A Novel Classification System for Prokaryotic DNA Methyltransferases Based on 3D Catalytic Domain Topology

AUTHORS

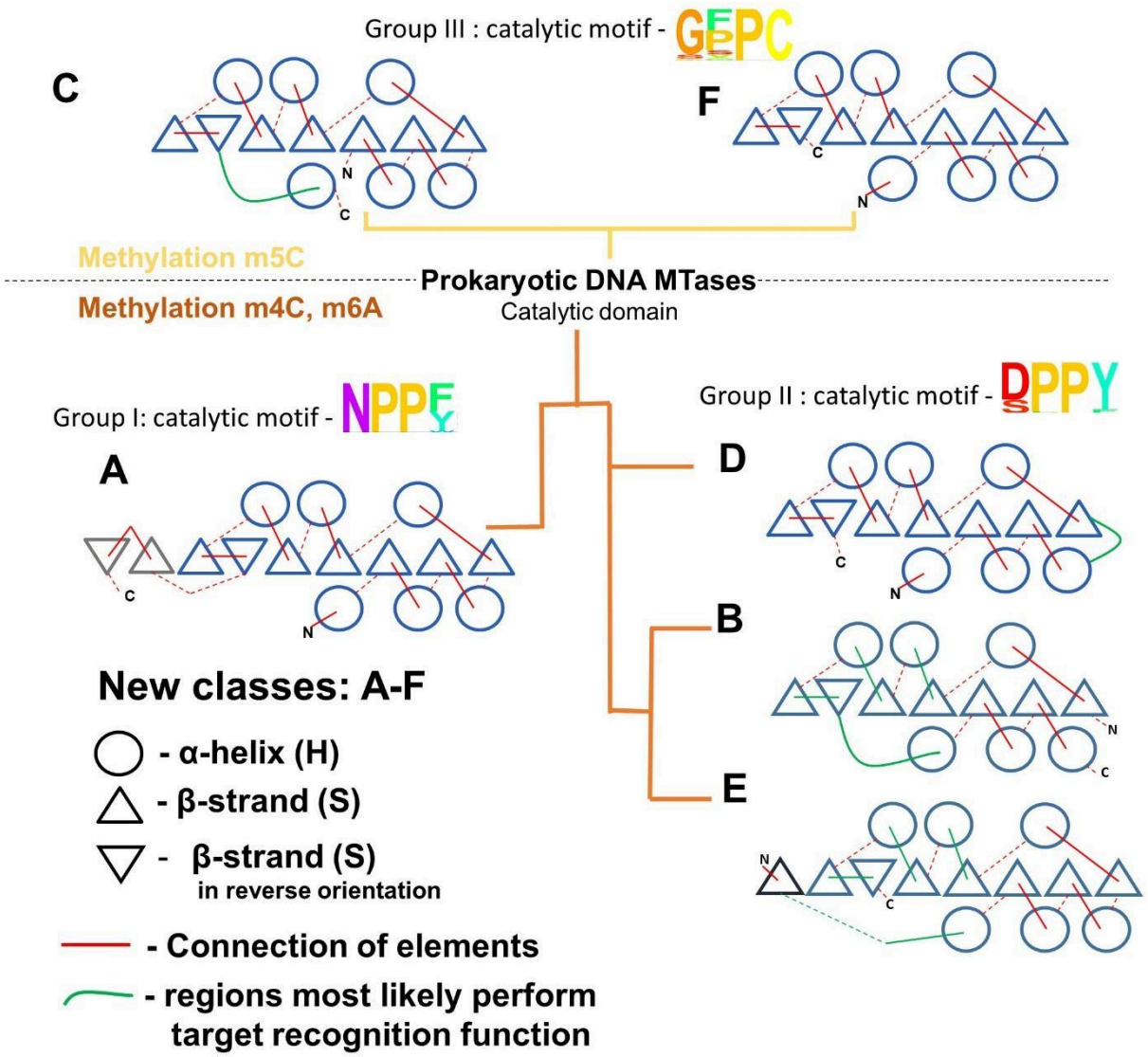
Maria Samokhina[1, 2]* and Alexeevski A.[1,2,3]

[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 database. 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

The discovery of DNA methylation in bacteria dates back to 1925, when Frederick Griffith first observed this phenomenon in Escherichia coli (E. coli) (1). Since then, researchers have been exploring the role of genome methylation and the enzymes responsible for this modification.

