3138–3145 *Nucleic Acids Research, 1999, Vol. 27, No. 15*

Exposition of a family of RNA m^{5}C methyltransferases from searching genomic and proteomic sequences

Ralph Reid, Patricia J. Greene and Daniel V. Santi*

Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143-0448, USA

Received March 26, 1999; Revised and Accepted June 8, 1999

ABSTRACT

The *Escherichia coli fmu* gene product has recently been determined to be the 16S rRNA m^{5}C 967 methyltransferase. As such, Fmu represents the first protein identified as an *S*-adenosyl-L-methionine (AdoMet)-dependent RNA m^{5}C methyltransferase whose amino acid sequence is known. Using the amino acid sequence of Fmu as an initial probe in an iterative search of completed DNA sequence databases, 27 homologous ORF products were identified as probable RNA m^{5}C methyltransferases. Further analysis of sequences in undeposited genomic sequencing data and EST databases yielded more than 30 additional homologs. These putative RNA m^{5}C methyltransferases are grouped into eight subfamilies, some of which are predicted to consist of direct genetic counterparts, or orthologs. The enzymes proposed to be RNA m^{5}C methyltransferases have sequence motifs closely related to signature sequences found in the well-studied DNA m^{5}C methyltransferases and other AdoMet-dependent methyltransferases. Structure–function correlates in the known AdoMet methyltransferases support the assignment of this family as RNA m^{5}C methyltransferases.

INTRODUCTION

RNA modifications have been well characterized (1), but relatively little is known about the enzymes that catalyze such modifications or about the functions of the modified residues. We have undertaken a program directed at identifying the enzymes responsible for formation of the modified nucleotides in RNAs. Recently, the *fmu* gene product (Fmu) was reported to catalyze formation of m^{5}C at position 967 of *Escherichia coli* 16S rRNA (2,3), and this represents the first RNA m^{5}C MTase whose amino acid sequence is known. [Fmu was renamed rrmB (2) and rsmB (3). To avoid confusion we have returned to the older designation.] Tscherne and co-workers (3) also reported searching GenBank and finding proteins similar to Fmu in 13 organisms. They suggested that three of these proteins might be RNA m^{5}C MTases.

In the present work, we have used the Fmu sequence as an initial probe in an iterative search of all available sequence databases, and we have identified more than 55 open reading frames (ORFs) that encode proteins that are homologous to Fmu. These homologs have sequence motifs which are closely related to signature sequences found in the well-studied DNA m^{5}C MTases, and these motifs allow us to identify the Fmu homologs as putative RNA m^{5}C MTases and to assign some structure–function relationships. Further, from homology comparisons within the family, we are able to classify these m^{5}C RNA MTases into eight subfamilies, of which at least five are likely to represent orthologs.

MATERIALS AND METHODS

This study used databases of deposited and undeposited sequences reported up to February 22, 1999. Using Fmu as a starting probe, Blast 2.0 and PSI-Blast (4) were used in iterative searches through NCBI databases of deposited protein sequences translated from DNA. Usual search parameters included an EXPECT value of 1000 and a minimum word length of 2 to pick up weak similarities. Filtering was usually turned off. Various similarity matrices were used of both PAM and BLOSUM types. Gap penalties were also varied, with low penalties chosen to enhance detection of homologs with insertions or deletions relative to the probes. We collected a large set of deposited protein sequences which were homologous to Fmu, and these sequences were used as probes in further iterative searches which were widened to include other databases available through NCBI. Preliminary sequence data were obtained from The Institute for Genomic Research website at http://www.tigr.org . These data come from many different sequencing projects and are made available through TIGR (http://www.tigr.org/tdb/mdb/mdb.html).

In certain eukaryotic cases, sequences from EST databases were collected by BLAST searches of type tblastn, and assembled into partial ORFs using the putative m^{5}C MTase sequences as guides. This procedure was done iteratively to produce substantial approximates of theoretical ORFs.

In the eubacterial and archaebacterial cases, undeposited genomic sequencing data available for Blast searches were used to collect homologous patches of translated sequence which were assembled into larger putative protein fragments. As in the eukaryotic case, these patches were added to our set of homologs, and the procedure was iterated.

