CLASSIFICATION OF PROKARYOTIC DNA METHYLTRANSFERASES BY TOPOLOGY AND SEQUENCE SIMILARITY

AUTHORS

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

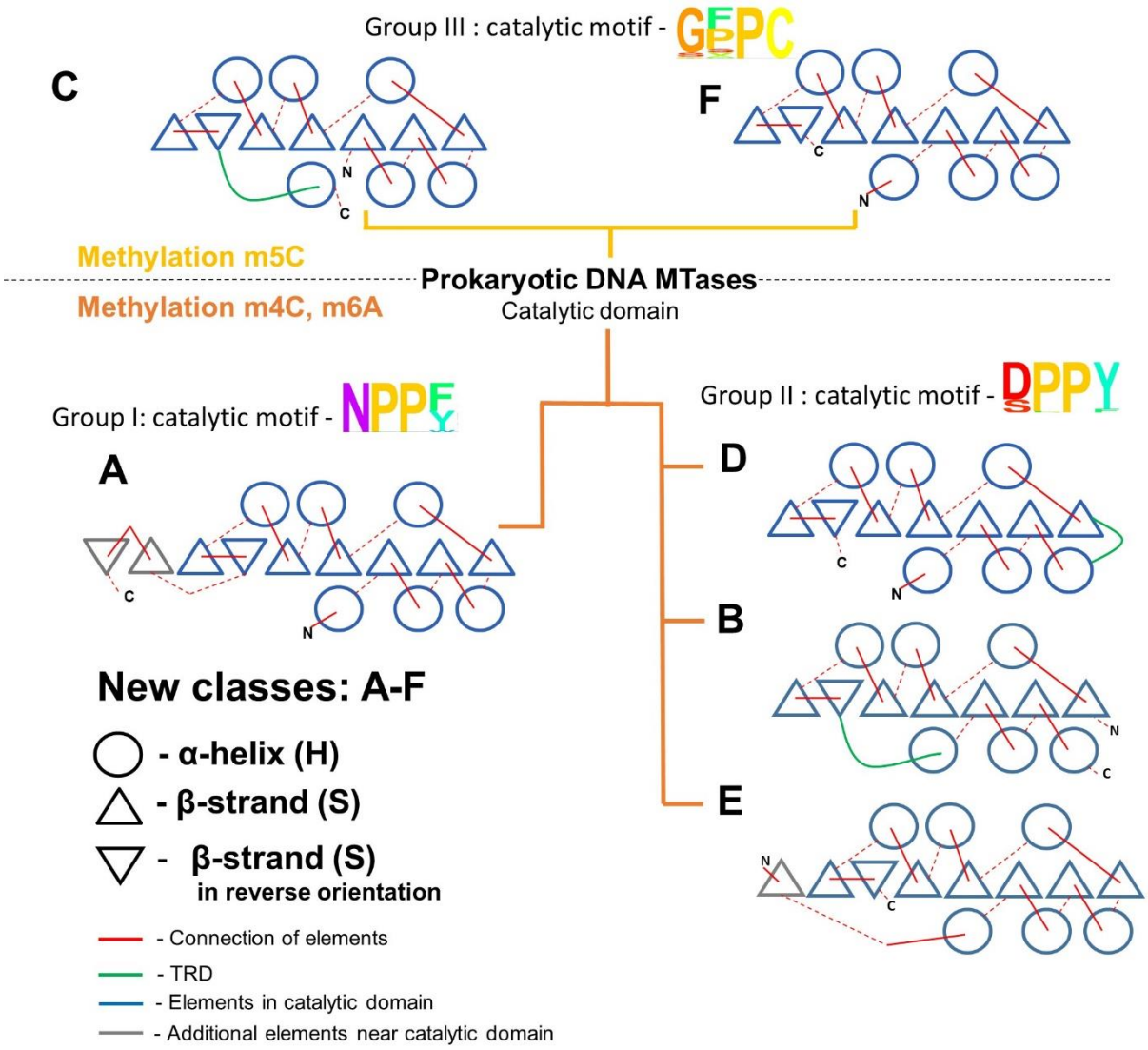
[1] Belozersky Institute of Physical and Chemical Biology, Lomonosov Moscow State University, Moscow, Russia

[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.