Initially, DNA methyltransferases (MTases) were identified as components of restriction-modification systems, where they played a crucial role in protecting bacterial genomes from foreign DNA (1). However, it soon became clear that these enzymes were involved in many other biological processes as well.

Today, we know that DNA methylation plays a vital role in various cellular processes, including immune mechanisms, cell cycle regulation, gene expression, and replication (2). In bacteria, DNA methylation helps to protect against foreign DNA, regulate gene expression, and ensure proper cell division (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 server 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 November 16, 2023, the REBASE database includes a list of 33 DNA methyltransferases (MTases) with known 3D protein structures. Among these, 32 structures were resolved using X-ray diffraction techniques, while one structure (RM.DrdV) was determined through electron microscopy (refer to Table S1). 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. In our search using UniProt services, we did not identify any additional prokaryotic MTases with known 3D structures.

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 versions v206, v309 and v409. 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 and another specifically comprising unique MTases from the "gold standard," 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 v206, while other sets of MTases were employed to validate the classification outcomes.

UniProt Source of MTase Sequences. On November 12, 2023, we conducted an advanced search in the UniProt database and identified 187,702 annotated MTases. 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.

We verified that all MTases with known 3D structures listed in UniProt are also included in the compilation from REBASE, confirming the accuracy and consistency of our structural data on methyltransferases.

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.

We used 11 profiles from the SUPFAM database, which are based on the 3D structures of prokaryotic DNA methyltransferases (MTases). These profiles all belong to the S-adenosyl-L-methionine-dependent methyltransferases superfamily. Notably, we selected gold standard proteins that did not exhibit any significant hits with the SUPFAM profiles. Instead, they showed matches with three profiles from Pfam: Dam, MT-A70, and EcoRI_methylase. The complete list of profiles can be found in Supplementary Material S5.

Among MTase sequences from the 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).

Domains and domain architectures detection. To build domain architectures, we used HMM profiles from DB Pfam. The hits were sorted in order from N-end to C-end.

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

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

The catalytic domains of methyltransferases (MTases) typically exhibit a structure comprising seven β-strands arranged with α-helices on either side of the β-sheet. The orientation of the strands is as follows: (↓↑↓↓↓↓↓) (12). In this arrangement, the first strand is oriented anti-parallel (↓), the second strand runs parallel to the first (↑), and the subsequent five strands are organized in an anti-parallel fashion (↓↓↓↓↓).

In our analysis, we examined the topologies of all 33 MTases with known three-dimensional structures (Table S1a). For each structure, we identified the catalytic domain as a distinct part of the protein globule, displaying 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 referred to as motif IV, and the S-adenosylmethionine SAM-binding motif (sam-motif), known as motif I, as previously described (10).

Our analysis confirmed the conservation of the Rossmann fold architecture across all MTases. In the catalytic domains of the 33 MTases, the β-sheet consisting of seven strands is consistently surrounded by three helices above and three helices below the sheet. We introduced designations for all conserved secondary structure elements, which we will reference in subsequent discussions (Fig. 1a). We also evaluated seven structures of eukaryotic MTases (Table S2), which likewise exhibited the Rossmann fold architecture.

Despite the highly conserved architecture of MTases, the topologies—including the connectivity of conserved secondary structure elements—vary among them. Based on the topology of the catalytic domains, all 33 MTases with known three-dimensional structures can be categorized into four groups: A, B, C, and D, each characterized by distinct catalytic domain topologies (Fig. 1b, Table S1). In the B, C, and D topologies, the cat-motif follows the sam-motif (sam-cat), while in topology B, the cat-motif precedes the sam-motif (cat-sam). Additionally, Helix Hd1 is positioned differently across the topologies: it is N-terminal in topologies A and D, C-terminal in topology C, and centrally located in the polypeptide chain in topology B.

The likely positioning of the target recognition domain (TRD) also varies by topology. In the B, C, and D topologies, the TRD is located between different consecutive secondary structure elements of the catalytic domain. In contrast, MTases classified under topology A are primarily derived from type I restriction-modification (R-M) systems, which include a separate DNA recognition protein, the S-subunit. Other MTases with topology A contain the TRD as a separate domain located

C-terminal to the catalytic domain, likely resulting from the fusion of M and S genes. Furthermore, topology A is characterized by the presence of two extra β-strands.

