structural organization and exquisite recognition specificity are indispensible model systems for investigating site-specific DNA-protein interaction. A prerequisite for a somewhat detailed study of any enzyme is the knowledge of its primary structure, hence cloning and sequencing of genes are necessary steps in that direction. Moreover, some efforts were directed at identifying the modification and target recognition domains (TRDs) of MTases on the basis of sequence data only. Sequence comparisons revealed a number of highly conserved regions to exist among cytosine-C5 MTases [9,10]. A limited degree of similarity was found among adenine MTases as well [11−14], however, no clear structural similarity between the sequences of these two types has been detected.

In this paper we present data on sequencing and analysis of the DNA region coding for the MTases M.*Cfr*9I and M.*Mva*I. Both enzymes are components of restriction-

 

modification systems of bacterial strains *Citrobacter freundii* RFL9 and *Micrococcus varians* RFL19, respectively; they transfer a methyl group onto the 4-amino moiety of the second cytosine residue of the recognition sequences (CCCGGG [4,15] and CCWGG [16]). The primary structure comparison revealed a strong similarity of M.*Cfr*9I to another m4C-forming MTase, M.*Pvu*II. The multiple sequence comparison revealed two types of homology regions common for adenine and cytosine amino MTases, which were found to have analogues within cytosine-C5 MTase sequences [10]. These data provided a basis for global alignment and classification of MTases.

## MATERIALS AND METHODS

A portion of DNA sequencing was performed according to the modified method of Maxam and Gilbert [17]. The plasmids were cleaved with an appropriate restriction endonuclease (Fermentas, Vilnius), end-labeled with $^{32}P$ (Izotop, Tashkent) and used for sequencing by the solid-phase chemical modification procedure. The dideoxynucleotide chain-termination procedure was also used; double-stranded, supercoiled plasmid DNA was used as the template [18]. Primers for sequencing were produced in our laboratory on 'GENE ASSEMBLER' (Pharmacia, Sweden) using methylphosphoramidite chemistry. These were deblocked and purified by reversed-phase HPLC. The sequencing reactions were carried out with Sequencing kit (Pharmacia, Sweden) and [$\alpha$-$^{33}P$]dATP (Izotop, Leningrad). The reactions were resolved by electrophoresis on wedge-shaped gels.

Computer programs DM5, NUCALN, PCSEARCH [19] were used to manage the sequence data.

Sequences for the following bacterial MTases which are known to form m5C were used: *Bsp*RI (recognition sequence GGCC) [20], *Bsu*RI (GGCC) [21], *Dde*I (CTNAG) [22], *Hha*I (GCGC) [23], *Eco*RII (CCWGG) [8], *Sin*I (GGWCC) [24], *Msp*I (CCGG) [25]. Multispecific MTases Phi3T (GGCC, GCNGC) [26] and SPR (GGCC, CCGG, CCWGG) [27] are encoded by *Bacillus* phages. Sequences were used for the following bacterial MTases which catalyze the formation of m6A: *Eco*RI (GAATTC) [28], *Eco*RV (GATATC) [29], *Hha*II (GANTC) [30], *Dpn*II (GATC) [31], *Pae*R7 (CTCGAG) [32 ], *Pst*I (CTGCAG) [33], *Taq*I (TCGA) [34], *Eco*dam (GATC) [35], *Hinf*I (GANTC) [36], *Eco*P1 (AGACC) and *Eco*P15 (CAGCAG) [37], *Eco*KI (AACN$_6$GTGC) [38]. T4dam (GATC) [39] and *Cvi*BIII (TCGA) [40] are MTases encoded by the phage T4 and *Chlorella* virus respectively. The sequences for MTase *Eco*57I and for the cognate restriction endonuclease with the property to modify the target sequences (CTGAAG) [41] is a personal communication of R.Vaišvila.

A newly developed procedure was used for multiple sequence alignments. It is based on using a pairwise amino acid scoring matrix and a 'sliding window' [42] sequence comparison algorithm. The first string of a defined length L of the first protein is compared to every such span of all other sequences resulting in extraction the matches with a similarity score not less than a preselected criterion K. The score is obtained by summation of L values resulting from the matched amino acid pairs (250PAMs mutability matrix [43] as well as an arbitrary selected structure-based amino acid matrix were used). If the matches occur in not less than N sequences the program sends all these patterns to output. All other consecutive strings of first sequence are processed in the same manner. Thus, the procedure extracts the consensus and groups all sequences under investigation on the basis of their similarity to the first protein.

To test the significance of similarity between two sequences the program RDF [44] was used.

All programs were run on TDK 286 computer with 80287 math co-processor installed.

