CLASSIFICATION OF PROKARYOTIC DNA METHYLTRANSFERASES BY TOPOLOGY AND SEQUENCE SIMILARITY

AUTHORS

Maria Samokhina[1,*], Rusinov I.[1] and Alexeevski A.[1,2,5]

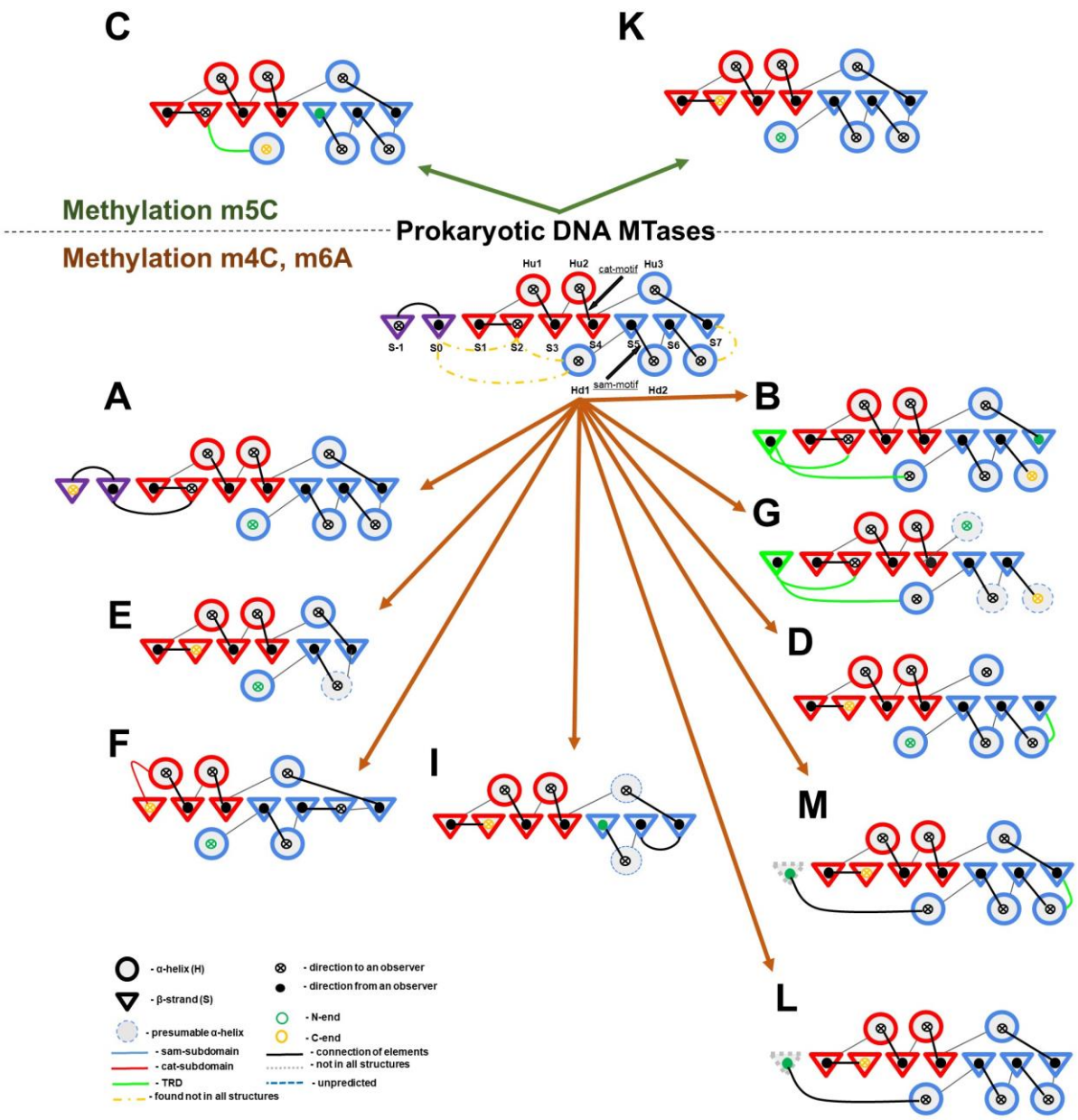
[1] Belozersky Institute of Physical and Chemical Biology, Lomonosov Moscow State University, Moscow, Russia

[2] Faculty of Bioengineering and Bioinformatics, Lomonosov MSU, Moscow, Russia

[3] Research Institute for System Analysis, RAS, Moscow, Russia

* Email: mashila6799@gmail.com

GRAPHICAL ABSTRACT



ABSTRACT

Prokaryotic DNA methyltrasferases (MTases) are essential enzymes within restriction-modification systems, preventing self DNA cleavage by restriction enzymes. They also participate in the regulation of gene expression, DNA repair, DNA replication and other processes.

We propose a sequence-based homology classification of prokaryotic MTases, addressing a deficiency that has existed for over 20 years in the previous classification of MTases by Malone, Bujnicki into groups α, β, γ, etc.

We defined 11 MTases classes on catalytic domain 3D topology by analysis available 3D structures and certain predicted 3D structures of MTases from AlfaFold DB. For each class we have selected from domain family's databases HMM-profiles, that predicts in the hits the topology class on the base of sequence similarity with catalytic and SAM-binding subdomains. Based on the set of non-predicted MTases from DB REBASE we developed the algorithm for MTase topology classification and location conservative catalytic and SAM-binding motifs. The algorithm was applied to all available MTases. Topology classes are further subdivided into subclasses with the same domains architectures according Pfam DB.

Homology classification allows to recover the MTases evolution as well as MTase annotations in existing and new sequences data sets.

INTRODUCTION

DNA methylation of C5 cytosine in Bacteria was discovered in 1925 (see the review (1)). Thus, DNA methylation has almost a century of history! Since its discovery, much has been learned about the role of genome methylation and DNA methylation enzymes (2, 3).

EC classification and history

According to the catalyzed chemical reactions, DNA methyltransferases (MTases) belong to three groups of Enzyme Classification: m5C MTases, producing C5-methylcytosine (EC 2.1.1.37); m6A MTases that produce N6-methyladenine (EC 2.1.1.72); m4C MTases producing N4-methylcytosine (EC 2.1.1.113) (4). All known MTases explore S-adenosyl-L-methionine cofactor (SAM, or AdoMet) as a donor of methyl group, except synthetic SAM modifications used in experiments (5).

MTases of all three EC-groups are ubiquitously among Bacteria and Archaea, with a prevalence of m6A MTases (6).

In Eukaryotes m5C MTases methylating cytosines within symmetric dinucleotide CG are prevalent. The roles of the methylation in Mammalia and Humans are well studied in various circumstances up to human heaths and diseases (7).

N6-methyladenine in Eukaryotes have also been detected, primarily in unicellular organisms (8, 9) in 1970th. Later N6-methyladenine were detected in Chlamydomonas (10) , Drosophila (11) and C. elegans (12). In the last decade on the basis of developed new methods for detection of m6A in genomes new hits of these modifications were found in a number of other organisms including mammals and humans (13). Also "writers", proteins producing N6-methyladenine in genome, "readers", proteins accepting methylation signal and "erasers", proteins removing this modification were determined, but many questions about the role of m6A modifications in human genome remain unsolved (13).

N4-methylcytosine modifications are not detected in eukaryotic genomes yet. There is only one published work that claims and proves the existence of N4-methylcytosine modifications in genomes of two closely related populations of Adineta vaga rotifers (14). The intriguing feature of rotifers is horizontal gene transfer to their genomes from DNA of bacteria and other habitats of a microbiota (15). Additional confirmations of m4C in eukaryotes are expected.

In this work we focused on prokaryotic MTases. Their first discovered function was participating in restriction-modification (R-M) systems (16). m5C MTases were components of those systems.

Then m6A MTases were also discovered within bacterial R-M systems (17). Later m6A modifications were studied in dam gene, participated in bacterial genome replication and mismatches repair (18).

The m4C modification was proven in 1983 (19) for MTase BcnI from the eponymous R-M system. Two years later all three EC-groups of DNA methylation were identified and studied in DNA in thermophilic bacteria (20).

Important step in the field of MTases was the classification of restriction enzymes, based on unified nomenclature for R-M components (21). R.J. Roberts in Cold Spring Harbor Laboratory (U.S.A.) laid the foundations of the collection of prokaryotic restriction enzymes and DNA methyltransferases and other R-M components. He learned about new restriction enzymes through intensive communication with specialists and personal research. At first, these data were published in NAR articles with large tables (22, 23). The tables were regularly

updated with new discovered restriction endonucleases in the consequent NAR publications. They were shared with specialists with e-mail and by personal requests (24).

The collection was named REBASE, - famous database collecting all data on R-M systems from at least 1970th till now. In the Internet era, REBASE became available through anonymous ftp (1994) (25) and www (1996) (26).

REBASE MTase classification

In the REBASE DBs MTases are classified according to the R-M systems it belongs to. Namely, type I MTases, type II MTases, type III MTases; MTases that additionally have endonuclease activity - RM proteins according REBASE nomenclature, are either IIG RM proteins or type IV RM proteins. All prokaryotic MTases that do not belong to any restriction-modification system are considered in the REBASE as orphan MTases.