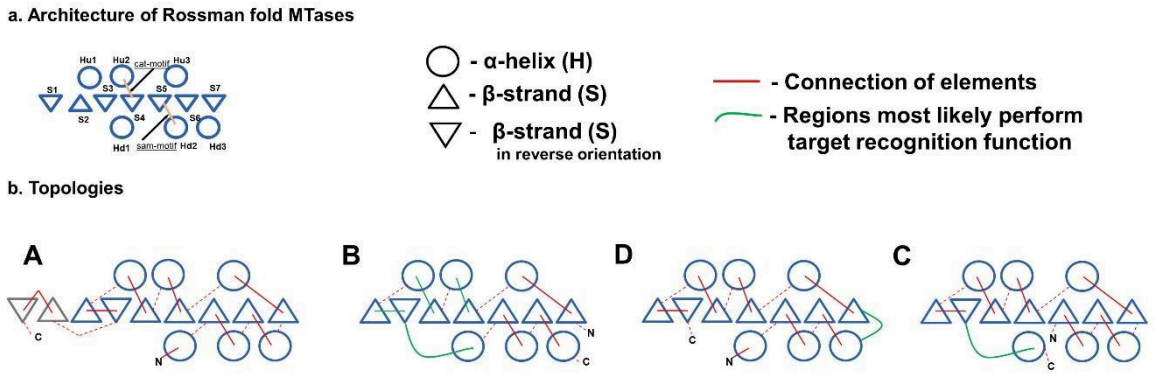


Figure 1. Groups of topological schemes of all MTases with known 3D structures.

MTase HMM profiles

We primarily selected profiles from the SUPFAM database, corresponding to the 3D structures of MTases. Each SUPFAM profile is constructed based on a single protein PDB 3D structure and the alignments of its close homologs, defined by reliable sequence similarity.

Of the 33 available MTase structures, only 11 were found in the SUPFAM database. We selected 11 profiles that were constructed on these 11 structures. All of these profiles belong to the S-adenosyl-L-methionine-dependent methyltransferases superfamily within SUPFAM.

Each profile corresponds to one of four topologies: A (#prof), B (∧), C (∧), or D (∧). This allows us to mark specific regions within the profile, corresponding to six key regions within each profile:

- Four regions linked to secondary structure elements
- Two regions associated with the catalytic and SAM-binding motifs

These six regions collectively cover the catalytic domain in topologies B, C, and D, and the main part of topology A, except for the S-1-S0 strands (as illustrated in Figure 2).

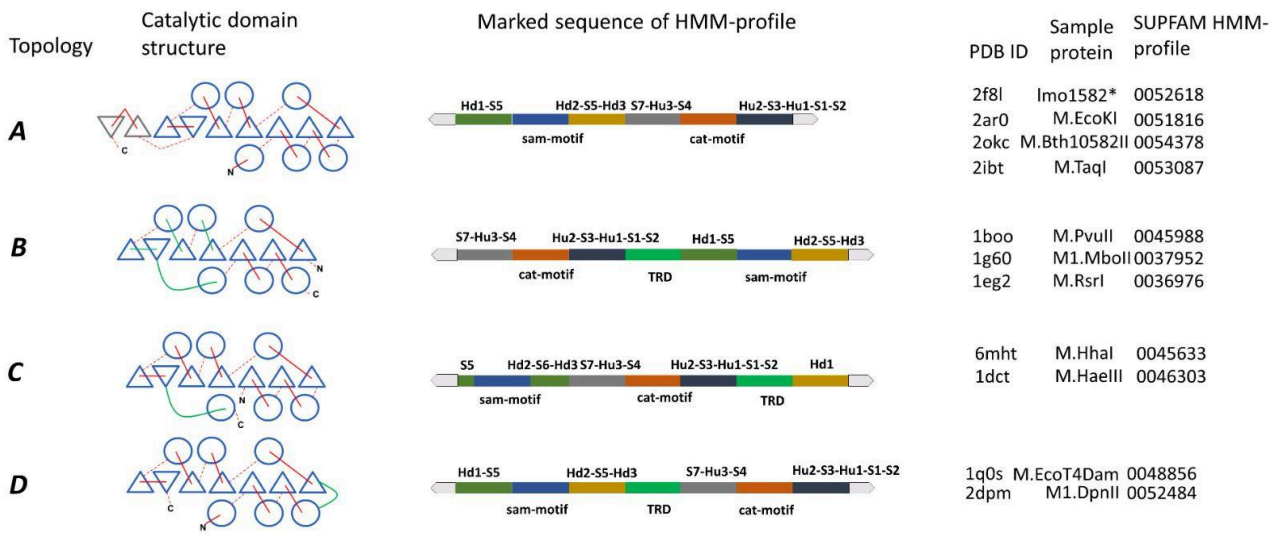


Figure 2. Marked MTase HMM-profiles.

