CLASSIFICATION OF PROKARYOTIC DNA METHYLTRANSFERASES BY TOPOLOGY AND SEQUENCE SIMILARITY

AUTHORS

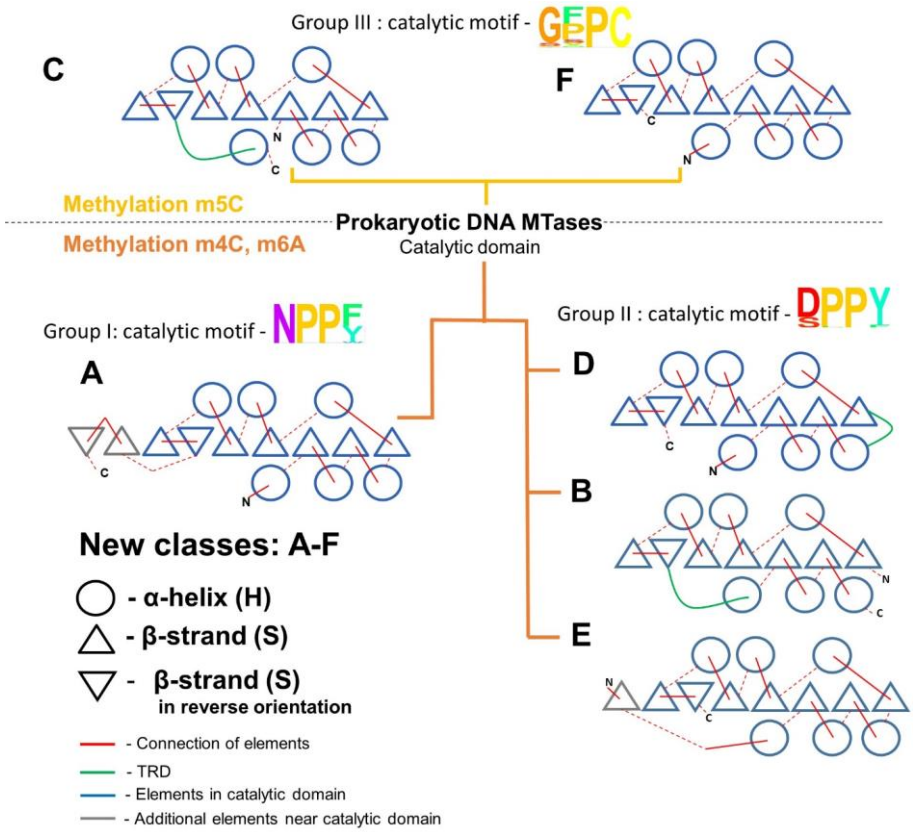
Maria Samokhina[1, 2]* and Alexeevski A.[1,2,3]

[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 methyltransferases (MTases) are essential enzymes that play a crucial role in restriction-modification systems, preventing the degradation of self DNA by restriction enzymes. In addition, they contribute to the regulation of gene expression, DNA repair, DNA replication, and other processes.

We propose a novel classification of prokaryotic MTases that addresses a longstanding deficiency in the previous classification scheme developed by Malone and Bujnicki. The latter classification, which grouped MTases into categories such as α, β, γ, etc., has been in use for over 20 years.

Our approach involves identifying 6 classes of MTases based on the 3D topology of the catalytic domain. This was achieved by analyzing the available 3D structures and certain predicted 3D structures of MTases from the AlfaFold DB. The catalytic domains of each class of MTase exhibit homology not only in structure but also in conserved motifs. To develop this classification algorithm, we utilized a set of MTases from the gold standard REBASE database.

Our new classification system facilitates the recovery of MTase evolution and MTase annotations in both existing and new sequence data sets. This is particularly important for understanding the complex relationships between MTases and their roles in various biological processes.

INTRODUCTION

DNA methylation of C5 cytosine in bacteria was first discovered in 1925, marking nearly a century of exploration into the role of genome methylation and the DNA methylation enzymes responsible for this modification. These enzymes, known as DNA methyltransferases (MTases), were initially identified as components of restriction-modification systems [1] and later recognized for their involvement in various biological processes, including other bacterial immune mechanisms, cell cycle regulation, gene expression, and replication [2-3].

MTases can be classified based on two distinct approaches: functional classification and sequence similarity classification.

Functional Classification: This method categorizes MTases according to the specific chemical reactions they catalyze, taking into account the type of DNA modification (e.g., the addition of a methyl group to a specific nucleotide base) and the resulting products. According to these chemical reactions, DNA methyltransferases fall into three enzyme classification groups [4]:

- **C5 MTases**: Produce C5-methylcytosine (EC 2.1.1.37)
- **N6 MTases**: Produce N6-methyladenine (EC 2.1.1.72)
- **N4 MTases**: Produce N4-methylcytosine (EC 2.1.1.113)

Today, three distinct types of DNA methylation are recognized: m6A (N6-methyladenine), m4C (N4-methylcytosine), and m5C (C5-methylcytosine) [5-6]. MTases from all three groups are found in both bacteria and archaea, with N6 MTases being the most prevalent [7]. In eukaryotes, C5 MTases that methylate cytosines within symmetric dinucleotide CG are predominant [8].