Conserved Motifs classification of Malone (27) and Bujnicki (28)

In 1995, Malone, based on the order of three parts (catalytic, SAM-binding and DNA-binding) in 42 MTases (known at that time), proposed a classification into 6 groups (Fig.1, adopted from (28)), according to all possible permutation of these three parts. Later these three parts were called catalytic subdomain (we will call below cat-subdomain), SAM-binding subdomain (we will call below sam-subdomain), target recognition domain (TRD). The groups were designated α, …, ζ. Representatives of only 3 of them (α, β, γ) were known at that time. In amino acid sequence alignments of MTases, conserved motifs in the catalytic domain were identified. Motifs are designated by numbers: I-VIII and X (27, 28). Motifs IV-VIII are assigned to the catalytic subdomain, motifs X and I-III are included in the SAM-binding subdomain. Conservative motifs were found separately in each group. According to the descriptions in the article (28), finding each of the motifs in sequence was difficult for the authors. The motif alignments presented in the article confirm the difficulty of finding conserved motifs, with the exception of two motifs IV (catalytic) and I (SAM-binding), which have the same conservative columns in all groups. No conserved motifs were found in the DNA-binding region, which is the most variable.

In (28), m5C MTases were separated from the gamma group into a separate group due to a cyclic rearrangement that moved the X motif from the N-terminus to the C-terminus. Two MTases have also been described that represent the zeta group, and one that appears to

represent the delta group, although it lacks the TRD subdomain. Currently, 5 proteins from the zeta group have been found: M.BssHII, M.Alw26I, M2.Eco31I, M.Esp3I and M2.BsaI (29). The zeta group includes m5C MTase proteins. As suggested in (28) TRD in these proteins is located before the catalytic domain.

α: N – X-I-II – TRD – III-IV-V-VI-VII-VIII – C

β: N – III-IV-V-VI-VII-VIII – TRD – X-I-II – C

γ: N – X-I-II-III-IV-V-VI-VII-VIII – TRD – C

δ: N – III-IV-V-VI-VII-VIII-X-I-II – TRD – C

ε: N –TRD – III-IV-V-VI-VII-VIII-X-I-II – C

ζ: N –TRD – III-X-I-II-IV-V-VI-VII-VIII – C

C5: N – I-II-III-IV-V-VI-VII-VIII – TRD – X – C

Figure 1. Adopted from (28). Final classification of MTases according to catalytic motifs, proposed in (28). Blue indicates motifs from the SAM-binding subdomain, red indicates motifs from the catalytic subdomain.

It is impossible to determine the number of proteins classified according to conservative motifs, because this classification is based on a small number of proteins (less than 100 sequences) and there are no methods for detecting all motifs and DNA-binding domains described in (27, 28). Of the 9 motifs described in the motif classification, in the modern data set, motif I (also called the SAM-binding motif or sam-motif) and motif IV (also called the catalytic motif or cat-motif) are sometimes detected in DB REBASE, but in a large number of REBASE MTases there is no any detected motifs.

MTases functions in prokaryotes

MTases are not only components of the protective immune systems of bacteria, but also participate in the regulation of gene expression, control of replication and repair of chromosomal DNA and other processes (30).

Dam is a collective name of MTases that transfer a methyl group from SAM-molecule to the N6 position of the adenine residues in the site 5′ -GATC- 3′ (31). Dam was initially discovered

in the bacteria Escherichia coli, where it was shown to be involved in methyl-directed DNA mismatch repair (MMR) and regulation of chromosome replication (31, 32).

CcrM MTases transfer a methyl group from SAM-molecule to the N6 position of the adenine residues in the site 5′ -GANTC- 3′ (33). Evidence suggests that CcrM is involved in the regulation of the cell cycle of C. crescentus by regulating the expression of cell division genes (34).

Dcm MTases can methylate the second cytosine at its 5'CCAGG 3' and 5'CCTGG 3' sites. Methylated 5'CAG3' sequences represent mutation hotspots. After deamination, 5-methylcytosine is a thymine and is not removed by uracil-N-glycosylase, so GC to AT mutations are common at 5' CCAGG 3' sites (30).

In the work we propose a new two-step homology classification of prokaryotic MTases. The first level of the classification is based on topologies of catalytic domains and conserved motifs. Second level is based on domain architectures of MTases. We have constructed our classification on the set of non-predicted (non-P) MTase from DB REBASE. REBASE is the most complete source of MTase sequences. We studied all available 3D structures and predicted 3D structures of MTases from AlfaFold DB. We split MTase to 11 classes according to their 3D topology. MTases from each topology have unique characteristic motifs that confirm the common evolutionary origin of each class. We have selected HMM profiles that could locate catalytic domains. We have described MTases with big loops catalytic domain. We provide algorithms for MTase classification and searching conservative motifs I and IV based on HMM profiles.

MATERIAL AND METHODS

Materials

MTase 3D structures. The REBASE list of the MTases with known 3D protein structures, downloaded on November 16 2023, incuded 33 MTases, 32 structures were solved by X-ray diffraction and one (RM.DrdV) by Electron microscopy (table S1a). Among them 5 MTases includes not only the MTase catalytic domain but also the catalytic domains of restriction endonuclease. Such proteins are marked RM in REBASE. We didn't find any other prokaryotic MTases with known 3D structures using the Uniprot services, see below. Predicted Structures of several MTases were taken from DB AlphaFold (https://alphafold.ebi.ac.uk/).

REBASE source of MTase sequences. The REBASE database as of June 2, 2022 contained 499,311 proteins. From these, we selected prokaryotic proteins that, according to REBASE, have methyltransferase activity, a total of 230,873 prokaryotic MTases. For all 230873 MTases, available information about MTases was downloaded from REBASE: methylation type, type of R-M system to which it belongs (type I-IV, Orphan), links to sources of information about MTase. 230873 MTases fall into 133990 clusters of 100% sequence identity (supplementary materials S2a-b). In REBASE MTases, like other proteins, are divided into biochemically verified (non-predicted) and computationally predicted. The names of the predicted MTases contain the suffix "P" (35). From 133990 representatives of clusters with 100% identical sequences, we selected 7017 sequences that make up the sample of non-P (non-predicted) MTases (supplementary materials S2c). If there was a non-P MTase in the cluster, then they took it.

Uniprot source of MTase sequences. On November 12, 2023 we found in Uniprot DB 187702 annotated MTases using advanced search. The request included restriction by taxonomy - prokaryotes: (eubacteria OR archaea) AND Enzyme Commission number - (EC 2.1.1.113 OR EC 2.1.1.72 OR EC 2.1.1.37). In addition to entry standard information (identifiers AC, ID, gene names; protein name; status: reviewed/unreviewed; organism; length) output table includes also Pfam and Supfam domains; REBASE index; PDB codes; taxonomic lineage. We verified that all MTases with known 3D structure according to the Uniprot annotation are in the list downloaded from the REBASE (*MTase 3D structures*).

Methods

Topology classification of MTases. Architecture and topology of all 3D structures were done by one of the authors visually using standard software PyMOL (https://pymol.org/).

Soft for REBASE data downloading and analyzing. We have developed a set of scripts for downloading and primary processing the data available on REBASE FTP server (http://rebase.neb.com/rebase/rebase.ftp.html). The processing consists in changing the format of data to generally accepted fasta and TSV; correction of some protein annotations and assigning individual proteins to R-M systems based on REBASE protein names nomenclature; and clustering of identical proteins. The scripts are available in the repository https://github.com/isrusin/rebase.

HMM-profiles search. In protein families DBs Pfam, Supfam, CDD, PANTHER and others there are a number of protein domain families, whose annotations include terms like "DNA methyltransferase domains". Each domain family is provided with a model, allowing to detect domains of a family within an input sequence.

In this work we use HMM profiles from DBs SUPFAM (https://supfam.mrc-lmb.cam.ac.uk/) and Pfam, which services are now takeover into Interpro DB (https://www.ebi.ac.uk/interpro/).

Hits of all profiles from the SUPFAM database in a sample of 133990 MTases from REBASE were determined. To do this, all HMM profiles of SUPFAM models were downloaded from the SUPFAM DB version 1.75 to the local server. The search for HMM profiles was carried out by the hmmsearch program from the hmmer package version 3.3.2 for all proteins from REBASE. If several profiles from the same superfamily were found in one place of the same sequence, then only the profile with the highest weight was retained.

Hits of all profiles from the Pfam DB in a sample of 133990 from REBASE were identified. To do this, all HMM profiles of curated Pfam-A domain families, version 34, were downloaded from the Pfam database to the local server. The search for HMM profiles was carried out by the hmmsearch program from the hmmer package version 3.3.2 using all proteins from REBASE. If several profiles from the same clan were found in one place of the same sequence, then only the profile with the highest weight was retained.

Domains and domain architectures detection. To build domain architectures, we used HMM profiles from DB Pfam. To do this, after searching for profiles in the tested proteins, we retained the best hits by weight that passed the threshold of e-value 0.01. Next, the hits were sorted in the order in which they appeared in the MTase sequence. In cases where the profiles overlapped with the hits of catalytic profiles, the location of such a profile and the subdomains of the catalytic domain in the protein was analyzed, depending on this, the profile was placed either before the catalytic or after the catalytic domain in the domain architecture.

Detecting MTase profiles. For selection of cat-domain profiles we primarily have considered models from DB SUPFAM (https://supfam.mrc-lmb.cam.ac.uk/) as those HMM profiles are

based on the three-dimensional structures of proteins. Each SUPFAM model corresponds to a certain protein with known 3D structure. We selected those models that correspond to MTase structures. Using 3D structures we marked in Supfam models regions corresponding to cat-subdomains, sam-subdomains and cat-motifs, sam-motifs (table S3a). Thanks to this profile marking, we identified among the profile hits - complete hits that cover both cat- and sam- subdomains in the order defined in the profile marking and the distance between estimated borders of subdomains is less than 20 aa.

From all Supfam models we selected 12 MTase profiles that locate complete catalytic domain in the largest number of non-P MTases, and adding other profiles does not increase the number of MTase hits in the sample non-P MTases.