*- Predicted MTases that lack matches in REBASE.

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).

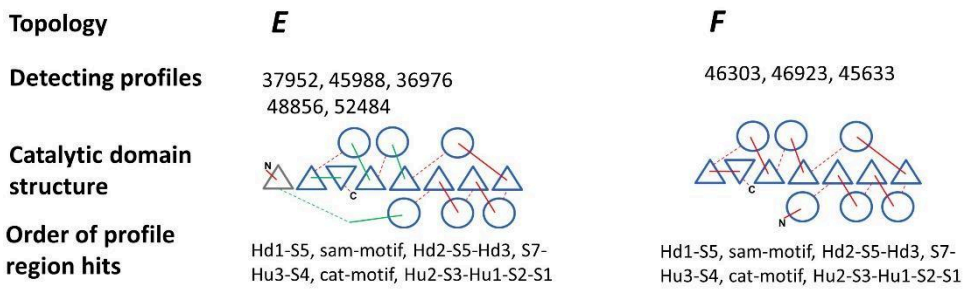


Figure 3. Two additional topologies of MTases. For topology E, loops that may be involved in DNA recognition are marked in analogy with topology B.

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 big subdomain located between the HD3 and S7 regions, which is

present in topology D. Additionally, 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, comprises the MTase M2.Eco31I (15), which has experimentally confirmed m5C activity. Although the topology of this group resembles that of topology D, it is characterized by profiles associated with topology C. A notable feature of topology F is that the Hd1 helix is located at the N-terminus, and there is no Target Recognition Domain (TRD) present within the catalytic domain. This observation is supported by both predicted structural analysis and the arrangement of the mapped profile regions

In summary, we identified a total of six groups of MTases, each characterized by unique topologies and structural features.

Classification algorithm

We have developed an automated algorithm capable of detecting the topology of MTases based on region hits from selected profiles within MTase sequences. The algorithm functions by initially identifying profiles that can detect specific motifs, along with at least two regions corresponding to secondary structure elements, as outlined in Table 1.

Once these profiles are selected, the algorithm prioritizes the profile that identifies the greatest number of secondary structure elements. Utilizing this prioritized profile along with the order of detected regions, the algorithm then classifies the MTase into its corresponding topology class.

For more information and access to the implementation of this algorithm, you can see supplementary materials S7 or visit our GitHub repository at https://github.com/MVolobueva/MTase-classification.git or explore our user interface at https://mtase-pipeline-6g1yfq9ugw8.streamlit.app/.

Table 1. Comparison topologies with detected profile regions.

Topology	Detection profile	Order of region hits	Samples of incomplete acceptable orders of region hits
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 *(Distance between sam-motif and cat-motif less 100 aa)*	sam-motif, Hd2-S5-Hd3, cat-motif, Hu2-S3-Hu1-S1-S2
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

The results of our MTase topology detection algorithm are presented in Table 2 and supplementary materials S7. Our analysis indicates that among the gold standard unclassified proteins, only two proteins have been confirmed to be MTases (M.Pbo13188I and M.TliII), both of which belong to topology D.

Table 2. Algorithm classification results.

	REBASE v206		**REBASE v309**		
Class	**Among unique gold standard MTases**	**Among all sequences**	**Among unique gold standard MTases**	**Among all sequences**	**Uniprot**
A	952	81311	951	97724	79494
B	275	61403	275	73642	25459
C	193	53326	190	65306	30462
D	181	46662	181	56162	29302
E	7	1498	7	1751	1513
F	2	357	2	515	86
several catalytic domains	15	1523	15	1874	1172

MTase classes

To determine whether the group of MTases with a particular topology is monophyletic or paraphyletic, we analyzed the catalytic motifs in the gold standard MTases for each topology. Our analysis revealed that the 6 MTase topologies can be grouped into 3 distinct groups based on the sequence of their catalytic motifs (Figure 4).

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.

Each topology has been assigned to a single class. The 6 classes can be further grouped into 3 superclasses based on the similarity of their catalytic motifs.

