ISSN 0006-2979, Biochemistry (Moscow), 2025, Vol. 90, No. 4, pp. 502-512 © Pleiades Publishing, Ltd., 2025.
Published in Russian in Biokhimiya, 2025, Vol. 90, No. 4, pp. 559-570.

Restriction–Modification Systems Specific toward GGATC, GATGC, and GATGG. Part 1. Evolution and Ecology

Sergey Spirin[1,2,3,a]*, Ivan Rusinov[1], Olga Makarikova[4], Andrei Alexeevski[1,3], and Anna Karyagina[1,5,6]

[1]*Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, 119234 Moscow, Russia*

[2]*Higher School of Economics National Research University, 109028 Moscow, Russia*

[3]*NRC "Kurchatov Institute" - SRISA, 117218 Moscow, Russia*

[4]*Moscow Institute of Physics and Technology, 117303 Moscow, Russia*

[5]*Gamaleya National Research Center for Epidemiology and Microbiology, Ministry of Healthcare of the Russian Federation, 123098 Moscow, Russia*

[6]*All-Russia Research Institute of Agricultural Biotechnology, 127550 Moscow, Russia*

[a]*e-mail: sas@belozersky.msu.ru*

Received January 21, 2025
Revised March 19, 2025
Accepted March 25, 2025

Abstract—The article presents the results of studies on the evolution of proteins from restriction–modification systems consisting of restriction endonucleases with the REase_AlwI family domain and either two DNA methyltransferases, each with the MethyltransfD12 family domain, or a single DNA methyltransferase with two domains of this family. It was found that all such systems recognized one of the three DNA sequences, namely GGATC, GATGC or GATGG. Based on the sequence similarity, restriction endonucleases of these systems could be attributed to three clades that unambiguously corresponded to the RM system specificity. The DNA methyltransferase domains of these systems were classified into two groups based on sequence similarity, with the two domains of each system belonging to different groups. Within each group, the domains were attributed to three clades according to their specificity. An evidence of multiple interspecific horizontal transfer of entire restriction-modification systems has been found, as well as the transfer of individual genes between the systems (including the transfer of one of DNA methyltransferases accompanied by changes in its specificity). Evolutionary relationships of DNA methyltransferases from the studied systems with other DNA methyltransferases, including orphan DNA methyltransferases, have been revealed.

DOI: 10.1134/S0006297925600115

Keywords: restriction–modification system, molecular evolution, DNA methyltransferase, restriction endonuclease, horizontal gene transfer

INTRODUCTION

Restriction–modification (RM) systems are defence systems of prokaryotes that protect these organisms against introduction of foreign DNA, in particular DNA of bacteriophages [1]. RM systems are traditionally classified into several types [2], of which Type II is the most studied. Each Type II system contains genes coding for proteins with two enzymatic activities: a restriction endonuclease (REase) that recognizes and hydrolyzes a specific DNA sequence, and at least one DNA methyltransferase (MTase) that methylates host DNA within the target sequence, thus preventing its recognition by the REase. MTases methylate DNA either by cytosine bases to form C^5-methylcytosine (5mC) or N^4-methylcytosine (4mC), or by

* To whom correspondence should be addressed.

adenine bases to form N^6-methyladenine (6mA). Most Type II RM systems recognize palindromic sequences, although some of them (subtype IIA) have asymmetric recognition sites. Typically, subtype IIA systems contain not one, but two MTases. In the case when a subtype II RM system has only one MTase, this enzyme contains two catalytic centres and is a fusion of two MTase proteins. Two MTases are necessary to provide methylation of both DNA strands in the asymmetric site, which excludes the appearance of unmodified sites after replication [3, 4].

REases and MTases constituting RM systems can belong to different protein families (according to sequence homology). Earlier [5], we classified RM systems containing one REase and two MTases based on the catalytic domain families identified in them by the Pfam database tools [6] and investigated in detail the evolution of RM systems consisting of one REase with the NOV_C family domain and two 5mC MTases, each with the DNA_methylase family domain. We showed that these systems might have descended from a single ancestral system of the same composition. Horizontal gene transfer of entire systems has played a major role in their evolution. We also observed the evidence of relatively rare gene exchange events between the systems.

Here, we investigated the evolution of RM systems consisting of a REase with an RE_AlwI family domain and either two 6mA MTases, each with a MethyltransfD12 family domain, or a single fusion MTase with two MethyltransfD12 family domains. All systems with this domain composition, for which their specificity has been determined, belonged to subtype IIA and recognized one of the three DNA sequences: GGATC/GATCC, GATGC/GCATC or GATGG/CCATC. And *vice versa*, almost all RM systems with such confirmed specificities had this domain composition, with very few exceptions.

MATERIALS AND METHODS

The composition of the studied systems and the sequences of MTases and REases were extracted from REBASE [7], v.303 of 28.02.2023. The evolutionary domains in protein sequences were identified according to the Pfam database v.35 [6]. Protein sequences were aligned with Muscle [8]. Phylogenetic trees were inferred with FastME [9], rooted at the midpoint, and visualised with MEGA [10] or iTOL service [11]. Clustering of protein sequences was performed with CD-HIT [12].

The list of MTases with confirmed methylation sites was compiled from MTases satisfying the following two conditions: (i) their names in REBASE did not end with the letter ‘P’ (in REBASE, ‘P’ at the end of protein name means ‘putative’); (ii) the page for the MTase on the REBASE website stated that the nucleotide sequence methylated by the enzyme has been confirmed by the host genome sequencing using the PacBio technology.

MTases homologous to MTases from the studied RM systems were found among the sequences of MTases with confirmed methylation sites using BLASTP; the E-value threshold was 0.001. The search for REases homologous to the studied REases was performed in the same way among REases belonging to RM systems containing MTases with confirmed methylation sites.

The contrast, i.e., the ratio of the observed number of sites in the genome to the expected one, was calculated using the formula proposed by Burge et al. [13] (see also [14]).