Supfam models hits in non-P REBASE sample cohort revision showed that not all MTases form non-P REBASE sample were covered by complete hits of Supfam models. We added profiles from DB Pfam: all from Rossman clan and two Pfam profiles (Dam, EcoRI_methylase) that are not involved in any clan and annotated as domains with DNA methylation function.

All these Pfam HMM profiles were searched in non-P REBASE MTases. Pfam profiles without hits were excluded as well as profiles whose all hits were covered by hits of 12 SUPFAM HMM profiles. Only 3 Pfam profiles (MT-A70, Dam, EcoRI_methylase) detected new non-P hits. They were added to the 12 SUPFAM profiles. Thus 15 profiles were used for MTases classification.

Topology classes assignment. For Supfam profiles topology class A-D were assigned according the topology class of the corresponding to Supfam profile 3D structure of MTase. Predicted 3D structures from the AlphaFold DB were used to assign topology class to MT-A70, Dam, EcoRI_methylase profiles (see table table S1b). Using predicted 3D structures from Alphafold DB we marked in the 3 Pfam models regions corresponding to cat-subdomains, sam-subdomains and cat-motifs, sam-motifs (see table S3b).

Representative proteins of topology classes. For all MTases, we determined the topology using expert evaluation of available and predicted structures, as well as alignment of proteins with proteins with accessible structure. For each topology, we compiled a list of representatives for which the alignment according to the detecting MTase profile does not contain insertions

or deletions more than 20 aa. The number of all and representative MTases of each topology and selection criteria are given in supplementary table S4.

Alignments of representative proteins. Alignments for all classes except I were based on the alignment of representative MTases according to the detecting MTase profile. From the alignment of the protein according to the detecting profile, the regions corresponding to the SAM-binding and catalytic subdomains and the regions corresponding to motif I and IV were cut out using profile marking. Scripts for obtaining protein profile alignment and cutting out regions are shown on https://github.com/MVolobueva/MTase-classification.git and described in the file Classification_pipeline.ipynb.

The coordinates of the regions aligned to the profile and the aligned sequences of all non-P MTases are given in the supplementary materials S5a. From the files, lines corresponding to representatives of the topologies were selected and subdomain alignments were constructed for each topology using Jalview program.

For class I alignment, an alignment of full-length sequences for MTase with DPPY catalytic motifs and an alignment of full-length sequences for MTase with DPPH catalytic motifs were constructed using Jalview program.

All alignments are given in the supplementary materials S5b.

RESULTS

Topological classification of MTases on the base of MTses' 3D structures

All known DNA methyltransferases belong to class I methyltransferases, i.e. have Rossmann fold (36). Catalytic domains (cat-domains) of MTases have seven stranded β--sheet with α-helices on both sides of the β--sheet. Strands have the following orientation relative to each other: (↓↑↓↓↓↓) (37), despite observed permutation of catalytic and SAM-binding subdomains (28).

We have analysed the topology of all 33 MTases with known 3D structures. Primary sequence identifiers we took from REBASE because all MTases with known 3D structures are presented in this database.

In each structure we identified a cat-domain. Within the cat-domain we identified the following subdomains: catalytic subdomain (cat-subdomain), SAM-binding subdomain (sam-subdomain) and DNA-binding domain (TRD), if presented. It was done as in earlier works (28). Cat- and sam-subdomains are easily distinguished as visually separated parts of the globule (Fig. 1c-f). TRD are represented by either a large loop (Fig. 1d) or separated part of the globule (Fig. 1e-f).

Within the cat-subdomain we localized the catalytic motif (cat-motif) also known as motif IV and within the sam-subdomain we localized the SAM-binding motif (sam-motif) also known as motif I (28).

We confirmed the conservation of architectures of the Rossman fold in MTases. In all 33 structures of MTases cat-domains' β--sheet of seven strands is presented; three helices are presented on the 'catalytic' side of the sheet, and also three helices are presented on the 'SAM-binding' side (Figure 1A). Figure 1b introduces denotations of all conserved secondary structure elements, which we use below. Cat-subdomain consists of 4 conserved α- helices (Hd1,Hd2,Hd3,Hu3) and 3 β- strands (S5,S6,S7); it is shown in red. Cat-motif is located in the loop between ss-elements S5 and Hd2 in all structures. Sam-subdomain consists of 4 α- helices (Hd1,Hd2,Hd3,Hu3) and 3 β- strands (S5,S6,S7) ; it is shown in blue. Sam-motif location is also conserved, it is located in the loop between conserved ss-elements S4 and Hu2.

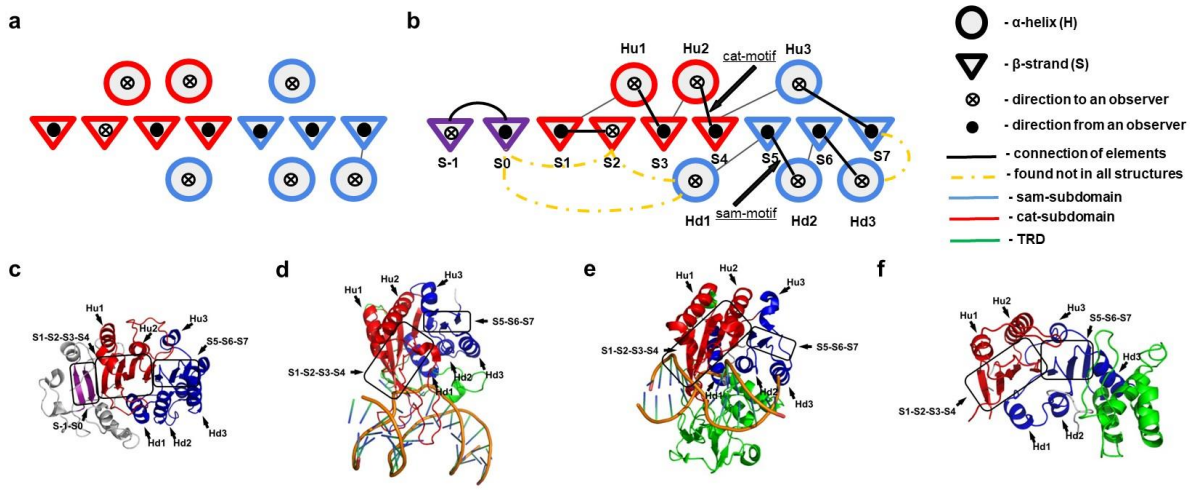


Figure 1. Common architecture of cat-domains of MTases with known 3D structure. a. Scheme of conserved core of MTase Rossmann fold architecture, α-helices (H) and β-strands

(S). b. Scheme of all ss-elements found in all MTase cat-domains 3D structures and connections between its. Figures c – f demonstrate ribbon representation of cat-domain of particular MTases. Colors corresponds to colors of ss-elements in figure b. TRD domains are shown in green. Black frames highlight the characteristic ss-elements of MTase cat-domain. c. The cat-domain of the MTase M.EcoKI (pdb code 2AR0). d. The cat-domain of the MTase M.CcrMI (pdb code 6PBD). e. The cat-domain of the MTase M.HhaI (pdb code 5CIY). F. The cat-domain of the MTase M1.DpnII (pdb code 2DPM).

Despite the extremely conserved MTases architecture, the topology, including even connectivity of conserved ss-elements, varies. By topology of cat-domains all 33 MTases with known 3D structures fall into four groups A (fig.1C), B (fig.1D), C (fig.1E) and D (fig.1F) with different topologies of the MTase cat-domain (Fig. 2, table S1a). In topologies B, C, and D cat-motif follows sam-motif (sam-cat), whilst in B cat-motif precedes sam-motif (cat-sam). Helix Hd1 is N-terminal in topologies A and D, C-terminal in C and it is in the center of a polypeptide chain in topology B.

Topologies also differ in position of the target recognition domain (TRD). In B, C, D topologies, the TRD forms a subdomain between different consecutive ss-elements of a catalytic domain. MTases with topology A are mainly from type I R-M systems, which include separate DNA recognition protein S-subunit. Other MTases with topology A contain the TRD as a separate domain C-terminal to a cat-domain. They may have appeared due to the fusion of M and S genes, as fusion of MTase from type I R-M systems and TRD with S-subunit are shown to have MTase activity (38). In topology A there are also two extra β-strands which we include in cat-subdomain. In several structures from topology B there is one extra β-strand which we include in TRD. In other structures from topology B there is an element that may be an extra β- strand, but proximity to TRD violates secondary structure prediction.

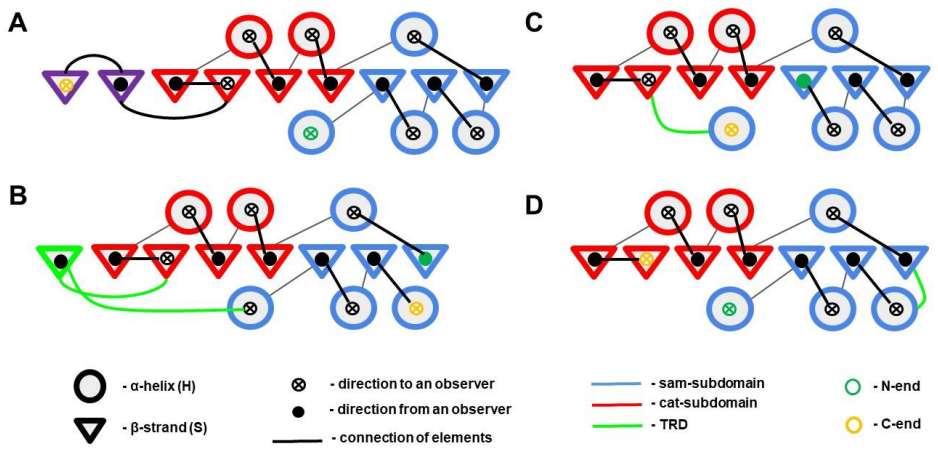


Figure 2. Groups of topological schemes of all MTases with known 3D structures.