*To whom correspondence should be addressed at: Department of Biochemistry, University of California, San Francisco, CA 94143-0448, USA. Tel: +1 415 476 1740; Fax: +1 415 476 0473; Email: santi@socrates.cgl.ucsf.edu

Table 1. Putative RNA m^{5}C MTase ORFs

Organism	I (Fmu)	II (P120)	III (Yebu)	IV	V	VI	VII (Ncl1)	VIII
I.Eubacteria								
Proteobacteria								
Actinobacillus actinomyces[a]	U[b]	-	-	-	-	-	-	-
Haemophilus influenzae[c]	P44788[d]	-	-	-	-	-	-	-
Pseudomonas aeruginosa[a]	U	-	-	-	-	-	-	-
E. coli[c]	P23866	-	2829645	-	-	-	-	-
Yersinia pestis[a]	U	-	-	-	-	-	-	-
Coxiella burnetii[a]	P45679	-	-	-	-	-	-	-
Vibrio alginolyticus	Fmu/v[e]	-	-	-	-	-	-	-
Neisseria gonorrhoeae[a]	U	-	-	U	-	-	-	-
Neisseria meningitidis[a]	U	-	-	U	-	-	-	-
Campylobacter jejuni[a]	-	-	-	-	U	-	-	-
Gram Positives								
B.subtilis[c]	3915867	-	-	-	-	-	-	-
Staphylococcus aureus[a]	U	-	-	-	-	-	-	-
Streptococcus pneumoniae[a]	U	-	U	-	-	-	-	-
Streptococcus pyogenes[a]	U	-	U	-	-	-	-	-
Lactococcus lactis[a]	P72943	-	-	-	-	-	-	-
Enterococcus faecalis[a]	U	-	-	-	-	-	-	-
Clostridium acetobutylicum[a]	U	-	-	-	-	-	-	-
Mycobacterium tuberculosis[c]	2829532	-	-	-	-	-	-	-
Other phyla								
Synechocystis[c]	P72943	-	-	-	-	-	-	-
Deinococcus radiodurans[a]	U	-	-	-	-	-	-	-
Porphyromonas gingivalis[a]	-	-	U	-	-	-	-	-
Thermotoga maritima[a]	U	-	-	-	-	-	-	-
II.Eukaryotes								
human	-	P46087	-	-	-	-	E[b]	E
mouse	-	A48998	-	-	-	-	E	E
D. melanogaster		E					4185892	E
C. elegans[c]	-	3886025	-	-	-	-	-	-
S. cerevisiae[c]	-	P40991	-	-	-	-	586408	1730712
S. pombe	-	3810844	-	-	-	-	2465159	2414617
III Archaebacteria								
Methanococcus jannaschii[c]	-	-	-	-	2500953	-	-	
Archaeoglobus fulgidus[c]	-	-	-	2648195	2649749	-	-	-
Pyrococcus horikoshii[c]	-	-	-	3131344	2648496 3131814	3130211	-	-
Pyrococcus furiosus[a]	-	-	-	3131110 2 U	2696448 2 U	U	-	-

[a]Undeposited genomic sequences from http://www.tigr.org . Funding sources and institutes supplying sequence are listed in http://www.tigr.org/tdb/mdb/mdb.html

[b]U, ORFs compiled from undeposited genomic sequencing data (2 U indicates there are two ORFs). E, ORFs compiled from the EST division of GenBank/EMBL/DDBJ. Most of these ORFs are incomplete.

[c]Complete genomic sequence is available.

[d]All numbers in cells are accession numbers for GenBank/EMBL/DDBJ databases.

[e]BAA31227 and BAA31226 are accession numbers are for fmu and fmv which have been combined to yield a single ORF as found in *E.coli*.

The collection of homologs was used to create a database that was used as the object of Blast searches using individual homologs as probes. Analysis of this set of protein sequences allowed division into subfamilies of closer relatives. The extended homologies within these subfamilies were used to further expand fragmentary individual sequences within our database in continued iterations of the procedure described above.

Alignment of each subfamily was initiated using the Pileup program inside GCG v.9.1 (5) and was then further adjusted. Alignments between subfamilies were initiated using conserved sequence patches in each subfamily and then adjusted.

