BMC Evolutionary Biology

Research article

Sequence permutations in the molecular evolution of DNA methyltransferases

Janusz M Bujnicki

Address: Bioinformatics Laboratory, International Institute of Molecular and Cell Biology, Trojdena 4, 02-109 Warsaw, Poland

E-mail: jamb@wp.pl

Published: 12 March 2002

BMC Evolutionary Biology 2002, **2**:3

Received: 2 November 2001
Accepted: 12 March 2002

This article is available from: http://www.biomedcentral.com/1471-2148/2/3

Abstract

Background: DNA methyltransferases (MTases), unlike MTases acting on other substrates, exhibit sequence permutation. Based on the sequential order of the cofactor-binding subdomain, the catalytic subdomain, and the target recognition domain (TRD), several classes of permutants have been proposed. The majority of known DNA MTases fall into the α, β, and γ classes. There is only one member of the ζ class known and no members of the δ and ε classes have been identified to date. Two mechanisms of permutation have been proposed: one involving gene duplication and in-frame fusion, and the other involving inter- and intragenic shuffling of gene segments.

Results: Two novel cases of sequence permutation in DNA MTases implicated in restriction-modification systems have been identified, which suggest that members of the δ and ζ classes (M.*Mwo*I and M.*Tvo*ORF1413P, respectively) evolved from β-class MTases. This is the first identification of the δ-class MTase and the second known ζ-class MTase (the first ζ-class member among DNA:m^4C and m^6A-MTases).

Conclusions: Fragmentation of a DNA MTase gene may result from attack of nucleases, for instance when the RM system invades a new cell. Its reassembly into a functional form, the order of motifs notwithstanding, may be strongly selected for, if the cognate ENase gene remains active and poses a threat to the host's chromosome. The "cut-and-paste" mechanism is proposed for β-δ permutation, which is non-circular and involves relocation of one segment of a gene. The circular β-ζ permutation may be explained both by gene duplication or shuffling of gene fragments. These two mechanisms are not mutually exclusive and probably both played a role in the evolution of permuted DNA MTases.

Background

DNA of prokaryotic and eukaryotic cells and their viruses is often modified by methylation, carried out by S-adenosyl-L-methionine (AdoMet)-dependent DNA methyltransferases (MTases). Since a particular nucleotide sequence may exist in its methylated or unmethylated form, methylation can be regarded as an increase of the information content of DNA, which serves a wide variety of biological functions. In Eukaryota, DNA methylation plays a role in crucial regulatory processes, such as regulation of gene expression, embryonic development, genomic imprinting, and carcinogenesis (reviewed in ref. [1]). In Prokaryota, DNA methylation can be involved in DNA mismatch repair, regulation of gene expression, and control of timing of DNA replication (reviewed in ref. [2]). However, the majority of prokaryotic MTases are paired with a restriction endonuclease of cognate sequence specificity, together forming restriction-modification (RM) systems. RM systems are thought to serve as defense mechanisms that protect the cell against invasion of foreign genetic elements such as phages and plasmids [3]. It has been also suggested that RM systems are maintained in evolution because they participate in generating bacterial diversity by promoting homologous recombination [4] or because they act as as "selfish" genetic elements that undergo extensive horizontal transfer [5]. These three hypotheses are contrasting, but not mutually exclusive.

MTases can be divided into three different groups on the basis of the chemical reactions they catalyze: generating N6-methyladenine (m^6A), N4-methylcytosine (m^4C), and C5-methylcytosine (m^5C). It has been suggested that m^4C and m^6A MTases (collectively termed "N-MTases") may be more closely related to each other than to m^5C MTases [6]. Nevertheless, subsequent analyses showed that the relationships between these groups of proteins are quite complicated and their evolution may have involved several independent conversions of the reaction specificity [7–9]. Amino acid sequence alignments of DNA MTases revealed several conserved motifs, of which I-VIII and X are common to most subfamilies, and a region of essentially higher variability [6,10]. Based on the results of X-ray crystallography of members of all three groups and structure-based multiple sequence alignment, motifs IV-VIII were assigned to the active-site subdomain, motifs X and I-III to the AdoMet-binding subdomain, and the variable region was recognized as a separate domain, implicated in recognition of the target sequence (and termed TRD for target-recognition domain) (reviewed in refs. [11,12]). Structural studies on m^5C, m^4C, and m^6A MTases demonstrated that the TRDs of these proteins are structurally dissimilar and most likely were acquired in independent gene fusion events [11].

DNA MTases have been subdivided into 6 classes (α, β, γ, ζ, and the hypothetical δ and ε ; Figure 1a) according to the possible linear arrangements of three modules: the AdoMet-binding subdomain, the active site subdomain, and the variable TRD [6]. All of the enzymes preserve the same spatial arrangement of the motifs and α, δ, and ε form a circularly permuted set as does β, γ, and ζ (Figure 1b). The majority of DNA N-MTases fall into the α, β, and γ classes, with no *bona fide* γ-m^4C MTases known. M.*Ngo*MXV [13] and its close homolog M.*Lmo*A118I [14] are the only experimentally characterized m^4C MTases, whose architecture is very similar to γ-m^6A MTases. Nevertheless, these remarkably small MTases (153 aa), as well as their uncharacterized homologs identified in sequence databases, lack both the classical TRD and the region corresponding to motif X and therefore can be regarded as "minimal" members of the family [13]. Another group of small (~175 aa) MTases, which are similar to γ-m^6A MTases but lack motif X and the TRD, are the a typical Dam (m^6A) MTases encoded by several phages, including HP1, VT-2, and T1 [15].

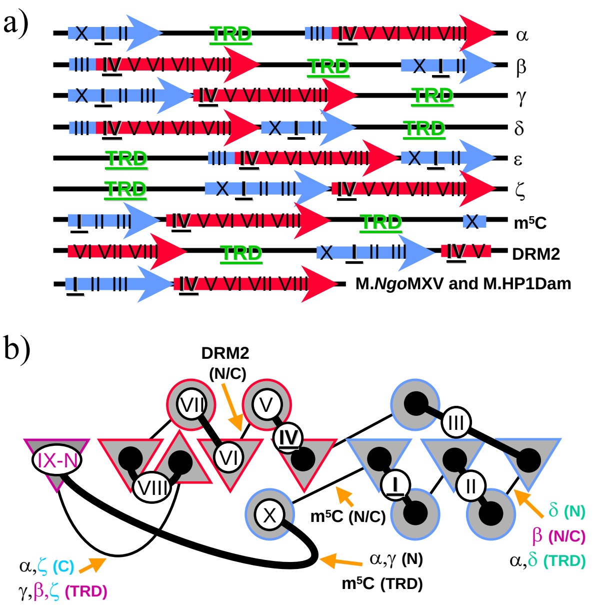

Figure 1
Conserved fold and variable topology of the MTase domain, **a)** The linear organization of six classes of amino-MTases (α-ζ) postulated by Malone et al. [6], m^5C MTases (the prevailing archetypal topology labeled as m^5C, and the two minor classes as ζ and DRM2), and the minimal MTases lacking the discrete TRD. The AdoMet-binding region is shown as a blue arrow, the catalytic region is shown as a red arrow, conserved motifs are labeled accordingly **b).** The topology diagram: triangles represent β-strands, circles represent a- and 3_{10}-helices, connecting lines represent loops; the thick lines correspond to the loops on the catalytic face of the protein, which harbor residues involved in cofactor binding and catalysis and most likely in DNA-binding. Elements forming the AdoMet-binding pocket are colored blue; elements forming the target base-binding/catalytic pocket are colored red. Circled Roman numerals represent nine motifs, the key motifs I and IV are shown in bold and underlined. Orange arrows show the topological breakpoints (N/C for generation of N- and C-termini) and sites of TRD insertion characteristic for the individual classes of MTases. The elements characteristic for the β-class MTases are shown in magenta, the corresponding elements found in M.*Mwo*I (δ-class) and TVN1413 (ζ-class) are shown in green and cyan, respectively.

