Genomics 93 (2009) 350–357

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Phylogenetic analysis of the eukaryotic RNA (cytosine-5)-methyltransferases

Athanasia Pavlopoulou [a,b], Sophia Kossida [a,*]

[a] Biomedical Research Foundation of the Academy of Athens, Department of Biotechnology, Bioinformatics and Medical Informatics Team, Soranou Efesiou 4, 11527 Athens, Greece
[b] Department of Pharmacy, School of Health Sciences, University of Patras, GR-26500 Rion-Patras, Greece

ARTICLE INFO

Article history:
Received 1 February 2008
Accepted 10 December 2008
Available online 23 January 2009

Keywords:
RNA methylation
RNA (cytosine-5)-methyltransferases
Eukaryotes
Phylogeny
Fmu
NOP2
NCL1
YNL022c
RCMT9

ABSTRACT

RNA (cytosine-5)-methyltransferases (RCMTs) have been characterized both in prokaryotic and eukaryotic organisms. The RCMT family, however, remains largely uncharacterized, as opposed to the family of DNA (cytosine-5)-methyltransferases which has been studied in depth. In the present study, an *in silico* identification of the putative 5-methylcytosine RNA-generating enzymes in the eukaryotic genomes was performed. A comprehensive phylogenetic analysis of the putative eukaryotic RCMT-related proteins has been performed in order to redefine subfamilies within the RCMT family. Five distinct eukaryotic subfamilies were identified, including the three already known (NOP2, NCL1 and YNL022c), one novel subfamily (RCMT9) and a fifth one which hitherto was considered to exist exclusively in prokaryotes (Fmu). The potential evolutionary relationships among the different eukaryotic RCMT subfamilies were also investigated. Furthermore, the results of this study add further support to a previous hypothesis that RCMTs represent evolutionary intermediates of RNA (uridine-5)-methyltransferases and DNA (cytosine-5)-methyltransferases.

Introduction

The addition of a group at the fifth carbon of a pyrimidine base is a common form of posttranslational RNA modification [1]. This RNA modification, which occurs mainly in rRNA and tRNA, has played a critical role in the RNA-to-DNA world transition [2]. The transfer of a methyl group from the *S*-adenosyl-L-methionine (AdoMet) to the fifth carbon of a cytosine base is specifically catalyzed by the enzyme family of RNA (cytosine-5)-methyltransferases (henceforth referred to as RCMTs). These enzymes are composed of a common catalytic domain which exhibits the Rossman-fold. RCMTs consist of eight signature sequence motifs, also present in the well-characterized DNA (cytosine-5)-methyltransferases (DNMTs). These eight motifs along with the highly conserved motif N1 compose the consensus core region. Auxiliary N-terminal and C-terminal extensions, variable both in size and sequence, are also present in the RCMTs [3].

Most of the enzymes that methylate the carbon-5 of pyrimidines, such as DNMTs, thymidylate synthase, and bacteriophage deoxycytidine hydroxymethylase, use an invariant cysteine residue in motif IV. However, in RCMTs an unrelated invariant cysteine residue within a ThrCys dipeptide in motif VI serves as the catalytic nucleophile [4], similar as in RNA (uridine-5)-methyltransferases (RUMTs) [5].

In brief, the catalytic mechanism of the RCMTs involves an attack by the thiolate of the motif VI Cys on the position 6 of the target cytosine base to form a covalent link, thereby activating carbon-5 for methyl group transfer. Following addition of the methyl group, motif IV Cys acts as a general base in the β-elimination of the proton from the methylated cytosine ring. The free enzyme is restored and the methylated product is released [6].

According to Bujnicki et al. [7], RCMTs share characteristics with both the RUMTs which originated in the "RNA world" and DNA:m^5C MTases which succeeded RUMTs in the "DNA world." Therefore, the RCMT family of enzymes represents an evolutionary intermediate of the two latter families of enzymes. RCMTs rather evolved from RUMTs by acquiring a second Cys residue in motif IV. This new Cys residue became the main catalyst in DNMTs and the original Cys in motif VI got lost [7].

RCMTs have been characterized in all three primary phylogenetic Domains of Life: Archaea, Eubacteria and Eukarya [8]. According to Reid et al. [8], RCMTs are divided into eight subfamilies (RCMT1-RCMT8), including five (RCMT1, RCMT3-6) which are strictly 'prokaryotic', i.e. without any identified eukaryotic orthologues and three strictly 'eukaryotic' subfamilies (RCMT2, RCMT7, RCMT8), which are named after the yeast founding members NOP2, NCL1 and YNL022c, respectively.

The purpose of this study was to identify *in silico* putative RCMT-related proteins in the eukaryotic genomes, to re-define eukaryotic RCMT subfamilies, to investigate the potential evolutionary relationship among these subfamilies, as well as to further test the hypothesis that RCMTs are evolutionary intermediates of RUMTs and DNMTs. The genomic organization of the *RCMT*-related genes was also investigated in this study.

* Corresponding author. Fax: +302106597545.
E-mail address: skossida@bioacademy.gr (S. Kossida).