Here, we propose a novel classification of prokaryotic MTases that addresses the limitations of 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.

We identified six structural classes (A-F) 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 AlphaFold DB. The catalytic domains of each class of MTase exhibit homology not only in structure but also in conserved motifs. Based on this classification, we developed an algorithm for the automatic annotation of MTase genes and assignment of their structural class.

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 at the C5 position of cytosine was first discovered in bacteria in 1925, marking nearly a century of research into genome methylation and the enzymes responsible for this modification. These enzymes, DNA methyltransferases (MTases), were initially identified as components of restriction-modification (R-M) systems [1] and have since been recognized for their roles in diverse biological processes, including other bacterial immune mechanisms, cell cycle regulation, gene expression, and DNA replication [2,3].

MTases are typically classified using three complementary approaches: functional classification, classification by association with restriction-modification systems, and sequence similarity classification.

Functional Classification categorizes MTases according to the specific chemical reactions they catalyze and the resulting products. This approach defines three major enzyme classes [4]:

- **C5-MTases**: Produce C5-methylcytosine (m5C; EC 2.1.1.37)
- **N6-MTases**: Produce N6-methyladenine (m6A; EC 2.1.1.72)
- **N4-MTases**: Produce N4-methylcytosine (m4C; EC 2.1.1.113)

These three types of DNA methylation—m6A, m4C, and m5C—are now well-established [5,6]. MTases from all three classes are found in bacteria and archaea, with N6-MTases being

the most prevalent [7]. In eukaryotes, C5-MTases that methylate cytosine in CG dinucleotides are predominant [8]. The quantitative distribution of these MTase types across domains of life, based on UniProt data, is presented in Table 1.

Table 1. Distribution of DNA methyltransferases in UniProtKB*.

Classification	Swiss-Prot (%)			TrEMBL (%)		
	Bacteria	Archaea	Eukaryota	Bacteria	Archaea	Eukaryota
C5-MTases (EC 2.1.1.37)	60 (36.6%)	4 (25.0%)	34 (91.9%)	42,384 (23.0%)	1,372 (12.4%)	12,672 (93.7%)
N4-MTases (EC 2.1.1.113)	16 (9.8%)	6 (37.5%)	0 (0%)	14,833 (8.1%)	4,463 (40.4%)	138 (1.0%)
N6-MTases (EC 2.1.1.72)	88 (53.7%)	6 (37.5%)	3 (8.1%)	127,496 (69.0%)	6,063 (54.8%)	292 (2.2%)
Total	164 (100%)	16 (100%)	37 (100%)	184,713 (100%)	11,061 (100%)	13,531 (100%)

**Data extracted from UniProtKB on 25 August 2025. Percentages represent the proportion within each domain and database.*

Classification by association with R-M systems groups MTases based on their role within the restriction-modification machinery, a key prokaryotic defense system. MTases are integral components of Type I, II, and III R-M systems. In Type II systems, the MTase is typically a standalone enzyme, while in Type I and III systems, it acts as a subunit within a multi-enzyme complex [9]. A notable variant is Type IIG systems, where the genes for the restriction enzyme and the MTase are fused, resulting in a single polypeptide chain (RM proteins) [10]. Furthermore, a significant number of "orphan" MTases exist that are not associated with any restriction enzyme and are involved solely in regulatory functions, such as controlling gene expression and DNA replication [11].

Sequence similarity classification was introduced in 1995 when researchers identified nine conserved motifs in known MTases. Based on the order of these motifs and the position of the target recognition domain (TRD), MTases were categorized into four groups (α, β, γ, and a group for C5-MTases) [12]. This classification was later formalized into an evolutionary model involving circular permutations, which proposed six possible classes (α, β, γ, δ, ε, ζ) by accounting for different linear arrangements of the conserved subdomains. Among these, the α, β, γ, and ζ classes were confirmed in experimentally characterized MTases, while the δ and ε classes remained hypothetical [13].