## RESULTS

*Nucleotide sequence analysis*

The cloned inserts carrying *cfr9IM* and *mvaIM* genes have been sequenced by both solid-phase chemical cleavage method [17] and dideoxynucleotide chain-termination approach [18]. The sequence for both strands has been determined (fig.1). M.*Cfr*9I deduced is a 300 amino acid protein of 33.8 kD. The coding region for M.*Mva*I encodes a protein of 454 amino acids and the predicted molecular weight of 53.1 kD. Open reading frames of both MTases are in good agreement with the results of deletion mapping experiments (not shown). The description of cloning and sequencing procedures as well as detailed analysis of gene organization of *Cfr*9I and *Mva*I restriction-modification systems will be published elsewhere.

*Amino acid sequence homologies*

Both predicted protein sequences were compared to the sequences of known restriction-modification enzymes. The first aim of the analysis was to compare the three m4C-forming MTases. The pairwise randomization tests have given the following RDF scores: *Cfr*9I−*Pvu*II-10 s.d.; *Pvu*II−*Mva*I-3.5 s.d.; *Cfr*9I−*Mva*I-2.5 s.d. This proves that despite unquestionable relatedness this group of sequences is quite diverse. All types of homology searches we carried out showed strong overall similarity of M.*Cfr*9I to M.*Pvu*II (fig.2). Among them there are four regions with varying degrees of similarity. Three of them are found in the amino-terminal half, while the fourth region covering over 45 amino acids is located near the carboxy-terminus. The homology regions are contiguously arranged, thus the MTase sequences can be readily aligned.

The application of the newly developed procedure for triple sequence homology searches proved that M.*Mva*I is somewhat different from the above MTases, since it has just two short regions of strong similarity with the aligned MTases (fig.2). A notable feature of M.*Mva*I is the opposite order of the domain arrangement, i.e. these domains are naturally swapped as compared to M.*Cfr*9I and M.*Pvu*II.

In order to detect common patterns in all MTase sequences two general approaches were used. The comparison of the conserved patterns of cytosine amino MTases to those known to date of adenine and cytosine-C5 MTases [10−14] was one of them. On other cases our comparison procedure for a larger group of MTase sequences was applied.

The consensus of the first conserved pattern found in the m4C-specific MTases is TSPPY (fig.2), while adenine MTases have the sequence (D,N)PPY instead [11−14]. Search for such a pattern among the invariant motifs of cytosine-C5 MTases [9,10] revealed a relationship with the fourth conserved motif (fig.3, domain **II**). There are several differences among these patterns as well as a number of invariant positions. In all cases the pattern is preceeded with a hydrophylic residue (fig.3, relative position **n**) which is predominantly D, followed by a triplet of hydrophobic amino acids (r.p. **o,p,q**). Such remarkable similarities in sequence parameters cannot be accidental. In other words, we believe that all three types of MTases possess a related structural building block.

The second conserved pattern of the cytosine-N4 MTases also appeared to have analogies in the MTase sequences of both other types. Clearly evident similarity was observed to