RNA sequence data were obtained from the rRNA WWW server at the University of Antwerp, Belgium (http://rrna.uia.ac.be). The phylogenetic assignments and terminology used (e.g. 'epsilon division of Proteobacteria') are those available on the NCBI web site (http://www.ncbi.nlm.nih.gov/Taxonomy).

RESULTS AND DISCUSSION

We used the sequence of Fmu as an initial probe in an iterative search of the available sequence databases. In the databases of finished sequences (GenBank/EMBL/DDBJ), we found a family of 27 ORFs with deduced proteins homologous to Fmu, and in the undeposited genomic sequencing data and EST sequences we found more than 30 additional Fmu homologs (Table 1). To obtain the latter, sequences were identified as putative m^{5}C MTase fragments and assembled into larger ORFs for incorporation into the analysis of the family of Fmu homologs. The translated sequences of the complete and incomplete ORFs are available at http://www.sacs.ucsf.edu/home/SantiLab/m5c.html . As this work progressed, we performed comparisons of the Fmu homologs with the more extensively studied DNA m^{5}C MTases and other nucleic acid MTases. This analysis revealed that the Fmu homologs possess a

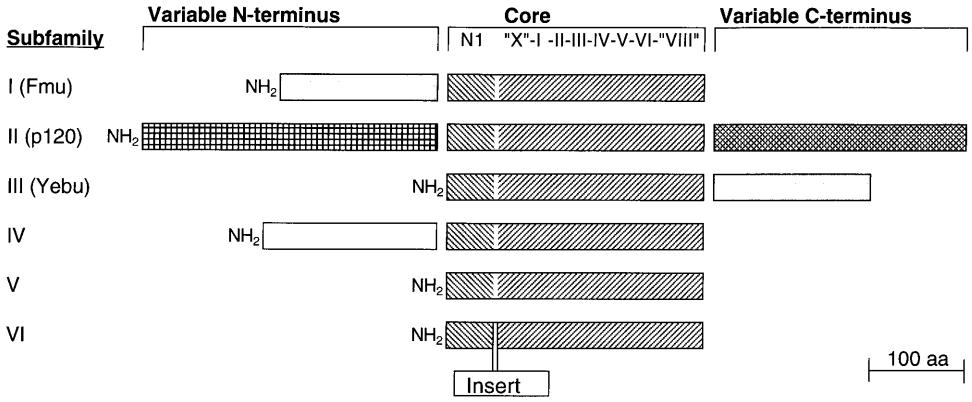


Figure 1. Arrangement of conserved regions in subfamilies of the putative RNA m^5C MTase family. A core region of about 270 amino acids is conserved throughout subfamilies I–VI of this family of putative AdoMet-dependent MTases, with extensive, but lower, homology in subfamilies VII and VIII (not shown). The core is divided into a small N-terminal region of about 40–50 amino acids and a larger C-terminal region of about 220–230 amino acids. The larger region contains conserved motifs which appear to be homologous to those seen in a wide range of AdoMet-dependent MTases.