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 Restriction Enzyme Database (REBASE) [14]. 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 this work, we present a new classification system for prokaryotic MTases based on the structure of their catalytic domain, leveraging a contemporary dataset from REBASE. We developed an algorithm for the automatic annotation of MTase genes and simultaneous determination of their structural class. We also provide a web application for visualizing classification results and, based on our findings, propose a plausible evolutionary model for prokaryotic MTases.

MATERIALS AND METHODS

Materials

MTase 3D Structures. We retrieved a list of 38 MTases with experimentally determined 3D structures from REBASE (as of June 3, 2025; see 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 REBASE.

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: the **REBASE MTase Dataset**, a collection of 174,874 unique protein sequences annotated as MTases in REBASE, and the **REBASE Gold Standard**, a subset of 1,678 unique MTases with experimentally confirmed enzymatic activity (refer to supplementary material S3). For the purpose of building our

classification system, we utilized the Curated REBASE Gold Standard, while the REBASE MTase Dataset was employed to validate the classification outcomes.

UniProt Source of MTase Sequences. On February 9, 2025, we conducted an advanced search in the UniProt Knowledgebase and identified 183,588 annotated MTases (supplementary material S4). This collection is hereafter referred to as the **UniProt MTase Dataset**. 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 entry 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 enzyme numbers, PDB codes, and taxonomic lineage.

No additional prokaryotic MTases with known 3D structures were found in UniProt that were not listed in REBASE. Among the eukaryotic MTases, we found seven proteins for which 3D structures are known (Table S1b).

Methods

Topology classification of MTases. Topology was determined by analyzing the order of secondary structure elements. Information about secondary structure was obtained from visualizations and structural data presented in the RCSB PDB (rcsb.org) and AlphaFold DB (alphafold.ebi.ac.uk) databases.

HMM-profiles search. In protein family DBs Pfam, SUPFAM, CDD, PANTHER and others there are a number of protein domain families, whose annotations include terms like "DNA methyltransferase domain". 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 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 5 families within the S-adenosyl-L-methionine-dependent methyltransferases superfamily in SUPFAM that are built exclusively on prokaryotic DNA

MTases. From these 5 families, we took all available profiles, totaling 12 HMM profiles. Additionally, we included 3 profiles from Pfam that detected MTases in the Gold Standard dataset not identified by the SUPFAM profiles.

Among Gold Standard sequences without hits from these selected profiles, M.BceJII showed similarity to known MTases. For this sequence and its homologs, we built a new HMM profile (New_MTase_profile) using the hmmbuild program (Supplementary Material S5b).

In total, 16 profiles were selected for classification: 12 from SUPFAM, 3 from Pfam, and 1 newly constructed profile (Supplementary Material S5a).

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 (↓↑↓↓↓↓↓) sandwiched between α-helices [15]. 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 convenience of subsequent analysis, we introduced a unified nomenclature system for secondary structure elements, defining a standard view in which the catalytic domain is oriented such that the six parallel β-strands run away from the viewer, and the single antiparallel strand faces the viewer. In this orientation, three α-helices are positioned above the β-sheet (labeled Hu1-Hu3) and three below it (Hd1-Hd3). β-strands are numbered S1 to S7 from left to right (Figure 1a).

Topological classification of prokaryotic MTases on the base of 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 domain. Based on the **topology** (the sequential order of secondary structure elements) of their catalytic domains, the 38 known MTase structures form four distinct groups, designated as **Topologies 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].