A

```
ACTAATTCAGCAATTAATTGATCAAATTTTTGACTTCTCGTCATAACTATTCCTCCTTGTTTTTTGATATGATAG
GGCATATACTATATATTACTGAATAATAAAAAGGTTTTTGTGTACTAATAGTGCACAAAAACCAAAGGAGACAAA

ATGGAATATTTAAATGATAAAGATCAACATTTAATTGATAAATTATCAAAAAAGATTAATGATAATAATCAATAT
 M  E  Y  L  N  D  K  D  Q  H  L  I  D  K  L  S  K  K  I  N  D  N  N  Q  Y

CTTGGTTTCTTAAATACAAATACTAAAGAATTAACTCATAGATATCATATTTATCCTGCTATGATGATCCCTCAA
 L  G  F  L  N  T  N  T  K  E  L  T  H  R  Y  H  I  Y  P  A  M  M  I  P  Q  50

TTGGCTAAAGAATTCATTGAATTAACTCAACAAGTAAAACCAGAAATCAAAAAATTATATGATCCTTTTATGGGC
 L  A  K  E  F  I  E  L  T  Q  Q  V  K  P  E  I  K  K  L  Y  D  P  F  M  G

TCTGGTACTTCTTTAGTAGAAGGACTTGCACATGGGTTGGAAGTATATGGAACAGATATAAATCCTCTATCACAA
 S  G  T  S  L  V  E  G  L  A  H  G  L  E  V  Y  G  T  D  I  N  P  L  S  Q 100

ATGATGAGTAAAGCTAAAACTACTCCTATAGAACCTTCAAAGTTGTCGAGAGCTATTTCAGATCTTGAATATTCT
 M  M  S  K  A  K  T  T  P  I  E  P  S  K  L  S  R  A  I  S  D  L  E  Y  S

ATAAGAGAAATGACAATTCTGTATCATGAGGGGAATTATAAAATAAGCAACCTTCCTGATTTTGATAGAATAGAT
 I  R  E  M  T  I  L  Y  H  E  G  N  Y  K  I  S  N  L  P  D  F  D  R  I  D 150

TTTTGGTTTAAAGAAGAAGTTATAATAAGTCTACAGTTAATAAAAAATTGCATAAATGAGTTTATAGAAGATGAT
 F  W  F  K  E  E  V  I  I  S  L  Q  L  I  K  N  C  I  N  E  F  I  E  D  D

TTGAAAACGTTCTTCATGGCAGCATTTAGTGAAACAGTTAGGCATGTTTCAAATACTCGTAATAATGAATTTAAA–
 L  K  T  F  F  M  A  A  F  S  E  T  V  R  H  V  S  N  T  R  N  N  E  F  K 200

CTGTATAGAATGGCACCTGAAAAATTAGAAATATGGAATCCGAATGTAACTGAAGAATTTTTAAAGAGAGTATAC
 L  Y  R  M  A  P  E  K  L  E  I  W  N  P  N  V  T  E  E  F  L  K  R  V  Y

AGAAATGAATTAGGCAATATGGATTTCTATAGACAACTTGAAAATGTAGGAAATTACTCGCCTAAAACTATAATA
 R  N  E  L  G  N  M  D  F  Y  R  Q  L  E  N  V  G  N  Y  S  P  K  T  I  I 250

AATAAGCAAAGCAACATAAAACTTCCAGAGGAATTTAAAGATGAAATGTTCGATATTGTAGTTACTTCTCCACCA
 N  K  Q  S  N  I  K  L  P  E  E  F  K  D  E  M  F  D  I  V  V  T  S  P  P

TATGGTGATAGTAAAACAACTGTAGCCTATGGGCAATTTTCAAGATTGTCCGCTCAATGGTTGGATCTGAAAATA
 Y  G  D  S  K  T  T  V  A  Y  G  Q  F  S  R  L  S  A  Q  W  L  D  L  K  I 300

GATGATGAGACTAAAATAAATCAATTAGATAATGTGATGCTTGGTGGAAAAACAGATAAAAATATTATTGTTAAT
 D  D  E  T  K  I  N  Q  L  D  N  V  M  L  G  G  K  T  D  K  N  I  I  V  N

GATGTGTTAGAATATCTCAATTCTCCAACGTCGAAATCAGTATTTAATTTAATAAGTCATAAAGATGAAAAAAGA
 D  V  L  E  Y  L  N  S  P  T  S  K  S  V  F  N  L  I  S  H  K  D  E  K  R 350

GCACTAGAAGTTCTTCAATTTTATGTTGATTTGGATAAATCTATTAAAGAAACTACAAGAGTGATGAAGCCCGAG
 A  L  E  V  L  Q  F  Y  V  D  L  D  K  S  I  K  E  T  T  R  V  M  K  P  E

TCATATCAATTTTGGGTAGTAGCTAATAGAACAGTAAAAATGATCAGTATACCAACTGATATTATAATTTCTGAG
 S  Y  Q  F  W  V  V  A  N  R  T  V  K  M  I  S  I  P  T  D  I  I  I  S  E 400

TTATTTAAAAAGTATAATGTTCATCATTTATATAGTTTCTATAGGAAAATCCCTAATAAACGTATGCCTTCAAAA
 L  F  K  K  Y  N  V  H  H  L  Y  S  F  Y  R  K  I  P  N  K  R  M  P  S  K

AATTCTCCTACTAATAAAATAGGTAATCATTCTGTTACCATGACTTCTGAGATTATATTAATGCTAAAAAATTAC
 N  S  P  T  N  K  I  G  N  H  S  V  T  M  T  S  E  I  I  L  M  L  K  N  Y 450

ATTAATAAAAGCTGATCTTCTTCAATCATGCTTACA
 I  N  K  S  *
```