Sequence Similarity Classification: This classification was introduced in literature in 1995, when researchers identified nine conserved motifs within the sequences of MTases known at that time. Based on the order of these motifs and the location of Target Recognition Domain (TRD), the MTases were categorized into three groups (α, β, γ) [9]. In 2002, a fourth class was identified specifically for C5 MTases, leading to an evolutionary model of MTases that proposed their evolution occurred via cyclic permutations [10].

Since 2002, the number of known MTase sequences has expanded dramatically, growing from fewer than 100 to over ten thousand, with this information aggregated in the REBASE database [11]. However, determining the precise number of proteins classified according to conserved motifs is challenging due to limitations in detection methods for all defined motifs and TRD. Among the nine motifs described, motif I (the SAM-binding motif or sam-motif) and motif IV (the catalytic motif or cat-motif) can be found within some entries in REBASE, yet many MTases have no detected motifs at all.

In our work, we propose a new classification system for prokaryotic MTases based on the structure of their catalytic domain, utilizing a modern MTase dataset from the REBASE database. To facilitate this classification, we developed an algorithm for the automatic annotation of MTase genes, allowing for simultaneous determination of enzyme class. Additionally, we created a web application to visualize the classification results. Based on these findings, we also propose a plausible evolutionary model for prokaryotic MTases.

MATERIAL AND METHODS

Materials

MTase 3D Structures. As of June 3, 2025, the REBASE database includes a list of 38 prokaryotic DNA methyltransferases (MTases) with known 3D protein structures (refer to Table S1a). Notably, five of these MTases contain not only the methyltransferase catalytic

domain but also the catalytic domains of restriction endonucleases. These dual-function proteins are designated with the prefix "RM" in the REBASE database.

For further analysis, predicted structures of several MTases were obtained from the AlphaFold database (https://alphafold.ebi.ac.uk/), which provides insights into the potential structural conformations of these enzymes.

MTase Sequences from REBASE. We obtained MTase sequences from REBASE v408. The data underwent preprocessing to cluster identical sequences, as detailed in supplementary material S2. This process resulted in two distinct lists: one of unique MTases (174874 proteins) and another specifically comprising 1678 unique MTases from the "gold standard" (1678 proteins), which indicates evidence of their enzymatic activity (refer to supplementary material S3). For the purpose of building our classification system, we utilized unique MTases from the gold standard in REBASE v408, while other sets of MTases were employed to validate the classification outcomes.

UniProt Source of MTase Sequences. On February 9, 2025, we conducted an advanced search in the UniProt database and identified 183588 annotated MTases (supplementary material S4). This search was performed with specific parameters, including a focus on prokaryotic taxonomy (eubacteria OR archaea) and targeted Enzyme Commission numbers: EC 2.1.1.113, EC 2.1.1.72, and EC 2.1.1.37. The output from the query included standard entry information such as identifiers (AC, ID), gene names, protein names, status (reviewed/unreviewed), organism, and sequence length. Additionally, the results provided details about Pfam and Supfam domains, REBASE indices, PDB codes, and taxonomic lineage.

No prokaryotic MTases with known three-dimensional structures were found in Uniprot that were not listed in REBASE. Among the eukaryotic MTases, we found seven proteins for which three-dimensional structures are known (Table S1b).

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/).

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 1.75 (https://supfam.mrc-lmb.cam.ac.uk/) and Pfam version 34, which services are now takeover into Interpro DB (https://www.ebi.ac.uk/interpro/). The search for all HMM profiles was carried out by the hmmsearch program from the hmmer package version 3.3.2 for all proteins from REBASE.

For classification we selected 12 profiles from SUPFAM and 3 from Pfam. All profiles in SUPFAM are divided into superfamilies and families. We selected 12 profiles from the superfamily *S-adenosyl-L-methionine-dependent methyltransferases*, which includes 5 families built on prokaryotic DNA MTases; these families included 12 profiles. Among the gold standard proteins were those for which no SUPFAM profile hits were found. These proteins contained findings of 3 profiles from Pfam. These 3 profiles were also included in the list for classification. Totally 15 profiles were selected (supplementary material S5a).

Among MTase sequences from gold standard without hits of selected profiles, one sequence M.BceJII was found to be similar to MTase. For this sequence, we found homologs and built new HMM-profile *New_MTase_profile* using hmmer program (supplementary material S5b).

Classification and annotation algorithm. The classification and annotation algorithm was written in Python. Web visualization of the algorithm was developed using the streamlit framework.