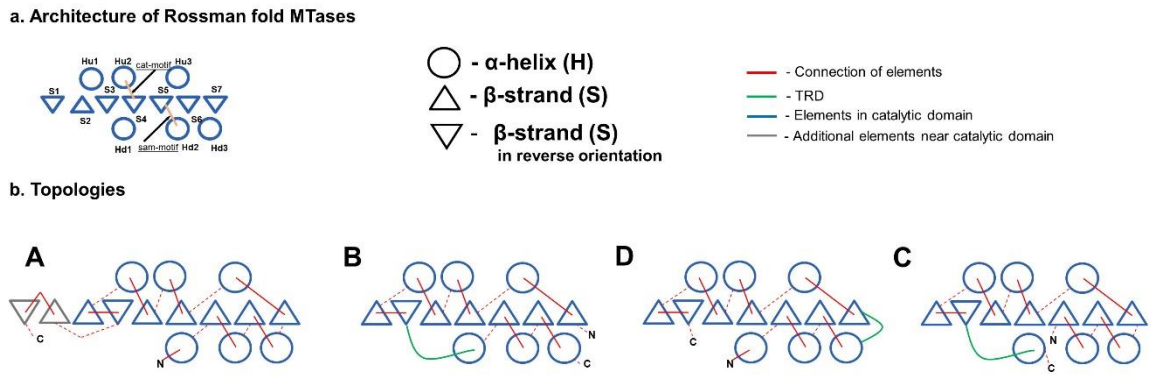


Figure 1. Topological schemes of MTase catalytic domains with experimentally determined structures.

Topology 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 topology 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 topology follows the consensus sequence NPPF/NPPY.

All RM proteins (RM.BpuSI, RM.DrdV, RM.LlaGI, RM.LlaBIII, and RM.MmeI) belong exclusively to Topology A, supporting their evolutionary relationship. Notably, Topology A MTases lack a TRD (target recognition domain) within the catalytic domain. Instead, the target site is recognized by another subunit of the active complex, called S-subunit, which is sometimes C-terminally joined to the MTase.

All 19 members of this topology 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 (loops,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.

Topology B comprises 10 MTases with unique structural features. Unlike Topology A, these proteins exhibit an altered motif order along the polypeptide chain: the catalytic motif (consensus DPPY/SPPF) precedes the SAM-binding motif. A hallmark of this topology is the presence of an embedded TRD (target recognition domain) between β-strand S2 and α-helix Hd1. Some members also feature an additional strand S8 and a strand S0 located within the TRD.

Topology 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 topologies. Unlike Topologies A and B, the catalytic domains in Topology 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. 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.

Topology 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 topology shares similarities with Topology 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 four members of Topology 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 SUPFAM profiles from the SCOP superfamily of S-adenosyl-L-methionine-dependent methyltransferases (MTases), which includes five families of prokaryotic DNA MTases. Each profile was assigned to a topology 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 (and table S6a).

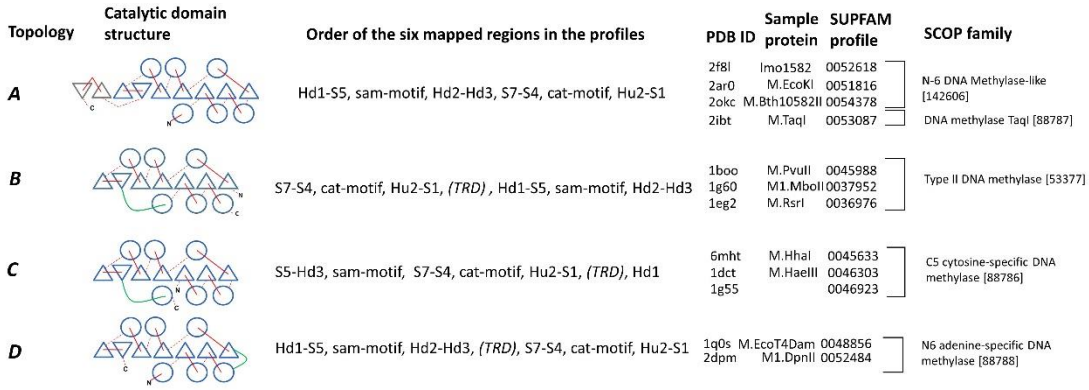


Figure 2. Division of 12 SUPFAM profiles into 4 groups by topology.

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. Notably, topology D was frequently detected not only by its specific D-topology profiles but also by two separate B-topology profile hits: one identifying the catalytic domain segment preceding the TRD insertion and another detecting the segment following the TRD.