RESULTS

In this work, we investigated RM systems containing REases of the Pfam family called ‘AlwI restriction endonuclease’ (Pfam ID RE_AlwI, Pfam AC PF09491) and either two MTases of the family ‘D12 class N^6 adenine-specific DNA methyltransferase’ (MethyltransfD12, PF02086) or one MTase with two MethyltransfD12 family domains. A total of 493 such systems with two MTases and 227 such systems with one MTase were found in REBASE v.303 (see Online Resource 1 for the list of identified systems). All these systems belonged to Type II. The genes for most of them were located on chromosomes; only 36 out of these 720 RM systems were encoded in plasmids. For some of the identified systems, REBASE indicated one of the three recognition sites: GGATC (or GATCC on the other chain), GATGC (GCATC), or GATGG (CCATC). Out of 97 RM systems for which these sites were confirmed with PacBio, 91 had the domain composition listed above (a REase with the RE_AlwI family domain and two MethyltransfD12 family domains in one or two MTases). Among the six exceptions (see Online Resource 2), three RM systems consisted of one REase with the RE_AlwI domain and one MTase with the MethyltransfD12 domain (according to REBASE), but next to the genes coding for these proteins there was a gene coding for an MTase with the MethyltransfD12 domain, so we assumed that the composition of these RM systems in REBASE was defined incorrectly. In the remaining three cases, some of the three typical domains was not detected in proteins of RM systems.

Phylogeny of REases and MTases. The phylogenetic tree inferred from the full-length REase sequences contained three clades, and REases with the same annotated specificity always belonged to the same clade. The phylogenetic tree presented in Fig. 1a

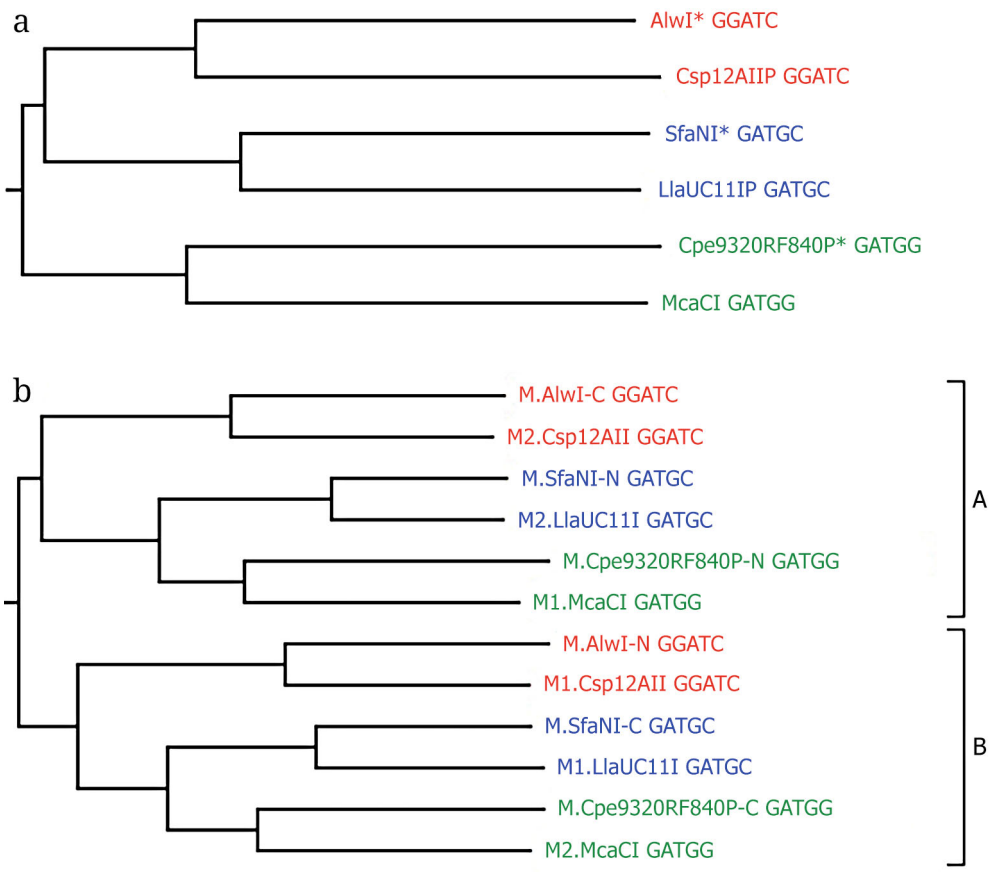

Fig. 1. Phylogenetic trees for REs and MTases from three RM systems with two MTases and three systems with a single (fusion) MTase. The recognition sequence (according to REBASE) is shown after the system name. a) Phylogenetic tree of REases. Enzymes from RM systems with the fusion MTase are marked with asterisks. b) Phylogenetic tree based on the MTase domain sequences. Domains from individual MTases are named according to the corresponding MTases from REBASE; 'N' and 'C' designate N- and C-terminal domains, respectively, of the fusion MTases; A and B denote two MTase groups.

for six REases illustrates a general trend, so we were able to predict with a high reliability the specificity of all systems studied in our work.

Below, we refer to RM systems specific toward GGATC/GATCC, GATGC/GCATC, and GATGG/CCATC as 'red', 'blue', and 'green', respectively (Fig. 1). Only the GGATC, GATGC, and GATGG variants were used to denote the recognition sites; the choice between the two complementary variants was determined by the specificity of the two MTases predicted by us [see Part 2. Functionality and Structure, Biochemistry (Moscow), vol. 90, issue 4]. It should be noted that REases from RM systems with the two-domain (fusion) MTase (marked with asterisks in Fig. 1a) or two individual MTases were not distinguished on the tree.

To study the evolution of MTases, we aligned separately N- and C-terminal domains from the fusion MTases to the MTase domains from RM systems with two MTases. The phylogenetic trees inferred from the resulting alignments showed the two MTases from the same RM system have evolved independently. Thus, the tree of the MTase domains branched into two clades, and for each of the studied systems, the two MTase domains of same system always belonged to different clades. Figure 1b illustrates this observation for the MTases of the same six RM systems. Henceforward, we will refer to MTases of the upper and lower clades (Fig. 1b) as group A and B MTases, respectively.