Some MTases from REBASE have predicted 3D structures in AlphaFold DB. In addition to the already known 4 topologies, 7 new topologies were found (Fig. 3, table S1b). It should be understood that structure predictions may differ from the actual structure. Only topologies K, L and M includes all MTase Rossman fold canonical ss-elements shown on figure 1A. Topology K is similar to topology A, but does not contain additional two β- strands. In several predicted structures from topologies L (M1.BcnI, M.ThaIV, M.CagIX) and M (M.PciI, M1.BtsI) there are extra β-strand as in topology B. Topologies E, G and I have conserved topology of cat-subdomains and a reduced set of ss-elements of sam-subdomains. Topology F has conserved architecture, but significant differences in topologies of both cat-

and sam-subdomains.

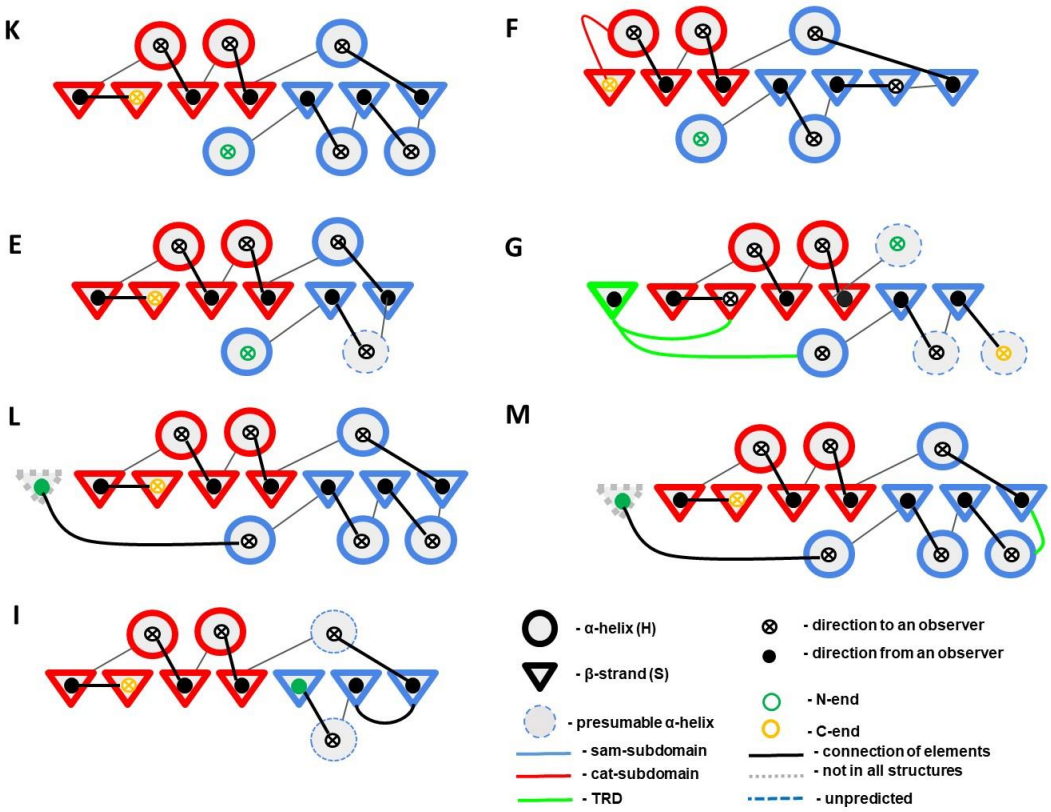


Figure 3. Groups of topological schemes of all MTases with predicted 3D structures in DB AlphaFold.

MTase HMM profiles

We have selected existing HMM profiles that annotate cat-domain in MTases. Those HMM profiles we have called cat-profiles. We have adopted cat-profiles for topology groups detection. For selection of cat-profiles we have considered models from DB SUPFAM (https://supfam.mrc-lmb.cam.ac.uk/) as those HMM profiles based on the three-dimensional structures of proteins. A Supfam model corresponds to a certain protein with known 3D structure (model MTase). Also we considered the HMM profile from DB Pfam (http://pfam.xfam.org/) as those profiles are based on the alignment of the protein and its reliable homologs. We adopted already selected existing HMM profiles of DNA methyltransferase catalytic domains from DBs for topology groups detection.

From the S-adenosyl-L-methionine-dependent methyltransferases superfamily (SF53335) in DB SUPFAM we selected all 5 families that correspond to DNA methyltransferases. Each

family includes several models, totalling 12 HMM profiles for 5 families. For each selected 12 HMM profiles we assigned a topology group by topology of model MTase (table S6.a). Cat-profiles from families Type II DNA methylase, C5 cytosine-specific DNA methylase, N6 adenine-specific DNA methylase were assigned to topologies B, C, D accordingly. Cat-profiles from families N-6 DNA Methylase-like and DNA methylase TaqIwere assigned to topology A. The difference between two families from topology A lies in the fact that the family DNA methylase TaqI is built on MTases that contain additional C-terminal TRD domain, family N-6 DNA Methylase-like is built on MTases from RM systems I and do not contain TRD domain and contain additional N-terminal non-catalytic domain.

All 12 cat-profiles were searched in non-P REBASE MTases. Among 7017 non-P MTases there were MTases that did not contain hits of these cat-profiles. Because of this we decided to add profiles from DB Pfam: Rossman clan and two Pfam profiles (Dam, EcoRI_methylase) that are not involved in any clan and annotated as domains with catalytic function. All these Pfam HMM profiles were searched in non-P REBASE MTases. Pfam profiles without hits were excluded. All hits of profiles from the Rossman clan were covered by hits of 12 SUPFAM HMM profiles, except 3 profiles (MT-A70, Dam, EcoRI_methylase) which detect MTases with specific topology of catalytic domain. To correspond profiles MT-A70, Dam and EcoRI_methylase to topology types, predicted 3D structures from Alphafold DB were used (table S6.b). Selected 15 profiles (12 from DB SUPFAM and 3 from DB Pfam) we will call cat-profiles.

As a result each cat-profile was assigned to one of the MTase topologies: A, B, C, D, E, F, G (table 2). Surprisingly there were no profiles with topology K, L, M and I. 6954 out of 7017 selected sequences had at least one hit from 12 SUPFAM profiles. 6802 out of 7017 selected sequences had at least one hit from Pfam cat-profiles. Among proteins without cat-profile hits there were no sequences that would be confirmed MTases or would have predicted MTase fold.

Table 2. Correspondence of cat-profiles and topological groups. Selected cat-profiles are shown in bold.

Topology group	SUPFAM family	SUPFAM ID	Number of SUPFAM hit**	Model Pfam ID***	Pfam ID	Number of Pfam hit**
A	*N-6 DNA Methylase-like*	**0051816** **0052618** **0054378**	*4267* *5756* *4773*	*N6_Mtase* *MTS* *Methyltransf_31* *RrnaAD* *UPF0020* *Eco57I*	*PF02384* *PF05175* *PF13847* *PF00398* *PF01170* *PF07669*	*3537* *1010* *348* *11* *1301* *1056*
	DNA methylase TaqI, N-terminal domain	**0053087**	*5709*	*MmeI_Mtase* *Methyltransf_5* *Methyltransf_25* *Methyltransf_11* *Methyltransf_23* *PCMT*	*PF20473* *PF01795* *PF13649* *PF08241* *PF13489* *PF01135*	*807* *1* *148* *43* *84* *1*
B	*Type II DNA methylase*	**0036976** **0037952** **0045988**	*3752* *2616* *2585*	*N6_N4_Mtase*	*PF01555*	*1718*
C	C5 cytosine-specific DNA methylase, DCM	**0045633** **0046303** **0046923**	*695* *567* *568*	*DNA_methylase*	*PF00145*	*629*
D	*N6 adenine-specific DNA methylase, DAM*	**0048856** **0052484**	*1646* *1867*	*MethyltransfD12*	*PF02086*	*1043*
E				**Dam**	PF05869	11
F				**EcoRI_methylase**	PF13651	23
G		0036976*		***MT-A70***	PF05063	35

		0037952* 0045988*				

* SUPFAM cat-profiles from group B detect only cat-subdomain in MTase sequences.

** Among non-P MTases with e-value threshold 0.01.

*** Profiles that had at least one hit with the highest weight among the profiles of the Rossman clan.

Marking of 15 HMM profiles. Our goal was to adopt selected HMM profiles for detection cat-domain and its topology in MTases. As seen from table 2, the number of cat-profiles hits exceeds the number of MTases. There are several reasons for this. The first one is that several cat-profiles could have hits in the same MTases. Cat-profiles from the same groups have a comparable number of hits as they are constructed on homologous proteins. The second reason is multidomain MTases that have 2 cat-domains or have additional non-catalytic domains that also are detected by cat-profile. The third reason is permuted proteins. In this case one cat-profile may have several hits in one MTase. The first hit detects one part of the cat-domain, the second hit detects the other part of the cat-domain. The fourth reason is large insets between ss-elements in a cat-domain. In this case also one cat-profile may have several hits in one MTase.