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.12.004

Table 1
Phylogenetic distribution of the eukaryotic RCMTs

Taxonomy/Species	RCMT1 (Fmu)	RCMT2 (NOP2)	RCMT7 (NCL1)	RCMT8 (YNL022c)	RCMT9	Total RCMTs
1. Metazoa						
1.1. Chordata						
1.1.1. Mammalia						
Homo sapiens (Human)		1	1	1		3
Mus musculus (Mouse)		1	1	1		3
1.1.2. Aves						
Gallus gallus (Chicken)		1	1	1		3
1.1.3. Amphibia						
Xenopus tropicalis (Frog)		1	1	1		3
1.1.4. Actinopterygii						
Danio rerio (Zebrafish)		1	1	1		3
1.1.5. Cephalochordata						
Branchiostoma floridae (Lancelet)		1	1	1		3
1.1.6. Urochordata						
Ciona intestinalis		1	1			2
1.2. Nematoda						
Caenorhabditis elegans		1	1	1		3
1.3. Echinodermata						
Strongylocentrotus purpuratus (Sea urchin)		1		1		2
1.4. Cnidaria						
Nematostella vectensis (Sea anemone)		1	1	1		3
1.5.Annelida						
Helobdella robusta (Leech)		1	1	1		3
1.6 Mollusca						
Lottia gigantea (Snail)		1	1	1		3
1.7 Arthropoda						
Drosophila melanogaster		1	1	1		3
2. Viridiplantae						
Arabidopsis thaliana	1	2	2	1	1	7
Oryza sativa (Rice)	1	2	2	1	1	7
Zea mays (Maize)	1[t]	1, 1[t]	1[t]	1[t]	1	6
Sorghum bicolor					1	1
Chlamydomonas reinhardtii		1	1	1		3
3. Fungi						
3.1 Dikarya						
3.1.1.Ascomycota						
Saccharomyces cerevisiae		1	1	1		3
Schizosaccharomyces pombe		1	1	1		3
3.1.2 Basidiomycota						
Cryptococcus neoformans		1	1	1		3
3.2 Microsporidia						
Encephalitozoon cuniculi		1	1	1		3
3.3 Fungi incertae sedis						
Rhizopus oryzae		1	1	1		3
4. Alveolata						
Cryptosporidium parvum		1	1	1		3
Paramecium tetraurelia		1		1	1	3
Plasmodium falciparum		1		1		2
Tetrahymena thermophila			1	1	1	3
Theileria parva		1	1		1	3
Toxoplasma gondii			1			1
5. Choanoflagellidae						
Monosiga brevicollis		1	1[t]	1		3
6. Cryptophyta						
Guillardia theta		1				1
7. Diplomonadida						
Giardia lamblia		1	1	1		3
8. Entamoebidae						
Entamoeba histolytica		1	1	1		3
9. Euglenozoa						
Trypanosoma cruzi		1	1	1	1	4
10. Heterolobosea						
Naegleria gruberi		1	1	1		3
11. Mycetozoa						
Dictyostelium discoideum		1		1	1	3
12. Parabasalidae						
Trichomonas vaginalis		1	1			2
13. Stamenopiles						
Phytophthora sojae		1	1	1		3
Thalassiosira pseudonana		1	1	1		3
14. Rhodophyta						
Cyanidioschyzon merolae		1	1	1		3

t: truncated.

Methods

Putative RCMT homologues search

The full length RCMT amino acid sequences identified by Reid et al. [8] were used as queries to search the publicly available non-redundant protein and genomic databases using BLASTp and tBLASTn [9]. The newly identified protein sequences were aligned with the known RCMT sequences and a phylogenetic tree was constructed. The sequences that did not group with the known RCMTs but formed their own subtree, were false positives which were omitted in the subsequent steps of the analysis. The candidate protein sequences were used as input in a Pfam [10] search in order to predict domain organization. Sets of amino acid sequences located in the same chromosome of the organism were considered redundant and only one sequence was used in the phylogenetic analysis. Partial protein sequences with a deletion that spanned more than 25% of the entire sequence (Table 1, indicated by the letter "t") were not included in the phylogenetic analyses.

Phylogenetic analyses

Multiple alignment of the retrieved protein sequences was performed using CLUSTALW [11]. The final sequence alignment was used to perform phylogenetic analysis employing the distance-based