Both A and B groups of MTases in Fig. 1 were subdivided into three clades based on the enzyme specificity. The phylogenetic tree for all MTases with the confirmed specificity mostly showed the same pattern, with one exception in group A (see "Gene recombination between RM systems" below).

Mutual arrangement of genes in RM systems. Table 1 contains data on the mutual arrangement of genes for all studied RM systems with the available relevant information in REBASE. In all the cases, the genes for A and B group MTases were located on the same DNA strand, although their order could be different. The REase gene was usually located on the same strand as the MTase genes, but with some exceptions.

Table 1 demonstrates that the 'red' and 'green' RM systems possessed a typical mutual genes

Table 1. Mutual arrangement of genes for RM systems recognizing GGATC, GATGC, and GATGG sequences

Gene order*	GATGG	GATGC	GGATC	GATGG	GATGC	GGATC
	Number of systems†			Number of REase clusters#		
ABR	**338**/0	**50**/**147**	0/0	**205**/0	**30**/**42**	0/0
ABr	0/0	0/7	0/0	0/0	0/5	0/0
BAR	0/0	0/0	7/**68**	0/0	0/0	6/**45**
BAr	0/0	0/0	0/1	0/0	0/0	0/1
BRA	2/–	7/–	0/–	2/–	3/–	0/–
BrA	0/–	1/–	0/–	0/–	1/–	0/–
RAB	1/1	13/3	0/0	1/1	6/2	0/0
rAB	0/0	**74**/1	0/0	0/0	**54**/1	0/0
Total	341/1	145/158	7/69	208/1	94/50	6/46

* A and B denote MTases or MTase domains (in two-domain MTases) from groups A and B, respectively; R and r denote REase genes with the same or opposite orientation, respectively, as the MTase genes.

† The numbers before and after the slash sign correspond to the numbers of systems or clusters with separate or fusion MTases, respectively; dash after the slash indicates that location of the REase gene between the MTase genes is impossible for the systems with fusion MTases. The most typical variants are highlighted in bold.

Number of clusters with 98% identity per 98% length of the shorter sequence for REases from RM systems with the corresponding gene arrangement.

arrangement characteristic for the system with each particular specificity. The typical sequence for the 'green" systems was (group A MTase) → (group B MTase) → (REase), with only four exceptions. Only one 'green' system had a fusion MTase. The typical gene sequence for 'red' systems was (group B MTase) → (group A MTase) → (REase), with only one exception. Most 'red' systems had fusion MTases. In contrast, no dominant location of the REase gene was found for the 'blue' systems, although genes for group A MTases often preceded genes for group B MTases (as in 'green' systems) and there were about as many 'blue' systems with the fusion MTases as with two individual enzymes. Analysis of REase phylogeny in the 'blue' RM systems (Fig. 2) showed that during the evolution of this group, fusion and separation of two MTases genes, as well as rearrangement of MTase and REase genes, have occurred repeatedly. The most parsimonious, and thus most probable, scenario is as follows. The order of the ancestral genes in the 'blue' systems was ABR, but during the evolution, the transfer of the REase gene has occurred several times with the formation of the RAB or rAB arrangements. The BRA arrangement was observed only in 'blue' systems that contained MTase A which was closer to the 'green' MTase A (see "Gene recombination between RM systems" below).

Distribution of RM systems across bacterial taxa. The studied RM systems were found to be represented in 11 phyla and 21 classes of bacteria, although their distribution across these taxa was very uneven (see Online Resource 1). Closely related RM systems were found in bacteria from different phyla, while poorly related systems were present in closely related bacteria (see phylogenetic trees for REases in Online Resource 3). For example, among 70% REase clusters with the GATGC specificity, there was a cluster that included RM systems from 14 bacterial species from six different phyla. For one of these species (the livestock pathogen *Mannheimia haemolytica*), 62 RM systems from different strains were represented in REBASE, while the remaining 13 species possessed one RM system each. For seven of these 13 RM systems, REase sequences demonstrated less than 10% difference between each other and with sequences from *M. haemolytica*. In the same cluster, two RM systems, whose proteins showed over 98% sequence similarity, have been found in unrelated human oral microflora bacteria, *Streptococcus oralis* (Bacillota) and *Fusobacterium nucleatum* (Fusobacteriota).

The distribution of RM systems across the strains within species with a large number of different strains with fully sequenced genomes (see Online Resource 4) demonstrated that for about a half of the species for which genomes of five or more strains had been completely sequenced (i.e., assembled up to whole chromosomes), the RM system with one of the studied