Most m^5C MTases resemble the N-MTases of the γ class, with the only difference being their motif X localized at the C-terminus instead of the N-terminus. However, a few m^5C MTases have been described with unusual permutations (Figure 1). To my knowledge, M.*Bss*HII is the only DNA MTase, for which the ζ architecture has been convincingly confirmed by experiment [16] and homology modeling [17]. In addition, a novel class of permuted m^5C MTases typified by DRM2 have been recently identified in plants; this prediction has also been supported by threading of the permuted sequence onto the common fold [18]. Some DNA MTases contain terminal extensions and various insertions [6], but it has never been demonstrated that they are related to known or predicted TRDs in other proteins.

DNA MTases are only one of the numerous families of remotely related enzymes that exhibit a common fold [11,19]. Other members of the superfamily methylate a variety of chemically diverse molecules, including various RNAs, proteins, lipids, small molecules, etc. While all members of the superfamily share the same structural core, only DNA MTases vary in the order of conserved motifs. The rest show the order X, I, II, III, IV, V, VI, VII, VIII and only protein-arginine MTases lack the last three motifs [19]. The question thus arises, why do DNA MTases, but not MTases in general, exhibit sequence permutations within the same structural framework?

Two models have been proposed to explain sequence permutations arising during the evolution of N-MTases [9,20]. Jeltsch argued that the process of domain permutation needs duplication and in-frame fusion of the MTase gene, producing one enzyme with two catalytic domains [20]. Subsequent introduction of a new start codon in the middle of the first gene copy and a stop codon at the equivalent position in the second gene copy would then result in a circularly permuted variant. For instance, the ζ- or β-like permutants could arise from a hypothetical tandem $\gamma\gamma$-class MTase. This model corresponds to the widely accepted concept that a permuted protein may arise naturally from tandem repeats by extraction of the C-terminal portion of one repeat together with the N-terminal portion of the subsequent repeat, so long as the protein's N and C termini are in close spatial proximity (reviewed in ref. [21]). Although the idea itself offers a plausible explanation for the origin of permutants within many protein families, the only known duplicated m^6A MTases are the type IIS enzymes of the $\alpha\alpha$-class, whose permutation would eventually produce enzymes of the δ or ε class that have not been identified to date. The TRDs of known MTases from β and γ classes are unrelated [22,23], hence it is unlikely that simple conversions "from $\gamma\gamma$ to β " or "from $\beta\beta$ to γ " have occurred in nature. Furthermore, the N- and C-termini of M.*Taq*I, the only γ-m^6A MTase whose 3D structure is known, are quite distant in space [22]. Still, this scenario may be valid for enzymes that have not been identified yet, or whose sequences have not been studied in enough detail.

It has been hypothesized (ref. [9]) that the permuted DNA MTase variants were generated by intra- or intergenic rearrangements of gene fragments (i.e. "module shuffling"; reviewed in refs. [24,25]) that left no evidence of duplication intermediates. However, only in one case (M.*Bss*HII) has it been possible to reconstruct a possible evolutionary history of shuffled fragments [17]. Moreover, no examples of N-MTases in different classes are known, whose TRDs are markedly similar. Hence, no convincing examples of permutations of an entire N-MTase molecule have been identified to date. The reported permutations in N-MTases concern only the segments within the catalytic domain, while the unrelated TRDs were acquired or evolved independently in distinct classes [20]. Identification of closely related DNA MTases with homologous TRDs that nevertheless lie in different classes might help decide between the two given hypotheses and shed light on how the different classes of DNA MTases arose.

Results

Candidates for N-MTases in the midst of the process of permutation were sought amongst all DNA MTases, whose sequences were available from REBASE [26]. Those sequences exhibiting deviations from the typical spacing between the conserved motifs or unusual extensions at the termini were chosen for detailed analysis. Within some extensions, several known or predicted DNA-binding domains were identified (to be published elsewhere), however only two pairs of MTase sequences were found that fit the criteria of similarity between the AdoMet-binding subdomain, the active site subdomain, and the TRD, without conservation of their co-linearity.

One MTase that emerged from the analysis was the M.*Mwo*I protein, which recognizes an interrupted palindrome GCNNNNNNNGC and methylates one of the cytosines in each strand, generating m^4C [27]. M.*Mwo*I (GenBank record 2961238) was earlier classified as a member of the β-class, however it exhibits quite unusual length of 668 aa, which is approximately twice the length of a typical β-class MTase [6], and it lacks the variable insertion between motifs VIII and X, corresponding to the TRD in β-class MTases [9].

BLAST searches and threading analysis (see Methods) revealed that the N-terminal part of M.*Mwo*I (aa 1–270) aligns very well with the catalytic and the AdoMet-binding subdomains of β-class MTases. However, in β-class MTases these subdomains are separated by the TRD, which in M.*Mwo*I is replaced by only a few residues (Figures 2a, 3a). Comprehensive results of the threading analysis of the M.*Mwo*I sequence are available at the URL [http://bioin-

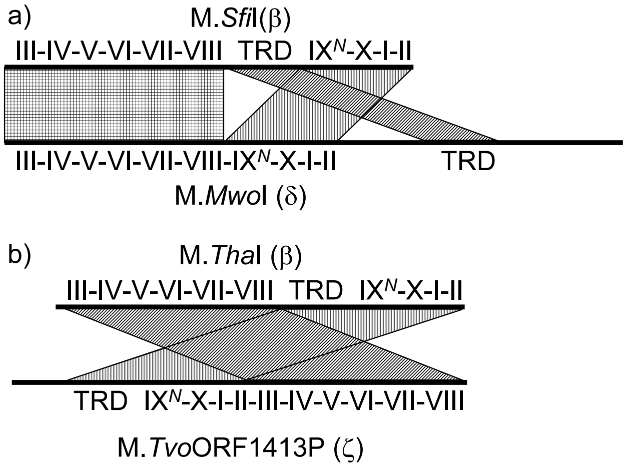
a)
M.*Sfi*I(β)
III-IV-V-VI-VII-VIII TRD IX^N-X-I-II
III-IV-V-VI-VII-VIII-IX^N-X-I-II TRD
M.*Mwo*I (δ)
b)
M.*Tha*I (β)
III-IV-V-VI-VII-VIII TRD IX^N-X-I-II
TRD IX^N-X-I-II-III-IV-V-VI-VII-VIII
M.*Tvo*ORF1413P (ζ)

Figure 2
Rearrangements in the primary sequence of the two pairs of permuted MTases **a)** Relationships between segments of M.*Sfi*I and M.*Mwo*I **b)** Relationships between segments of M.*Tha*I and M.*Tvo*ORF1413P.

fo.pl/meta/target.pl?id=4139] , the alignments of the N-terminal region are essentially identical to those reported previously [9]. Remarkably, The C-terminal region of M.*Mwo*I revealed no similarities to other sequences, with one prominent exception, namely the predicted TRD of the m^{4}C MTase M.*Sfi*I (GenBank entry 2761010) with the BLAST expectation (e) value of 10^{-3}. Further database searches using the sequence of M.*Sfi*I and the isolated fragments of putative TRDs confirmed that the three major subdomains of M.*Mwo*I and M.*Sfi*I exhibit significant sequence similarity, but the linear order of these elements differs between them. If this prediction is correct, M.*Mwo*I should be classified as the first member of the δ-class, rather than the β-class (Figures 1, 2a). It is noteworthy that using either PSI-BLAST or threading, no significant sequence or structural similarities of the TRD of M.*Mwo*I and M.*Sfi*I could be detected to TRDs of other MTases and generally to sequences of other proteins. The additional sequence region of M.*Mwo*I (aa 434–497), which may be regarded as a linker between the N-terminal catalytic domain and the newly identified C-terminal TRD, also showed no matches to any sequences in the database. Hence, the determination of which arrangement of subdomains, that of M.*Mwo*I or M.*Sfi*I, corresponds to the ancestral state must await discovery of their homologs.