RESULTS

The catalytic domains of MTases feature a conserved seven-stranded β -sheet (↓ ↑ ↓ ↓ ↓ ↓ ↓) flanked by α-helices on both sides [12]. In our study, we analyzed the spatial structures of prokaryotic (Table S1a) and eukaryotic (Table S1b) MTases and found that all examined structures retain this core architecture: a seven-β-strand sheet surrounded by six α-helices—three positioned above and three below the sheet.

For subsequent analysis convenience, we introduced a unified labeling system for secondary structure elements of MTases. β-strands were designated with a capital S (strand) and numbered 1 through 7 (from left to right). α-helices located above the β-sheet were labeled Hu (helix, upper) and numbered 1–3, while helices below the β-sheet were marked as Hd (helix, down) and similarly numbered 1 to 3 (Figure 1a).

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

As of July 3, 2025, 38 prokaryotic DNA MTases with resolved spatial structures are known. Among them, five are RM proteins, meaning they also contain an endonuclease subunit. Based on the topology of the catalytic domain — specifically, the order of secondary structure elements—these MTases are divided into four groups: A, B, C, and D (Figure 1b). For each structure, we identified the catalytic domain as a distinct part of the protein globule, exhibiting a "sandwich-like" architecture where the β-sheet is surrounded by α-helices. Within the catalytic domain, we localized: the catalytic motif (cat-motif), also known as motif IV, and the SAM-binding motif (sam-motif), or motif I, as previously described [10].

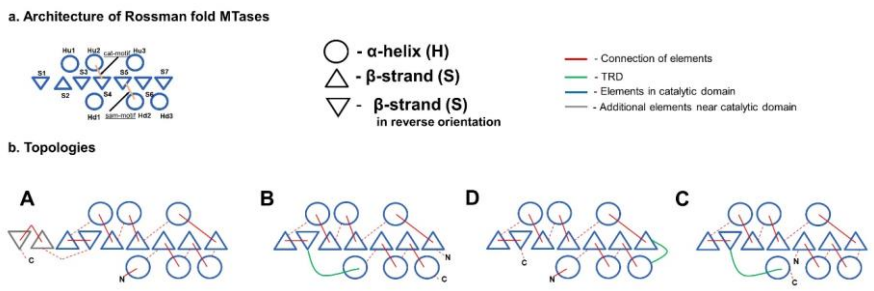


Figure 1. Groups of topological schemes of all MTases with known 3D structures.

Group A is the most abundant, comprising 19 out of the 38 studied MTases. The catalytic domain of these proteins begins with the Hd1 element and ends with the S2 element. A distinctive feature of this group is the presence of two additional β-strands (S0 and S(-1)) at the C-terminus, which are absent in other topologies. The catalytic motif in this group follows the consensus sequence NPPF/NPPY.

All RM proteins (RM.BpuSI, RM.DrdV, RM.LlaGI, RM.LlaBIII, and RM.MmeI) belong exclusively to class A, supporting their evolutionary relationship. Notably, class A MTases lack a TRD (target recognition domain) within the catalytic domain. Instead, the TRD is either present as an S-subunit positioned C-terminally as an independent domain, or encoded by a separate gene.

All 19 members of this group are multidomain proteins, meaning they contain additional functional domains alongside the catalytic domain. The size of the catalytic domain varies from 213 amino acid residues (M.TaqI) to 292 residues (RM.MmeI), with the exception of M.Sen23580I, which has a significantly larger catalytic domain (361 residues). Its structure features unique extended insertions (loop) (35 and 42 residues long, respectively) between secondary structure elements (Hd1-loop-S5 and Hu3-loop-S4). However, these insertions do not disrupt the overall domain topology.

Group B comprises 10 MTases with unique structural features. Unlike Group A, these proteins exhibit an inverted motif order along the polypeptide chain: the catalytic motif (consensus DPPY/SPPF) precedes the SAM-binding motif. A hallmark of this group is the presence of an embedded TRD (target recognition domain) inserted between β-strand S2 and α-helix Hd1. Some members also feature an additional strand S8 and a strand S0 located within the TRD. This group displays diversity in methylation specificity: 8 out of 10 MTases catalyze m6A formation (including M.CcrMI and M.EcoP15I), M.PvuII uniquely modifies m4A.

Group C consists of seven MTases, all exhibiting m5C methylation type. These proteins are characterized by a unique catalytic motif (GPPC/GFPC) not found in other groups. Unlike Groups A and B, the catalytic domains in Group C are the most extended, ranging from 312 (M.HhaI) to 441 (M.SflTDcmP) amino acid residues. This size variation is primarily due to the C-terminal TRD domain.

Six of the seven members (M.HhaI, M.HaeIII, M.EcoKO157ORF1953P, M.MpeI, M.SflTDcmP, and M.HpyAXI) are compact single-domain proteins. The exception is M.MpeI, which contains a small (≈50 aa) additional N-terminal domain of unknown function. A distinctive feature of these catalytic domains is the strictly conserved position of the TRD, which is always located between structural elements S2 and Hd1.