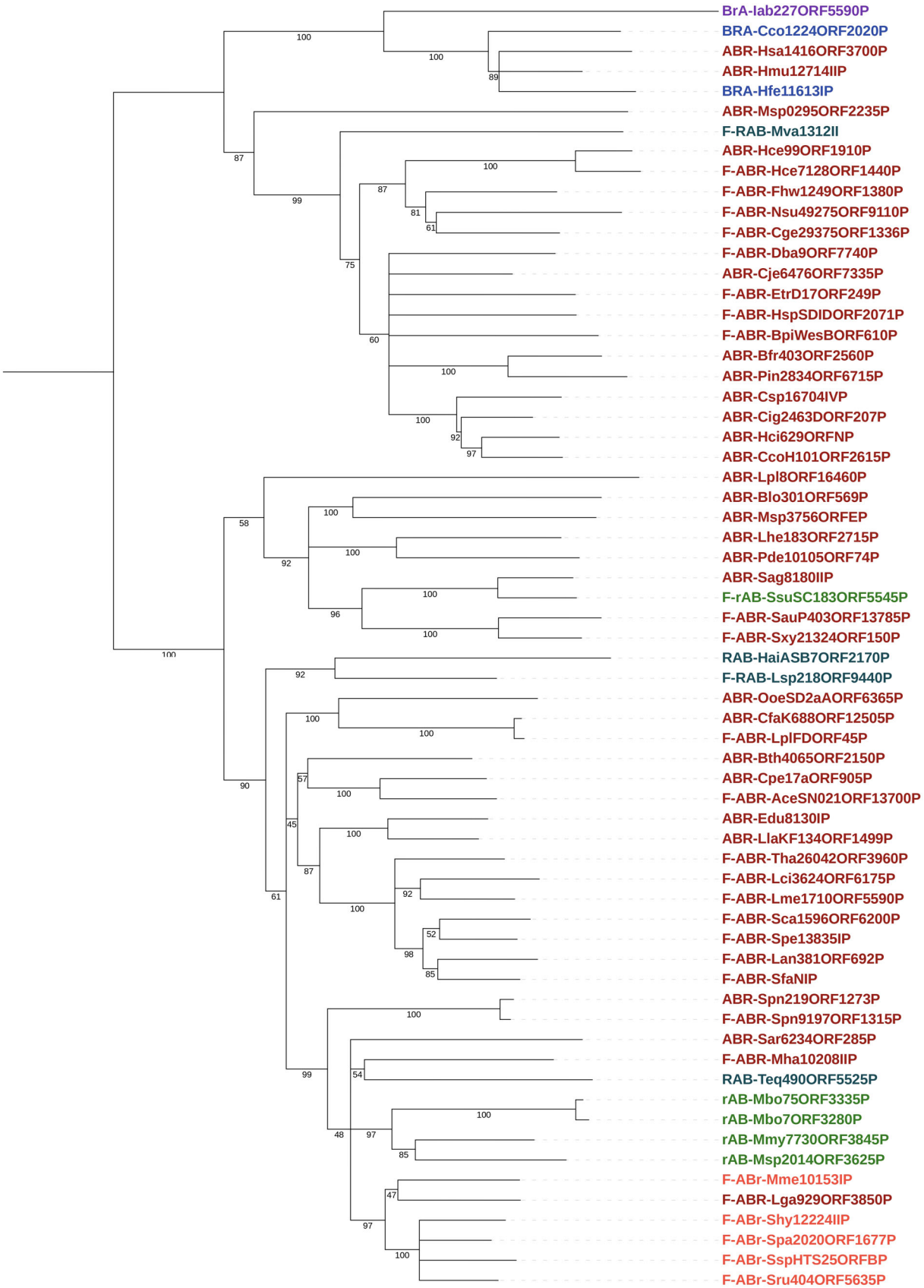

Fig. 2. Phylogenetic tree of REases from RM systems with the GATGC specificity (with the gene order). Cluster representatives were selected based on 70% sequence identity. Letter F at the beginning of the name indicates fusion MTase; gene order is shown with three letters (similar to the Table 1). The numbers on the branches indicate the bootstrap support.

specificities was found in the genome of only one strain. Among genomes of 274 *Streptococcus pneumoniae* strains, RM systems with the GCATC specificity were found in 15 genomes (5.5%). For only ten species, such systems were found in half or more of the strains, and in all these species except one (*M. haemolytica*), RM systems from different strains were highly similar to each other (>85% identity of complete protein sequence). As for *M. haemolytica*, this was true for all but one strain, which contained the Mha10208II system relatively unrelated to the others (25 and 59% identity of the REase and the fusion MTase, respectively, with the corresponding enzymes from other strains).

Avoidance of RM system recognition sites in genomes. A prolonged presence of an active RM system often leads to the avoidance of its recognition site in the host genome [15, 16]. Three bacterial species (*Bacteroides caccae*, *Campylobacter upsaliensis*, *Kingella kingae*) containing RM systems specific toward GATGG and one species (*Helicobacter cinaedi*) with the GATGC-specific system demonstrated a significant underrepresentation of the corresponding recognition sites in their genomes, i.e., the ratio of the observed number of sites to the expected one was less than 0.8. The underrepresentation of these sites was found in the genomes of all strains of these species, although in the case of *K. kingae* and *H. cinaedi*, the genes for the RM system were found in less than a half of the strains. In particular, genes for RM systems with the GATGC specificity were identified in three out of 12 genomes of *H. cinaedi* strains, while other two genomes contained genes for RM systems with the GATGG specificity. No avoidance of GATGG was observed in any *H. cinaedi* genome. An example of the opposite situation is *Campylobacter concisus*, in which two (out of 16) strains carried genes for the GATGC-specific RM systems, while the genes for the GATGG-specific systems were absent in all the strains; at the same time all 16 genomes demonstrated a strong underrepresentation of GATGG, but not significant deviations of the GATGC frequency from the expected one. The average contrasts across all genomes of each species (i.e., the ratios of observed word counts to the expected ones) for the words GATGC, GATGG, and GGATC are available in the Supplementary materials (Online Resource 4; columns G, H, and I).

Gene recombination between RM systems (horizontal transfers of RM system components). We found evidence of horizontal gene transfer between the RM systems with different specificities. In particular, we identified several related RM systems in which REases and MTases B were similar to the corresponding enzymes in the RM systems with the GATGC specificity, but their MTases A were more similar to MTases A of the RM systems with the GATGG specificity. Among others, such RM systems included Cup11541IV, Hfe11613I, and Hmu12714II, whose specificity toward GATGC has been confirmed with PacBio. It is likely that the ancestor of these RM systems had once acquired an 'alien' group A MTase, which has changed its specificity from GATGG to GATGC. This change in the specificity could occur either simultaneously with a mutation in the MTase gene (which led to changes in the recognition DNA sequence) or through a temporary weakening of the MTase specificity, for example, to GATGS. The molecular basis of changes in the enzyme specificity is discussed in Part 2. Functionality and Structure, Biochemistry (Moscow), vol. 90, issue 4.