M.*Sfi*I recognizes the sequence GGCCNNNNNGGCC, which belongs to a broader set of sequences recognized by the GCNNNNNNNGC-specific M.*Mwo*I. It is not known, how these enzymes recognize such a lengthy sequence with a non-specific spacer. Nonetheless, it can be imagined that the TRD of M.*Sfi*I evolved from the TRD of M.*Mwo*I by acquisition of new contacts to bases outside and inside the GC pair (N->G)GC(N->C)NNNNN(N->G)GC(N->C) or conversely, the stringent DNA-recognition specificity of the M.*Sfi*I-like TRD was relaxed to give rise to the less specific M.*Mwo*I. In the absence of protein-DNA co-crystal structures for the β-class of MTases and lack of suitable structural templates for modeling the TRD structure in M.*Sfi*I and M.*Mwo*I, prediction of the detailed protein-DNA contacts is unfeasible. However, I hope that the finding reported herein will prompt mutagenesis experiments – it is tempting to speculate that swapping the predicted TRDs between M.*Mwo*I and M.*Sfi*I will result in an exchange of specificities.

A second MTase that came from the initial screen was M.*Tvo*ORF1413P, interpreted as a member of the γ-class in ReBase [26] [http://rebase.neb.com/rebase/enz/M.TvoORF1413P.html], but exhibiting an extension of over 150 aa located N-terminally to motif X instead of the C-terminal extension after motif VIII required by the structure of the γ-class. In a BLAST search initiated with the M.*Tvo*ORF1413P sequence, M.*Tha*I was reported as the best hit, with a highly scored alignment of the AdoMet-binding region (e-value 5*10^{-14}) and a quite poorly scored (0.049) alignment of the catalytic region. However, M.*Tha*I is a member of the β-class and these two regions of similarity are swapped in the primary sequences of the two MTases (Figures 2b, 3b). BLAST searches initiated with the N- and C-terminal parts of M.*Tvo*ORF1413P showed that its C-terminal region scores better (e-value 6*10^{-4}) when aligned with the catalytic domain of another CGCG-specific β-m^{4}C MTase, M.*Tma*I (GenBank record 4980829), a close relative of M.*Tha*I (data not shown).

Threading analysis of the M.*Tha*I sequence revealed its perfect compatibility with the M.*Rsr*I and M.*Pvu*II structures (results are available at [http://bioinfo.pl/meta/target.pl?id=4134] , allowing homology modeling of the M.*Tha*I structure (Figures 3b, 4). These results reveal that the TRD of M.*Tha*I is shorter than TRDs of M.*Rsr*I and M.*Pvu*II, which may be an indication that in this enzyme some DNA-binding residues migrated to other loops [9]. On the other hand, threading of M.*Tvo*ORF1413P revealed that its C-terminus corresponds to the last β-strand of the common MTase core (motif VIII; Figure 1), strongly arguing against its assignment to the γ-class, which requires the presence of the TRD C-terminal to this element (threading results are available at [http://bioinfo.pl/meta/target.pl?id=4133], [http://bioinfo.pl/meta/target.pl?id=4144] and [http://bioinfo.pl/meta/target.pl?id=4145] with the three entries corresponding to the full length sequence, the N-terminal part and the C-terminal parts, respectively). Instead, all threading algorithms reported that the N-terminus of M.*Tvo*ORF1413P matches perfectly the additional β-strand (motif IX-N) in M.*Rsr*I and M.*Pvu*II and this region, along with the predicted

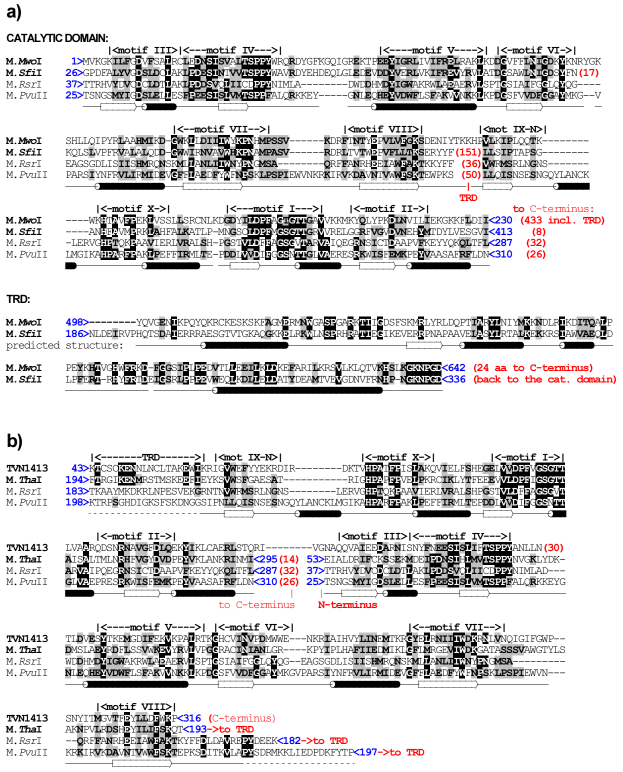

Figure 3
Sequence alignment of β-MTases with the permuted MTases: **a)** M.*Mwo*I (δ-class) **b)** M.*Tvo*ORF1413P (ζ-class). Blue numbers indicate the sequence coordinates (residue numbers). Red numbers in parentheses indicate the size of terminal extensions or loops, which were omitted for clarity. Identical residues are highlighted in black, conservative substitutions are highlighted in gray. Conserved motifs are labeled according to the nomenclature described by Malone et al [6]. Secondary structure elements shown below the alignment were deduced from the M.*Pvu*II coordinates or predicted for the TRD of M.*Sfi*I/M.*Mwo*I.

TRD[,] aligns quite well with motif IX-N and the TRD of M.*Tha*I (Figure 3b). It is noteworthy, that for the core regions, the predicted secondary structure agreed very well both between M.*Tha*I and M.*Tvo*ORF1413P, and between these MTases and the experimentally determined structures of M.*Rsr*I and M.*Pvu*II (for details see the above mentioned links to MetaServer results). For the 42 N-terminal residues of M.*Tvo*ORF1413P no similarity to known sequences or structures could be demonstrated, and modeling based on N-terminally extended threading alignments resulted in misfolded structures; it is therefore possible that this region forms an elaboration of the common fold, which is unique to M.TvoORF1413P.

The MTase activity of M.*Tvo*ORF1413P remains to be demonstrated. However, its close homolog has been recently identified, which exhibits a genuine DNA:m^{4}C MTase activity (Drs. M.A. Abdurashitov and S.K. Degtyarev, personal communication). M.*Bst*F5I-4, whose sequence remains unpublished, is evidently homologous to M.*Tvo*ORF1413P over the region including the predicted N-terminal TRD, as well as motifs IX-N, X, and I-VIII (BLAST e-value $3*10^{-20}$, 26% identical plus 23% conservatively substituted residues; with 60% identical residues in the predicted TRD, i.e. the 10 aa loop preceding motif IX-N; data not shown). It cannot be ruled out that the small TRD of M.*Tha*I, M.*Bst*F5I-4, and M.*Tvo*ORF1413P harbors only a fraction of specificity determinants and that other loops on the catalytic face of the protein contribute to specific DNA recognition. Nevertheless, according to the classical definition of the TRD (the variable region between motifs VIII and X [6,28]), the presented results of sequence analysis and structure prediction suggest that the common ancestor of ζ-class MTases M.*Bst*F5I-4 and M.*Tvo*ORF1413P evolved from M.*Tha*I (β-class member) by sequence permutation.