To adopt HMM models for assigning MTase to topology group and for detection of regions of a cat-domain we marked up in cat-profiles regions of sam-subdomains and cat-subdomains, sam-motifs and cat-motifs using available or predicted structures of model MTases. As cat-motif we consider four amino acids in S5-Hd1 loop, as sam-motif we consider 9 amino acids in S4-Hu2 loop. The marking of cat-profiles presented at figure 4 and table S7.

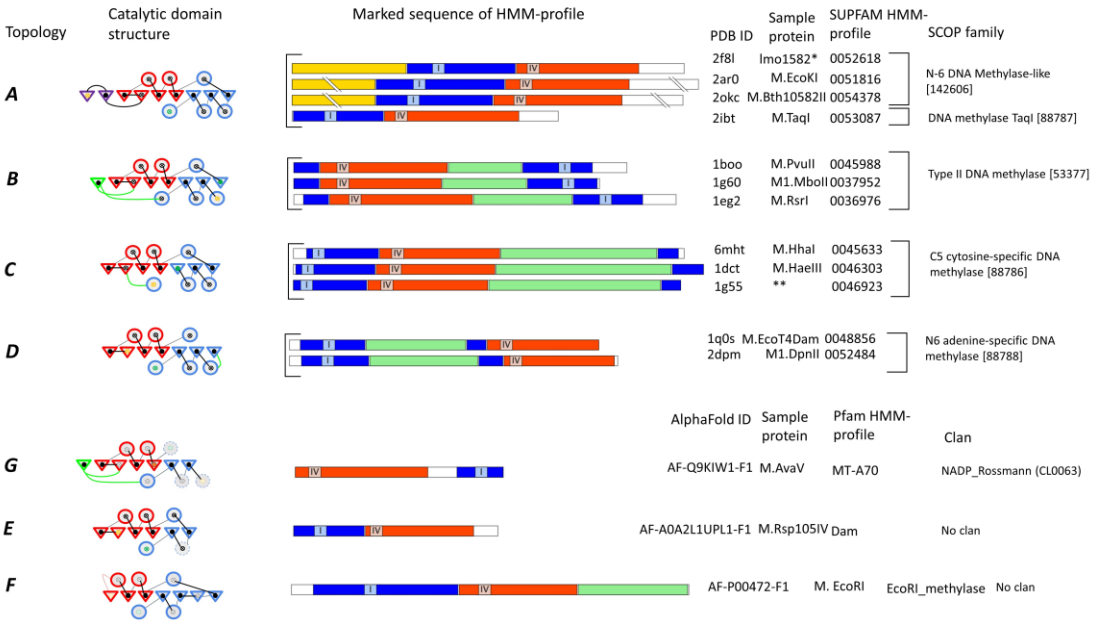


Figure 4. Marking of cat-profiles. Blue colour indicates sam-subdomain, red colour indicates cat-subdomain, green colour indicates TRD sequence.

MTases with two cat-profiles hits. We have discovered MTases in which a single cat-domain is detected by two cat-profile hits. After marking the profiles, we analyzed the location of the subdomain in all non-P MTases. MTases with with two cat-profiles hits are divided into 2 groups: (1) MTases in which there is a large insert inside the cat-domain, (2) MTases in which the topology is permuted compared to the detection cat-profile.

Topologies K, L, M and I were discovered for MTases in which the topology is permuted compared to the detecting cat-profile (table 3). The difference between classes L and M is that in class L the distance is less than 20, in class M it is more than 20. In class I, cat-profiles detect motifs and several amino acids around the motifs.

Table 3. Topology permuted compared to the detecting cat-profile.

Topology	Detecting cat-profiles	Elements detecting by first cat-profile hit	Elements detecting by second cat-profile hit	Sample protein
K	0045633 0046303 0046923	Hd1	S5-Hd2-S6-Hd3-S7-Hu3-S4-Hu2-S3-Hu1-S2-S1	M2.Eco31I
L	0036976 0037952 0045988	Hd1-S5-Hd2-S6-Hd3-S7-Hu3	S4-Hu2-S3-Hu1-S2-S1	M1.BcnI

M	0036976 0037952 0045988	Hd1-S5-Hd2-S6-Hd3-S7-Hu3	S4-Hu2-S3-Hu1-S2-S1	M1.BtsI
I	0036976 0037952 0045988	sam-motif	cat-motif	M.BceJII

Classification of non-P MTases

6913 non-P MTases have been classified to 11 topologies by expert assignment of available or predicted 3D structure and sequence similarity with known MTases (table 4, table S8a). In 67 non-P MTases only one subdomain has been found (table S8b). Among proteins not assigned to any topology there has been no sequences that would be confirmed MTases or would have predicted MTase fold (table S8c).

We have analysed hits of regions (cat-subdomain, sam-subdomain, cat-motif, sam-motif regions) from marked HMM-profiles in non-P MTases. We have revealed that topology of each MTases could be detected through region hits in certain order from one cat-profile group in MTases sequences.

The marked cat-profiles turned out to be good tools for detecting the coordinates of cat- and sam-subdomains and cat- and sam-motifs in the MTase sequences. Cat-profiles detect cat-motifs 6948 and sam-motifs 6945 sequences from 7017 non-P MTases.

We constructed a representative alignment of subdomains by cat-profile for all classes (supplementary S5). We found that each class has its own unique characteristic set of motifs (table S9). This is especially noticeable in larger classes A, B, C, D. Class A has a unique set of motifs that does not intersect with other topologies, except classes E and F. Classes C and K have intersected sets of motifs, also classes C and K are detected by profiles from group C that indicate the common evolutionary origin of these classes and the permutation event in their evolution. Classes B, L and M that often one extra β- strand (S0) are detected by cat-profiles from group B and have intersected sets motifs. Classes B, L and M are likely to have common evolutionary origin too. Class D is similar to class L and has similar motifs, but does not contain extra β-strand (S0), so this class is likely to have occurred as a result of the loss of extra β-strand (S0). Cat-subdomain of class G is similar to class B, and for both classes cat-domains are detected by profiles from class B, but class G has reduced sam-subdomains that are detected only by cat-profiles from group G. Class I contains MTases where cat-subdomain is detected by cat-profiles from group B, while the sam-subdomain is reduced.

The presence of hits of cat-profiles from the same topological group as well as the uniqueness of the set of motifs in MTases from one class confirms a common evolutionary origin for each class.

Table 4. Classification of all non-P MTases.

Topology/ Class	Detecting cat-profiles	Connectivity of ss-elements	Most common cat-motifs
A	51816 52618 54378 53087	Hd1-S5-Hd2-S6-Hd3-S7-Hu3-S4-Hu2-S3-Hu1-S2-S1-S-1-S0	NPPF, NPPY
B	36976 37952 45988	S7-Hu3-S4-Hu2-S3-Hu1-S2-S1-(S0)-TRD-Hd1-S5-Hd2-S6-Hd3	SPPY, DPPY
D	48856 52484	Hd1-S5-Hd2-S6-Hd3-TRD-S7-Hu3-S4-Hu2-S3-Hu1-S2-S1	DPPY
C	45633 46303 46923	S5-Hd2-S6-Hd3-S7-Hu3-S4-Hu2-S3-Hu1-S2-S1-TRD-Hd1	GFPC, GPPC
K	45633 46303 46923	Hd1-S5-Hd2-S6-Hd-S7-Hu3-S4-Hu2-S3-Hu1-S2-S1	GPPC
I	*36976** *37952** *45988**	-S4-Hu2-S3-Hu1-S2-S1-	DPPY, DPPH
E	Dam	-S4-Hu2-S3-Hu1-S2-S1-	NPPY
F	EcoRI_methylase	-S4-Hu2-S3-Hu1-S2-S1-	NPPF
G	MT-A70 *36976** *37952** *45988**	-S4-Hu2-S3-Hu1-S2-S1-	DPPW
L	36976 37952 45988	(S0)-Hd1-S5-Hd2-S6-Hd3-S7-Hu3-S4-Hu2-S3-Hu1-S2-S1	HPPY, SPPY, DPPY
M	36976 37952 45988	(S0)-Hd1-S5-Hd2-S6-Hd3-TRD-S7-Hu3-S4-Hu2-S3-Hu1-S2-S1	HPPY, SPPY, DPPY

Classification algorithm

We have developed an automatic algorithm that detects the topology of MTase by region hits of cat-profiles in MTase sequences (https://github.com/MVolobueva/MTase-classification.git or https://mtase-pipeline-6g1yfq9ugw8.streamlit.app/). Using the algorithm we assigned almost all non-P MTases to one of the topology groups. We have used this algorithm for classification of all MTases (non-P and predicted). Results of classification

algorithm presented in table 5 and supplementary material S11. The class I is not determined by the algorithm, because this small class lacks good alignment.

Table 5. Classification of all MTases by expert assignment and algorithm assignment.

Class	**Number of non-P MTases** Expert assignment	**Number of non-P MTases** Algorithm assignment	**Number of all MTases** Algorithm assignment
A	3725	3716	44481
B	1471	1465	26266
D	896	906	14387
C	514	499	22817
K	3	3	66
I	7	-	-
E	11	11	2446
F	21	21	480
G	31	31	1361
L	10	10	389
M	130	120	2105
Two domains			
C/A	24	24	166
B/D	1	1	8
A/K	4	4	10
D/D	36	22	182
C/B	4	4	32
C/C	2	2	16
Totally	6913	6839	115982

Comparison of new classification with previous classification

We have compared the developed classification with existing ones (table 6). MTases from class C and K have m5C methylation type. MTases from the P-M I system are found only in class A, MTases from the P-M III system are found only in class B.

The comparison with the classification by motifs (28) shows that classes A, L, K, F correspond to class γ, classes M and D correspond to class α. The K class was previously allocated to a

separate class ζ*. However, there is no experimental evidence that class ζ is different from class γ.