We also found evidence of repeated exchange of components within each of the three groups of RM systems with different specificity. For example, for the three 'blue' RM systems (Teq529ORF1000P, Mha10208II, and Mva1312II), the REs from the first two systems were similar to each other (the score of the local alignment constructed by the *water* program of the EMBOSS package with default parameters was 1502) but significantly less similar to the REase from the third system (scores of 464.5 and 501, respectively). MTases from the first and third systems were close to each other (score, 1850), whereas MTase M.Mha10208II was more distant from them (scores, 1512 and 1574). This favours the suggestion that the ancestor of one of these systems was formed from REase and MTase genes that had originated from different RM systems with the same specificity (GATGC). Several more similar examples can be given for RM systems with each of the three specificities. However, a detailed analysis of the frequency of such events was obstructed by the poor quality of the MTase phylogeny reconstruction: the bootstrap support for many branches of the MTase phylogenetic tree was below 50%.

Evolutionary relationships of RM systems specific toward GGATC, GATGC, and GATGG and RM systems with homologous MTases. It appeared interesting to study the evolution of systems specific toward GGATC, GATGC, and GATGG in the context of evolution of RM systems with other specificities and homologous MTases and REs. The phylogenetic trees of MTases with the recognition sites confirmed with PacBio and close to the studied group A and B MTases (see Materials and Methods) are shown in Fig. 3.

A large number of MTases specific toward GATC were related to group B MTases, but did not form a monophyletic group within them. Some of these MTases formed sister with 'red' group B MTases (with the GGATC specificity), while others, including the well-studied M.EcoKDam, occupied a more basal position. Special attention should be paid to M.EcoT4Dam. The specificity of M.EcoT4Dam toward GATC has not been confirmed by the PacBio sequencing, but we added it to our set because this enzyme

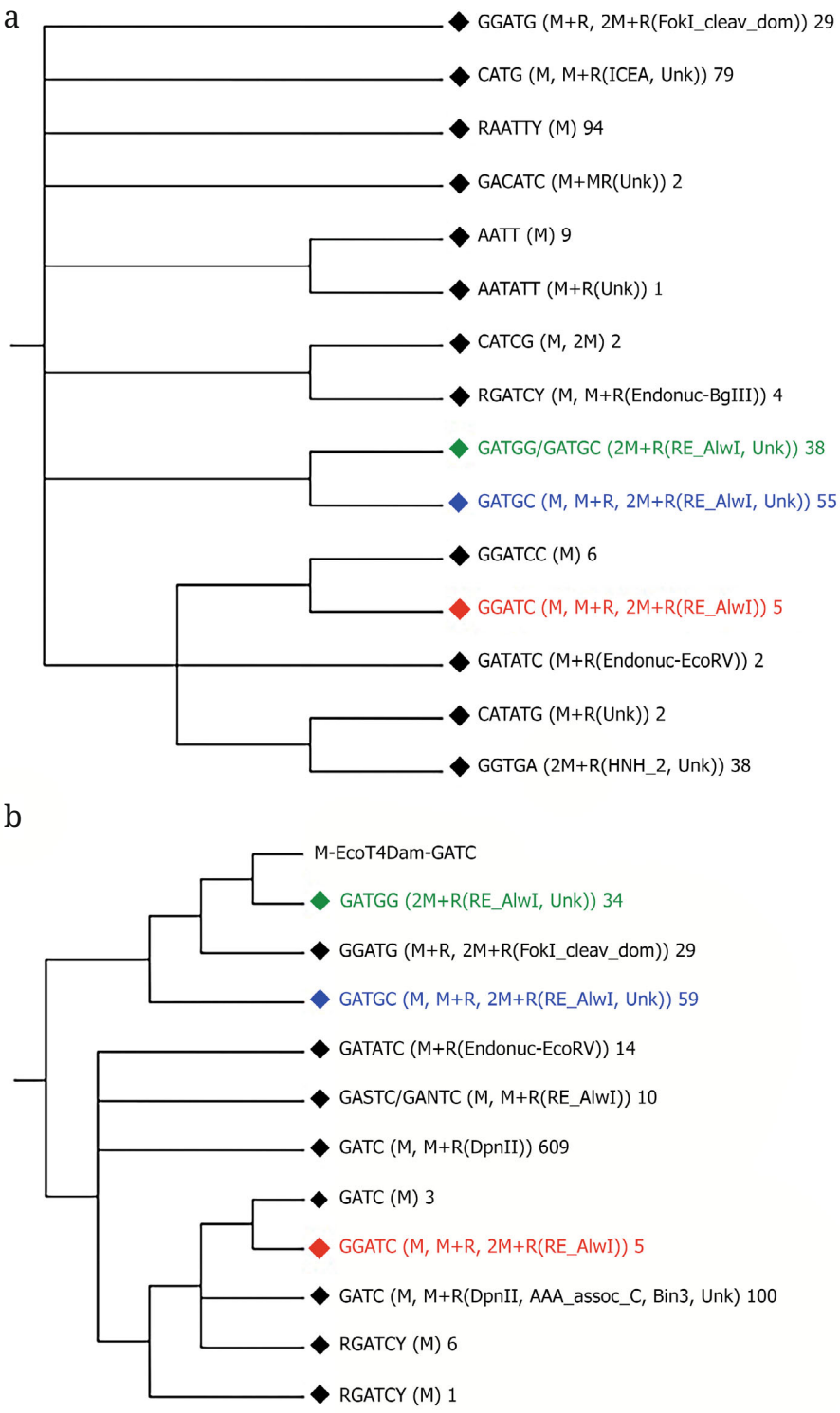

Fig. 3. Phylogenetic trees of group A (a) and group B (b) MTases of MethyltransfD12 family from the RM systems specific toward GGATC, GATGC, and GATGG and related enzymes. Branches with the bootstrap support less than 40% were removed. Monophyletic groups of MTases with the same specificity were combined and indicated with diamonds; the recognition site is given for each of these groups followed by parentheses with variants of RM system composition, where M is for MTase, 2M is for two MTases, R is for REase, and MR for fusion bifunctional protein containing MTase and REase domains. For RM systems with REases, the identifiers of catalytic domain families found in the REases are given in the second set of parentheses; Unk means that no domains were identified by Pfam profiles in the REase sequence. The number of systems for which the recognition site has been experimentally confirmed by the PacBio technology is given after the system composition. Group B MTase includes M.EcoT4Dam MTase, whose specificity toward GATC has not been confirmed by PacBio.