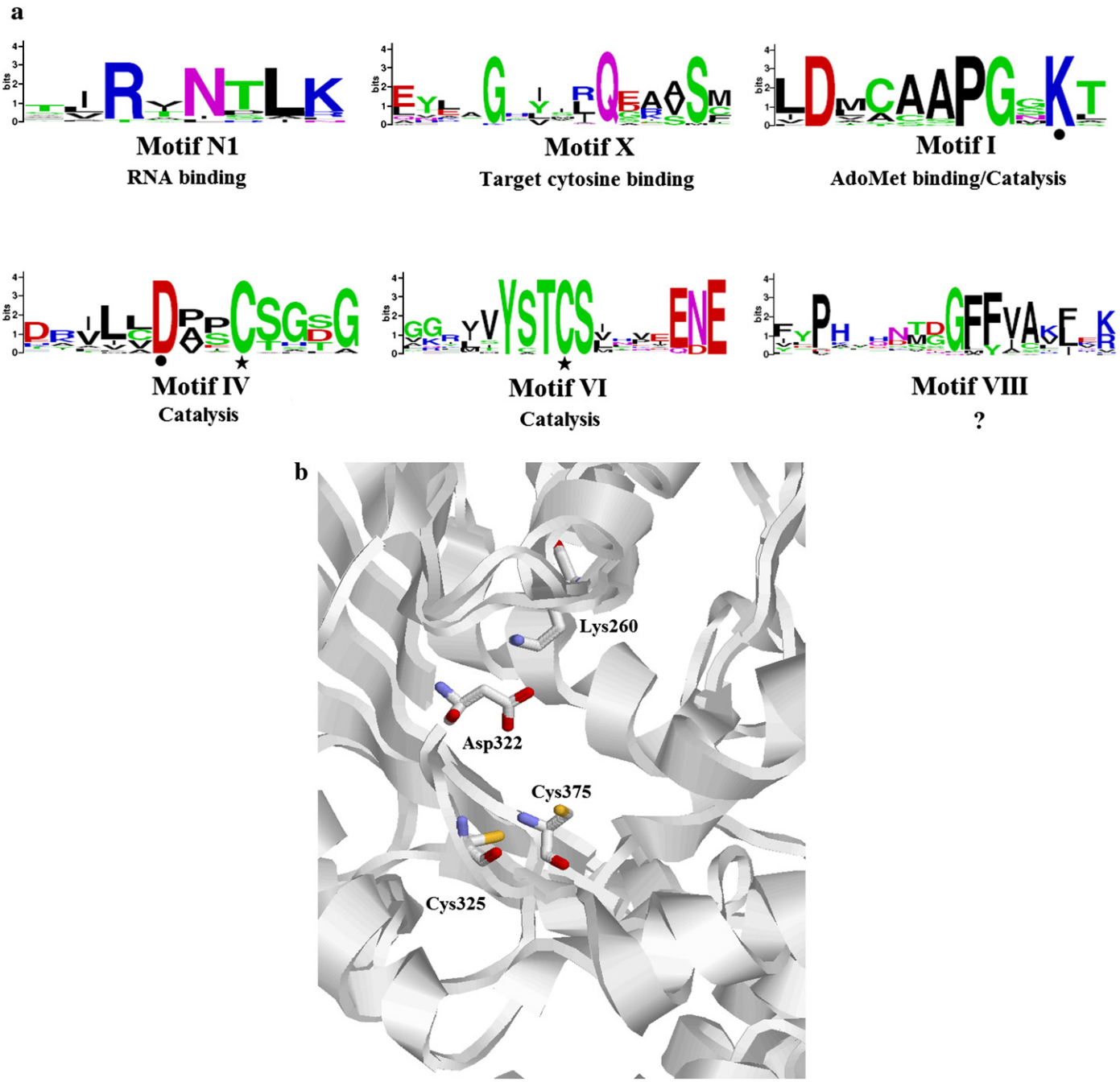


Fig. 1. (a) Sequence logo of the six prime motifs present in RCMTs. The name and the proposed function of each of the motifs are indicated below. The question mark indicates that a function has not been assigned to this motif. The conserved cysteine residues are represented by asterisks. The dots indicate the two invariant residues which are indispensable for 5-methylcytosine RNA methyltransferase activity. (b) Crystal structure of the *E.coli* Fmu. Sheets are represented as arrows and helices as spirals. The amino acid residues shown to be critical for methyltransferase activity are represented as wireframes and labeled. The two invariant cysteine residues are appropriately positioned so nucleophilic attack and β elimination take place on the same side of the cytosine ring. This figure was prepared using RasMol (http://www.umass.edu/micro/rasmol).

neighbor-joining method implemented in the MEGA 4 program [12] and the maximum-likelihood method implemented in the PHYML program [13]. PHYML is based on the nearest-neighbor-interchange heuristic and is an accurate method for reconstructing large phylogenies [13]. The JTT model was used to estimate the number of amino acid substitutions/site [14]. Bootstrap analysis (500 replicates) was performed to test the robustness of the inferred trees. Phylogenetic analysis was also performed using the Bayesian method implemented in MrBayes 3 under a mixed model of amino acid substitution [15]. MrBayes 3 was run for 1 million generations, and a 50% majority rule consensus tree was constructed.

Logo sequence analysis

Sequences of the six catalytic motifs of the eukaryotic RCMT-like proteins presented in Fig. 3 with less than 90% identity were used as input to logo [16]. The sequences were manually edited for insertions or gaps. The height of each letter is proportional to the frequency of the corresponding amino acid, and the letters are ordered so that the most frequent is on the top.

Results

The *in silico* identification and phylogenetic analysis of putative RCMT-like proteins (Table 1) in the eukaryotic genomes is presented below.

Sequence analysis

Putative eukaryotic RCMT-related proteins were initially identified by searching the publicly available databases for sequences possessing the six prime signature motifs present in RCMTs as described by Reid et al. [8]. This process was reiterated until no new putative proteins could be found, suggesting that a comprehensive representation of the RCMT family was achieved. A set of 301 putative RCMT-related proteins was obtained in this way. If any RCMT-related proteins were not detected, they are most likely remote homologues of the sequences analyzed in this study.

A logo sequence analysis was also generated in order to determine the consensus sequence of each of the six prime motifs (Figs. 1a and b). In motif N1, the conserved residue Arg (R) is positioned close to the AdoMet cofactor, suggesting a possible role for N1 in binding the RNA phosphate backbone [6]. Motif X possesses a conserved Gln (Q) which presumably makes contact with the target base. Motif I is traditionally thought to be involved in AdoMet-binding. However, according to crystal structure and homology modeling, the invariant residue Lys (K) was shown to be located in the active site raising the possibility that Motif I is involved in the catalysis, as well [7]. Motif IV contains the invariant and essential amino acid Cys which is totally conserved both in DNA:m^5C MTases and RNA:m^5U MTases [4]. In addition, the invariant acidic residue Asp(D) could possibly serve a function similar to that of Glu(E) in Motif VI of DNA:m^5C MTases in targeting cytosine binding [7]. Motif VI contains the second invariant cysteine residue which is the principal catalyst in RCMTs [4]. A function has not been assigned to Motif VIII yet. Further structural and biochemical studies are needed to elucidate the function of this prime motif.

Phylogenetic analyses reveal five subfamilies of eukaryotic RCMT-like proteins

The final multiple sequence alignment of the retrieved set of 301 sequences was used to construct a neighbor-joining phylogenetic tree (see Methods). Five highly resolved monophyletic branches without erratic groupings were distinguished which allowed the preliminary classification of the candidate proteins into five groups (Fig. 2). A sample of 122 RCMT protein sequences (Table 1, Fig.S1, Table S2) identified within the main eukaryotic taxonomic divisions (according to the NCBI taxonomy database [17] was chosen for further phylogenetic analysis by employing the maximum-likelihood (ML) and Bayesian analysis methods (see Methods). Both methods provided phylogenetic trees (Fig. 3 and data not shown) with similar topologies; the differences occurred mainly in the poorly supported regions of the trees. Four selected prokaryotic RCMT-related proteins that corres-pond to the subfamilies RCMT1,-3,-4 and -5 were also included in the phylogenetic analysis in order to gain an understanding of the evolutionary relationships between the eukaryotic proteins and their prokaryotic relatives (Fig. 3). The prokaryotic 5-methyluridine rRNA generating enzyme sequence, RumA [18], was selected as an outgroup based on the tentative phylogenetic model proposed by Bujnicki et al. [7] (Fig. 3).