Table 6. Comparison of the new classification with the previous classifications.

Class	Methylation type	R-M type	Subtype
A	m4C, m6A	I, II	γ
B	m4C, m6A	II, III	β
D	m4C, m6A	II	α
L	m4C, m6A	II	γ
M	m4C, m6A	II	α
C	m5C	II	C5
K	m5C	II	ζ*, γ
E	m6A	II	
F	m6A	II	γ
G	m6A	II	
I	m4C, m6A	II	γ

* - Buinitsky (27) classified zeta into a separate class

Large loops in catalytic domains

In some MTases we have found large inserts (loops more than 20 a.a.) between ss-elements of cat-domain. We have named loops by ss-elements around them (table 7 and figure 6). Location of these loops does not correspond to the previously described location of the TRD (table 4). These insertions were detected by cat-profile alignment with MTase sequences. Non-P MTases with large loops are listed in table S12.

The large loops do not affect the fold of the catalytic domain. In MTases from class A large loops may be more than 100 a.a (figure 6A). It is likely that the large loops from the cat-domain may play a role similar to TRD. For example, there are small loop S3-Hu1 (less than 20 a.a.) and large S4-Hu2 (more than 20 a.a.) in MTase M.CcrMI from class B. These two loops bind to DNA according to 3D structure and might stabilise the protein-DNA complex (Figure 6B). In MTases M.MpeI from class C there is loop S7-Hu3 that is located near binded DNA (Figure 6C). This loop could potentially stabilise the protein-DNA complex too.

Table 7. Large loops in MTase classes.

Class	Large loops	Sample proteins
A	Hd1-S5, S7-Hu3, Hu3-S4, S4-Hu2	BloAII, Cco280IV, BbuB31I, Aod1I
B	S4-Hu2, S3-Hu1	M.AatII, M.AbaZI, M.CcrMI
C	Hd2-S6, S6-Hd3, S7-Hu3, Hu3-S4, S4-Hu2, S3-Hu1, Hu1-S1	M.BspRI, M.TspMI, M.BlmNCII, M.ZmaV, M.SghWII, M.AvaII, M.BsrFI

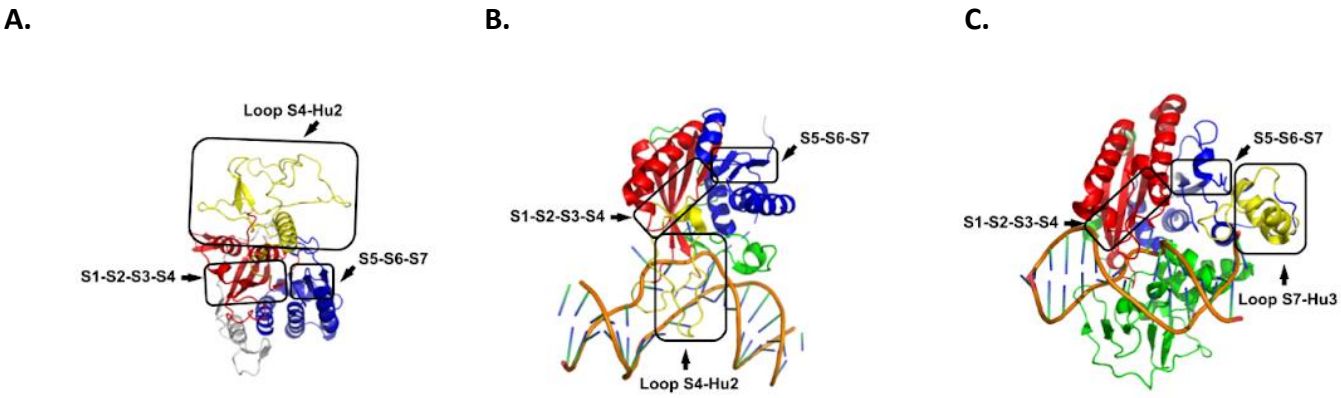


Figure 6. MTases with large loops in cat-domains. The cat-subdomains are red, the sam-subdomains are blue, TRDs are green, large loops are yellow. A. Predicted structure of MTase M.CmuRB25I (AF-W0LFE9-F1 in DB Alphafold). B. Structure of MTase M.CcrMI (pdb code 6PBD) C. Structure of MTase M.MpeI (pdb code 4DKJ).

Non-catalytic domains in MTases

We have built domain architectures for all non-P MTases from DB REBASE (table S13), using profiles from DB Pfam. We have called profiles non-catalytic if they had hits in non-P MTases, but they were not assigned to cat-profiles. Non-catalytic profiles are divided into groups by domain functions (table 8). The function was experimentally established for profile groups ER, TRD, HTH and ATPase. Profiles with predicted function are assigned to groups N-terminal and C-terminal based on their location relative to cat-domain. Domains TRD and ATPase are found in class A. HTH, ER domains are common to many classes. Multidomain proteins were not found only in E, F and I classes. In MTases with two cat-domains there TRD occurs in class A/K, HTH occurs in class D/D. The most diverse domain architectures are found for classes A and B (figure 7).

Table 8. Non-catalytic domain groups.

Domain group	Determination method	Function
MT	Hits of cat-profiles	Catalytic
ER	Hits of profiles MmeI_N and profiles from clan PPDEXK	Endonuclease
TRD	Hits of profiles MmeI_TRD and profiles from clan TRD	DNA-binding
HTH	Hits of profiles from clan HTH	Regulatory
ATPase	Hits of profiles from clan P-loop_NTPase	Helicase
N-terminal	All profile finds not assigned to clans that occur before the catalytic domain	Function is undefined, sometimes there is function prediction
C-terminal	All profile finds not assigned to clans that occur before the catalytic domain	Function is undefined, sometimes there is function prediction

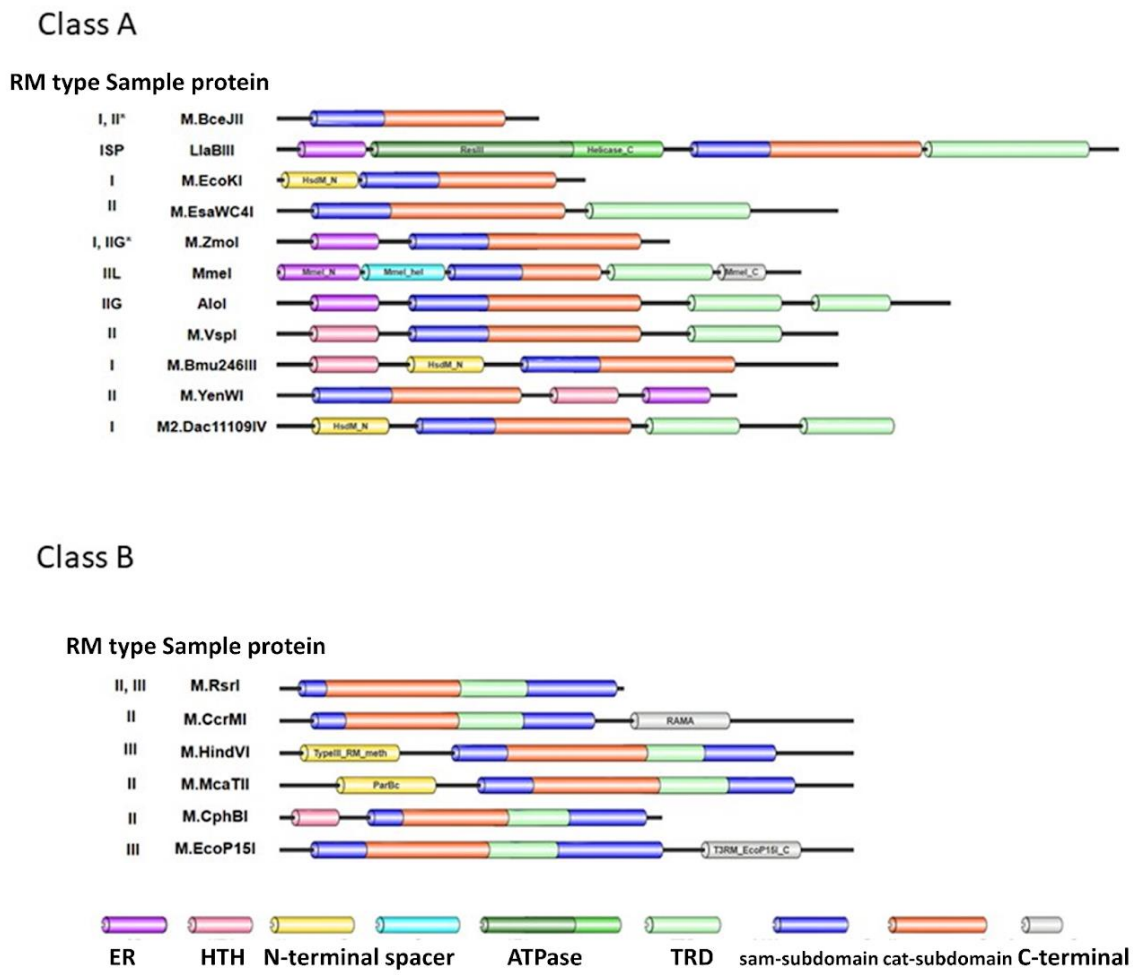


Figure 7. Most common domain architectures of MTases from class A and B. The black line indicates the polypeptide chain of MTase. The colored cylinders on the black line indicate domains in MTase.