Group D is represented by four structurally characterized MTases: M.EcoKDam, M1.DpnII, M.EcoT4Dam, and M.CbeI (253 aa). In terms of secondary structure organization (Hd1-S5-Hd2-S6-Hd3-TRD-S7-Hu3-S4-Hu2-S3-Hu1-S1-S2), this class shares similarities with Group A, but differs by the absence of additional C-terminal β-structures (S0 and S(-1)), which are characteristic of the latter.

The catalytic domain of these proteins contains a compact TRD domain (60-80 aa) integrated between helix Hd3 and β-strand S7, ensuring structural compactness while maintaining functional activity. All three members of Group D are single-domain proteins. M.EcoKDam, M.EcoT4Dam, and M1.DpnII mediate m6A methylation, M.CbeI is specific for m4C modification. Notably, all three retain the conserved catalytic motif (DPPY).

MTase HMM profiles

We selected 12 profiles from the superfamily of S-adenosyl-L-methionine-dependent methyltransferases (MTases), which includes five families of prokaryotic DNA MTases. Each profile was assigned to a topology group based on its corresponding Protein Data Bank (PDB) structures. To aid in the detection of the catalytic domain using the selected profiles, we mapped six key regions within each profile: four linked to secondary structure elements and two associated with the catalytic and SAM-binding motifs as illustrated in Figure 2 (table S6a).

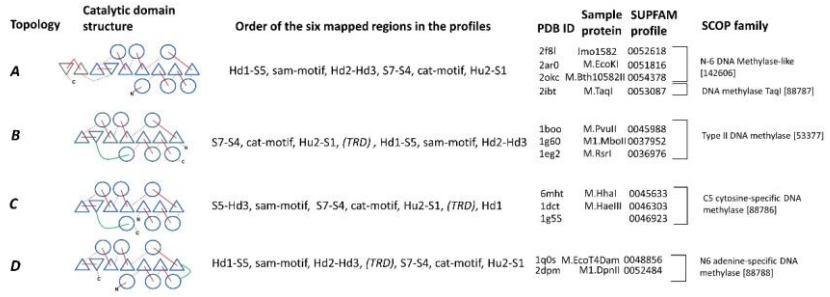


Figure 2. Division of 12 SUPFAM profiles into 4 groups by topology.

In our profile search conducted on the gold standard MTases, we found that most of these proteins conformed to one of four established topologies based on the sequences of the profile region hits. However, we identified two additional groups of MTases with topologies that differ from the previously established four, as determined by the location of the mapped regions. These two groups include proteins with experimentally confirmed MTase activity. Furthermore, the predicted protein structures of these groups display a standard MTase architecture (see Figure 3)

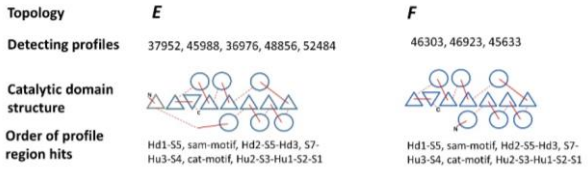


Figure 3. Two additional topologies of MTases.

The first group, referred to as **topology E**, includes the experimentally confirmed MTases M1.BcnI [13] and M4.BstF5I [14]. MTases from this group are detected by profiles associated with topologies B and D. Although the order of regions and motifs in these enzymes is similar to that of topology D, members of topology E lack the TRD between the Hd3 and S7 regions, which is present in topology D. In topology D, the distance between the sam-motif and the catalytic motif exceeds 150 amino acids; however, in topology E, this distance is less than 100 amino acids. Notably, the predicted structures of M1.BcnI and M4.BstF5I reveal the

presence of an extra S0 beta strand, which may contribute to the unique characteristics of this group.

The second group, designated as **topology F**, includes the M2.Eco31I MTase [15] with experimentally confirmed m5C activity. While this topology resembles topology D, it exhibits profiles associated with topology C, and its catalytic motif (GPPC) matches those of group C. A key feature of topology F is the N-terminal positioning of helix Hd1 and the absence of a TRD within the catalytic domain. These observations are supported by both predicted structure analysis and the mapped regions of profile alignments.

In summary, we identified a total of six groups of MTases (A, B, C, D, E, F), each characterized by unique topologies and structural features.

Classification algorithm

We developed an automated classification algorithm that identifies MTase topologies by detecting conserved sequence regions using 12 predefined SUPFAM profiles. The analysis begins with a comprehensive scan of input sequences against our profile library, which is specifically designed to recognize: (i) catalytic motifs (cat-motif), (ii) SAM-binding motifs (sam-motif), and (iii) characteristic secondary structure elements. For a profile alignment to qualify as a valid region detection, it must satisfy three quality criteria: (1) the alignment must cover >40% of the reference profile's residues, (2) the ratio of insertion positions to aligned residues must be <2.5, and (3) for functional motifs (sam-motif and cat-motif), the alignment must cover >75% of the motif's reference positions with no more than one insertion.