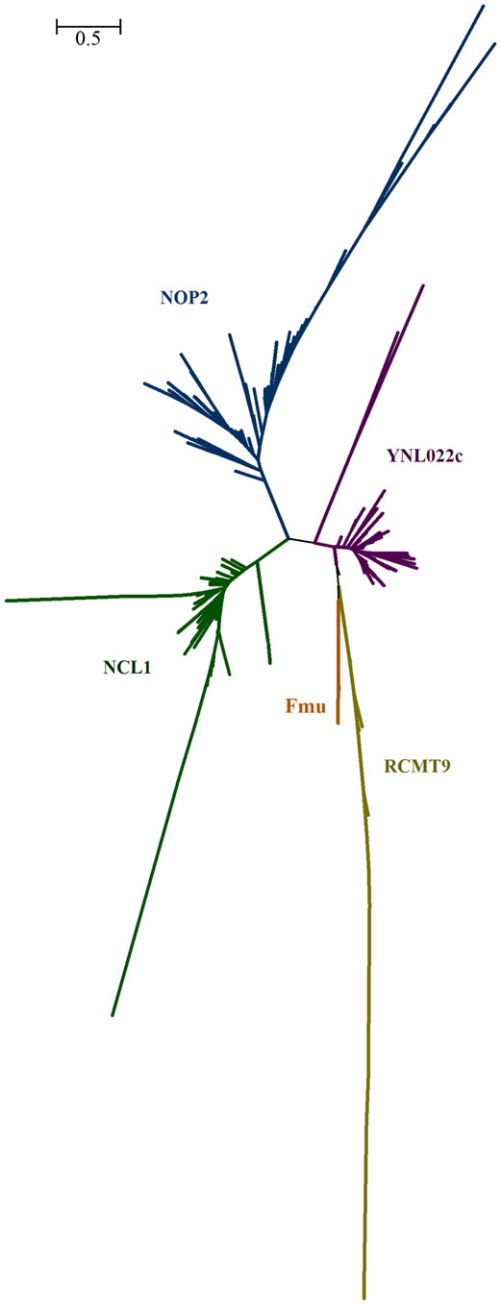
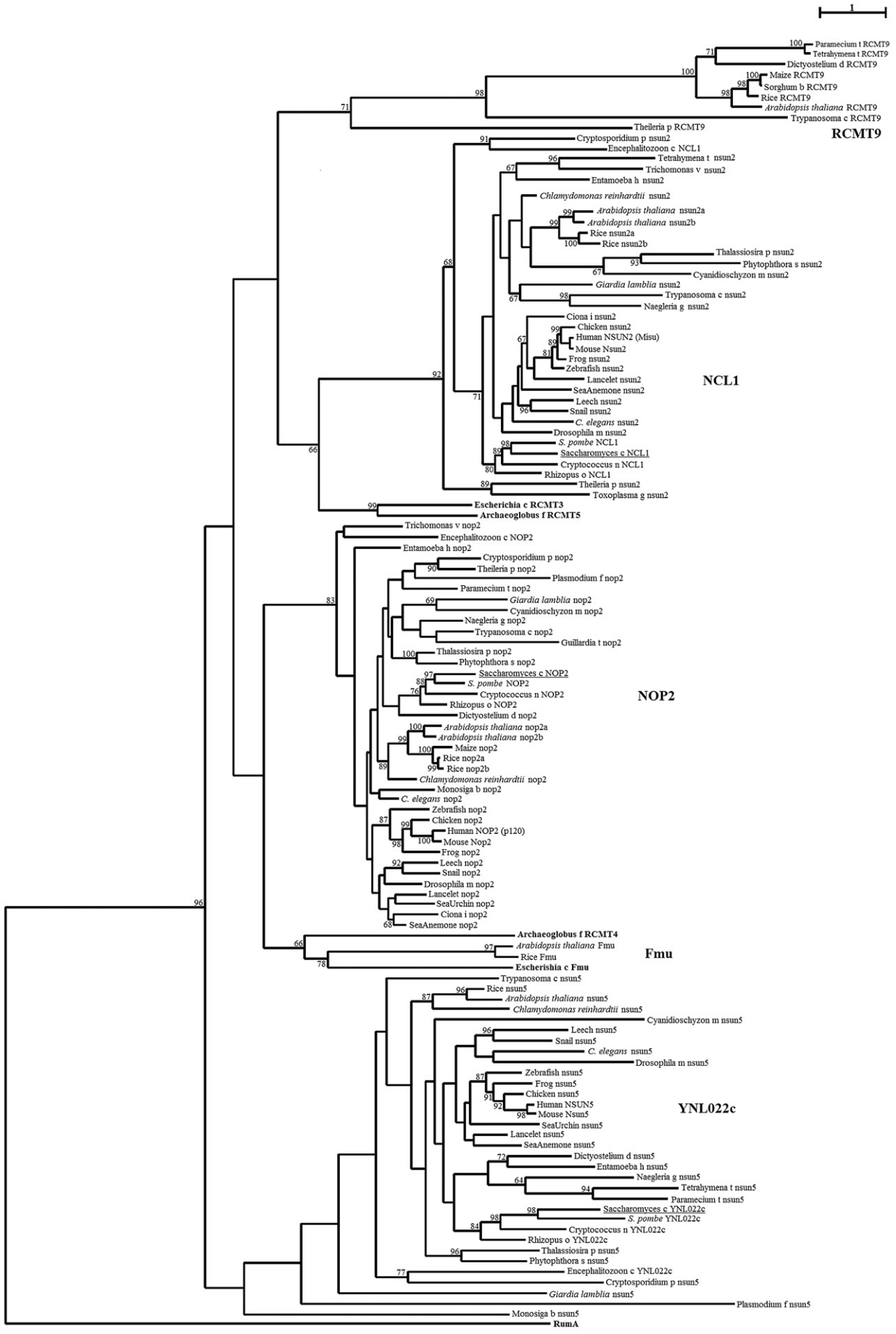


Fig. 2. Phylogenetic analysis of the eukaryotic RCMTs by employing the neighbor-joining method. The scale bar at the upper left indicates the length of amino acid substitutions/site.


1
Paramecium t RCMT9
Tetrahymena t RCMT9
Dictyostelium d RCMT9
Maize RCMT9
Sorghum b RCMT9
Rice RCMT9
Arabidopsis thaliana RCMT9
Trypanosoma c RCMT9
Theileria p RCMT9
RCMT9
Cryptosporidium p nsun2
Encephalitozoon c NCL1
Tetrahymena t nsun2
Trichomonas v nsun2
Entamoeba h nsun2
Chlamydomonas reinhardtii nsun2
Arabidopsis thaliana nsun2a
Arabidopsis thaliana nsun2b
Rice nsun2a
Rice nsun2b
Thalassiosira p nsun2
Phytophthora s nsun2
Cyanidioschyzon m nsun2
Giardia lamblia nsun2
Trypanosoma c nsun2
Naegleria g nsun2
Ciona i nsun2
Chicken nsun2
Human NSUN2 (Misu)
Mouse Nsun2
Frog nsun2
Zebrafish nsun2
Lancelet nsun2
SeaAnemone nsun2
Leech nsun2
Snail nsun2
C. elegans nsun2
Drosophila m nsun2
S. pombe NCL1
Saccharomyces c NCL1
Cryptococcus n NCL1
Rhizopus o NCL1
Theileria p nsun2
Toxoplasma g nsun2
NCL1
Escherichia c RCMT3
Archaeoglobus f RCMT5
Trichomonas v nop2
Encephalitozoon c NOP2
Entamoeba h nop2
Cryptosporidium p nop2
Theileria p nop2
Plasmodium f nop2
Paramecium t nop2
Giardia lamblia nop2
Cyanidioschyzon m nop2
Naegleria g nop2
Trypanosoma c nop2
Guillardia t nop2
Thalassiosira p nop2
Phytophthora s nop2
Saccharomyces c NOP2
S. pombe NOP2
Cryptococcus n NOP2
Rhizopus o NOP2
NOP2
Dictyostelium d nop2
Arabidopsis thaliana nop2a
Arabidopsis thaliana nop2b
Maize nop2
Rice nop2a
Rice nop2b
Chlamydomonas reinhardtii nop2
Monosiga b nop2
C. elegans nop2
Zebrafish nop2
Chicken nop2
Human NOP2 (p120)
Mouse Nop2
Frog nop2
Leech nop2
Snail nop2
Drosophila m nop2
Lancelet nop2
SeaUrchin nop2
Ciona i nop2
SeaAnemone nop2
Archaeoglobus f RCMT4
Arabidopsis thaliana Fmu
Rice Fmu
Escherishia c Fmu
Fmu
Trypanosoma c nsun5
Rice nsun5
Arabidopsis thaliana nsun5
Chlamydomonas reinhardtii nsun5
Cyanidioschyzon m nsun5
Leech nsun5
Snail nsun5
C. elegans nsun5
Drosophila m nsun5
Zebrafish nsun5
Frog nsun5
Chicken nsun5
Human NSUN5
Mouse Nsun5
SeaUrchin nsun5
Lancelet nsun5
SeaAnemone nsun5
YNL022c
Dictyostelium d nsun5
Entamoeba h nsun5
Naegleria g nsun5
Tetrahymena t nsun5
Paramecium t nsun5
Saccharomyces c YNL022c
S. pombe YNL022c
Cryptococcus n YNL022c
Rhizopus o YNL022c
Thalassiosira p nsun5
Phytophthora s nsun5
Encephalitozoon c YNL022c
Cryptosporidium p nsun5
Giardia lamblia nsun5
Plasmodium f nsun5
Monosiga b nsun5
RumA