B

```
                                GATTTAAAAGTTGTAGGTTGTTGCATGTCTGCATTGTGCGTGAGGAATATTT

ATGCCAAGTAAAAAGAGTAGTTCGCCGCTGAGTGTTGAGAAACTTCATCGTTCTGAGCCCTTGGAGTTGAACGGA
 M  P  S  K  K  S  S  S  P  L  S  V  E  K  L  H  R  S  E  P  L  E  L  N  G

GCTACCCTTTTTGAAGGTGATGCTCTGTCAGTATTGAGGAGACTTCCGAGCGGCTCAGTTCGGTGCATCGTCACT
 A  T  L  F  E  G  D  A  L  S  V  L  R  R  L  P  S  G  S  V  R  C  I  V  T  50

TCTCCGCCATACTGGGGGCTACGTGATTACGGCATAGATGAACAAATCGGTTTAGAAAGTAGCATGACTCAGTTT
 S  P  P  Y  W  G  L  R  D  Y  G  I  D  E  Q  I  G  L  E  S  S  M  T  Q  F

TTAAATCGTCTTGTTACGATCTTTTCTGAGGCGAAACGTGTATTGACTGACGACGGAACGCTATGGGTTAACATT
 L  N  R  L  V  T  I  F  S  E  A  K  R  V  L  T  D  D  G  T  L  W  V  N  I 100

GGTGATGGATATACAAGCGGAAATCGCGGGTATAGAGCTCCTGATAAGAAAAATCCGGCACGAGCTATGGCTGTT
 G  D  G  Y  T  S  G  N  R  G  Y  R  A  P  D  K  K  N  P  A  R  A  M  A  V
```

```
CGCCCGGATAGGCCAGAAGGACCAAAACCGAAGGATCTGATTGGGATTCCTTGGCGGTTAGCGTTCGCTTTGCAA
 R  P  D  R  P  E  G  P  K  P  K  D  L  I  G  I  P  W  R  L  A  F  A  L  Q 150

GAAGATGGGTGGTACCTACGAAGCGACATTGTTTGGAATAAACCTAACGCGATGCCTGAAAGTGTAAAAGACCGG
 E  D  G  W  Y  L  R  S  D  I  V  W  N  K  P  N  A  M  P  E  S  V  K  D  R

CCTACCCGTTCTCATGAGTTCCTTTTTATGCTGACCAAATCAGAGAAATATTATTACGATTGGGAAGCGGTGAGA
 P  T  R  S  H  E  F  L  F  M  L  T  K  S  E  K  Y  Y  Y  D  W  E  A  V  R 200

GAAGAAAAAGATAGCGGAGGTTTCAGAAATCGACGCACAGTATGGAATGTAAATACGAAACCTTTTGCAGGGGCT
 E  E  K  D  S  G  G  F  R  N  R  R  T  V  W  N  V  N  T  K  P  F  A  G  A

CATTTCGCAACATTCCCAACGGAGCTAATTCGTCCATGCATCTTGGCATCCACGAAACCTGGTGATTACGTATTA
 H  F  A  T  F  P  T  E  L  I  R  P  C  I  L  A  S  T  K  P  G  D  Y  V  L 250

GATCCATTCTTCGGCTCTGGTACTGTAGGCGTTGTATGCCAGCAGGAAGACCGCCAGTATGTTGGTATTGAACTC
 D  P  F  F  G  S  G  T  V  G  V  V  C  Q  Q  E  D  R  Q  Y  V  G  I  E  L

AATCCAGAATATGTTGATATAGCTGTAAATCGTTTGCAGGGTGAGGATACAAATGTGATAAGGATCGCGGCAGCA
 N  P  E  Y  V  D  I  A  V  N  R  L  Q  G  E  D  T  N  V  I  R  I  A  A  A 300

TGACTAATAAAATAGTTTTC
 *
```

**Figure 1.** DNA sequence of genes with predicted amino acid sequence for: **A**—*Mva*I methylase, **B**—*Cfr*9I methylase.

the I conserved region [11] of a group of adenine MTases, *Eco*dam, T4dam, *Dpn*II and *Eco*RV, possessing great overal similarity to each other. Inspection of conserved patterns of m5C-forming MTases [9,10] also revealed an apparent similarity of the first of them to this invariant region of cytosine-N4 MTases (see fig.3, domain **I**).

Application of our procedure for multiple homology detection to other adenine MTases resulted in extraction of analoguous motifs from the sequences of M.*Hinf*I, M.*Eco*PI, M.*Eco*P15 (fig.3). Detection of such pattern in the sequences of the rest adenine MTases appeared to be more complex. Only at lower stringency the programm yielded the patterns with however several deviations from the consensus. Thus, there is a poorer conservation at r.p.**d** and the lack of invariant F at r.p.**e** (fig.3, domain **I**). On the other hand, the F residue occures at r.p.**k** of the pattern. The other two structural circumstances witnessing to nonaccidental occurence of the extracted regions is the global preceeding of these patterns by a couple of hydrophobic amino acids (r.p.**a**,**b**) as well as the constant location of these regions before the second conserved domain.

## DISCUSSION

Thus, all MTase sequences analysed possess two necessary conserved patterns. The described structural relationship among MTases of various types of specificity could not be accidental and most probably it means of their common ancestry as well as of the generality in mechanisms by which these enzymes catalyze methyl transfer from AdoMet onto DNA.

The homology analysis data presented provides a basis for global alignment and classification of MTases. All sequences under consideration can be divided into four groups on the basis of the amino acid occuring at r.p.s of the homology domain **II**. Thus, all sequences that have S at this position are of m4C specificity, whereas those which carry G are m5C-specific ones. Adenine MTases have either D or N residue here. One of the most apparent differences between these two families is that the 'N-sequences' have the first domain deviating from the consensus, while 'D-sequences' are much more invariant in this respect (fig.3). As can be seen from the alignment scheme (fig.4) 'N'+'G' *versus*