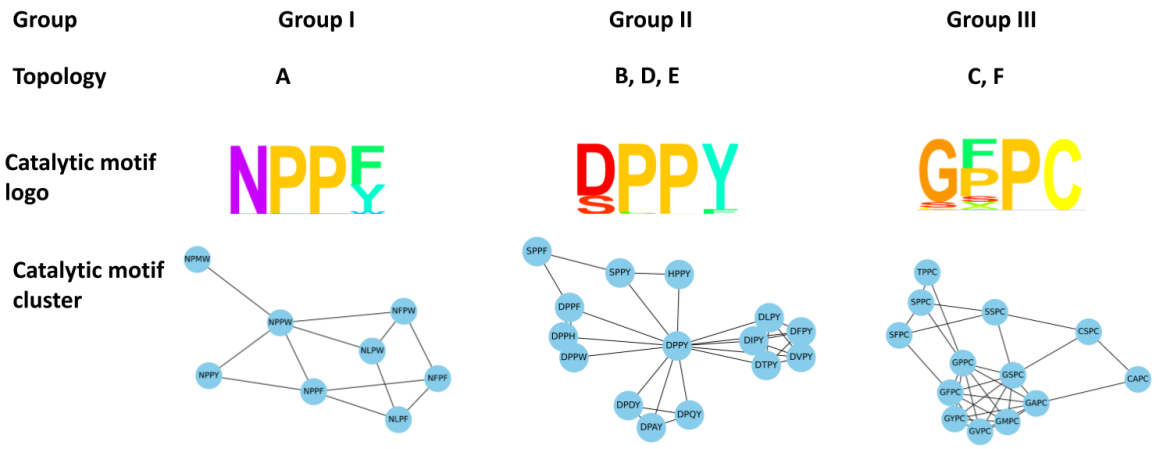


Figure 4. Division of 6 classes into 3 groups according to the similarity of catalytic motifs. The Catalytic Motif Logo displays the most conserved positions in a catalytic motif within all sequences of the same group. The catalytic motif cluster lists all unique catalytic motif sequences within one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

Large loops in catalytic domains

We constructed alignments for each MTase class, as detailed in Supplementary Materials S9. 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 3 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 3. 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

We have conducted a comprehensive analysis of available and predicted structures of MTases from REBASE, and subsequently classified these enzymes into six distinct classes. Classes A, B, C, and D are based on the availability of three-dimensional structures, which exhibit a characteristic fold comprising seven beta-strands surrounded by six alpha-helices. In contrast, Classes E and F share the same fold, but their topology is constructed using predicted structures. We propose that each class constitutes a monophyletic group, as MTases within a class exhibit identical topological arrangements and similar catalytic motifs. Furthermore, based on the similarity of these catalytic motifs, we have grouped the classes into three distinct categories: Group I, Group II, and Group III.

Comparison of new classification with previous classification

In Table 3, 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 4. 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 2) 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 (28) 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 5. 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*

Large insertion within MTase catalytic domain

Large insertions were found in the protein sequences of several MTase classes, which do not correspond to the previously described TRD regions. These insertions were identified through expert analysis of protein sequence alignments within each class. Notably, these large loops do not appear to disrupt the overall structure of the catalytic domain, as supported by both available and predicted structural data.

These large loops likely play a crucial role in MTase function, potentially affecting the enzyme's specificity and activity. For example, in MTase M.CcrMI from class B, four loops (S4-Hu1, S5-Hu1, S1-S2 and S2-Hu3 (TRD)) have been shown to recognize specific residues in the target DNA strand (16). Similarly, MTases from class E also possess four loops (S4-Hu1, S5-Hu1, S1-S2 and S0-Hu3 (analog of a loop S2-Hu3 in class B)) that are likely involved in DNA recognition. Additionally, MTase M.CcrMI features an independent non-specific DNA binding domain at its C-terminus, which may also be considered part of the TRD region (16).

This study highlights the complexity of TRD regions, which may consist of multiple elements , a single large loop or independent domain. For instance, MTases from class A have a TRD as an independent domain, while those from classes B and E possess TRDs made up of several elements. In contrast, MTases from classes D and C present TRDs as single large loops. Moreover, some MTases from class B, in addition to having four loops, also contain a large loop (S2-Hu3) that may serve as an additional DNA-binding region. In MTase M.EcoP15I, for example, a portion of the S2-Hu3 loop interacts with the DNA backbone, inducing conformational changes in the targeted modified region (17). The presence of large loops across almost all classes of MTase indicates that TRDs comprised of multiple elements are a common feature.