Discussion

Sequence analysis resulted in identification of two novel cases of sequence permutation in DNA MTases, and demonstration for the very first time, that DNA:m^{4}C MTases of different classes may exhibit significant sequence similarity not only in the catalytic domain, but also in the TRD. This finding suggests that the analyzed gene pairs diverged relatively recently, permitting a test of the hypothesis that the observed rearrangements occurred according to the

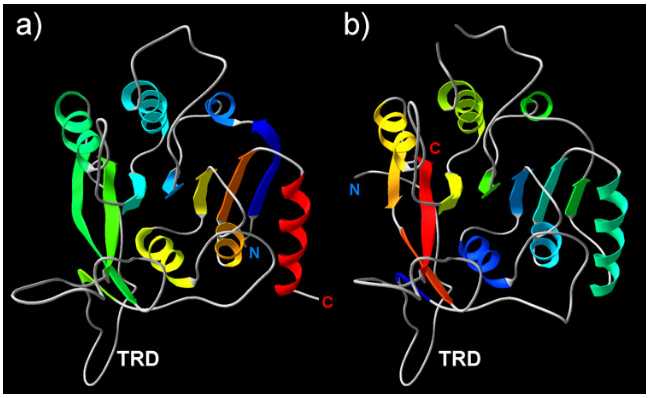

Figure 4
Cartoon diagrams depicting the structure of homology models of **a)** M.*Tha*I **b)** M.*Tvo*ORF1413P. Secondary structure elements are colored according to the linear order – from blue (N-terminus) to red (C-terminus). The termini of the models are labeled; the N-terminal 51 aa of M.*Tha*I and 38 aa of M.*Tvo*ORF1413P, as well as an insertion comprising aa 195–220 of M.*Tvo*ORF1413P were not modeled.

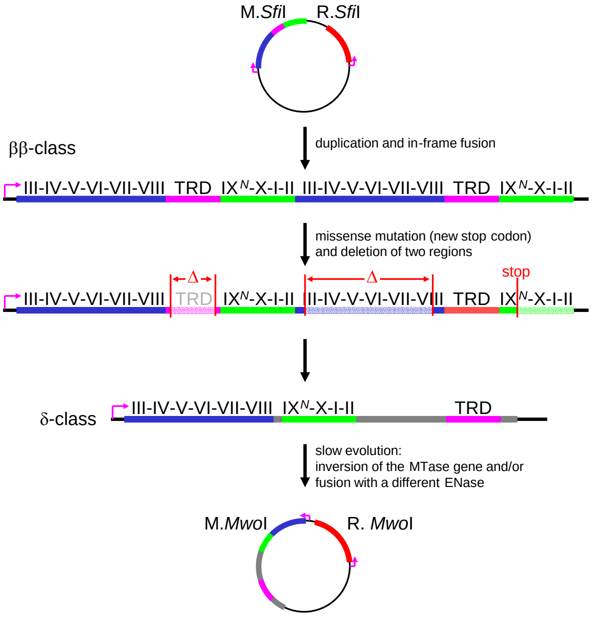
M.*Sfi*I
R.*Sfi*I
ββ-class
duplication and in-frame fusion
III-IV-V-VI-VII-VIII TRD IX^N-X-I-II III-IV-V-VI-VII-VIII TRD IX^N-X-I-II
missense mutation (new stop codon)
and deletion of two regions
Δ
Δ
stop
III-IV-V-VI-VII-VIII TRD IX^N-X-I-II III-IV-V-VI-VII-VIII TRD IX^N-X-I-II
δ-class
III-IV-V-VI-VII-VIII IX^N-X-I-II TRD
slow evolution:
inversion of the MTase gene and/or
fusion with a different ENase
M.*Mwo*I
R. *Mwo*I

Figure 5
Diagram illustrating a putative evolutionary history of the M.*Sfi*I-M.*Mwo*I gene pair according to the "permutation-by-duplication" mechanism. Conserved motifs and the TRD are labeled. Protein sequence regions corresponding to the AdoMet-binding subdomain, the catalytic subdomain, the TRD, and non-conserved segments of unknown function are depicted in green, blue, magenta and grey, respectively. Grey labels indicate non-functional regions that are presumably able to fold in solution and do not interfere with the folding of the functional unit. Red vertical bars denote ends generated by new start or stop codons or the boundaries of intragenic deletions. Deleted segments are marked with "Δ". Putative promoters are indicated as pink arrows.

"permutation-by-duplication" model [20] or to the alternative model, involving intragenic relocation of gene segments. If sequences resembling fragments of one of the DNA MTase analyzed herein were identified in its own neighborhood, this would provide strong evidence that gene duplication occurred. It would also suggest that this particular MTase is a permuted version of its homolog, whose neighborhood is free from duplicated fragments, rather than the opposite.

Regrettably, the neighborhood of the *Sfi*I and *Mwo*I RM systems is unknown, however the context of *Tha*I RM system and M.*Tvo*ORF1413P can be analyzed using the complete genome sequences of *Thermoplasma acidophilum*[29] and *T. volcanium*[30], respectively. The genome sequences of both *Thermoplasma* species flanking M.*Tha*I and M.*Tvo*ORF1413P (10 000 base pairs in each direction) were compared using the BLAST-family programs at the level of DNA and putative translations in all open reading frames (with stop codons translated as missing characters, e.g. "X"). No evidence of sequences similar to the genes encoding these two MTases were found, except for another putative DNA:m^{4}C MTase M.*Tvo*ORF1416P located 2640 bp 5' to M.*Tvo*ORF1413P. M.*Tvo*ORF1416P is a typical α-class member and exhibits significantly higher similarity to α-MTases such as M.*Psp*GI (BLAST e-value 2*10^{-73}) or M.*Mva*I (e-value 3*10^{-67}) than to M.*Tvo*ORF1413P (insignificant e-value 5.3). Therefore, the two MTases should be regarded as remote homologs of each other and as members of different phylogenetic lineages [9] that met rather accidentally in the *T. volcanium* chromosome rather than as products of recent duplication of one gene.

The lack of evidence supporting the gene duplication mechanism in the case of MTases from *Thermoplasma* is not entirely convincing, especially since no direct evidence supporting the alternative "cut-and-paste" mechanism can be provided by sequence analysis. Hence, the events leading to M.*Sfi*I(β)-M.*Mwo*I(δ) and M.*Tha*I- (β) M.*Tvo*ORF1413P (ζ) permutations were reconstructed based on both mechanisms (Figures 5, 6, 7, and 8). For the sake of simplicity, it was assumed that the unique δ and ζ-class members were in these cases generated by permutation of common β-class members, however an analogous reconstruction could be carried out assuming the opposite directionality of rearrangements, leading to similar conclusions.

Figures 5 and 6 shows possible histories leading to both permutations according to the gene duplication mechanism. The order of the sequence motifs remains conserved between M.*Sfi*I and M.*Mwo*I, but these two MTases differ in that a large segment bearing the TRD appears in the middle of the former but in the C-terminus of the latter, hence permutation of these two proteins is not circular (Figure 5). To produce M.*Mwo*I, duplication of M.*Sfi*I would have to be followed by both creation of a novel stop codon to eliminate the region encoding the "new" catalytic subdomain of the C-terminal repeat and deletion of the regions corresponding to the "old" TRD and the "new" AdoMet-binding subdomain. It is quite unlikely that all these changes occurred in a single event, and their occurrence in a series of steps would seem inevitably to produce a nonfunctional intermediate, which would require two steps to regain activity. If these changes did occur gradually, the product of gene duplication, in which one repeat retained the AdoMet-binding subdomain but lost the catalytic subdomain (or conversely), would expose the hydrophobic core of the remaining nonfunctional subdomain to the solvent. Folding and enzymatic function of such "1 & $^1/_2$" mutant would be probably heavily compromised. The function of M.*Sfi*I and M.*Mwo*I is to protect the chromosome from being cleaved by the cognate ENase. Hence, it seems rather unlikely that the