From all qualifying profiles, the algorithm selects the optimal candidate through a two-tiered selection process. First, it identifies profiles detecting the maximum number of regions for a given sequence. When multiple profiles detect an equal number of regions, the algorithm prioritizes the profile demonstrating the highest average alignment coverage across all its detected regions.

The final topology assignment is determined by analyzing the order of regions in the selected profile, as detailed in Table 1. To account for evolutionary divergence and partial sequence data, the algorithm permits classification based on incomplete region sets, requiring detection of at least four regions including both functional motifs. For instance,

Class A MTases can be confidently identified with just four detected regions (Hd1-S5, sam-motif, Hd2-S5-Hd3, and cat-motif) when these meet all quality thresholds.

Validation against reference datasets from REBASE v408 and UniProt (Table 2) demonstrated robust performance, with Class A emerging as the most prevalent topology. The algorithm also successfully identified sequences containing multiple catalytic domains, though these were excluded from class-specific enumerations.

The complete implementation, including the profile library and quality filtering parameters, is publicly available through our GitHub repository (https://github.com/MVolobueva/MTase-classification.git). For researcher convenience, we also provide a user-friendly web interface (https://mtase-pipeline.streamlit.app/) that enables sequence submission and topology visualization. Comprehensive details on profile-specific region orders and acceptable partial detection patterns are provided in Table 1 and Supplementary Materials S7, offering complete transparency for method replication and application.

Table 1. Comparison topologies with detected profile regions.

Topology	**Detection profile**	**Order of region hits in profile**	**Samples of incomplete acceptable by algorithm orders of region hits detected by profile**
A	51816, 52618, 54378, 53087	Hd1-S5, sam-motif, Hd2-S5-Hd3, S7-Hu3-S4, cat-motif, Hu2-S3-Hu1-S1-S2	Hd1-S5, sam-motif, Hd2-S5-Hd3, cat-motif
B	37952, 45988, 36976	S7-Hu3-S4, cat-motif, Hu2-S3-Hu1-S1-S2, TRD, Hd1-S5, sam-motif, Hd2-S5-Hd3	cat-motif, Hu2-S3-Hu1-S1-S2, TRD, Hd1-S5, sam-motif
C	46303, 46923, 45633	S5-Hd3, sam-motif, TRD, S7-Hu3-S4, cat-motif, Hu2-S3-Hu1-S1-S2, Hd1	sam-motif, TRD, S7-Hu3-S4, cat-motif, Hu2-S3-Hu1-S1-S2
D	48856, 52484	Hd1-S5, sam-motif, Hd2-S5-Hd3, TRD, S7-Hu3-S4, cat-motif, Hu2-S3-Hu1-S1-S2	sam-motif, Hd2-S5-Hd3, cat-motif, Hu2-S3-Hu1-S1-S2
E	37952, 45988, 36976, 48856, 52484	Hd1-S5, sam-motif, Hd2-S5-Hd3, S7-Hu3-S4, cat-motif, Hu2-S3-Hu1-S1-S2	sam-motif, Hd2-S5-Hd3, cat-motif, Hu2-S3-Hu1-S1-S2

		(Distance between sam-motif and cat-motif less 100 aa)	
F	46303, 46923, 45633	Hd1-S5, sam-motif, Hd2-S5-Hd3, S7-Hu3-S4, cat-motif, Hu2-S3-Hu1-S1-S2	Hd1-S5, sam-motif, cat-motif, Hu2-S3-Hu1-S1-S2

Table 2. Algorithm classification results.

	REBASE v408		**Uniprot**
Class	**Among unique gold standard MTases**	**Among all sequences**	
A	958	58672	79494
B	277	33243	25459
C	192	32425	30462
D	184	20150	29302
E	7	1076	1513
F	2	87	86
several catalytic domains	15	1191	1172

MTase classes

To determine whether MTases with a given topology form a monophyletic or paraphyletic group, we analyzed the catalytic motifs of gold-standard MTases for each topology. Our analysis revealed that the six MTase topologies can be divided into six distinct classes, which cluster into three broader groups based on their catalytic motif signatures (Figure 4, Table S8).

Group I consists of MTases with topology A, which all possess a catalytic motif that begins with the amino acid N and terminates with one of the amino acids Y, F, or W. Group II includes MTases with topologies B, D, or E, whose catalytic motifs initiate with any of the amino acids D, S, or H and conclude with one of the amino acids Y, F. Group III comprises MTases with topologies C or F, whose catalytic motifs terminate with the sequence PC.