Overall, this study suggests that large loops may play a significant role in MTase function, potentially influencing the enzyme's specificity and activity. Understanding the roles of these loops is crucial for designing MTases with tailored recognition activities, which could have important implications for various biotechnological applications. Future investigations into the structure, function, and regulation of these loops are essential for realizing the full potential of MTase engineering.

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.

Notably, two MTases were not successfully classified by our algorithm due to the presence of large Target Recognition Domains (TRDs) within their catalytic domains. This likely interfered with the detection of secondary structural elements that our algorithm relies on for classification. Furthermore, large inserts within catalytic domains also hindered the detection of these elements by the selected profiles. This highlights a limitation of our algorithm in cases where TRDs are significantly large, suggesting that further refinement may be necessary to improve detection capabilities for MTases with complex structures.

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.

Unclassified REBASE sequences

Within the unclassified sequences from the gold standard dataset, we identified four groups exhibiting predicted structures similar to the Rossman fold, albeit lacking several key secondary structure elements. Notably, all sequences within these groups contain the catalytic motif, though the region typically associated with the SAM-binding motif is significantly reduced. MTases from three of these groups were detected using profiles from Pfam, while for the fourth group, we generated new MTase profiles. These four profiles were further mapped across six regions, and we designated these groups as G-J (see Figure 5). However, we did not find any experimental evidence to support the idea that these sequences possess MTase activity. This raises important questions regarding the nature of these sequences: do they represent additional classes of MTases, or are they simply annotation errors? Further investigation is required to clarify this issue.

Table 6. Group of sequences from gold standard with non-MTase architecture.

Group	Detecting profile	Sample sequence	Predicted structure of sample sequence
G	MT-A70 from Pfam	M.AvaV	AF-Q9KIW1-F1
H	Dam from Pfam	M.Rsp105IV	AF-A0A2L1UPL1-F1
I	EcoRI_methylase from Pfam	M. EcoRI	AF-P00472-F1

J	New-MTase-profile	M.BceJII	AF-B4EFP2-F1

Additionally, we discovered other unclassified sequences that were annotated as MTases but were, in fact, fragments of MTases. This finding underscores the importance of thorough curation and annotation of genomic data to ensure an accurate representation of biological function.

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 B and E form a clade because the MTases in these groups feature TRDs comprised of four loops within their catalytic domains. In contrast, MTases from topology D possess a single large loop that serves as the TRD. 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 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.

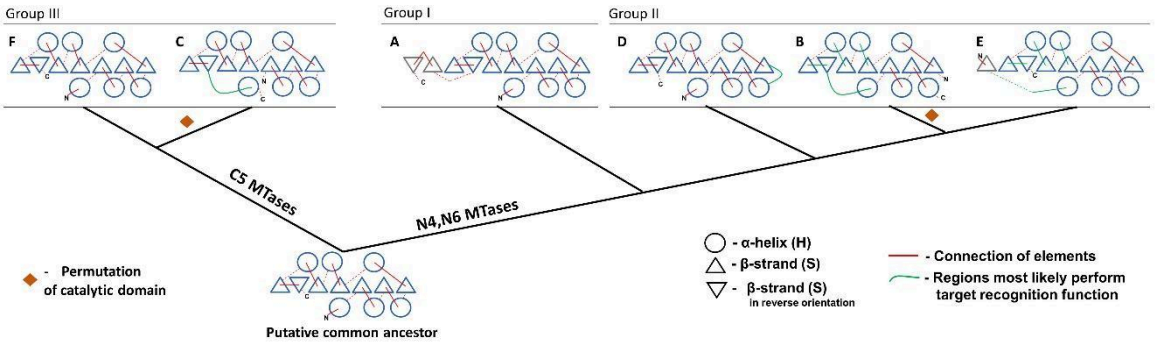


Figure 5. Presumable MTase evolution.

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 superclasses, 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-6g1yfq9ugw8.streamlit.app/.

SUPPLEMENTARY DATA

Supplementary Data are available https://drive.google.com/drive/folders/1A3dkZcIimKS23ro0VvFaTd2qtGeBFBfO?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. 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.

17. Gupta,Y.K., Chan,S.-H., Xu,S.-Y. and Aggarwal,A.K. (2015) Structural basis of asymmetric DNA methylation and ATP-triggered long-range diffusion by EcoP15I. *Nat. Commun.*, **6**, 1–10.