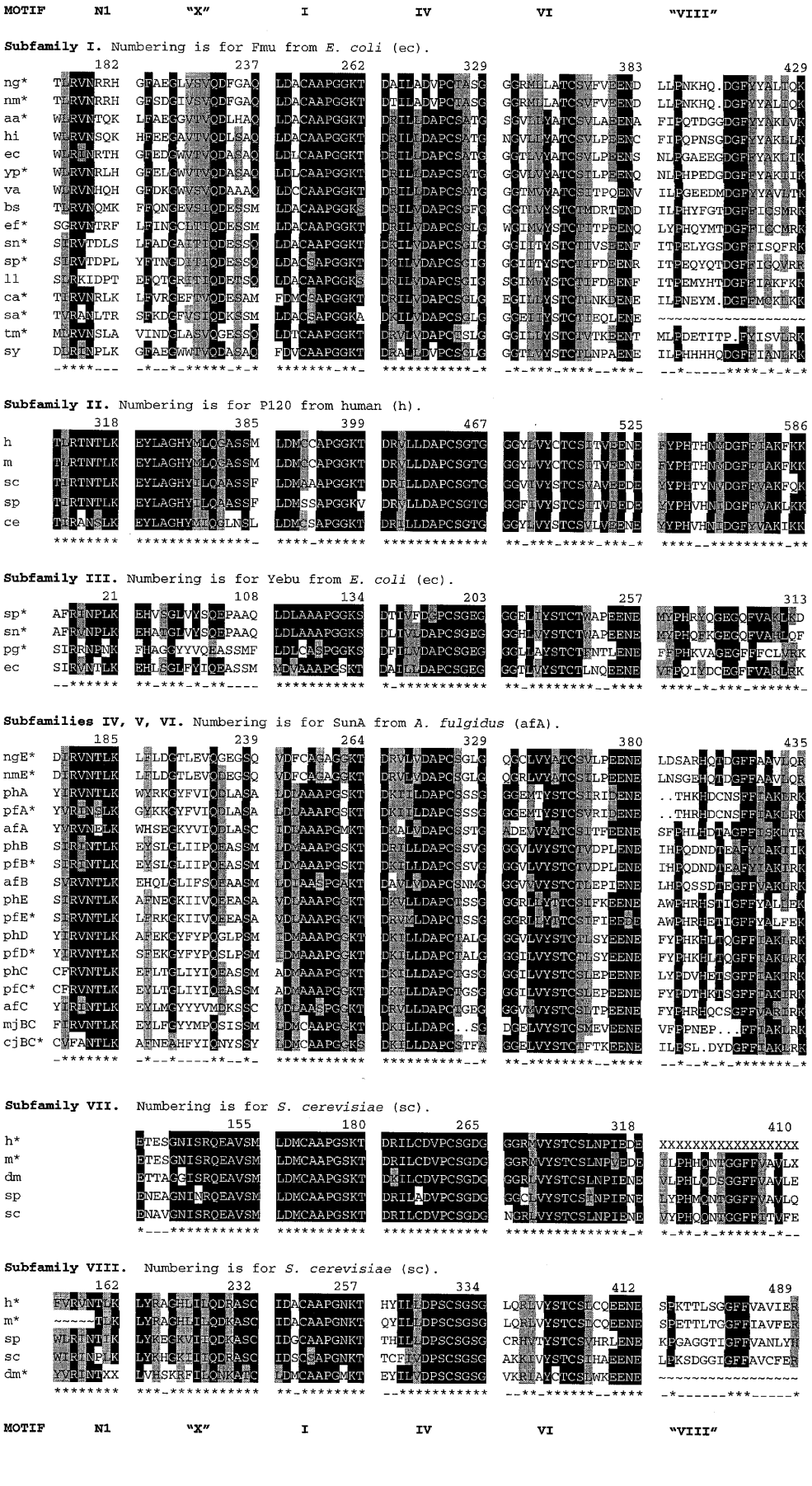
conserved core with sequence motifs that have homology to conserved signature motifs described in DNA MTases and other *S*-adenosyl-L-methionine (AdoMet)-dependent MTases (6–12). The established structure–function correlates of these motifs in the DNA MTases enabled provisional assignment of m^5C MTase function to this family of RNA modifying enzymes. In the following, we first describe the overall family of putative RNA m^5C MTases obtained by homology searches using Fmu as an initial probe, followed by a comparison of conserved sequence motifs to the signature sequences of other AdoMet MTases.

Subfamilies of RNA m^5C MTases

Based on the linear arrangement of conserved motifs and sequence homology levels, the family of RNA m^5C MTases was subdivided into eight subfamilies (I–VIII). Figure 1 shows the arrangement of conserved regions in subfamilies I–VI. All subfamilies have a conserved core region consisting of a small N-terminal region of ~40–50 amino acids and a larger C-terminal region of about 220–230 amino acids which contains signature sequence motifs (Fig. 2). The subfamilies are distinguished by N- or C-terminal extensions or, in the case of subfamily VI, an insertion into the core sequence. These extensions of the core sequence have regions of homology within a subfamily, but vary between subfamilies (Fig. 3). The RNA substrate recognition domains may exist in these non-core extensions, or they may exist in thus far unrecognized proteins that serve as RNA-binding subunits for these putative RNA m^5C MTases. We propose that members of subfamilies I, II, VII, VIII and probably III, are made up of direct genetic counterparts or orthologs.