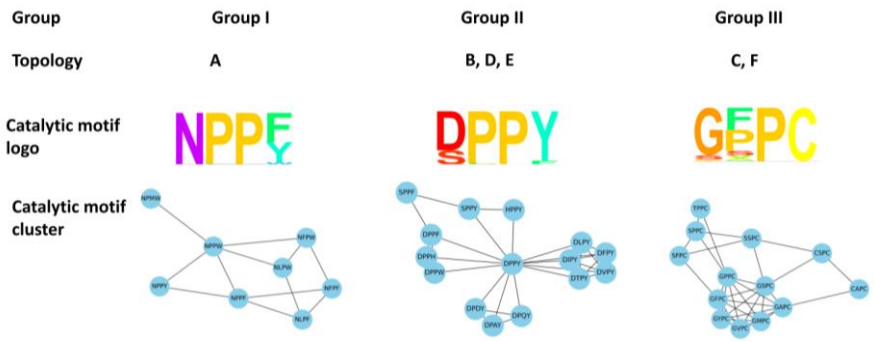


Figure 4. Division of 6 classes into 3 groups according to the similarity of catalytic motifs. The Catalytic Motif Logo displays the most conserved positions in a catalytic motif within all sequences of the same group. The catalytic motif cluster lists all unique catalytic motif sequences within one group of MTases. An edge is drawn between two catalytic motifs in a cluster if the motifs differ by one amino acid.

Algorithm-unclassified sequences

Among the unclassified sequences from the gold standard dataset (1,678 MTases), our algorithm failed to classify 44 sequences. Detailed analysis revealed that 12 sequences were misannotated as MTases (Table S9a), as they lacked both catalytic and SAM-binding motifs and displayed non-Rossmann fold predicted structures. Another 11 sequences represented MTase fragments (Table S9b), containing only one of the two characteristic motifs. Two MTases (M.Pbo13188I and M.TliII) were identified as class D members (Table S9b) but contained unusually large TRD domains (>300 aa), preventing proper detection of their catalytic domains by our algorithm.

The remaining 19 proteins were divided into four subgroups (G, H, I, J) based on sequence similarity, catalytic motif conservation, and predicted structural topology (Table S9d-g). MTases from three of these groups were identified using existing Pfam profiles (*MT-A70, Dam, EcoRI_methylase*), while for the fourth group we developed novel MTase-specific profiles *New_MTase_profile* (see materials and methods). All four profiles were systematically mapped across six conserved regions (Table S6b) using the same methodology employed for SUPFAM profile. Through comparative analysis of catalytic motifs and structural topologies, we classified subgroup G as class B, subgroups H and I as class A, and subgroup J as class D

(Table 3), consistent with their respective domain architectures and sequence of catalytic motifs.

Among these 19 unclassified sequences, experimental validation of methyltransferase activity has been conclusively demonstrated only for M.SsoI (100% identical to M.EcoRI at the amino acid level). This protein exhibited a predicted structural topology most similar to class A, though lacking two additional β-strands (S0 and S(-1)) and showing replacement of the canonical α-helix Hd3 between strands s6 and s7 with a β-strand - a feature requiring experimental verification as it might represent either a genuine structural peculiarity or prediction artifact.

Table 3. Group of sequences from gold standard with non-MTase architecture.

Subgroup	Class	Detecting profile	Sample sequence	Catalytic motif	Predicted structure of sample sequence
G	B	MT-A70 from Pfam	M.AvaV	DPPW	AF-Q9KIW1-F1
H	A	Dam from Pfam	M.Rsp105IV	NPPY	AF-A0A2L1UPL1-F1
I	A	EcoRI_methylase from Pfam	M.EcoRI	NPPF	AF-P00472-F1
J	E	New-MTase-profile	M.BceJII	DPPH	AF-B4EFP2-F1

Large loops in catalytic domains

We constructed alignments for each MTase class, as detailed in Supplementary Materials 10. Upon analyzing these alignments, we observed that secondary structure elements form conservative columns, which are flanked by large insertions (>20 amino acids) in some

MTases. Table 4 provides a list of sample proteins with these large inserts. Structural analysis, including both available and predicted data, reveals that these inserts are loops that do not disrupt the overall fold of the enzyme. Notably, in MTases from class A, these loops can be quite extensive, exceeding 100 amino acids in length. Each insert is designated by the secondary structure elements that flank its location.

Table 4. Sample of large insets in MTase classes.

Topology	Large insets (more 20 a.a.)	Sample MTase
A	Hd1-S5	Ssp6803IV
	S5-Hd2	M.BspD4I
B	S4-Hu2	M.AatII
C	Hd2-S6	M.BspRI
D	S4-Hu2	M.AvaI
E	S4-Hu2	M.HtuII

DISCUSSION

Добавлено примечание ([MB1]):

Добавлено примечание ([MB2R1]):

Добавлено примечание ([MB3R1]):

Our structural and sequence analysis of MTases from REBASE has led to their classification into six distinct topological classes (A–F). Classes A–D, supported by experimentally determined structures, exhibit the canonical Rossmann-like fold with seven β-strands and six α-helices. While Classes E and F share this overall architecture, their classification relied on predicted structures due to the current lack of experimental data. The conserved topology and catalytic motifs within each class strongly suggest their monophyletic origin. Furthermore, based on catalytic motif signatures, these classes cluster into three evolutionary groups (I–III), supporting functional and structural relationships between diverse MTase families.