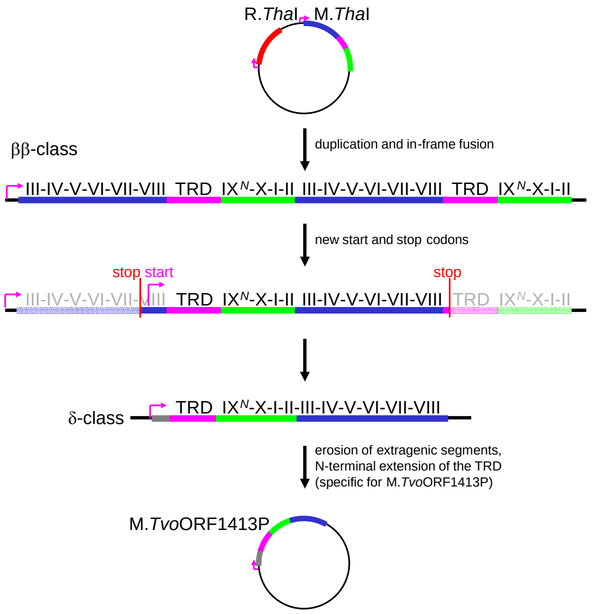
R.*Tha*I
M.*Tha*I
ββ-class
duplication and in-frame fusion
III-IV-V-VI-VII-VIII TRD IX^N-X-I-II III-IV-V-VI-VII-VIII TRD IX^N-X-I-II
new start and stop codons
stop start
stop
III-IV-V-VI-VII-VIII TRD IX^N-X-I-II III-IV-V-VI-VII-VIII TRD IX^N-X-I-II
δ-class
TRD IX^N-X-I-II-III-IV-V-VI-VII-VIII
erosion of extragenic segments,
N-terminal extension of the TRD
(specific for M.*Tvo*ORF1413P)
M.*Tvo*ORF1413P

Figure 6
Diagram illustrating a putative evolutionary history of the M.*Tha*I-M.*Tvo*ORF1413P gene pair according to the "permutation-by-duplication" mechanism. Labels follow the legend to Figure 5.

host cell would survive such a series of unlikely events passing through functionally compromised intermediates.

Compared to M.*Sfi*I and M.*Mwo*I, evolution of M.*Tvo*ORF1413P from M.*Tha*I seems more likely (Figure 6), since in this "classical" case of circular permutation requires only removal of the terminal regions by formation of new start and stop codons. However, in this case deletion of terminal subdomains, or their large parts, would also have to be concurrent, otherwise nonfunctional intermediates could arise, leading to cell death due to insufficient protection against the cognate ENase. If only one of the repeats in the original tandem ββ fusion protein is damaged, deletion of the remaining nonfunctional part would most likely restore the highly active, single copy version of the parent MTase. It is noteworthy that the tandem duplication mechanism offers no stage, at which evolutionary pressure would result in optimization of a poorly active intermediate specifically towards the permuted version.

It can be argued that owing to the selection pressure provided by the ENase, accidental damage to one of the would result in generation of such suppressor mutants, which exhibit a compensatory deletion in the other repeat, resulting in permutation. This argument pertains to both M.*Sfi*I-M.*Mwo*I and M.*Tvo*ORF1413P-M.*Tha*I pairs. However, it seems that the most effective "suppression" could be achieved by Another problem with the scenario involving tandem ββ MTase fusion is that such fusions have never been reported to occur. In the X-ray structure of the β-class member M.*Rsr*I, the two identical subunits make quite extensive contacts (a loss of 1799.3 Ang**2 of solvent accessible surface area per chain upon complex formation; see URL: [http://pdb-browsers.ebi.ac.uk/pdb-bin./macmol.pl?fi1ename=leg2] , suggesting that the dimeric structure of this MTase is biologically relevant [23]. The TRD and the active site reside on opposite sides of the M.*Rsr*I monomer, but the unique dimeric configuration brings the TRD of one subunit near the active site of the other, indicating that dimerization may be required for recognition and methylation to occur. The N- and C-termini of M.*Rsr*I are located close to each other in the monomer (8.8 A), but the C-terminus of one monomer is located on the opposite side of the dimer in respect to the N-terminus of the other monomer, separated by a distance of 74 A in a straight line. If the configuration observed in the crystal structure of M.*Rsr*I is representative for other members of the β-class that use two cooperating MTase domains, covalent joining of the termini of these domains would require a very long linker peptide, looping around the dimer. Hence, tandem fusion seems disrupting for cooperation of two β-class MTases within the dimer.

An alternative "cut-and-paste" mechanism (Figures 7, 8), inspired by the observed genomic rearrangements associated with the presence of restriction endonucleases [5,31], involves generation of a functional gene from fragments. In one scenario (Figures 7 and 8, left panels), the fragments may be generated due to combined action of various endo- and exonucleases that partially degrade the DNA fragment encoding the MTase gene, thereby producing recombingenic ends. If degradation occurs, at least two copies of the MTase gene must be present in the cell in order to reconstruct all important regions. Another scenario (Figures 7 and 8, right panels) involves precise action of a sequence-specific endonuclease, which fortuitously cleaves the MTase gene in the regions corresponding to linkers between the TRD and the two subdomains of the catalytic domain. In this case, one or more copies of the MTase gene may be present in the cell. These two scenarios are not mutually exclusive, provided that the fragments resulting from any type of cleavage or degradation span all regions necessary for the MTase activity, and are able to recombine with each other or use "sticky ends" to ensure ligation.

The "cut-and-paste" mechanism (Figures 7, 8) differs from the "permutation-by-duplication" mechanism (Figures 5, 6) in that it involves a momentary stage at which there are