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).

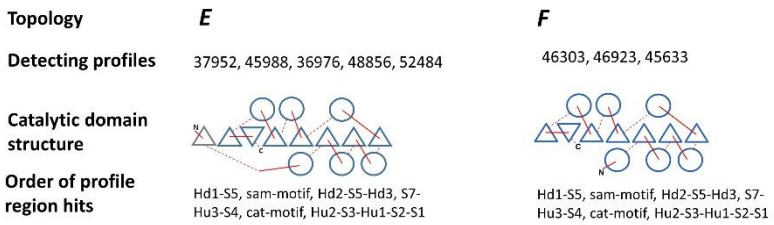


Figure 3. Two additional topologies of MTases.

The first group, referred to as **Topology E**, includes the experimentally confirmed MTases M1.BcnI [16] and M4.BstF5I [17]. Similar to topology D, topology E is detected by the same set of B- and D-associated profiles but lacks the TRD insertion between Hd3 and S7 regions. This structural difference is reflected in the spacing between functional motifs: while topology D exhibits a distance exceeding 150 amino acids between the sam-motif and catalytic motif, this interval is shortened to less than 100 amino acids in topology E. Notably, the predicted structures of M1.BcnI and M4.BstF5I reveal an additional S0 beta strand, which may contribute to the structural stabilization of the catalytic domain in the absence of the TRD insertion.

The second group, designated as **Topology F**, includes the M2.Eco31I MTase [18] 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 fact that its target recognition domain (TRD) is not inserted within the catalytic domain, as is characteristic of topology D. Instead, the TRD is likely positioned as an independent N-terminal module. These observations are supported by both predicted structure analysis and the mapped regions of profile alignments.

In summary, based on the identified topologies, we introduce a system of six structural classes (A-F) for MTases, where each class is defined by a unique topology of the catalytic domain.

Classification algorithm

We developed an automated classification algorithm that identifies MTase classes 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 each query sequence, a global alignment to the full-length profile was generated. Subsequently, segments of this alignment corresponding to each mapped region within the profile were extracted and evaluated against a set of quality criteria: (1) the alignment segment must cover more than 40% of profile positions within the region; (2) the ratio of insertion positions to aligned

residues within the segment must be less than 2.5; and (3) for functional motifs (sam-motif and cat-motif), the segment must cover more than 75% of the motif's reference positions with no more than one insertion. These thresholds were empirically determined to maximize the number of correctly classified MTases from the REBASE Gold Standard dataset, effectively distinguishing true positive hits from spurious alignments while maintaining sensitivity for divergent sequences. Only segments fulfilling all applicable criteria were accepted as evidence for the presence of that specific profile region in the query sequence.

This core algorithm was subsequently extended by incorporating three Pfam profiles and an additional novel profile to classify sequences that remained unassigned after the initial round (see Section "Unclassified Sequences").

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.

We clustered the SUPFAM profiles into three profile groups based on which structural classes they detect:

- Profile Group I (detects Class A): P profiles 51816, 52618, 54378, 53087.
- Profile Group II (detects Classes B, D, E): P profiles 37952, 45988, 36976, 48856, 52484.
- Profile Group III (detects Classes C, F): P profiles 46303, 46923, 45633.

Profiles within the same group often identify the same class, acting as alternatives for one another. For example, both profile 52484 and profile 45988 can be the optimal for different sequences of Topology D.

The final class assignment for a query sequence is therefore determined by the group of the top-performing profile and the order of its detected regions, as detailed in Table 2. To account for evolutionary divergence and partial sequences, 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 3) 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 the profile-specific order of the regions and acceptable partial detection patterns are provided in Table 2 and Supplementary Materials S7.

Table 2. Profile groups, region order, and structural class definitions.