Among unclassified sequences, we identified four subgroups (G–J) that initially did not fit our classification system. Through comprehensive analysis of sequence similarity, catalytic motifs, and predicted topologies, we assigned subgroup G to Class B, subgroups H and I to Class A, and subgroup J to Class D, consistent with their domain organization and motif conservation. Three of these subgroups were detected using existing Pfam profiles, while the fourth required the development of a new MTase-specific profile.

Our classification system successfully captures the majority of known MTase diversity while highlighting exceptions that may represent structural innovations or transitional forms. Future structural studies of these outliers will further refine our understanding of MTase evolution and functional diversity.

Comparison of new classification with previous classification

In Table 5, we present the classification of well-known MTases. As evident from the table, the names Dam, Dcm, and CcrM correspond to proteins from different classes, which exhibit distinct topologies.

Table 5. Classes of famous MTases.

MTases	REBASE name	Methylation type	Methylation site	Class
Dam	BcgI	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	F
	M.BsaHI	m5C	GRCGYC	C
	M.AvaII	m5C	GGWCC	C
CcrM	M.CcrMI	m6A	GANTC	B
	M.BslI	m4C	CCNNNNNNNGG	B

We have compared our newly developed classification system with existing classification schemes (Table 6) to assess their congruence. Our results indicate that MTases from Class C and F, Group III, exhibit the m5C methylation type. Furthermore, we found that MTases from the P-M I system are exclusively found in Class A, whereas MTases from the P-M III system are exclusively found in Class B.

Comparison with the classification system based on motifs [10] reveals that Classes A, E, and F correspond to the γ class, while Class D corresponds to the α class. Notably, MTases from the F class also exhibit similarities with Class ζ*. However, given the absence of experimental evidence supporting the distinctness of Class ζ from Class γ, we propose that the F class may be reclassified as part of the γ class.

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

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

** - Buinitsky classified zeta into a separate class [10]*

Large insertion within MTase catalytic domain

Large insertions not corresponding to previously described TRD domains have been identified in protein sequences of several MTase classes. These insertions were revealed through analysis of amino acid sequence alignments within each class. The extended loops appear to preserve the overall catalytic domain structure, as confirmed by structural data.

These insertions likely play crucial functional roles in MTase activity and specificity. For instance, in class B MTase M.CcrMI, four loops participate in specific DNA recognition, and the enzyme additionally possesses a C-terminal DNA-binding domain [16]. Similarly, class E MTases contain four loops at equivalent positions that may be also involved in sequence recognition by analogy with M.CcrMI.

Our analysis reveals the complex organization of TRD domains, which may consist of: multiple structural elements (M.CcrMI), a single large loop (classes C and D), an independent domain (class A).

These findings demonstrate that large loops can significantly modulate MTase function. Understanding their roles is essential for engineering MTases with tailored properties. Further investigation of these structural elements will advance protein engineering applications in this field.

MTase classification algorithm

We have developed an algorithm that leverages the findings of selected profiles from the SUPFAM database to classify MTases. The algorithm detects the coordinates of secondary structure elements and motifs by transforming the coordinates from the mapped profile to the alignment between the sequence and profile.

The primary limitation of the algorithm is its ability to detect topology only in sequences with a single catalytic domain. When the algorithm encounters a sequence containing two or more catalytic motifs and two or more SAM-binding motifs, it refers to this sequence as belonging to a group with multiple catalytic domains. However, this approach does not provide information about the topology of the individual catalytic domains within the sequence.

To identify the topologies of the catalytic domains in such sequences, it is necessary to first divide the sequence into two or more separate sequences, each containing only one catalytic domain. This process, known as sequence fragmentation or splitting, allows the algorithm to analyze each catalytic domain independently and determine its specific topology.

The potential applications of our algorithm extend to the detection of DNA MTases in newly sequenced genomes and metagenomes. However, it is essential to recognize that some protein and RNA MTases may exhibit similarities to DNA MTases in terms of sequence and topology, which could lead to incorrect classification. Consequently, there may be instances where our algorithm detects MTases with alternative substrates, underscoring the need for further validation and refinement.

Presumable evolution model of prokaryotic DNA MTases

We propose that Groups I, II, and III represent distinct branches of the DNA MTase evolutionary tree (see Fig.5). Our analysis suggests that Group III, which includes C5 MTases, is the earliest to diverge from the common ancestor of DNA MTases. This inference is supported by the observation that the catalytic motifs in Groups I and II exhibit greater similarity compared to those in Group III. The presence of N4 and N6 MTases across all classes in Groups I and II indicates that these methylation types could evolve into one another over time.

Classes A, E, and F MTases display similar catalytic domain topologies; however, A and E have additional elements that differentiate them, while still maintaining their classification in distinct groups. Based on these observations, we propose that the common ancestor of DNA MTases likely possessed a topology akin to that of Class F, which is the simplest among the three. Furthermore, we suggest that this ancestral form did not include a TRD within its catalytic region.