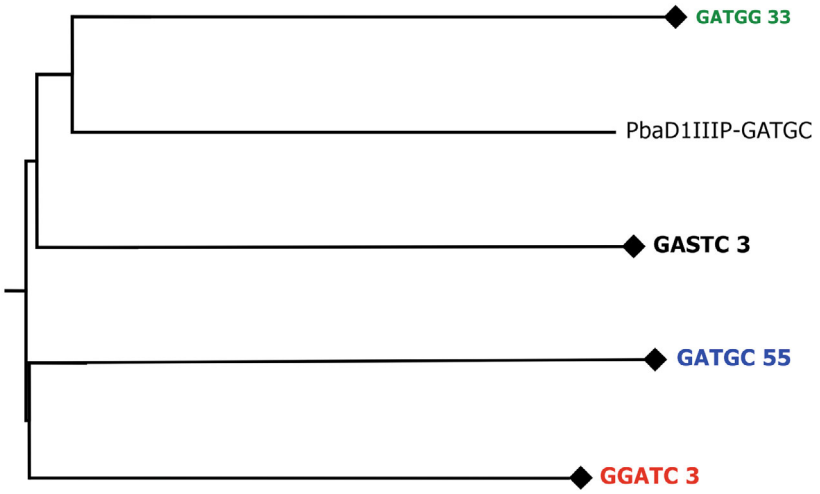

Fig. 4. Phylogenetic tree of REases close to REases specific toward GGATC, GATGC, and GATGG. The clades consisting of REases of the same specificity are indicated by diamonds; the recognition sequence is followed by the number of REases of this clade belonging to RM systems with MTases, whose specificity was confirmed by PacBio.

has been extensively studied (in particular, a number of structures of its complexes with DNA were solved). M.EcoT4Dam appeared to be closer to the 'green' and 'blue' group B MTases, whereas all MTases with the GATC specificity confirmed by PacBio were closer to the 'red' ones. A significant fraction of MTases with the GATC specificity that were close to M.EcoKDam [clade 'GATC (M, M+R(DpnII)) 609' in Fig. 3b] belonged to the resident Dam-MTases according to the terminology of [17], i.e., represented orphan (not included in the RM systems) MTases of Gammaproteobacteria with the vertical inheritance, as their phylogeny coincided with the phylogeny of the bacteria themselves.

The phylogenetic tree of REases close to REases from the studied RM systems is shown in Fig. 4. Beside the studied REases, it included only REases with the GASTC specificity. The REases specific towards all four sites were approximately equidistant from each other. The REase PbaD1IIIP, a member of the M.PbaD1III system with one MTase, whose specificity against GATGC was confirmed by PacBio-confirmed but which contained only one MethyltransfD12 family domain, was an exception that did not cluster with other REases of the same specificity.

DISCUSSION

REases and MTases with the same set of domains can have different specificity. The phylogenetic trees for REases and group B MTases (Fig. 3b) exhibited clearly distinguishable clades corresponding to the three specificities (GGATC, GATGC, and GATGG). The phylogenetic tree for group A MTases had a small fraction of 'blue' MTases that formed a clade within 'green' MTases, but the remaining group A MTases were also well separated by the specificity (Fig. 3a). In the REase phylogenetic tree, the 'red', 'blue', and 'green' clades were almost equidistant from each other, while MTases with the GATGC specificity ('blue') were significantly closer to MTases with the GATGG specificity ('green') than to enzymes with the GGATC specificity ('red') (Figs. 1 and 3). Based on these data, as well as joint trees including MTases with other specificities (Fig. 3), we can conclude that the RM systems containing REases with the RE_AlwI family domain and MTases with two MethyltransfD12 family domains have appeared at least twice in the evolution.

MTases of RM systems specific toward GATGG and GATGC were found to be closely related to MTases of the FokI-like systems with the GGATG specificity. The phage MTase M.EcoT4Dam and related MTases of other phages with the GATC specificity have probably originated from an ancestor common with group B MTases of such systems. Meanwhile, other MTases with the GATC specificity, in particular, Dam MTases from various *E. coli* strains (including M.EcoKDam), were closer to group B MTases with the GGATC specificity. It is likely that the specificity toward GATC has been acquired by MTases more than once.

Gene order and fusion/separation of genes. The systems with two MTases and one (fusion) MTase were approximately equally distributed among RM systems with the GATGC specificity ('blue'). The overall tree for REases from these and other systems (Fig. 2) shows that the fusion or separation of the MTase genes have occurred at least 10 times. These RM systems were also characterised by a different location of the REase gene relative to the two MTase genes (Table 1 and Fig. 2). Beside the main ABR variant (see Table 1), six more rare variants of the mutual arrangement of the three genes were observed. It should be noted that

the rAB variant, which was represented in 75 systems and 55 clusters based on 98% similarity of REase sequences, was found almost exclusively in *Mycobacterium bovis* strains. Systems with the rAB arrangement formed a single clade on the tree (Fig. 2). In this regard, the rAB order should be considered less common than the main ABR order.

The situation was different for RM systems with the other two specificities. Thus, only one RM system with a fusion MTase, Cpe832ORF840, was found among the systems specific toward GATGG; all other RM systems contained two single-domain MTases. This suggests that the fusion MTases with this specificity are less functional (or not functional at all) for some reason, so the fusion event has not been fixed in evolution. It is possible that some structural features of recognition of the GATGG site prevent proper binding of fusion MTases to DNA. In contrast, most RM systems with the GGATC specificity contained fusion MTases, except only seven systems that had separate MTases (see Table 1). Also, a non-standard REase gene location was less common in 'red' and 'green' systems than in 'blue' systems. Therefore, the gene mobility within an RM system is more common in the systems with the GATGC specificity. This is consistent with the fact that some RM systems with this specificity ('blue') contained group A MTases apparently 'borrowed' from RM systems with a different specificity ('green').