The inferred maximum-likelihood-based phylogenetic tree presented in Fig. 3 is overall well-resolved. The low support values in some deep-branching nodes suggest alternative branching. Five monophyletic clades are distinguished that correspond to the five eukaryotic subfamily lineages. These sublineages are named after the founding members of this family; the novel subfamily was arbitrarily named RCMT9 (Fig. 3).

Two plant RCMT homologues were identified in this work which grouped (albeit moderately) with the Fmu subfamily (Fig. 3) which until to date was considered strictly 'prokaryotic'. The founding member of this sublineage, Fmu, catalyzes the formation of 5-methylcytosine at position 967 of 16S rRNA in *E. coli* [19]. The Fmu sublineage has a widespread distribution in Eubacteria. It would be intriguing to speculate that these two eukaryotic species acquired the RCMT proteins by prokaryotes through horizontal gene transfer. Indeed, Eubacteria (especially Gram-negative bacteria) are common plant pathogens [20].

NOP2 is the largest eukaryotic subfamily in size with a wide distribution among Eukaryotes (Fig. 3). The founding member of the NOP2 sublineage, the *S.cerevisiae* NOP2 and its human homologue, the proliferation-association nucleolar antigen p120, are both involved in pre-RNA processing and in large ribosomal subunit assembly [21,22]. The yeast NOP2 was also found to be essential for cell growth and viability [23]. In addition, p120 is over-expressed in a variety of malignant tumors and it is associated with decreased cell growth [24,25], providing an attractive target for therapeutic interventions. NOP2 members possess a C-terminal and a large N-terminal extension, the latter of which contains nuclear and nucleolar localization signals [26].

The third eukaryotic subfamily, NCL1, is also broadly distributed in Eukaryotes (Fig. 3). The yeast NCL1, the prototypic member of this sublineage appeared at the nuclear periphery [27], where the splicing and processing of tRNA takes place [28]. The *S.cerevisiae* NCL1 enzyme displayed "multisite-specificity", since it was found to catalyze the formation of 5-methylcytosine at four distinct sites in tRNA. Unlike *NOP2*, the yeast *NCL1* was not essential for cell viability [29]. Interestingly, the human NCL1 homologue, Misu, is a potential tumor marker since it is upregulated in tumors and mediates the effects of Myc (a proto-oncogene) on cell proliferation [30].

The members of the YNL022c subfamily displayed a variation in the conserved ProCys dipeptide in Motif IV (Fig. 1b, ProSerCys). The cytosine-5 RNA methyltransferase activity of YNL022c has not been biochemically validated. However, GFP localization patterns revealed that YNL022c was closely associated with the nuclear envelope throughout the cell cycle [31], suggesting a possible role for Ynl022c in tRNA processing. Nevertheless, the widespread distri-bution of YNL022c-like proteins across the major taxonomic groups (Fig. 3) indicates that there is a strong and selective pressure to retain this enzyme. Future biochemical and genetic studies should be directed towards elucidating the biological function of this subfamily.

A novel sublineage, RCMT9, was also identified that appears to be paralogous to the NCL1. Members of this family exhibit a variation in the conserved ProCys dipeptide in Motif IV (Fig. 1b, GluCys). RCMT9 is under-represented since it is restricted to a few phyla, including higher plants, Alveolata, Euglenozoa and Mycetozoa. Notably, representatives of this lineage were not identified in the genomes of Fungi and Metazoa.

Phylogenetic distribution of putative RCMT-related proteins in eukaryotes

A total of 122 eukaryotic RCMT-related proteins were identified in the eukaryotic organisms listed in Table 1. A broad phylogenetic distribution of putative RCMT-related proteins was observed across the major eukaryotic taxa. RCMTs were identified in organisms which were shown to be DNMT-negative in a previous study by Ponger and Li [32]. In particular, RCMTs were identified in the Microsporidian *Encephalitozoon cuniculi* (3), *C.elegans* (3), Diplomonadida (3) and Cryptophyta (1) and even the basal eukaryote *Giardia lamblia* [33] (Table1). A greater number of RCMTs, compared to DNMTs, was detected in several eukaryotic organisms and in particular in species evolving at notoriously slower rates, such as Fungi (except Microsporidia), Animals, Stamenopiles and Ciliophora (*Tetrahymena thermophila*, *Paramecium tetraurelia*) [34] (Table 1). Of particular interest is the finding that RCMT-like proteins were detected in the parasitic eukaryotes *Theileria parva*, *Trypanosoma cruzi*, *Plasmodium falciparum*, *Entamoeba histolytica* and *Toxoplasma gondii*. These parasites probably acquired 5-methylcytosine RNA generating enzymes by their host organisms and maintained them in subsequent generations, suggesting that these enzymes may have an important function in their metabolism. RCMTs appeared to be present in several lower Metazoans (like sea urchin, sea anemone, leech, snail etc.), as well. Moreover, in some species like *Toxoplasma gondii, Guillardia theta* and *Sorghum bicolor*, RCMT representatives were found exclusively in certain subfamilies (Table 1).