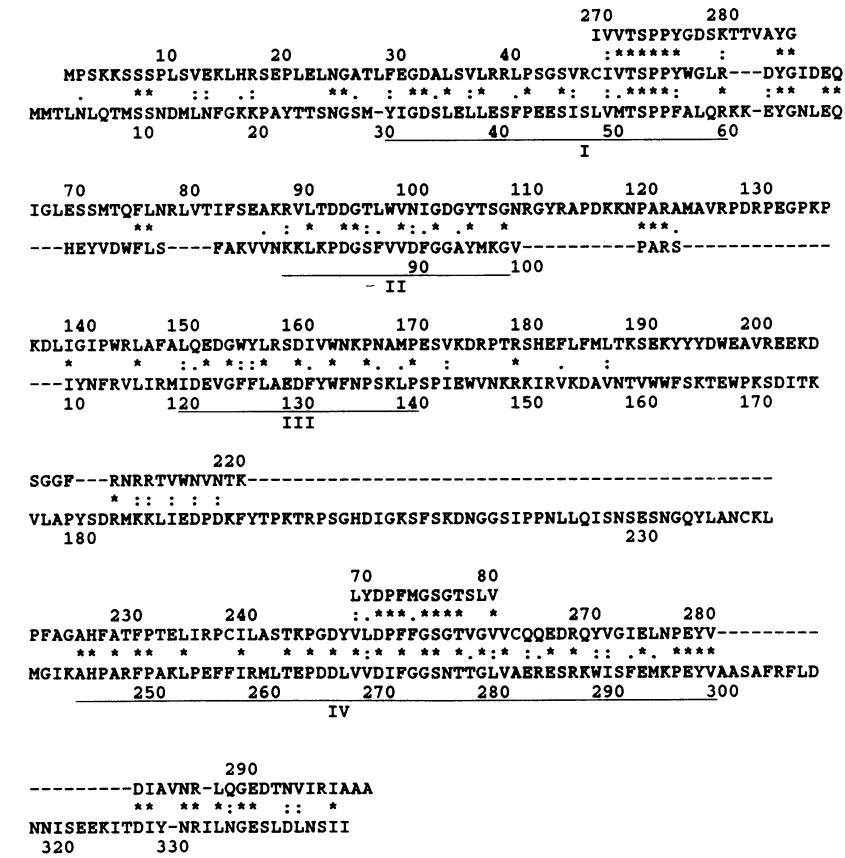


**Figure 2.** Alignment of *Mva*I (fragmental), *Cfr*9I (upper) and *Pvu*II (lower) DNA-methylase sequences. Designations: '*'—identity, ':'—functional similarity, '.'—functional compatibility. The regions of homology are underlined.

'S'+'D' groups do differ in a relative position of the homology regions within the whole sequence.

From our point of view the latter regularity might reflect a more general structural principle. It's already known that TRDs in m5C-generating MTases are contiguous sequences of 80−260 amino acids not overlapping with the invariant regions [43,44]. If this regularity is of general nature it means that the regions of homology **I** and **II** dissect all the sequences into three parts within which only the TRD could be located. On the other hand, the region of TRD location should be large enough to contain it. As can be judged from the alingment scheme (fig.4) the most suitable region for TRDs within the MTases of 'G' and 'N' families would be the C-terminal section, whereas for 'D' and 'S' families this most probably would be the middle section.

Indeed, there are several lines of evidence proving that in cytosine-C5 MTases the TRD is located after the **II** homology domain [10,45,46], as well as the data witnessing to its