The evolutionary advantage of fusion of two MTases is not obvious. In a standard situation, the substrate of these enzymes is a half-methylated DNA formed after replication, so that only one of the two MTases acts at each site. It is likely that a fusion MTase in a 'red' or 'blue' system is equally efficient compared to separate enzymes and has been fixed or lost in the neutral evolution.

Horizontal transfer and lifetime of RM systems in bacteria. Comparison of the REase phylogenetic tree with the taxonomy of host bacteria showed that some groups of these RM systems have been transferred from genome to genome quite often, including transfer between phylogenetically distant bacteria, and have been lost rather quickly. For example, the Kki66ORF4915P, Kki10529IP, Hpa4058ORF9600P, and Aur25976ORF26P systems (GATGG specificity) were very close to each other (>85% identity between REases); however, they were present in bacteria of two different classes: the first two systems were found in two strains of *K. kingae* from the class Betaproteobacteria, Hpa4058ORF9600P – in *Haemophilus parainfluenzae* from the class Gammaproteobacteria, and Aur25976ORF26P – in *Actinobacillus ureae*, also from the class Gammaproteobacteria, while no other similarly related RM systems are represented in REBASE, and the RefSeq protein database contains only one other REase with the same level of similarity, which is from *Pasteurella multocida* (Gammaproteobacteria). All four listed bacteria are human pathogens. Of the remaining RM systems, the three closest ones (over 40% REase sequence identity) belonged to three different bacterial phyla. This suggests that such systems are very easily transferable, including between unrelated bacteria, but the duration of their existence in the host is relatively short.

Another group (specific toward GATGC) included 16 RM systems from bacteria belonging to 14 different families and eight different classes; at the same time, 62 identical systems from different *M. haemolytica* strains belonged to the same group. The sequence identity of REs from different systems in this group was at least 75%. Analysis of fully assembled genomes of *M. haemolytica* strains showed that these systems were present in about half of the strains. Apparently, in *M. haemolytica*, such RM systems have become essential for some reason, whereas in other hosts, RM systems of this group are easily acquired and rapidly lost. It is interesting to note that all organisms hosting RM systems of this group, except *M. haemolytica*, belonged to the human microflora, pathogenic or normal, which may be related to the frequency of RM system exchange between them.

No other such closely related groups of RM systems from equally diverse bacteria have been found, but the fact that RM systems are generally present in a very small fraction of strains of each species also indirectly indicates relatively frequent acquisition and loss of RM systems.

Besides *M. haemolytica*, several other bacterial species contained studied RM systems in a significant proportion of strains. It is likely that in these species, such RM systems also fulfil essential functions, for example, a role in maintaining subspecies identity [18].

The lifetime of an RM system can be indicated by the avoidance of its recognition site in the genome [15, 16]. For example, the avoidance of the GATGC site and the absence of avoidance of the GATGG site in the genomes of all *H. cinaedi* strains (see Online Resource 4) may indicate that the systems specific toward GATGC had been acquired by this species a long time ago but have been then lost by most strains, whereas RM systems with the GATGG specificity have been acquired relatively recently. A strong underrepresentation of GATGG in the genomes of *C. concisus* strains may indicate a long-term presence of RM systems with this specificity in this bacterial species, although no such RM systems have been detected in the sequenced genomes of this species.

Evolution of RM system specificity. The phylogenetic tree of MTases specific toward GGATC, GATGG, and GATGC and closely related MTases (Fig. 3) suggests a rather complex pattern of evolution of the system specificity. It is likely that RM systems

with the GGATC specificity have been assembled from two MTases and a REase independently of systems with the GATGG and GATGC specificities. We can assume a parallel evolution of the two MTases and REases in most lineages of the 'blue' and 'green' systems. In other words, all these systems had a common ancestor, and no exchange of components have occurred in their evolution, with one exception. Systems with the GGATG specificity and the FokI_cleav_dom domains in REases have probably acquired their MTases from an ancestor common with the present-day RM systems specific toward GATGG and GATGC.

The assumption made in [19] about a possible origin of two MTases of the M.FokI family (GGATG specificity) by duplication of the gene of the ancestral MTase with the SSATSS recognition sequence does not fit well with the fact that the two MTases of such systems are less related to each other than each of them is to MTases with other specificities (Fig. 3). Most likely, MTases of this family have originated from MTases with other specificities and in the course of evolution, have undergone a change in the recognition sequence, which must have occurred in parallel in two MTases that methylate different DNA strands. Such change was possible if the genome contained an RM system with a broader specificity or through the weakening of the system own specificity. The same is true for MTases specific toward GATGG ('green') and GATGC ('blue'), which have a relatively recent common ancestor with M.FokI-like MTases. As for REases, they might have been acquired independently by RM systems with different specificities, since FokI REase contains a catalytic domain of a different family, and REases of the systems specific toward GATGG and GATGC, although homologous, are significantly less related than MTases of the same systems: the sequence similarity between them is about the same level as with the REases recognising GGATC (Figs. 1a and 4).

The evolution of MTases of all these systems is closely related to the evolution of Dam MTases. It is possible that all group B MTases have originated from orphan MTases that recognized GATC. In this case, the phage MTase M.EcoT4Dam and related phage MTases that also recognize the GATC site, appear to have evolved from an ancestor of group B MTases with the GATGC specificity and have undergone a reversal of specificity. This scenario is supported by the differences in the mechanisms by which M.EcoKDam and M.EcoT4Dam MTases recognize the same GATC site [20].

CONCLUSION

All REs of RM systems specific toward GGATC, GATGC, and GATGG are homologous to each other. The same is true for MTases of these systems, although some of RM systems include one fusion MTase with two catalytic domains, while others contain two MTases. The N- and C-terminal parts of the fusion MTases were compared with each other and with the sequences of separate MTases. Based on the sequence similarity, MTases were classified into two rather distant groups. Two separate MTases of the same system and two parts of fusion MTases always belong to two different groups. Analysis of phylogenetic trees inferred from REases and from each of the MTase groups showed that the fusion and separation of MTases have occurred repeatedly in the evolution.

We found the evidence of rearrangements of the gene structure of RM systems, as well as horizontal transfer of entire RM systems and individual genes to other RM systems. In particular, a group of RM systems contained MTases that have probably originated from an MTase with a different recognition specificity.