Class	**Profile Group**	**Detection profile**	**Order of region hits in profile**	**Samples of incomplete acceptable by algorithm orders of region hits detected by profile**
A	I	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	II	37952, 45988, 36976 48856, 52484	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	III	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	II	48856, 52484 37952, 45988, 36976	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	II	37952, 45988, 36976, 48856,	Hd1-S5, sam-motif, Hd2-S5-Hd3, S7-Hu3-S4, cat-motif,	sam-motif, Hd2-S5-Hd3, cat-motif,

		52484	Hu2-S3-Hu1-S1-S2 *(Distance between sam-motif and cat-motif less 100 aa)*	Hu2-S3-Hu1-S1-S2
F	III	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 3. Classification results of the MTase topology prediction algorithm.

Class	REBASE v408		UniProt MTase Dataset
	REBASE Gold Standard Dataset	REBASE MTase Dataset	
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 catalytic motifs

To assess whether the identified structural classes exhibit distinct sequence signatures in their catalytic pockets, we analyzed the catalytic motifs of gold-standard MTases from each class (Figure 4, Table S8).

We observed a consistent relationship between structural class and catalytic motif sequence:

- **Class A** (detected by Profile Group I) is predominantly associated with the motif N-P-P-(Y).
- **Classes B, D, and E** (detected by Profile Group II) most frequently contain motifs like (D/S/H)-P-P-(Y/F), such as D-P-P-Y or S-P-P-F.

- **Classes C and F** (detected by Profile Group III) are uniquely characterized by a conserved proline-cysteine dipeptide, forming motifs like G-P-P-C.

We also detected rare motif variants (e.g., NLPY) that deviate from the canonical signatures described above. The functional and evolutionary significance of these substitutions remains to be determined.

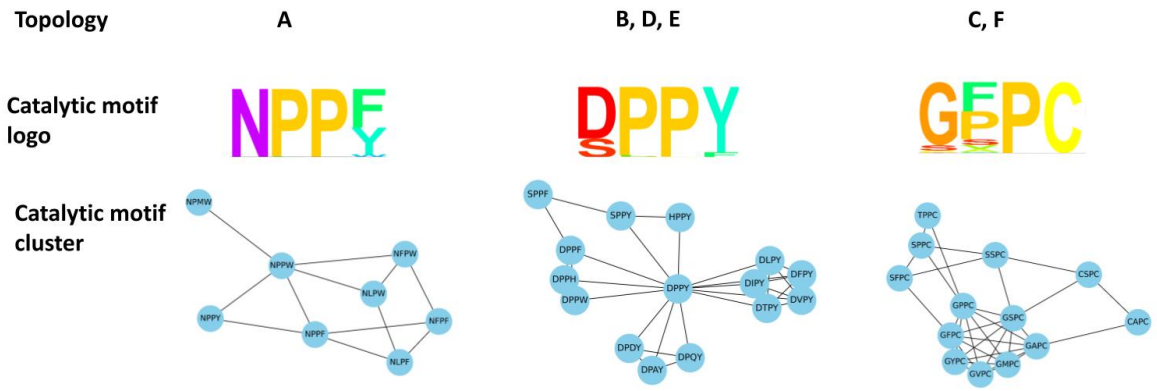


Figure 4. MTase catalytic motifs. The Catalytic Motif Logo displays the most conserved positions in a catalytic motif within all sequences of the same group. The catalytic motif cluster lists all unique catalytic motif sequences within one group of MTases. An edge is drawn between two catalytic motifs in a cluster if the motifs differ by one amino acid.

Unclassified sequences

Among the 1,678 sequences in the REBASE gold standard dataset, our algorithm failed to assign a structural class to 44 MTases. Detailed analysis revealed that 10 sequences were misannotated as MTases (Table S9a), as they lacked both catalytic and SAM-binding motifs and displayed non-Rossmann fold predicted structures. Another 13 sequences represented MTase fragments (Table S9b), containing only one of the two characteristic motifs. Two MTases (M.Pbo13188I and M.TliII) belong to 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). We adapted these profiles for use in

our algorithm by annotating the same set of six conserved regions within them (Table S6b). 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 4), 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 4. Classification of gold standard sequences with non-canonical MTase architecture.