DISCUSSION

We analysed available and predicted structures of MTases from REBASE and classified MTases into 11 classes. Each class has its unique set of cat-motifs and topology. To describe topologies, we used the terms subdomains introduced in the work (28). The structure of the cat-subdomain is the same in all MTases except class F; the structure and location of sam-subdomain varies greatly among different classes. Our classification covers more than 98% of non-P MTases from REBASE. Among the proteins not covered by classification, no MTases were found for which additional classes could be identified. All unclassified proteins are either annotation errors, fragments, or proteins that form complexes with MTases. Classes of

famous MTases are presented in table 9. As can be seen from the table, the names Dam, Dcm, CcrM include proteins from different classes with different topologies.

Table 9. Classes of famous MTases.

MTases	REBASE name	Methylation type	Methylation site	Class
Dam	Bcgl	m6A	CGATCNNNNTGC	A
	BclI	m6A	TGATCA	A
	DpnII	m6A	GATC	D
	M.XbaI	m6A	TCTAGATC	B
	M1.MboI	m6A	GATC	D
	M1.MboII	m6A	GAAGA	B
	M2.MboI	m6A	GATC	B
	M2.MboII	m6A	GAAGA	B
Dcm	M2.BsaI	m5C	GGTCTC	K
	M.BsaHI	m5C	GRCGYC	C
	M.AvaII	m5C	GGWCC	C
CcrM	M.CcrMI	m6A	GANTC	B
	M.BslI	m4C	CCNNNNNNNGG	B

We selected profiles from DB SUPFAM and Pfam that detect cat-domains and named them cat-profiles. All selected cat-profiles were divided into 7 groups by topology of the cat-domain they detected. Topologies A, B, C, D are detected by both profiles from DB SUPFAM and profiles from the DB Pfam. Moreover, the profiles from DB SUPFAM have hits in a larger number of MTases than the cat-profiles from DB Pfam. Topologies E, F, G are detected only by profiles from DB Pfam. There are no cat-profiles that detect the whole cat-domain in topologies I, L, M and K. Topologies I, L, M are detected by two hits of cat-profiles from group B, however their topologies differ from topology B. Also topology K is detected by two hits of cat-profiles by profiles from group C, but its topology differs from topology C.

For each class, we constructed alignments of representative MTases. It is impossible to construct alignments for all proteins from one class at once as one class may be highly divergent. In particular, there may be large loops in the cat-domain (see 'Large loops in catalytic domains'), making it impossible to construct a good alignment for such a large class A. Class D is probably a subgroup of class M. Class D consists of highly conserved proteins, and in some cases the assignment of a protein to class M or D is not obvious. The class I with reduced sam-subdomain includes 2 subclasses: sequences with cat-motif DPPY and sequences with cat-motif DPPH.

In non-P MTases cat-profiles almost always detect both cat-motif and sam-motif, with the exception of a few proteins that had a large insertion in the cat-domain. Cat-profiles detect motifs by taking into account the sequence information around the motif, allowing them to find motifs even in cases where the sequence of the motif differs significantly from previously described sequences. Compared to data from DB REBASE, which also implements a search for motifs based on similarity to previously described motifs, the method of searching motifs using cat-profiles detects motifs in a significantly larger number of MTases.

We developed an algorithm for assigning proteins to a class that uses hits of selected cat-profiles. Sometimes the results of the algorithm require expert clarification, especially in cases of classifying proteins from topologies B, D, L, M as those topologies could be detected by cat-profiles from group B and D. Also topology I could not be detected by algorithm as in this class B cat-profiles from group B detect only cat-subdomain.

We divided proteins into subgroups according to domain architectures. Domain architectures were built using Pfam profiles, because DB Pfam was found to contain a greater diversity of MTase non-catalytic domain profiles than DB SUPFAM. To construct domain architectures, profiles were grouped into clans to combine profiles based on homologous sequences. Most non-catalytic domains are characteristic of only one class. Common to all classes are HTH (regulatory domains) and ER (endonuclease domains). It is likely that the fusion of these domains to the MTase sequence occurred several times independently. The TRD region as a separate domain is found only in class A. Class A is not characterised by the presence of TRD region within cat-domain. Proteins with additional TRD subunit have probably been formed as a result of the fusion of MTase with the S subunit. The presence of

a separate subunit TRD makes proteins from class A the most common target of genetic engineers, since by replacing subunits you can change the specificity of DNA binding.

New classification allows to determine the type of MTase methylation and detect annotation errors in the type of methylation. In particular, for the M.AvaV protein, the m4C methylation type was indicated in DB REBASE. According to our classification, this protein belongs to the G class; this class corresponds to the m6A methylation type, which is consistent with literature data (39).

Large insertions were found in the proteins that did not coincide with the previously described location of TRD regions. Large inserts are located at the site of loops in cat-domain. These inserts may have a DNA-binding function. The folding of the cat-domain does not change in the presence of inserts. The presence of large insertions was detected by expert assessment of protein sequence alignments with cat-profiles.

It was previously assumed that the evolution of MTases occurred due to cyclic permutations (28). We constructed a putative evolutionary tree for MTase classes based on assumptions about a minimum number of evolutionary events (Fig. 8). The tree is confirmed by the similarity of cat-motifs. MTase with K topology was chosen as the ancestral structure, because similar topologies occur in C5 and N6, N4 MTases. Our proposed model for the evolution of MTases shows that cyclic permutations occurred independently at least twice, resulting in the appearance of the C and B classes. The formation of the D and M classes is associated not with cyclic permutations, but with the inclusion of a DNA-binding site inside the catalytic domain. In addition, the inclusion of large insertions into the catalytic domain has occurred many times during evolution within classes A, B, C (see 'Large loops in catalytic domains'). It is likely that the included inserts can also acquire the function of a DNA-binding domain. According to our model of evolution, the ancestral MTase had only sam-subdomain and cat-subdomain; binding to DNA occurred due to the formation of a complex with an additional S-subunit. MTases from class A and not from RM I presumably merged after fusion MTases with S-subunits. Classes D, M and C were formed due to the insertion of the DNA-binding domain within the cat- domain.

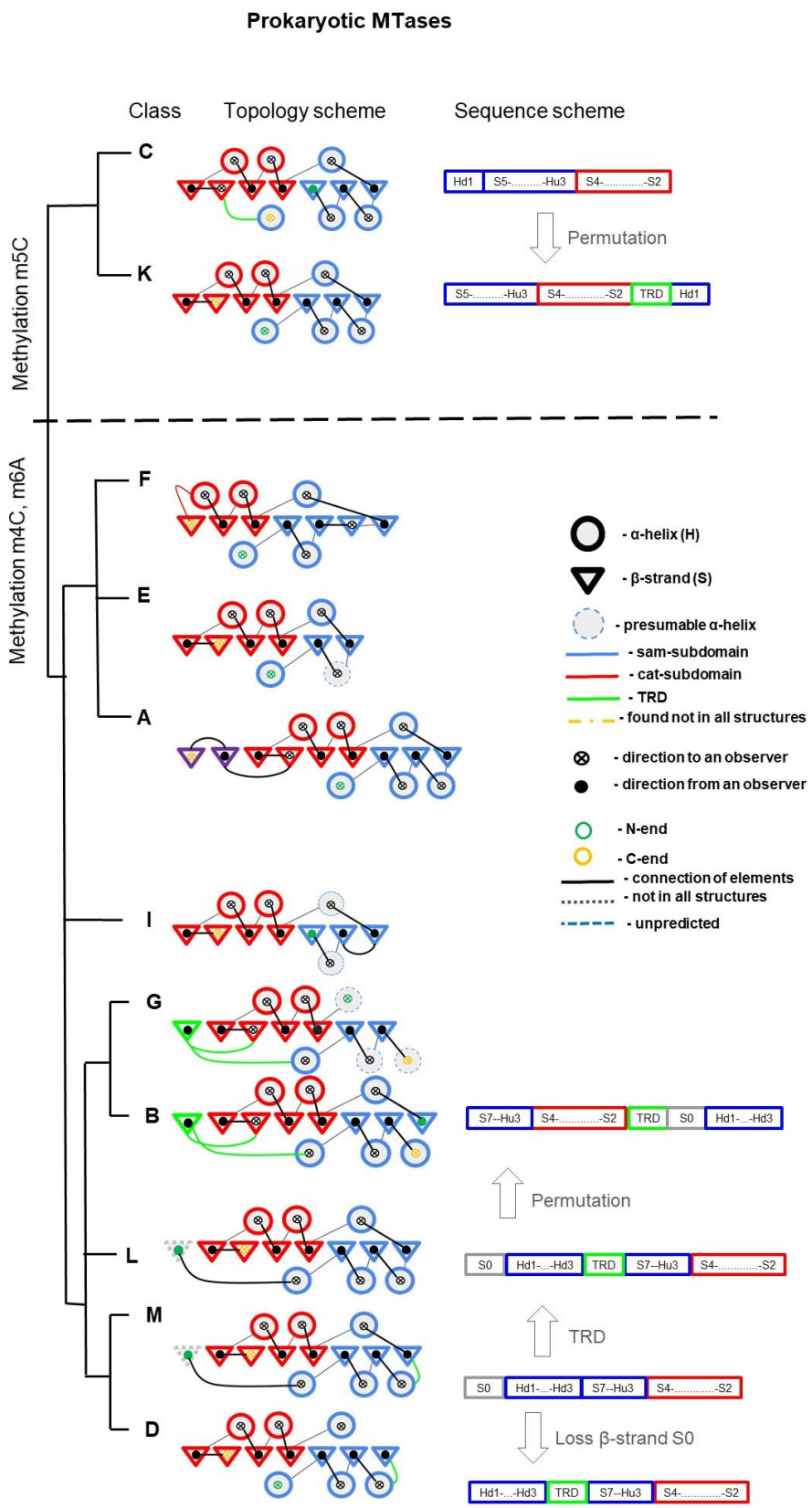


Figure D1. Presumable evolution of MTase classes. Blue colour indicates sam-subdomain, red colour indicates cat-subdomain, green colour indicates TRD sequence.