Most of the studied RM systems are represented in a very small fraction of strains of the corresponding bacterial species, indicating a high probability of their loss by their hosts. At the same time, a few groups of these RM systems might have become essential to their hosts because they are represented in the majority of strains.

Abbreviations. MTase, DNA methyltransferase; REase, restriction endonuclease; RM system, restriction–modification system.

Supplementary information. The online version contains supplementary material available at https://doi.org/10.1134/S0006297925600115.

Acknowledgements. The authors would like to thank Dr. A. V. Grishin for his help in preparing the text.

Contributions. S. Spirin, A. Alexeevski, and A. Karyagina developed the concept and supervised the study; S. Spirin, I. Rusinov, O. Makarikova, and A. Karyagina prepared the data, developed program support, and analyzed the results; S. Spirin and A. Karyagina wrote and edited the manuscript.

Funding. This work was supported by the Russian Science Foundation (project no. 21-14-00135).

Ethics approval and consent to participate. This work does not contain any studies involving human or animal subjects.

Conflict of interest. The authors of this work declare that they have no conflicts of interest.

REFERENCES

1. Williams, R. J. (2003) Restriction endonucleases: classification, properties, and applications, *Mol. Biotechnol.*, **23**, 225-244, https://doi.org/10.1385/mb:23:3:225.

2. Roberts, R. J. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes, *Nucleic Acids Res.*, **31**, 1805-1812, https://doi.org/10.1093/nar/gkg274.
3. Madhusoodanan, U. K., and Rao, D. N. (2010) Diversity of DNA methyltransferases that recognize asymmetric target sequences, *Crit. Rev. Biochem. Mol. Biol.*, **45**, 125-145, https://doi.org/10.3109/10409231003628007.
4. Vasu, K., and Nagaraja, V. (2013) Diverse functions of restriction-modification systems in addition to cellular defense, *Microbiol. Mol. Biol. Rev.*, **77**, 53-72, https://doi.org/10.1128/mmbr.00044-12.
5. Fokina, A. S., Karyagina, A. S., Rusinov, I. S., Moshensky, D. M., Spirin, S. A., and Alexeevski, A. V. (2023) Evolution of restriction–modification systems consisting of one restriction endonuclease and two DNA methyltransferases, *Biochemistry (Moscow)*, **88**, 253-261, https://doi.org/10.1134/S0006297923020086.
6. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., and Bateman, A. (2020) Pfam: the protein families database in 2021, *Nucleic Acids Res.*, **49**, D412-D419, https://doi.org/10.1093/nar/gkaa913.
7. Roberts, R. J., Vincze, T., Posfai, J., and Macelis, D. (2014) REBASE – a database for DNA restriction and modification: enzymes, genes and genomes, *Nucleic Acids Res.*, **43**, D298-D299, https://doi.org/10.1093/nar/gku1046.
8. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.*, **32**, 1792-1797, https://doi.org/10.1093/nar/gkh340.
9. Lefort, V., Desper, R., and Gascuel, O. (2015) FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program, *Mol. Biol. Evol.*, **32**, 2798-2800, https://doi.org/10.1093/molbev/msv150.
10. Kumar, S., Stecher, G., and Tamura, K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets, *Mol. Biol. Evol.*, **33**, 1870-1874, https://doi.org/10.1093/molbev/msw054.
11. Letunic, I., and Bork, P. (2021) Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation, *Nucleic Acids Res.*, **49**, W293-W296, https://doi.org/10.1093/nar/gkab301.
12. Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22**, 1658-1659, https://doi.org/10.1093/bioinformatics/btl158.
13. Burge, C., Campbell, A. M., and Karlin, S. (1992) Over- and under-representation of short oligonucleotides in DNA sequences, *Proc. Natl. Acad. Sci. USA*, **89**, 1358-1362, https://doi.org/10.1073/pnas.89.4.1358.
14. Rusinov, I. S., Ershova, A. S., Karyagina, A. S., Spirin, S. A., and Alexeevski, A. V. (2018) Comparison of methods of detection of exceptional sequences in prokaryotic genomes, *Biochemistry (Moscow)*, **83**, 129-139, https://doi.org/10.1134/S0006297918020050.
15. Karlin, S., Burge, C., and Campbell, A. M. (1992) Statistical analyses of counts and distributions of restriction sites in DNA sequences, *Nucleic Acids Res.*, **20**, 1363-1370, https://doi.org/10.1093/nar/20.6.1363.
16. Rusinov, I., Ershova, A., Karyagina, A., Spirin, S., and Alexeevski, A. (2015) Lifespan of restriction-modification systems critically affects avoidance of their recognition sites in host genomes, *BMC Genomics*, **16**, 1084, https://doi.org/10.1186/s12864-015-2288-4.
17. Brézellec, P., Hoebeke, M., Hiet, M. S., Pasek, S., and Ferat, J. L. (2006) DomainSieve: a protein domain-based screen that led to the identification of dam-associated genes with potential link to DNA maintenance, *Bioinformatics*, **22**, 1935-1941, https://doi.org/10.1093/bioinformatics/btl336.
18. Murray, N. E. (2002) 2001 Fred Griffith review lecture. Immigration control of DNA in bacteria: self versus non-self, *Microbiology*, **148**, 3-20, https://doi.org/10.1099/00221287-148-1-3.
19. Friedrich, T., Fatemi, M., Gowhar, H., Leismann, O., and Jeltsch, A. (2000) Specificity of DNA binding and methylation by the M.FokI DNA methyltransferase, *Biochim. Biophys. Acta*, **1480**, 145-159, https://doi.org/10.1016/s0167-4838(00)00065-0.
20. Horton, J. R., Liebert, K., Bekes, M., Jeltsch, A., and Cheng, X. (2006) Structure and substrate recognition of the *Escherichia coli* DNA adenine methyltransferase, *J. Mol. Biol.*, **358**, 559-570, https://doi.org/10.1016/j.jmb.2006.02.028.