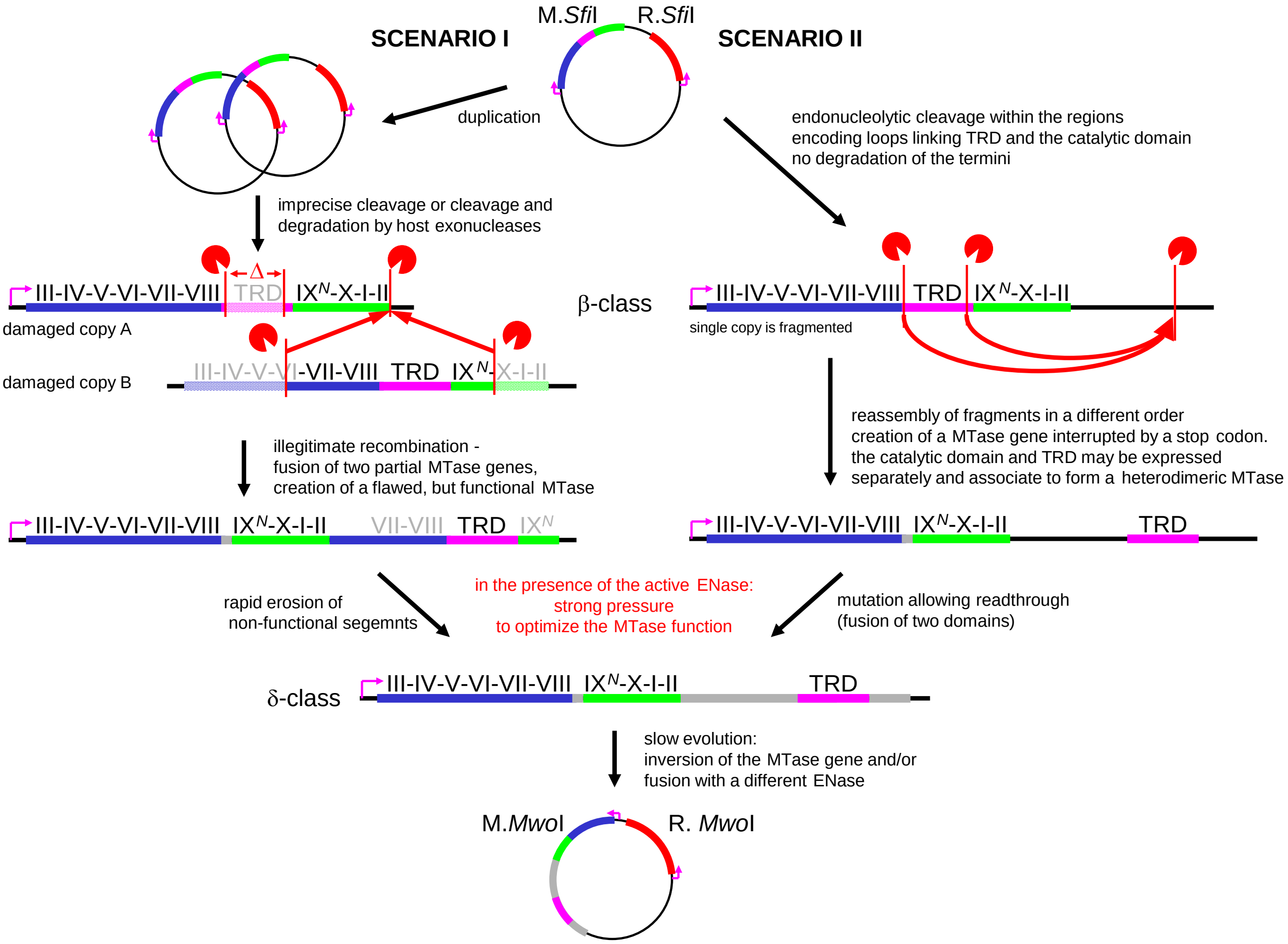
M.*Sfi*I
R.*Sfi*I
SCENARIO I
SCENARIO II
duplication
endonucleolytic cleavage within the regions
encoding loops linking TRD and the catalytic domain
no degradation of the termini
imprecise cleavage or cleavage and
degradation by host exonucleases
Δ
III-IV-V-VI-VII-VIII TRD IX^N-X-I-II
damaged copy A
β-class
III-IV-V-VI-VII-VIII TRD IX^N-X-I-II
single copy is fragmented
III-IV-V-VI-VII-VIII TRD IX^N-X-I-II
damaged copy B
illegitimate recombination -
fusion of two partial MTase genes,
creation of a flawed, but functional MTase
reassembly of fragments in a different order
creation of a MTase gene interrupted by a stop codon.
the catalytic domain and TRD may be expressed
separately and associate to form a heterodimeric MTase
III-IV-V-VI-VII-VIII IX^N-X-I-II VII-VIII TRD IX^N
III-IV-V-VI-VII-VIII IX^N-X-I-II TRD
rapid erosion of
non-functional segemnts
in the presence of the active ENase:
strong pressure
to optimize the MTase function
mutation allowing readthrough
(fusion of two domains)
δ-class
III-IV-V-VI-VII-VIII IX^N-X-I-II TRD
slow evolution:
inversion of the MTase gene and/or
fusion with a different ENase
M.*Mwo*I
R. *Mwo*I

Figure 7
Diagram illustrating a putative evolutionary history of the M.*Sfi*I-M.*Mwo*I gene pairs according to the "cut-and-paste" mechanism. Red vertical bars denote ends generated by endo- or exo-nucleolytic cleavage (red "PacMan" faces). Red arrows delineate points of insertion of gene fragments. Other labels follow the legend to Figure 5.

no intact, active MTase gene copies in the cell. The accompanying ENase gene might be fragmented as well, leading to elimination of the RM system from the cell. However, if the ENase remains active for a certain period of time, only those cells survive in which the MTase gene is restored from fragments. Such MTase may exhibit various deletions, duplications of certain regions and rearrangements, as long as these modifications allow the protein to provide protection against the ENase. The selection pressure will result in rapid optimization of the MTase function, most likely to the nearest maximum in the fitness landscape. With a certain probability, the permuted gene copy will arise, and under such "all or nothing" conditions, its sequence will be optimized towards the modification activity sufficient to protect the host's chromosome. It is worth mentioning that in the short term the ENase may remain active and provide selective pressure for restoring expression of the MTase even if its own gene has been destroyed.

A hybrid mechanism can be envisaged in which the complete, fully functional MTase gene undergoes recombination with a fragment of a MTase gene (not shown). This mechanism has limitations similar to that of the "permutation-by-duplication" mechanism in that the newly fused fragment must not compromise the function of the original protein. However, there is no specific reason why a part of the original domain should be deleted with a higher frequency than the new fragment and why the latter scenario should be selected for, unless the alternative fragments are not identical and the new fragment encodes a function, which may increase the fitness of the protein.

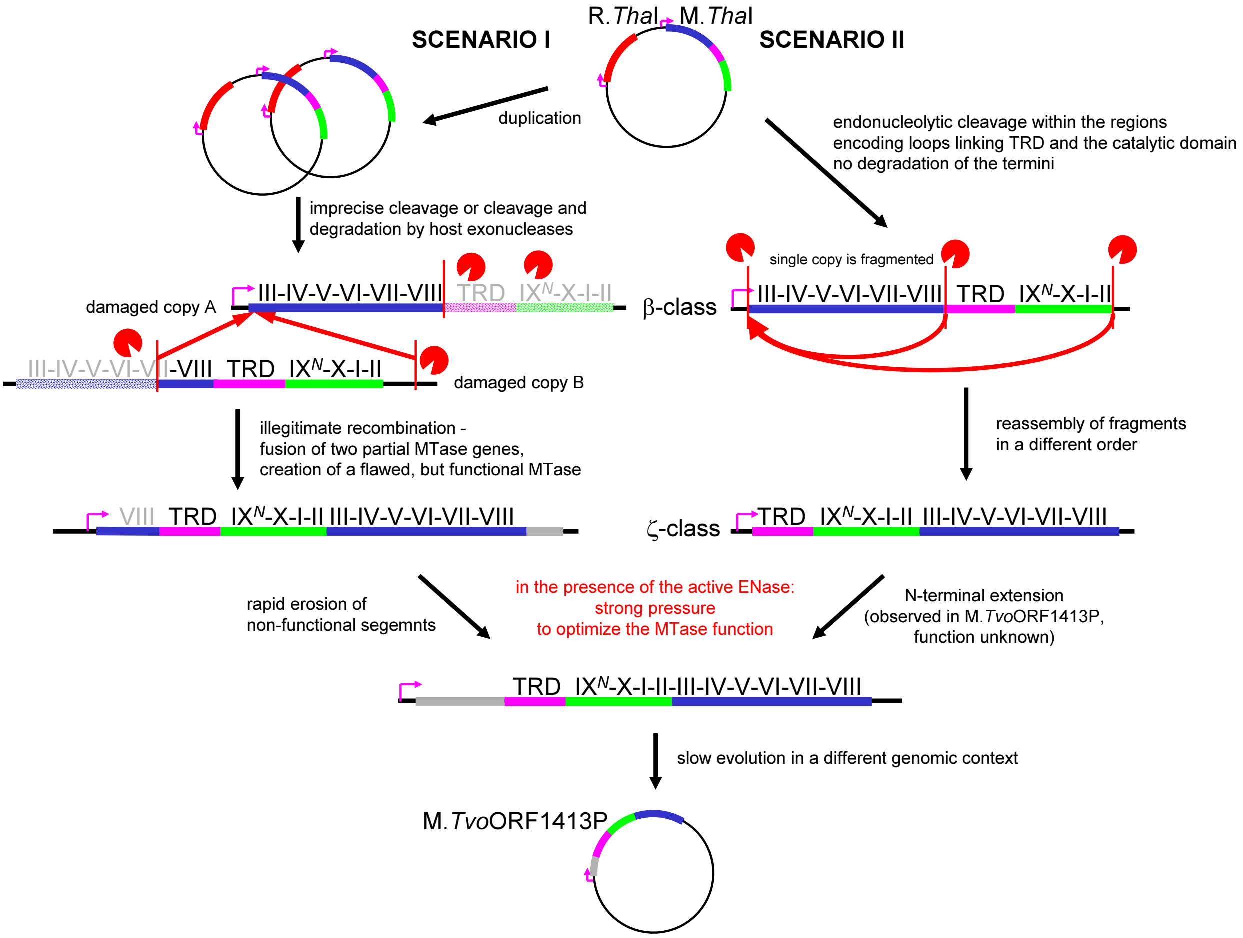
SCENARIO I
R.*Tha*I
M.*Tha*I
SCENARIO II
duplication
endonucleolytic cleavage within the regions
encoding loops linking TRD and the catalytic domain
no degradation of the termini
imprecise cleavage or cleavage and
degradation by host exonucleases
single copy is fragmented
damaged copy A
III-IV-V-VI-VII-VIII TRD IX^N-X-I-II
β-class
III-IV-V-VI-VII-VIII TRD IX^N-X-I-II
III-IV-V-VI-VII-VIII TRD IX^N-X-I-II
damaged copy B
illegitimate recombination -
fusion of two partial MTase genes,
creation of a flawed, but functional MTase
reassembly of fragments
in a different order
VIII TRD IX^N-X-I-II III-IV-V-VI-VII-VIII
ζ-class
TRD IX^N-X-I-II III-IV-V-VI-VII-VIII
rapid erosion of
non-functional segemnts
in the presence of the active ENase:
strong pressure
to optimize the MTase function
N-terminal extension
(observed in M.*Tvo*ORF1413P,
function unknown)
TRD IX^N-X-I-II-III-IV-V-VI-VII-VIII
slow evolution in a different genomic context
M.*Tvo*ORF1413P