DATA AVAILABILITY

Data available on github https://github.com/MVolobueva/MTase-classification.git or on web-interface https://mtase-pipeline-6g1yfq9ugw8.streamlit.app/.

SUPPLEMENTARY DATA

Supplementary Data are available https://drive.google.com/drive/folders/1A3dkZclimKS23ro0VvFaTd2qtGeBFBfO?usp=sharing.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGEMENTS

FUNDING

The work of the 3th and authors has been partly supported by FSI SRISA RAS (project FNEF-2024-0001, reg. No. 1023032100070-3-1.2.1)

CONFLICT OF INTEREST

REFERENCES

1. Mattei,A.L., Bailly,N. and Meissner,A. (2022) DNA methylation: a historical perspective. Trends Genet., 38, 676–707.
2. Anton,B.P. and Roberts,R.J. (2021) Beyond Restriction Modification: Epigenomic Roles of DNA Methylation in Prokaryotes. Annu. Rev. Microbiol., 75, 129–149.
3. Nasrullah, Hussain,A., Ahmed,S., Rasool,M. and Shah,A.J. (2022) DNA methylation across the tree of life, from micro to macro-organism. Bioengineered, 13, 1666–1685.
4. Ratel,D., Ravanat,J.-L., Berger,F. and Wion,D. (2006) N6-methyladenine: the other methylated base of DNA. Bioessays, 28, 309–315.
5. Rudenko,A.Y., Mariasina,S.S., Sergiev,P.V. and Polshakov,V.I. (2022) Analogs of -Adenosyl--Methionine in Studies of Methyltransferases. Mol. Biol., 56, 229–250.
6. Oliveira,P.H. and Fang,G. (2021) Conserved DNA Methyltransferases: A Window into Fundamental Mechanisms of Epigenetic Regulation in Bacteria. Trends Microbiol., 29, 28–40.

7. Greenberg,M.V.C. and Bourc'his,D. (2019) The diverse roles of DNA methylation in mammalian development and disease. Nat. Rev. Mol. Cell Biol., 20, 590–607.
8. Cummings,D.J., Tait,A. and Goddard,J.M. (1974) Methylated bases in DNA from Paramecium aurelia. Biochim. Biophys. Acta, 374, 1–11.
9. Gorovsky,M.A., Hattman,S. and Pleger,G.L. (1973) (6 N)methyl adenine in the nuclear DNA of a eucaryote, Tetrahymena pyriformis. J. Cell Biol., 56, 697–701.
10. Hattman,S., Kenny,C., Berger,L. and Pratt,K. (1978) Comparative study of DNA methylation in three unicellular eucaryotes. J. Bacteriol., 135, 1156–1157.
11. Zhang,G., Huang,H., Liu,D., Cheng,Y., Liu,X., Zhang,W., Yin,R., Zhang,D., Zhang,P., Liu,J., et al. (2015) N6-methyladenine DNA modification in Drosophila. Cell, 161, 893–906.
12. Greer,E.L., Blanco,M.A., Gu,L., Sendinc,E., Liu,J., Aristizábal-Corrales,D., Hsu,C.-H., Aravind,L., He,C. and Shi,Y. (2015) DNA methylation on N6-adenine in C. elegans. Cell, 161, 868.
13. Shen,C., Wang,K., Deng,X. and Chen,J. (2022) DNA N-methyldeoxyadenosine in mammals and human disease. Trends Genet., 38, 454–467.
14. Rodriguez,F., Yushenova,I.A., DiCorpo,D. and Arkhipova,I.R. (2022) Bacterial N4-methylcytosine as an epigenetic mark in eukaryotic DNA. Nat. Commun., 13, 1072.
15. Gladyshev,E.A., Meselson,M. and Arkhipova,I.R. (2008) Massive horizontal gene transfer in bdelloid rotifers. Science, 320, 1210–1213.
16. Arber W. Host-controlled modification of bacteriophage. Annu Rev Microbiol. 1965;19:365-78. doi: 10.1146/annurev.mi.19.100165.002053. PMID: 5318444.
17. Hattman,S., Schlagman,S., Goldstein,L. and Frohlich,M. (1976) Salmonella typhimurium SA host specificity system is based on deoxyribonucleic acid-adenine methylation. J. Bacteriol., 127, 211–217.
18. Smith,D.W., Garland,A.M., Herman,G., Enns,R.E., Baker,T.A. and Zyskind,J.W. (1985) Importance of state of methylation of oriC GATC sites in initiation of DNA replication in Escherichia coli. EMBO J., 4, 1319–1326.
19. Janulaitis,A., Klimasauskas,S., Petrusyte,M. and Butkus,V. (1983) Cytosine modification in DNA by BcnI methylase yields N4-methylcytosine. FEBS Lett., 161, 131–134.
20. Ehrlich,M., Gama-Sosa,M.A., Carreira,L.H., Ljungdahl,L.G., Kuo,K.C. and Gehrke,C.W. (1985) DNA methylation in thermophilic bacteria: N4-methylcytosine, 5-methylcytosine, and N6-methyladenine. Nucleic Acids Res., 13, 1399–1412.

21. Smith,H.O. and Nathans,D. (1973) Letter: A suggested nomenclature for bacterial host modification and restriction systems and their enzymes. J. Mol. Biol., 81, 419–423.
22. Roberts,R.J. (1980) Restriction and modification enzymes and their recognition sequences. Nucleic Acids Res., 8, r63–r80.
23. Brooks,J.E. and Roberts,R.J. (1982) Modification profiles of bacterial genomes. Nucleic Acids Res., 10, 913–934.
24. Roberts,R.J. and Macelis,D. (1991) Restriction enzymes and their isoschizomers. Nucleic Acids Res., 19 Suppl, 2077–2109.
25. Roberts,R.J. and Macelis,D. (1994) REBASE — restriction enzymes and methylases. Nucleic Acids Res., 22, 3628–3639.
26. Roberts,R.J. and Macelis,D. (1996) REBASE — Restriction Enzymes and Methylases. Nucleic Acids Res., 24, 223–235.
27. Malone,T., Blumenthal,R.M. and Cheng,X. (1995) Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyltransferases, and suggests a catalytic mechanism for these enzymes. J. Mol. Biol., 253, 618–632.
28. Bujnicki,J.M. (2002) Sequence permutations in the molecular evolution of DNA methyltransferases. BMC Evol. Biol., 2, 3.
29. Albert,P., Varga,B., Zsibrita,N. and Kiss,A. (2018) Circularly permuted variants of two CG-specific prokaryotic DNA methyltransferases. PLoS One, 13, e0197232.
30. Gao,Q., Lu,S., Wang,Y., He,L., Wang,M., Jia,R., Chen,S., Zhu,D., Liu,M., Zhao,X., et al. (2023) Bacterial DNA methyltransferase: A key to the epigenetic world with lessons learned from proteobacteria. Front. Microbiol., 14, 1129437.
31. Marinus,M.G. and Morris,N.R. (1973) Isolation of deoxyribonucleic acid methylase mutants of Escherichia coli K-12. J. Bacteriol., 114, 1143–1150.
32. Urig,S., Gowher,H., Hermann,A., Beck,C., Fatemi,M., Humeny,A. and Jeltsch,A. (2002) The Escherichia coli dam DNA methyltransferase modifies DNA in a highly processive reaction. J. Mol. Biol., 319, 1085–1096.
33. Reich,N.O., Dang,E., Kurnik,M., Pathuri,S. and Woodcock,C.B. (2018) The highly specific, cell cycle-regulated methyltransferase from Caulobacter crescentus relies on a novel DNA recognition mechanism. J. Biol. Chem., 293, 19038–19046.
34. Fioravanti,A., Fumeaux,C., Mohapatra,S.S., Bompard,C., Brilli,M., Frandi,A., Castric,V., Villeret,V., Viollier,P.H. and Biondi,E.G. (2013) DNA Binding of the Cell Cycle

Transcriptional Regulator GcrA Depends on N6-Adenosine Methylation in Caulobacter crescentus and Other Alphaproteobacteria. PLoS Genet., 9, e1003541.
35. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2015) REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res., 43, D298–9.
36. Abdelraheem,E., Thair,B., Varela,R.F., Jockmann,E., Popadić,D., Hailes,H.C., Ward,J.M., Iribarren,A.M., Lewkowicz,E.S., Andexer,J.N., et al. (2022) Methyltransferases: Functions and applications. Chembiochem, 23, e202200212.
37. Cheng,X. and Roberts,R.J. (2001) AdoMet-dependent methylation, DNA methyltransferases and base flipping. Nucleic Acids Res., 29, 3784–3795.
38. Chen,K., Roberts,G.A., Stephanou,A.S., Cooper,L.P., White,J.H. and Dryden,D.T.F. (2010) Fusion of GFP to the M.EcoKI DNA methyltransferase produces a new probe of Type I DNA restriction and modification enzymes. Biochem. Biophys. Res. Commun., 398, 254.
39. Matveyev,A.V., Young,K.T., Meng,A. and Elhai,J. (2001) DNA methyltransferases of the cyanobacterium Anabaena PCC 7120. Nucleic Acids Res., 29, 1491–1506.

==========

REBASE is curated database, all newly sequenced and automatically annotated proteins, putatively belonging to R-M systems, undergoes verification according to REBASE protocolsWe performed comparison of the set of Uniprot annotated prokaREBAEThis is why