| Domain No | | I | | | II | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Relative pos. | | ab cdefghi jkl | | | mn opq rstuvwxyz | | | | |
| *Cfr*9 I | 249 | VL DPTFGSG TVGV | 44 | SVR CIV TSPP | YWGLRD | | | | |
| *Pvu* II | 269 | VV DIFGGSN TTGL | 46 | SIS LVM TSPP | FALQRK | m4C | | | |
| *Mva* I | 69 | LY DPTMGSG TSLV | 267 | MFD IVV TSPP | YGDSKT | | | | |
| *Hha* II | 190 | IL DPASGGY *S*VFE | 30 | AVK IAF FDPN | YRGVLD | | | | |
| *Hinf* I | 196 | VL DPTFGTG TTGA | 29 | SID LIF ADPP | YFMNTD | | | | |
| *Eco*P1 | 432 | VL DFTAGSG TTAE | 116 | KVN MIY IDPP | YNTGKD | | | | |
| *Eco*P15 | 450 | IL DFTAGSG TTAH | 118 | KVN MIY IDPP | YNTGKD | | | | |
| T4dam | 28 | FV DLTCGGL *S*VSL | 164 | DGD FVY VDPP | YLITVA | | | | |
| *Dpn* II | 39 | YF EPTVGGG ALFF | 187 | TGD FVY FDPP | YIPLSE | | | | |
| *Eco*dam | 31 | LV EPTVGAG *S*VFL | 174 | DAS VVY CDPP | YAPLSA | | | | |
| *Eco*R V | 35 | WV EPTMGTG VVAF | 186 | RDD VVY CDPP | YIGRHV | | | | |
| | | | | | | m6A | | | |
| *Taq* I | 43 | VL EPACAHG PFLR | 98 | AFD LIL GNPP | YGIVGE | | | | |
| *Pae*R7 | 22 | LL EPSFGCG DFLL | 113 | QFD FVV GNPP | YVRPEL | | | | |
| *Cvi*BIII | 53 | IL EPSCGTG EIIS | 113 | KFD FIV GNPP | YVVRPS | | | | |
| r*Eco*57I | 388 | FA DIACGSG AFII | 519 | QFD VIV GNPP | YMATEH | | | | |
| *Eco*57I | 38 | IL EPSCGDG VFIQ | 109 | IFD GAL GNPP | FIRYQF | | | | |
| *Eco*R I | 112 | SV SSFCGDG DFRS | 133 | KSD IVV TNPP | FSLFRE | | | | |
| *Eco*K | 171 | VQ DPAAGTA GFLI | 260 | KAH IVA TNPP | FGSAAG | | | | |
| *Pst* I | 51 | LF ESYVGEH EILD | 145 | KYN KAI LNPP | YLKIAA | | | | |
| *Hha* I | 14 | FI DLFAGLG GFRL | 73 | HD ILC AGFPCQAFSI SGK | | | | | |
| *Eco*R II | 98 | FI DLFAGIG GIRK | 177 | HD VLL AGFPCQPFSLAGV | | | | | |
| *Msp* I | 105 | FI DLFSGIG GIRQ | 166 | HD ILC AGFPCQPFSHIGK | | | | | |
| *Dde* I | 3 | II DLFAGCG GFSH | 67 | VD GII GGPPCQGFSL SGN | | | | | |
| *Sin* I | 77 | AL SFFSGAM GLDL | 145 | ID LIM GGPPCQAFSTAGK | | | | | |
| Phi3T | 6 | VM SLFSGIG AFEA | 69 | FD LLT SGFPCPTFSVAGA | | m5C | | | |
| SPR | 6 | VM SLFSGIG AFEA | 69 | FD LLV GGSPCQSFSVAGH | | | | | |
| *Bsu*R I | 61 | VL SLFSGCG GLDL | 148 | CN LIL GGFPCPGFSEAGP | | | | | |
| *Bsp*R I | 60 | VL SLFCGAG GLDL | 147 | AN LVI GGFPCPGFSEAGP | | | | | |
| Rat glycine MTase | 255 | VQ EAFGGRC QHSV | | | | | | | |

**Figure 3.** Sequences within the homology domains of the DNA-methylase sequences. Different letter fonts correspond to different levels of amino acid conservation in the column. The number to the left of each sequence indicates the position of the first residue of the segment within the whole sequence.

location within the central section in several sequences of 'D' family [11,14,47]. Thus, from this we can hypothesize that the TRDs of M.*Cfr*9I and M.*Pvu*II are inside the middle section (fig.4) and most probably between the second and fourth homology regions (fig.2) where the similarity is the weakest. The prediction for non type II enzymes is a more complicated task.

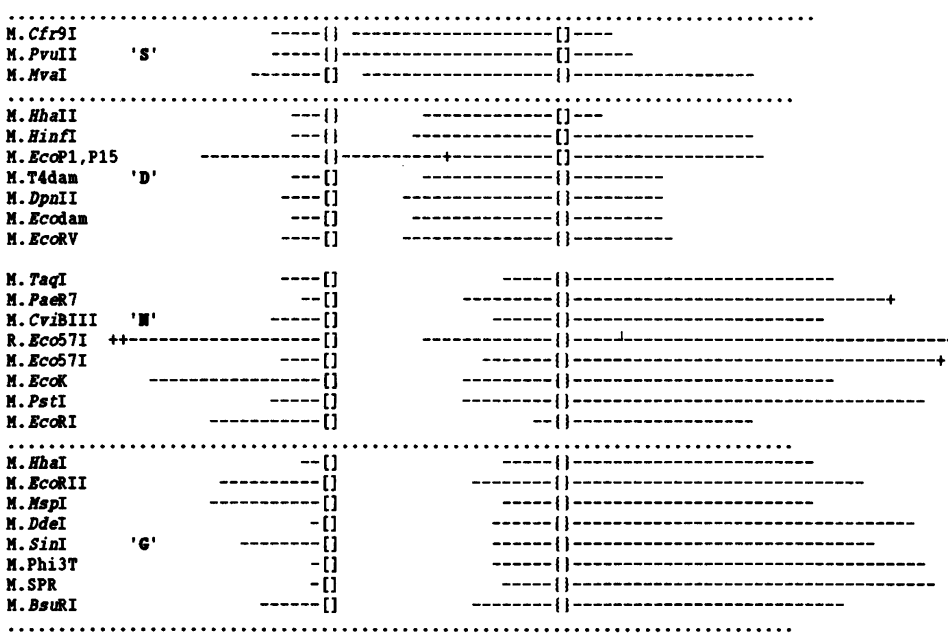


**Figure 4.** Alignment scheme of DNA-methylase sequences. All sequences are from amino to carboxy end. Each '−' represents ten amino acids, '+' stands for one hundred amino acids, while spaces indicate the gaps. '[ ]' and '{ }'stand for the first and the second homology domains, correspondingly. Boldface letters in single quotes indicate classificational families of DNA-methylases.