Our model indicates that the permutations of the catalytic domain occurred independently in Groups II and III, rather than as a result of cyclic permutations as previously assumed in work [10]. In future studies, we aim to investigate DNA MTases from eukaryotes and explore other substrates to further elucidate the evolutionary history of S-adenosyl-L-methionine-dependent MTases.

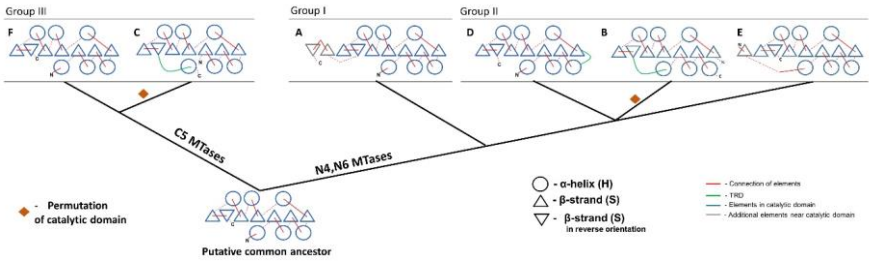


Figure 5. Presumable MTase evolution.

CONCLUSION

We have developed a novel classification system for prokaryotic DNA methyltransferases (MTases) into six distinct classes, each with a unique 3D topology. These classes are organized into three groups, each with its own distinct set of conservative catalytic motifs.

To develop this classification system, we designed an algorithm that not only classifies MTases into their respective categories but also enables the detection of MTase genes in genomes.

Overall, our novel classification system provides a powerful tool for understanding the diversity and evolution of prokaryotic DNA MTases, and its applications are likely to be far-reaching in fields such as genomics, evolutionary biology, and biotechnology.

DATA AVAILABILITY

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

SUPPLEMENTARY DATA

Supplementary Data are available https://drive.google.com/drive/folders/1MXtbj3CyQ_bj_bh-XdVDiYy0vtiZzSfZ.

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. 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.
3. Anton,B.P. and Roberts,R.J. (2021) Beyond Restriction Modification: Epigenomic Roles of DNA Methylation in Prokaryotes. Annu. Rev. Microbiol., 75, 129–149.
4. Poh,W.J., Wee,C.P.P. and Gao,Z. (2016) DNA Methyltransferase Activity Assays: Advances and Challenges. Theranostics, 6, 369–391.
5. Payelleville,A. and Brillard,J. (2021) Novel Identification of Bacterial Epigenetic Regulations Would Benefit From a Better Exploitation of Methylomic Data. Front. Microbiol., 12, 685670.
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. McIntyre,A.B.R., Alexander,N., Grigorev,K., Bezdan,D., Sichtig,H., Chiu,C.Y. and Mason,C.E. (2019) Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. Nat. Commun., 10, 579.
8. de Mendoza,A., Bonnet,A., Vargas-Landin,D.B., Ji,N., Li,H., Yang,F., Li,L., Hori,K., Pflueger,J., Buckberry,S., et al. (2018) Recurrent acquisition of cytosine methyltransferases into eukaryotic retrotransposons. Nat. Commun., 9, 1341.

9. 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.
10. Bujnicki,J.M. (2002) Sequence permutations in the molecular evolution of DNA methyltransferases. BMC Evol. Biol., 2, 3.
11. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2022) REBASE: a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res., 51, D629–D630.
12. Cheng,X. and Roberts,R.J. (2001) AdoMet-dependent methylation, DNA methyltransferases and base flipping. Nucleic Acids Res., 29, 3784–3795.
13. Merkiene,E., Vilkaitis,G. and Klimasauskas,S. (1998) A pair of single-strand and double-strand DNA cytosine-N4 methyltransferases from Bacillus centrosporus. Biol. Chem., 379, 569–571.
14. Golikova,L.N., Guturov,V.V., Evdokimov,A.A., Shchelkunov,S.N., Gonchar,D.A., Okhapkina,S.S., Degtiarev,S.K. and Netesova,N.A. (2002) [M.BstF5I-4, the forth DNA methyltransferase of BstF5I restriction-modification system from Bacillus stearothermophilus F5]. Bioorg. Khim, 28, 84–86.
15. Bitinaite,J., Mitkaite,G., Dauksaite,V., Jakubauskas,A., Timinskas,A., Vaisvila,R., Lubys,A. and Janulaitis,A. (2002) Evolutionary relationship of Alw26I, Eco31I and Esp3I, restriction endonucleases that recognise overlapping sequences. Mol. Genet. Genomics, 267, 664–672.
16. Horton,J.R., Woodcock,C.B., Opot,S.B., Reich,N.O., Zhang,X. and Cheng,X. (2019) The cell cycle-regulated DNA adenine methyltransferase CcrM opens a bubble at its DNA recognition site. Nat. Commun., 10, 1–9.