Figure 8
Diagram illustrating a putative evolutionary history of the M.*Tha*I-M.*Tvo*ORF1413P gene pair according to the "cut-and-paste" mechanism. All labels follow the legends to Figures 5 and 7.

The scenarios shown in Fig. 5a require that the TRD can function autonomously, outside of its original structural environment. For some TRDs, at least, this is the case (reviews: [2,12]). Not only is the movement of TRDs within an MTase plausible, but the exchange of TRDs between MTases can provide different specificities and thereby functional advantage [32]. Indeed, DNA:m^{5}C MTases that methylate more than one specific DNA target owing to the presence of several TRDs at various locations of the enzyme have been identified [33,34], and shuffling of TRDs have been suggested to occur among both mono- and multispecific DNA:m^{5}C MTases [35]. However, while the unrelated TRDs of many structurally characterized MTases form structurally autonomous domains, the TRDs of two members of the β-class, M.*Rsr*I [23] and M.*Pvu*II [36], form only an amendment of the common fold, which is quite unlikely to behave as an independently folded and functionally autonomous unit (review: [12]).

The putative TRD of M.*Tvo*ORF1413P and M.*Tha*I is most likely too small and too poorly structured to be regarded as an independent domain (Figures 3b, 4). However, the putative TRD of M.*Sfi*I and M.*Mwo*I is much longer and comprises at least four predicted helices (Figure 3a), therefore it cannot be excluded that it may form an independently folded, functional unit. It has been demonstrated that DNA:m^{5}C MTase M.*Aqu*I comprises two independent polypeptides corresponding to the catalytic domain and the TRD, which associate in solution to form a functional enzyme [37]. Correspondingly, one of the scenarios of evolution of the M.*Mwo*I enzyme involves temporary separation of the gene fragments encoding the catalytic domain and the TRD (Figure 7). According to this scenario, the linker region between the catalytic domain and the TRD originated from the initially non-coding sequence that initially separated the two functional units. It seems likely that covalent linkage of the two domains by the newly established linker increased the fitness of the rearranged MTase.

The "cut-and-paste" scenario offers an explanation for the sequence permutations being observed in DNA MTases and not among ENases. This results from the asymmetry of selection for restoring methylation function and restriction function after the corresponding genes are fragmented. In other words, the newly permuted proteins are probably poor enzymes, but the ENase provides strong selective pressure for optimization of the MTase function, while in the second case selective pressure is relatively weak (functional ENases probably provide only a minor selective advantage and they are not required for protection of the host "against" the MTase). However, it cannot be excluded that ENases are simply not amenable to any sequence permutation for structural reasons, since these proteins exhibit different fold than MTases.

Interestingly, sequence permutations have been observed only amongst DNA MTases [11], but not in other MTase families (enzymes acting on RNA, proteins, small molecules, etc.), despite their common structure. To date, no explanation has been offered for this peculiarity, even though it raised considerable interest in the field [11]. It is tempting to speculate that the rearrangements observed amongst DNA MTases but not other MTases are induced [5] by the increased exposure of their genes to the repertoire of various nucleases encoded by different hosts during horizontal transfer events common among the RM systems (J. Elhai, personal communication). High frequency of such events was inferred from sequence analyses [5,38]. If this hypothesis is correct, interaction with the mechanisms of defense against alien genetic elements encoded by various Prokaryotes may be also responsible for permutations of entire domains within type I and type III RM systems and a plethora of combinations of various domains in many "non-classical" RM system subtypes (review: [12]).

Summary

Analyses of the evolutionary scenarios presented herein favor the "cut-and-paste" mechanism or the hybrid mechanism (fusion of an intact MTase with the TRD) for the M.*Sfi*I(β)-M.*Mwo*I(δ) rearrangement and the original "permutation-by-duplication" mechanism or the "cut-and-paste" mechanism (rather than their hybrid) for the M.*Tha*I-(β) M.*Tvo*ORF1413P(ζ) rearrangement. Even though in these and probably other cases, certain scenarios may seem more likely than the other, none of them can be ruled out completely. The presented mechanisms are not mutually exclusive, and all have probably played significant roles in the generation of permuted MTases.

Methods

The PSI-BLAST algorithm [39] was used to search the non-redundant version of current sequence databases (nr) and the publicly available complete and incomplete genome sequences at the NCBI website [http://www.ncbi.nlm.nih.gov] . All genuine and putative N-MTase sequences available from REBASE [26] were submitted as queries with default parameters. Protein structure prediction was carried out using the MetaServer available at [http://bioinfo.pl/meta/] , which combines several secondary structure prediction and threading methods (ref. [40] and references therein). These threading methods compare the query sequence (the target) with a library of structures (templates) and return 10 alignments that scored best according to the implemented criterion of compatibility. The results are evaluated by the Pcons server [41], which compares the models and the associated scores and produces a ranking of potentially best predictions (target-template alignments). Based on the results produced by the MetaServer, homology modeling was carried using the SWISS-MODEL/PROMOD II server [42]. Model evaluation was carried out using the PROSA II [43] program integrated with PROMOD II, suggesting that the stereochemistry and energetic parameters of the models were acceptable.

Acknowledgements

I would like to thank Jeff Elhai, who as a non-anonymous reviewer provided extensive comments, and raised constructive criticism of the "cut-and-paste" mechanism, which lead to considerable improvement of the manuscript, Dr. Ichizo Kobayashi for critical reading and comments on the initial version of the manuscript and Drs. Murat Abdurashitov and Sergey Degtyarev for supplying unpublished data. This work was supported by KBN (grant 6P04 B00519).