Genomic organization of the RCMT-related genes

The genomic organization of the RCMT-related genes was analyzed by identifying the boundaries between the exons encoding the core (and most highly conserved) region of the RCMT-like proteins. In order to obtain the exon–exon junctions, tBLASTn [9] searches of the *Arabidopsis* and rice genomes were performed, by using the amino acid sequences of the core domains of the *Arabidopsis* and rice Fmu, nop2 and nsun5 proteins as seeds. In this way, the consecutive encoding exons were retrieved. As seen in Fig. 4, the proteins Fmu, nop2 and nsun5 have essentially the same splicing patterns in *Arabidopsis* and rice.

Discussion

Five RCMT subfamilies were identified in this study, four of which without any detected prokaryotic orthologues, suggesting that eukaryotes have evolved a 'sophisticated' RNA modifying machinery that specifically methylates different types of RNA. A plethora of putative RCMT-related proteins were identified in this study in more taxonomically diverged species and even within the same species compared to DNMTs. These findings corroborate the hypothesis postulated by Bujnicki et al. [7] that RCMTs are rather evolutionary intermediates of RUMTs and DNMTs. A 5-methylcytosine generating enzymatic family may have not emerged yet in some of these (particularly slow-evolving) organisms. However, experimental assays are necessary to test whether the cytosine-5 RNA methyltransferase activity of the newly identified proteins is functional or not.

The eukaryotic and prokaryotic RCMT homologues form their own separate strongly supported groups, suggesting that the series of evolutionary events that gave rise to the last common ancestor of the RCMT family must have occurred prior to the divergence between eukaryotes and prokaryotes (Fig. 3). After extensive database searches no eukaryotic counterparts were identified in the strictly prokaryotic subfamilies and *vice versa*, except the Fmu subfamily (Fig. 3) which hitherto was thought to be restricted to prokaryotes. Examination of the splicing patterns (Fig. 4) revealed that *Fmu* orthologues from different plants have virtually identical splice sites.

Fig. 3. Maximum-likelihood phylogenetic analysis of the RCMT proteins listed in Table 1. The prototypic members are underlined. The prokaryotic RCMT sequences are shown in boldface. RumA was used as an outgroup. The branch lengths indicate evolutionary distance. Bootstrap values higher than 60 are indicated at the nodes. The scale bar at the upper right represents the length of amino acid substitutions/position.

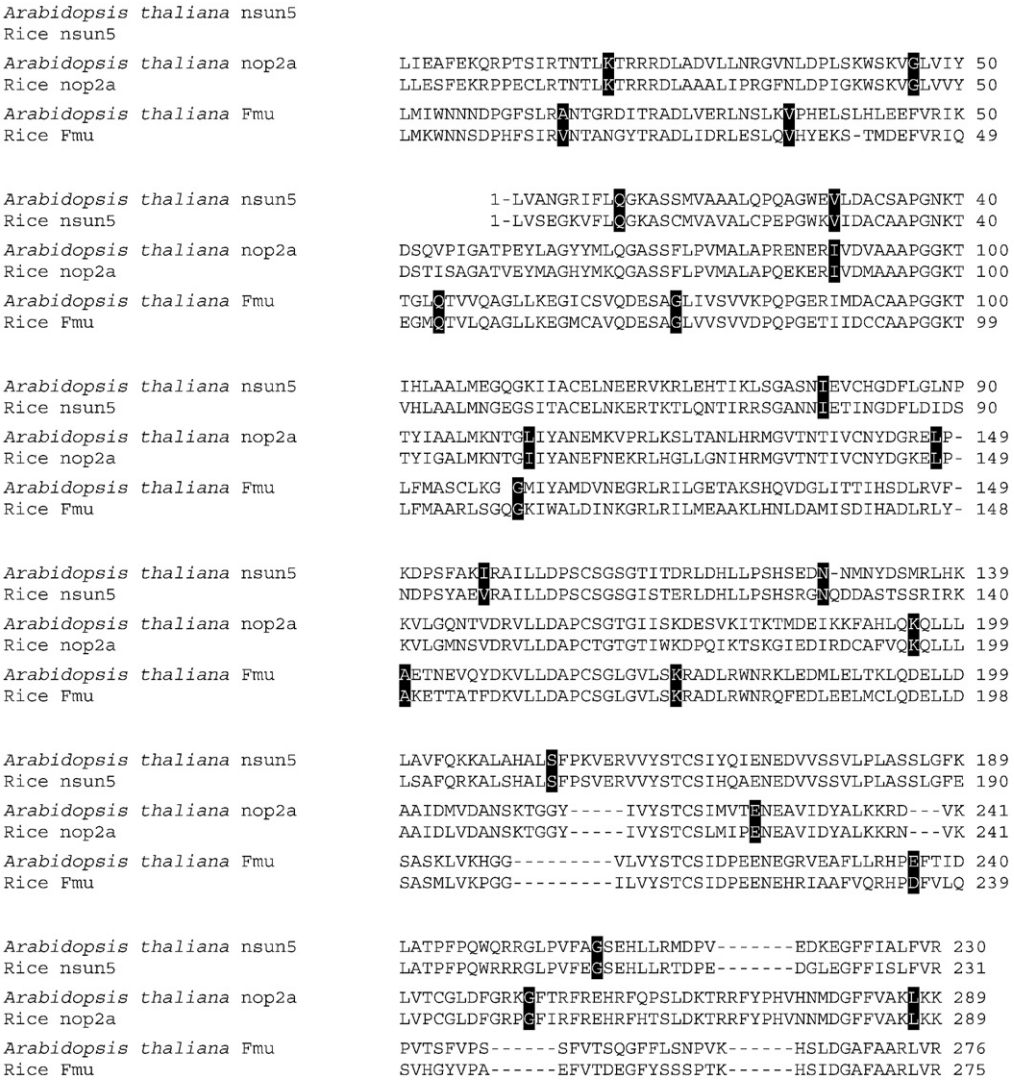


Fig. 4. Multiple alignment of the amino acid sequences corresponding to the core domain of the Fmu, nop2 and nsun5 proteins from *Arabidopsis* and rice. The sequences were aligned using CLUSTALW. The splice sites are denoted at the ends of the respective exons as white letters in a black background. The numbers refer to the amino acid positions with respect to the starting position of the core domain. Despite the splicing patterns are not conserved across Fmu, nop2 and nsun5, they are more conserved between the rice and *Arabidopsis* proteins of each of the three subfamilies.