Subgroup	Class	Detecting profile	Sample sequence	Catalytic motif	Predicted structure of sample sequence (AlphaFold2 Model)
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 multiple sequence alignments for each MTase class to identify conserved core elements and variable regions (Supplementary Materials 10). Manual analysis of these alignments revealed that while the core secondary structure elements (e.g., β-strands S1-S7,

α-helices Hd1-Hd3, Hu1-Hu3) are highly conserved, they are sometimes flanked by regions of high sequence variability. In some MTases these variable regions manifest as large insertions (>20 amino acids).

To investigate the structural nature of these insertions, we relied on AlphaFold2 predictions. Table 5 presents a non-exhaustive selection of MTases containing such inserts, illustrating their diversity in length and location within the catalytic domain. Examination of the predicted structures confirms that these insertions form surface-exposed loops that do not disrupt the conserved Rossmann-fold core of the catalytic domain. Notably, some inserts, particularly in Class A MTases, can be exceptionally long, exceeding 100 amino acids.

Table 5. Examples of MTases with large insertions in the catalytic domain based on AlphaFold2 predictions.

Class	MTase	Insert Position	Insert Size (aa)	AlphaFold2 Model
A	Ssp6803IV	Between Hd1 and S5	131	AF-Q6YRV4-F1
		Between Hd2 and S6	49	
		Between S7 and Hu3	111	
B	M.AatII	Between S4 and Hu2	55	AF-P70724-F1
		Between S1 and S2	29	
C	M.BspRI	Between Hd2 and S6	24	AF-P13906-F1
D	M.AvaI	Between S4 and Hu2	90	AF-P0A461-F1
E	M.HtuII	Between S4 and Hu2	70	AF-D2RV96-F1

Correlation between structural classes and R-M system types

To assess the biological relevance of our classification, we analyzed the association between structural classes and restriction-modification (R-M) system types using REBASE annotations (Table 6). The distribution revealed strong correlations: Class A contains all MTases from Type I and Type IIG systems, and all Type III system MTases belong exclusively to Class B. MTases from Type II systems were distributed across classes A, B, C, and D, reflecting their structural diversity. Orphan MTases were found primarily in classes B, C, and D, with the highest number in Class D.

Table 6. Association of MTase structural classes with R-M system types.

Class	Orphan M	Type I	Type II	Type IIG	Type III
A	2	733	110	113	0
B	13	0	166	0	99
C	30	0	206	0	0
D	47	0	141	0	0
E	0	0	7	0	0
F	0	0	2	0	0
B (Subgroup G)	0	0	8	0	0
A (Subgroup H)	2	0	2	0	0
A (Subgroup I)	0	0	4	0	0
E (Subgroup J)	1	0	2	0	0

DISCUSSION

We propose a novel classification system that groups prokaryotic DNA methyltransferases into six structural classes (A-F) based on the topology of their catalytic domain. The conserved topology and catalytic motifs within each class imply their monophyletic origin, suggesting that these structural classes seem to represent families in the evolution of DNA methyltransferases.

Our system successfully captures the broad diversity of known MTases while also identifying rare variants, such as subgroups G-J, which may represent structural innovations or transitional forms. Future structural studies of these outliers will be crucial to refine our understanding of MTase evolution and functional diversification.

Comparison of new classification with previous classification

We have compared our newly developed classification system with existing classification schemes (Table 7) to assess their congruence and biological relevance.

The functional correlations demonstrate the biological relevance of our classification. Classes C and F (Profile Group III) exclusively catalyze m5C methylation, while other classes mediate m4C or m6A modification. Furthermore, we observe strict associations with restriction-modification system types: all Type I systems belong to Class A, all Type III systems to Class B, and Type IIG systems exclusively to Class A. In contrast, Type II systems distribute across Classes A, B, C, and D, reflecting their broad structural diversity.