Subfamily I, which includes Fmu and putative orthologs, has 20 eubacterial members. Members of this family contain an N-terminal extension of ~220 amino acids (Fig. 3A). In addition to homology, strong evidence that members of subfamily I are orthologs is derived from a correlation between the presence of the gene and a C residue at the target site of small subunit RNA. The target of Fmu, C967 of *E.coli* 16S rRNA, is the fourth base from the 5' end of the loop of a conserved 4 bp stem–8 base loop found in all small subunit RNAs (helix 31 in *Escherichia coli* 16S rRNA) (13). All ORFs of subfamily I belong to species of Eubacteria that contain a C residue at the analogous site of their small subunit RNAs. In five other complete and two nearly complete eubacterial genomes we did not find putative Fmu orthologs, and in each case the nucleotide in the small subunit RNAs corresponding to position 967 of *E.coli* 16S rRNA was not a C (G in *Mycoplasma genitalium* and *Mycoplasma pneumoniae*, and A in *Helicobacter pylori*, *Treponema pallidum*, *Borrelia burgdorferi*, *Campylobacter jejuni and Porphyromonas gingivalis*). Together with certain sequence homologies, this correlation suggests that all members of subfamily I are Fmu orthologs that methylate the 5 carbon of a C residue at the fourth base of the loop of the

Figure 2. (Opposite) Signature motifs of the putative RNA m^5C MTases. Motifs assignments are as described (8). Identical residues are black with white lettering, gray shading indicates similar residues. Organisms are designated by the first initials of the genus and species name in lower case, except for *S.pneumoniae* and *Synechocystis*, which are sn and sy, respectively. * following the organism designation indicates that the amino acid sequence is a translation of an incomplete ORF. ~ indicates N- or C-terminus of the available sequence. X indicates incomplete internal sequence. Motifs shown are from 16 members from subfamily I, five members from subfamily II, and all members of the other subfamilies listed in Table 1. Subfamilies IV–VI are combined. In the databases, Sun is a common designation for Fmu homologs, SunA–SunE refers to the Fmu homologs in families IV–VI. In the organism designations for these families, A or E, subfamily IV; B, C or BC, subfamily V; D, subfamily VI.


MOTIF N1 "X" I IV VI "VIII"
Subfamily I. Numbering is for Fmu from E. coli (ec).
Subfamily II. Numbering is for P120 from human (h).
Subfamily III. Numbering is for Yebu from E. coli (ec).
Subfamilies IV, V, VI. Numbering is for SunA from A. fulgidus (afA).
Subfamily VII. Numbering is for S. cerevisiae (sc).
Subfamily VIII. Numbering is for S. cerevisiae (sc).
MOTIF N1 "X" I IV VI "VIII"