This finding further supports the hypothesis that the plant *FmuL* genes may have evolved from a common ancestor, most likely from a prokaryotic *Fmu* gene by horizontal gene transfer, as discussed earlier. Taken together, these observations lead to the suggestion that the eukaryotic RCMTs have rather evolved from RUMTs and not from 'prokaryotic' RCMTs. Moreover, the rRNA specific NOP2 members exhibit the highest overall similarity to the fellow rRNA modifying *E. coli* Fmu (Fig. 3), suggesting that the phylogenetic relationship among RCMTs likely correlates with their RNA target substrate specificity (e.g. rRNA).

In the reconstructed phylogenetic tree in Fig. 3, Viridiplantae, Fungi and Metazoa RCMT homologues appear to form their own distinct monophyletic groups, confirmed with relatively high bootstrap values. This observation suggests that the RCMT enzyme family has probably evolved prior to the emergence of the three kingdoms and then independently diversified within each one of them. On the contrary, the distribution of Protozoan RCMT homologues is quite erratic, since they do not form distinct groups—indicated by the low support values. The rice and *Arabidopsis* RCMT paralogues in the NOP2 and NCL1 branches must be products of a recent gene duplication. However, after extensive database searches, only one copy of *YNL022c-like* product was detected in the complete genome of rice and *Arabidopsis*, suggesting that gene duplication events have occurred in a lesser extent within the YNL022c sublineage. In all eukaryotic RCMT subfamilies, the monocot (rice, maize, *Sorghum bicolor*) and eudicot (*A. thaliana*) RCMT homologues form their own separate clades within the Viridiplantae branch, triggering the speculation that the series of the evolutionary events that gave rise to the individual subfamilies may have occurred after the monocot–eudicot divergence.

Of particular interest is the observation that in NCL1 subfamily, as opposed to NOP2 and YNL022c subfamilies, the topology of the Metazoan, Plant and Fungal branches (Fig. 3) is in agreement with the topology of the universal Tree of Life, according to which Metazoans are more closely related to Fungi and Viridiplantae to Protista [33]. This observation suggests that the NCL1 subfamily may have preceded the divergence of the Viridiplantae and Fungi/Metazoa, whereas the other two subfamilies followed that split.

Subfamily RCMT9 appears to be restricted to green plants and protozoa, the earliest-emerging branches. Given that the Opisthokonta (Fungi and Metazoa) radiated later in the biological evolution [35], three possible hypotheses may account for this observation: 1) RCMT9 proteins got lost along the evolutionary pathway leading to Fungi and Metazoa, 2) differences in the lifestyle or physiology of Opisthokonta led to the loss of this family, 3) representatives of this

subfamily may emerge later in the evolutionary path in fungal and metazoan species. Notably, the ciliated Alveolata *T. thermophila* and *P. tetraurelia*, group into their own strongly supported clade, apart from their fellow Alveolatum, *Theileria parva*. This discrepancy could be explained by the fact that ciliates are organisms evolving in a notoriously slow rate, and so they are more related to the slow-evolving plants [34]. On the other hand, *T. parva* appears to be more close to *T.cruzi*, an organism that belongs to the fast-evolving Euglenozoa [34]. The members of the RCMT9 sublineage may have a novel and unique nuclear function, such as a possible involvement in modifying other types of cellular RNA. It would be also tempting to speculate that, by virtue of homology to NCL1, RCMT9 may have evolved towards a function similar to NCL1, like methylating different sites of tRNA substrates. Further studies of the substrate specificity of RCMT9 are needed.

The microsporidian *Encephalitozoon cuniculi* clusters apart its fellow fungi in all sublineages and appears to be closer to parasitic organisms (Fig. 3). This peculiar grouping is probably due to the parasitic nature of *E. cuniculi* [36]. Moreover, *Monosiga brevicolis* groups with *C. elegans* in the NOP2 branch. This is expected since Choanoflagellates are among the closest unicellular relatives of animals.

In the light of the findings of the current work it would be intriguing to speculate that the eukaryotic RCMT family of enzymes is found in the evolutionary "crossroad", and it has likely presided on the RNA/DNA transitions. Moreover, the results of this study further expand recent theories [37,38] that the eukaryotic DNMTs were most probably derived from RCMTs than the prokaryotic restriction/modification DNA methyltransferases by horizontal transfer.

Acknowledgments

We are grateful to Constantinos D. Paliakasis for valuable discussions throughout this work. Special thanks to Ioannis Michalopoulos and Karin Söderman for technical support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.12.004.