Comparison with the motif-order classification [12, 13] demonstrates that class B and D correspond clearly to the β and α subtypes, respectively. However, class A highlights limitations of the previous system. According to that model, the γ subtype requires a C-terminally fused TRD. In class A, this applies only to a subset of MTases; the majority (particularly those from Type I R-M systems) lack a fused TRD and rely on a separate S-subunit for DNA recognition. Thus, class A contains both γ-like MTases and a distinct group with multi-subunit architecture that has no direct analogue in the previous classification system.

Class E represents a novel structural arrangement that is not accounted for by the previous classification system. Proteins like M1.BcnI lack a large inserted TRD domain, and their DNA recognition mechanism, likely mediated by surface loops, prevents their assignment to any of the known α, β, or γ subtypes.

Class F MTases, with their unique topology and m5C activity, align with the previously described but rare ζ subtype.

This detailed comparison demonstrates that our structural classification not only recapitulates the known groups (α, β, C5, ζ) but also resolves ambiguities and reveals previously unappreciated structural diversity, such as the subdivision of the γ-related forms and the identification of novel classes like E.

Table 7. Comparison of the new classification with the previous classifications.

New class	Profile Group	Methylation Type	R-M Type	Corresponding Subtype (Motif-based)	Notes on TRD
A	I	m4C, m6A	I, IIG, II	γ (partial)	Correspondence to γ is valid only for members with aC-terminally fused TRD.
B	II	m4C, m6A	II, III	β	
D	II	m4C, m6A	II	α	
E	II	m4C, m6A	II	(Unclear / Novel)	Lacks a large, inserted TRD; recognition likely via loops. Does not fit previous subtypes.
C	III	m5C	II	C5	
F	III	m5C	II	ζ	

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 retain 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 [19]. Similarly, class E

MTases contain four loops at equivalent positions that may be also involved in sequence recognition by analogy with M.CcrMI.

The data show the complex organization of TRD domains, which may consist of either multiple structural elements (M.CcrMI), or a single large loop (classes C and D), or 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.

A further limitation is the algorithm's inability to account for large insertions within the catalytic domain. These insertions can distort the predicted spacing between secondary structure elements and obscure domain boundaries, which may lead to misclassification or a failure to assign a topology.

Despite these limitations, the potential applications of our algorithm extend to the detection of DNA MTases in newly sequenced genomes and metagenomes. However, it should be considered 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.

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 detecting HMM-profiles and 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 benefit genomic annotation, evolutionary studies, and the engineering of novel MTases for biotechnological applications.

DATA AVAILABILITY

Data available on github https://github.com/MVolobueva/MTase-classification 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

M. Samokhina conceived the study, performed the formal analysis, developed the classification system, implemented the algorithm, conducted the investigation, and wrote the

original draft. A. Alexeevski supervised the research, provided conceptual guidance, acquired funding, and reviewed and edited the manuscript.

ACKNOWLEDGEMENTS

The authors thank Ivan Rusinov for his valuable recommendations on the manuscript content and his assistance with text refinement.

FUNDING

The work 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. Ershova AS, Rusinov IS, Spirin SA, Karyagina AS, Alexeevski AV. Role of Restriction-Modification Systems in Prokaryotic Evolution and Ecology. Biochemistry (Mosc). 2015 Oct;80(10):1373-86.
10. Furuta Y, Kobayashi I. Restriction-Modification Systems as Mobile Epigenetic Elements. In: Madame Curie Bioscience Database [Internet]. Austin (TX): Landes Bioscience; 2000-2013. Available from: https://www.ncbi.nlm.nih.gov/books/NBK63963/
11. Isaev A.B., Musharova O.S., Severinov K.V. Microbial Arsenal of Antiviral Defenses – Part I. Biochemistry (Moscow), 2021, Vol. 86, No. 3, pp. 319-337.
12. 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.
13. Bujnicki,J.M. (2002) Sequence permutations in the molecular evolution of DNA methyltransferases. BMC Evol. Biol., 2, 3.
14. 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.
15. Cheng,X. and Roberts,R.J. (2001) AdoMet-dependent methylation, DNA methyltransferases and base flipping. Nucleic Acids Res., 29, 3784–3795.
16. 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.
17. 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.
18. 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.

19. 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.