Another general question is the role these homology domains play in the methylation mechanism. It was earlier proposed that the conserved region **II** (fig.3) can serve as the AdoMet binding domain in adenine MTases [11,14]. We consider this idea not consistent enough with the detected relationship of this domain to the invariant motif of cytosine-C5 MTases (fig.3), which is thought to be directly involved in the modification of cytosine nucleus [48,23,10]. There are several differences among these sequences: the insert of a triplet C(P,Q)(A,G,S) (r.p.**v**,**w**,**x**) and conservation of residues at yet other positions. These differences might possibly reflect distinct chemistries used for the methyl transfer onto C5- *versus* exocyclic amine [49], while the similarity would indicate a common ancestry and a common function in general.

On the other hand, one could expect that the sequence patterns conserved in all types of MTases should be those involved in one of common functions: unspecific binding to DNA or the binding of AdoMet. The pattern **I** (fig.3) fulfils this requirement. Such a sequence motif occurs also in the sequence of Rat glycine-methylase [50] which probably has no DNA-binding capability. Hence, this domain most probably is involved in the methyl donor binding.

As far as structural features are concerned this domain fits the requirements for the AdoMet binding site well. The invariant F (r.p. **e**) could stack with the adenine ring of the AdoMet molecule while the hydroxyl-containing residue—D, E or S, (r.p. **c**) could form hydrogen or ionic bond with either $-S^{+}-$ or $-NH_3^{+}$ moiety.

From the mechanistic point of view it is clear that these domains should be tandemly arranged since the functions they perform are two consecutive steps in the methylation reaction. Another possibility is that both homology regions form a common three-dimensional domain which is responsible for both the discussed functions.

## ACKNOWLEDGEMENTS

The authors thank Z.Manelienè and J.Me skauskas for supplying our work with synthetic primers.

*To whom correspondence should be addressed

## REFERENCES

1. Razin, A., Cedar, H., Riggs, A.D. (1984) 'DNA Methylation: Biochemistry and Biological Significance'. Springer-Verlag, New York.
2. Adams, R.L.P., Burdon, R.H. (1985) 'Molecular Biology of DNA Methylation'. Springer-Verlag, New York.
3. Janulaitis, A.A., Klimasauskas, S., Petrusyte, M., Butkus, V. (1983) FEBS Lett. 161: 131–134.
4. Janulaitis, A., Stakenas, P., Petrusyte, M., Bitinaite, J., Klimasauskas, S., Butkus, V. (1984) Molec. Biol. (Moscow) 18: 115–129.
5. Ehrlich, M., Gama-Sosa, M.A., Carreira, L.H., Ljungdahl, L.G., Kuo, K.C., Gehrke, C.W. (1985) Nucl. Acids Res. 13: 1399–1412.
6. Ehrlich, M., Wilson, G.G., Kenneth, C.K., Gehrke, C.W. (1987) J. Bacteriol. 169: 939–943.
7. Butkus, V., Klimasauskas, S., Petrauskiene, L., Maneliene, Z., Lebionka, A., Janulaitis, A.A. (1987) Biochim. Biophys. Acta 909: 201–207.
8. Tao, T., Walter, J., Brennan, K.J., Cotterman, M.M., Blumenthal, R.M. (1989) Nucl. Acids Res. 17: 4161–4175.
9. Som, S., Bhagwat, A.S., Friedman, S. (1987) Nucl. Acids Res. 15: 313–332.
10. Posfai, J., Bhagwat, A.S., Posfai, G., Roberts, R.J. (1989) Nucl. Acids Res. 17: 2421–2435.
11. Lauster, R., Kriebardis, A., Guschlbauer, W. (1987) FEBS Lett. 220: 167–176.
12. Chandrasegaran, S., Smith, H.O. (1988) in Structure and Expression Volume 1: From Proteins to Ribosomes. eds. Sarma R.H. and Sarma M.H.. (Adenine Press) pp 149–156.
13. Narva,K.E., Van Etten J.L., Slatko, B.E., Benner, J.S. (1988) Gene 74: 253–259.
14. Guschlbauer, W. (1988) Gene 74: 211–214.
15. Butkus, V., Petrauskiene, L., Maneliene, Z., Klimasauskas, S., Laucys, V., Janulaitis, A.A. (1987) Nucl. Acids Res. 15: 7091–7102.
16. Butkus, V., Klimasauskas, S., Kersulyte, D., Vaitkevicius, D., Lebionka, A., Janulaitis, A.A.(1985) Nucl. Acids Res. 13: 5727–5746.
17. Chuvpilo, S.A., and Kravchenko, V.V. (1984) FEBS Lett. 1: 34–36.
18. Chen, E.Y., and Seeburg, P.H. (1985) DNA 2: 165–170.
19. Mount, D.W., Conrad, B. (1984) Nucl. Acids Res. 12: 811–818.
20. Posfai, G., Kiss, A. Erdei, S., Posfai, J., Venetianer, P. (1983) J. Mol. Biol. 170: 597–610.
21. Kiss, A., Posfai, G., Keller, C.C., Venetianer, P., Roberts, R.J. (1985) Nucl. Acids Res. 13: 6403–6420.
22. Sznyter, L.A., Slatko, B., Moran, L., O'Donnell, K.H., Brooks, J.E. (1987) Nucl. Acids Res. 15: 8249–8266.
23. Caserta, M., Zacharias, W., Nwankwo, D., Wilson,G.G., Wells, R.D. (1987) J .Biol. Chem. 262: 4770–4777.
24. Karreman, C., de Waard, A. (1988) J. Bacteriol. 170: 2527–2532.
25. Lin, P.M., Lee, C.H., Roberts R.J. (1989) Nucl. Acids Res. 17: 3001–3011.
26. Tran-Betcke, A., Behrens, B., Noyer-Weidner, M., Trautner, T.A. (1986) Gene 42: 89–96.
27. Posfai, G., Baldauf, F., Erdei, S., Posfai, J., Venetianer, P., Kiss, A. (1984) Nucl. Acids Res. 12: 9039–9049.
28. Greene, P.J., Boyer, H.W. (1981) Fed. Proc. 40: 293.
29. Bougueleret, L., Schwarzstein, M., Tsugita, A., Zabeau, M. (1984) Nucl. Acids Res., 12: 3659–3677.
30. Chandrasegaran S., Wu L.P., Valdo E., Smith, H.O. (1988) Gene 74: 15–21.
31. Mannareli, B.M., Balganesh, T.S., Greenberg, B., Springhorn, S.S., Lacks, S.A. (1985) Proc. Natl. Acad. Sci. USA 82: 4468–4472.
32. Theriault, G., Roy, P.H., Howard, K.A., Benner, J.S., Brooks, J.S., Waters, A., F., Gingeras, T.R. (1985) Nucl. Acids Res. 13: 8441–8461.