References

1. Vertino PM: **Eukaryotic DNA methyltransferases.** *In S-adenosylmethionine-dependent methyltransferases: structures and functions.* 1999, 341-372
2. Dry den DT: **Bacterial DNA methyltranferases.** *In S-Adenosylmethionine-dependent methyltransferases: structures and functions.* 1999, 283-340
3. Bickle TA, Kruger DH: **Biology of DNA restriction.** *Microbiol. Rev.* 1993, **57:**434-450
4. Arber W: **Genetic variation: molecular mechanisms and impact on microbial evolution.** *FEMS Microbiol. Rev.* 2000, **24:**1-7
5. Kobayashi I: **Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution.** *Nucleic Acids Res.* 2001, **29:**3742-3756
6. Malone T, Blumenthal RM, Cheng X: **Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyltransferases, and suggests a catalytic mechanism for these enzymes.** *J. Mol. Biol.* 1995, **253:**618-632
7. Jeltsch A, Christ F, Fatemi M, Roth M: **On the substrate specificity of DNA methyltransferases. Adenine-N6 DNA methyltransferases also modify cytosine residues at position N4.** *J. Biol. Chem.* 1999, **274:**19538-19544
8. Bujnicki JM: **Comparison of protein structures reveals monophyletic origin of the AdoMet-dependent methyltransferase family and mechanistic convergence rather than recent dif-**

ferentiation of N4-cytosine and N6-adenine DNA methylation. *In Silico Biol.* 1999, **1:**1-8 [http://www.bioinfo.de/isb/1999-01/0016/]

9. Bujnicki JM, Radlinska M: **Molecular evolution of DNA-(cytosine-N4) methyltransferases: evidence for their polyphyletic origin.** *Nucleic Acids Res.* 1999, **27:**4501-4509
10. Posfai J, Bhagwat AS, Posfai G, Roberts RJ: **Predictive motifs derived from cytosine methyltransferases.** *Nucleic Acids Res.* 1989, **17:**2421-2435
11. Fauman EB, Blumenthal RM, Cheng X: **Structure and evolution of AdoMet-dependent MTases.** *In S-Adenosylmethionine-dependent methyltransferases: structures and functions.* 1999, 1-38
12. Bujnicki JM: **Understanding the evolution of restriction-modification systems: clues from sequence and structure comparisons.** *Ada Biochim. Pol.* 2001, **48:**1-33
13. Radlinska M, Bujnicki JM, Piekarowicz A: **Structural characterization of two tandemly arranged DNA methyltransferase genes from *Neisseria gonorrhoeae* MS11: N4-cytosine specific M.*Ngo*MXV and nonfunctional 5-cytosine-type M.*Ngo*Morf2P.** *Proteins* 1999, **37:**717-728
14. Bujnicki JM, Radlinska M: **Cloning and characterization of M.*Lmo*A118I, a novel DNA:m4C methyltransferase from the *Listeria monocytogenes* phage A118, a close homolog of M.*Ngo*MXV.** *Ada Microbiol. Pol.* 2001, **50:**151-156
15. Radlinska M, Bujnicki JM: **Cloning of enterohemorrhagic *Escherichia coli* phage VT-2 Dam methyltransferase.** *Ada Microbiol. Pol.* 2001, **50:**157-163
16. Xu SY, Xiao JP, Posfai J, Maunus RE, Benner JS: **Cloning of the *Bss*HII restriction-modification system in *Escherichia coli: Bss*HII methyltransferase contains circularly permuted cytosine-5 methyltransferase motifs.** *Nucleic Acids Res.* 1997, **25:**3991-3994
17. Bujnicki JM: **Homology modelling of the DNA 5mC methyltransferase M.*Bss*HII. is permutation of functional subdomains common to all subfamilies of DNA methyltransferases?** *Int. J. Biol. Macromol.* 2000, **27:**195-204
18. Cao X, Springer NM, Muszynski MG, Phillips RL, Kaeppler S, Jacobsen SE: **Conserved plant genes with similarity to mammalian de novo DNA methyltransferases.** *Proc. Natl. Acad. Sci U.S.A.* 2000, **97:**4979-4984
19. Cheng X, Roberts RJ: **AdoMet-dependent methylation, DNA methyltransferases and base flipping.** *Nucleic Acids Res.* 2001, **29:**3784-3795
20. Jeltsch A: **Circular permutations in the molecular evolution of DNA methyltransferases.** *J. Mol. Evol.* 1999, **49:**161-164
21. Heinemann U, Hahn M: **Circular permutation of polypeptide chains: implications for protein folding and stability.** *Prog. Biophys. Mol. Biol.* 1995, **64:**121-143
22. Labahn J, Granzin J, Schluckebier G, Robinson DP, Jack WE, Schildkraut I, Saenger W: **Three-dimensional structure of the adenine-specific DNA methyltransferase M.*Taq*I in complex with the cofactor S-adenosylmethionine.** *Proc. Natl. Acad. Sci. U.S.A.* 1994, **91:**10957-10961
23. Scavetta RD, Thomas CB, Walsh MA, Szegedi S, Joachimiak A, Gumport RI, Churchill ME: **Structure of *Rsr*I methyltransferase, a member of the N6-adenine beta class of DNA methyltransferases.** *Nucleic Acids Res.* 2000, **28:**3950-3961
24. Heringa J, Taylor WR: **Three-dimensional domain duplication, swapping and stealing.** *Curr. Opin. Struct. Biol.* 1997, **7:**416-421
25. Lupas AN, Ponting CP, Russell RB: **On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?** *J Struct. Biol* 2001, **134:**191-203
26. Roberts RJ, Macelis D: **REBASE-restriction enzymes and methylases.** *Nucleic Acids Res.* 2001, **29:**268-269
27. Lunnen KD, Morgan RD, Timan CJ, Krzycki JA, Reeve JN, Wilson GG: **Characterization and cloning of *Mwo*I (GCN$_7$GC), a new type-II restriction-modification system from *Methanobacterium wolfei*.** *Gene* 2001, **77:**11-19
28. Lauster R, Trautner TA, Noyer-Weidner M: **Cytosine-specific type II DNA methyltransferases. A conserved enzyme core with variable target-recognizing domains.** *J. Mol. Biol.* 1989, **206:**305-312
29. Ruepp A, Graml W, Santos-Martinez ML, Koretke KK, Volker C, Mewes HW, Frishman D, Stocker S, Lupas AN, Baumeister W: **The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*.** *Nature* 2000, **407:**508-513
30. Kawashima T, Amano N, Koike H, Makino S, Higuchi S, Kawashima-Ohya Y, Watanabe K, Yamazaki M, Kanehori K, Kawamoto T, *et al*: **Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*.** *Proc. Natl. Acad. Sci U.S.A.* 2000, **97:**14257-14262
31. Chinen A, Uchiyama I, Kobayashi I: **Comparison between *Pyrococcus horikoshii* and *Pyrococcus abyssi* genome sequences reveals linkage of restriction-modification genes with large genome polymorphisms.** *Gene* 2000, **259:**109-121
32. Chinen A, Naito Y, Handa N, Kobayashi I: **Evolution of sequence recognition by restriction-modification enzymes: selective pressure for specificity decrease.** *Mol. Biol Evol.* 2000, **17:**1610-1619
33. Sethmann S, Ceglowski P, Willert J, Iwanicka-Nowicka R, Trautner TA, Walter J: **M.(phi)*Bss*HII, a novel cytosine-CS-DNA-methyltransferase with target- recognizing domains at separated locations of the enzyme.** *EMBO J.* 1999, **18:**3502-3508
34. Walter J, Trautner TA, Noyer-Weidner M: **High plasticity of multispecific DNA methyltransferases in the region carrying DNA target recognizing enzyme modules.** *EMBO J.* 1992, **11:**4445-4450
35. Bujnicki JM, Radlinska M: **Molecular phylogenetics of DNA SmC-methyltransferases.** *Acta Microbiol. Pol.* 1999, **48:**19-33
36. Gong W, O'Gara M, Blumenthal RM, Cheng X: **Structure of *Pvu*II DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment.** *Nucleic Acids Res.* 1997, **25:**2702-2715
37. Karreman C, de Waard A: **Agmenellum quadruplicatum M.*Aqu*I, a novel modification methylase.** *J. Bacteriol.* 1990, **172:**266-272
38. Jeltsch A, Pingoud A: **Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems.** *J. Mol. Evol.* 1996, **42:**91-96
39. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res.* 1997, **25:**3389-3402
40. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L: **Structure prediction Meta Server.** *Bioinformatics* 2001, **17:**750-751
41. J Lundstrom, L Rychlewski, JM Bujnicki, A Elofsson: **Pcons: A neural-network-based consensus predictor that improves fold recognition.** *Protein Sci* 2001, **10:**2354-2362
42. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling.** *Electrophoresis* 1997, **18:**2714-2723
43. Sippl MJ: **Recognition of errors in three-dimensional structures of proteins.** *Proteins* 1993, **17:**355-362