References

[1] J. Rozenski, P.F. Crain, J.A. McCloskey, The RNA Modification Database: 1999 update, Nucleic Acids Res. 27 (1999) 196–197.
[2] J.A. Martinez Gimenez, G.T. Saez, R.T. Seisdedos, On the function of modified nucleosides in the RNA world, J. Theor. Biol. 194 (1998) 485–490.
[3] F. EB, B. RM, C. X, Structure and evolution of AdoMet-dependent MTases, in: C. X, B. RM (Eds.), S-Adenosylmethionine dependent methyltransferases: structures and functions, World Scientific Inc, Singapore, 1999, pp. 1–38.
[4] Y. Liu, D.V. Santi, m5C RNA and m5C DNA methyl transferases use different cysteine residues as catalysts, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 8263–8265.
[5] J.T. Kealey, D.V. Santi, Identification of the catalytic nucleophile of tRNA (m5U54) methyltransferase, Biochemistry 30 (1991) 9724–9728.
[6] P.G. Foster, C.R. Nunes, P. Greene, D. Moustakas, R.M. Stroud, The first structure of an RNA m5C methyltransferase, Fmu, provides insight into catalytic mechanism and specific binding of RNA substrate, Structure 11 (2003) 1609–1620.
[7] J.M. Bujnicki, M. Feder, C.L. Ayres, K.L. Redman, Sequence–structure-function studies of tRNA:m5C methyltransferase Trm4p and its relationship to DNA:m5C and RNA:m5U methyltransferases, Nucleic Acids Res. 32 (2004) 2453–2463.
[8] R. Reid, P.J. Greene, D.V. Santi, Exposition of a family of RNA m(5)C methyltransferases from searching genomic and proteomic sequences, Nucleic Acids Res. 27 (1999) 3138–3145.
[9] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[10] A. Bateman, L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L. Sonnhammer, D.J. Studholme, C. Yeats, S.R. Eddy, The Pfam protein families database, Nucleic Acids Res. 32 (2004) D138–141.
[11] D.G. Higgins, CLUSTAL V: multiple alignment of DNA and protein sequences, Methods Mol. Biol. 25 (1994) 307–318.
[12] S. Kumar, K. Tamura, M. Nei, MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment, Brief Bioinform. 5 (2004) 150–163.
[13] S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst. Biol. 52 (2003) 696–704.
[14] D.T. Jones, W.R. Taylor, J.M. Thornton, The rapid generation of mutation data matrices from protein sequences, Comput. Appl. Biosci. 8 (1992) 275–282.
[15] F. Ronquist, J.P. Huelsenbeck, MrBayes 3: Bayesian phylogenetic inference under mixed models, Bioinformatics 19 (2003) 1572–1574.
[16] G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14 (2004) 1188–1190.
[17] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, L.Y. Geer, W. Helmberg, Y. Kapustin, D.L. Kenton, O. Khovayko, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, K.D. Pruitt, G.D. Schuler, L.M. Schriml, E. Sequeira, S.T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T.O. Suzek, R. Tatusov, T.A. Tatusova, L. Wagner, E. Yaschenko, Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 34 (2006) D173–180.
[18] S. Agarwalla, J.T. Kealey, D.V. Santi, R.M. Stroud, Characterization of the 23 S ribosomal RNA m5U1939 methyltransferase from Escherichia coli, J. Biol. Chem. 277 (2002) 8835–8840.
[19] J.S. Tscherne, K. Nurse, P. Popienick, H. Michel, M. Sochacki, J. Ofengand, Purification, cloning, and characterization of the 16S RNA m5C967 methyltransferase from Escherichia coli, Biochemistry 38 (1999) 1884–1892.
[20] J.R. Alfano, A. Collmer, Bacterial pathogens in plants: life up against the wall, Plant Cell 8 (1996) 1683–1698.
[21] B. Hong, J.S. Brockenbrough, P. Wu, J.P. Aris, Nop2p is required for pre-rRNA processing and 60S ribosome subunit synthesis in yeast, Mol. Cell. Biol. 17 (1997) 378–388.
[22] W.C. Gustafson, C.W. Taylor, B.C. Valdez, D. Henning, A. Phippard, Y. Ren, H. Busch, E. Durban, Nucleolar protein p120 contains an arginine-rich domain that binds to ribosomal RNA, Biochem. J. 331 (Pt 2) (1998) 387–393.
[23] E. de Beus, J.S. Brockenbrough, B. Hong, J.P. Aris, Yeast NOP2 encodes an essential nucleolar protein with homology to a human proliferation marker, J. Cell. Biol. 127 (1994) 1799–1813.
[24] L. Perlaky, R.K. Busch, H. Busch, Combinatorial effects of monoclonal anti-p120 antibody (MAbp120), liposomes and hyperthermia on MCF-7 and LOX tumor cell lines, Oncol. Res. 8 (1996) 363–369.
[25] K. Sato, T. Nishi, H. Takeshima, M. Kochi, J. Kuratsu, N. Masuko, Y. Sugimoto, Y. Yamada, Y. Ushio, Expression of p120 nucleolar proliferating antigen in human gliomas and growth suppression of glioma cells by p120 ribozyme vector, Int. J. Oncol. 14 (1999) 417–424.
[26] B.C. Valdez, L. Perlaky, D. Henning, Y. Saijo, P.K. Chan, H. Busch, Identification of the nuclear and nucleolar localization signals of the protein p120. Interaction with translocation protein B23, J. Biol. Chem. 269 (1994) 23776–23783.
[27] P. Wu, J.S. Brockenbrough, M.R. Paddy, J.P. Aris, NCL1, a novel gene for a non-essential nuclear protein in Saccharomyces cerevisiae, Gene 220 (1998) 109–117.
[28] E. Bertrand, F. Houser-Scott, A. Kendall, R.H. Singer, D.R. Engelke, Nucleolar localization of early tRNA processing, Genes Dev. 12 (1998) 2463–2468.
[29] Y. Motorin, H. Grosjean, Multisite-specific tRNA:m5C-methyltransferase (Trm4) in yeast Saccharomyces cerevisiae: identification of the gene and substrate specificity of the enzyme, RNA 5 (1999) 1105–1118.
[30] M. Frye, F.M. Watt, The RNA methyltransferase Misu (NSun2) mediates Myc-induced proliferation and is upregulated in tumors, Curr. Biol. 16 (2006) 971–981.
[31] A. Brachat, N. Liebundguth, C. Rebischung, S. Lemire, F. Scharer, D. Hoepfner, V. Demchyshyn, I. Howald, A. Dusterhoft, D. Mostl, R. Pohlmann, P. Kotter, M.N. Hall, A. Wach, P. Philippsen, Analysis of deletion phenotypes and GFP fusions of 21 novel Saccharomyces cerevisiae open reading frames, Yeast 16 (2000) 241–253.
[32] L. Ponger, W.H. Li, Evolutionary diversification of DNA methyltransferases in eukaryotic genomes, Mol. Biol. Evol. 22 (2005) 1119–1128.
[33] F.D. Ciccarelli, T. Doerks, C. von Mering, C.J. Creevey, B. Snel, P. Bork, Toward automatic reconstruction of a highly resolved tree of life, Science 311 (2006) 1283–1287.
[34] H. Philippe, P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M. Muller, H. Le Guyader, Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions, Proc. Biol. Sci. 267 (2000) 1213–1221.
[35] M. Medina, Genomes, phylogeny, and evolutionary systems biology, Proc. Natl. Acad. Sci. U. S. A. 102 (Suppl. 1) (2005) 6630–6635.
[36] R. Upadhya, H.S. Zhang, L.M. Weiss, System for expression of microsporidian methionine amino peptidase type 2 (MetAP2) in the yeast Saccharomyces cerevisiae, Antimicrob. Agents Chemother. 50 (2006) 3389–3395.
[37] M.G. Goll, F. Kirpekar, K.A. Maggert, J.A. Yoder, C.L. Hsieh, X. Zhang, K.G. Golic, S.E. Jacobsen, T.H. Bestor, Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2, Science 311 (2006) 395–398.
[38] A. Pavlopoulou, S. Kossida, Plant cytosine-5 DNA methyltransferases: structure, function, and molecular evolution, Genomics 90 (2007) 530–541.