33. Walder, R., Walder, J., Donelson, J. (1984) J. Biol. Chem. 259: 8015–8026.
34. Slatko, B.E., Benner, J.S., Jager-Quinton, T., Moran, L.S., Simcox, T.G., Vancott, E.M., Wilson, G.G. (1987) Nucl. Acids Res. 15: 9781–9796.
35. Brooks, J.E., Blumenthal,R.M., Gingeras, T.R. (1983) Nucl. Acids Res. 11: 837–851.
36. Chandrasegaran, S., Lunnen, K.D., Smith, H.O., Wilson, G.G. (1988) Gene, 70: 387–392.
37. Humbelin, M., Suri,B., Rao, D.N., Hornby D.P., Eberle, H., Pripfl, T., Kenel, S., Bickle, T.A. (1988) J. Mol. Biol. 200: 23–29.
38. Loenen, W.A.M., Daniel, A.S., Braymer, H.D., Murray, N.E. (1987) J. Mol. Biol., 198: 159–170.
39. MacDonald, P.M., Mosig, G.(1984) EMBO J. 3: 2863–2871.
40. Narva, K.E., Wendell, D.L., Skrdla, M.P., VanEtten, J.L. (1987) Nucl. Acids Res. 15: 9807–9823.
41. Petrusyte, M.,Bitinaite, J., Menkevicius, S., Klimasauskas, S., Butkus, V., and Janulaitis, A. (1988) Gene 74: 89–91.
42. Argos, P., McCaldon, P. (1988) Genetic engineering 10: 21–65.
43. Dayhoff, M.O., Barker, W.C., Hunt, L.T. (1984) Methods in Enzymology, 91: 524–545.
44. Lipman, D.J., Pearson, W.R. (1985) Science 227: 1435–1441.
45. Trautner, T.A., Balganesh, T.S., Pawlek, B. (1988) Nucl. Acids Res. 16: 6659–6658.
46. Wilke, K., Rauhut, E., Noyer-Weidner, M., Lauster, R., Pawlek, B., Behrens, B., Trautner, T.A. (1988) EMBO J. 7: 2601–2609.
47. Miner, Z., Schlagman, S., Hattman, S. (1988) Gene 74: 275–276.
48. Wu, J.C., Santi, D.V. (1987) J. Biol. Chem. 262: 4778–4786.
49. Pogolotti, A.L., Ono, A., Subramaniam, R., Santi, D.V. (1988) J. Biol. Chem. 263: 7461–7464.
50. Ogawa, H., Konishi, K., Takata, Y., Nakashima, H., Fijioka M. (1987) Eur. J. Biochem. 168: 141–151.

**This article, submitted on disc, has been automatically converted into this typeset format by the publisher.**