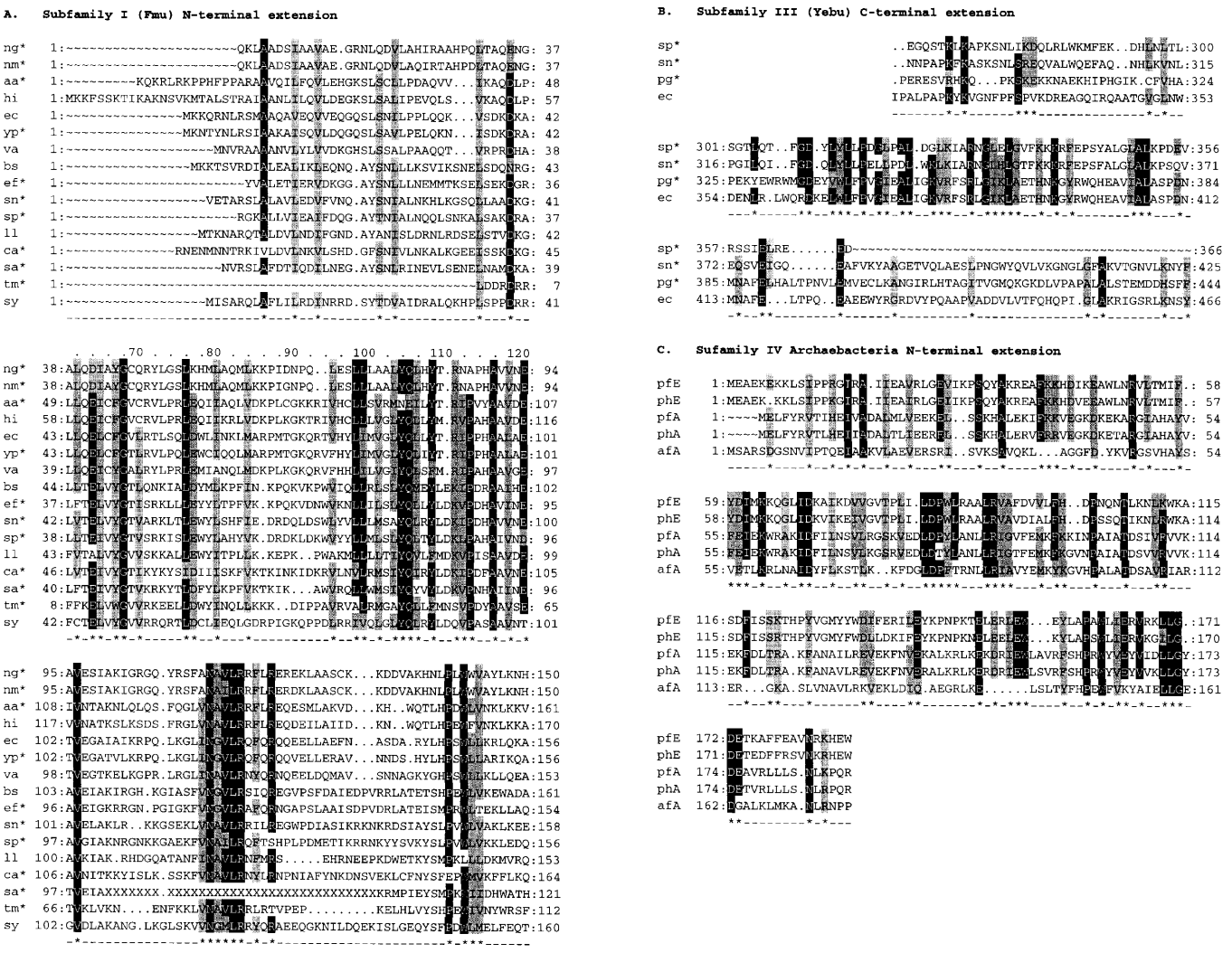


Figure 3. Conserved extensions of three subfamilies of the putative RNA m^5C MTases. Numbering is shown for each amino acid sequence. ~ and X as in Figure 2. (**A**) N-terminal extension of selected members of subfamily I. (**B**) Conserved region of the C-terminal extension of members of subfamily III. (**C**) N-terminal extension of members of subfamily IV.

conserved 4 bp stem–8 base loop of small subunit RNAs. Although some Eukaryotes and Archaebacteria have a C in the corresponding position, the known modification patterns in this region are very different from those found in Eubacteria (14,15), and no putative orthologs of Fmu were found in these phyla.

Subfamily II possesses several putative orthologs in eukaryotes; six ORFs are listed in Table 1. Members of this family have a C-terminal extension in addition to a large N-terminal extension which has been shown to contain nuclear and nucleolar localization motifs (16). The proteins from human, mouse and yeast have been studied without knowledge that they might be RNA m^5C MTases. Previously, the AdoMet-binding motifs and the nucleolar localization of these proteins led workers to propose that they may be RNA methylating enzymes (17), possibly nucleotide 2'-O-MTases (18). Because of the signature motifs described below, we now propose that these ORFs encode RNA m^5C MTases rather than 2'-O-MTases. The human protein in this subfamily, P120, has attracted interest as a tumor marker (19–22).

Subfamily III, including YebU from *E.coli*, contains four eubacterial members that we predict to be orthologs. These ORFs encode the core plus a conserved C-terminal extension (Fig. 3B). Since rRNA of *E.coli* contains only three m^5C residues, and Fmu methylates C967 of 16S rRNA, YebU from *E.coli* may be responsible for producing one or both of the remaining m^5C residues of rRNA: C1407 of 16S rRNA or C1962 in 23S rRNA.

Subfamily IV contains five members from Archaebacteria which have in common a conserved set of N-terminal extensions different from the set found in the Fmu subfamily (Fig. 3C). Two Archaebacteria, *Pyrococcus horikoshii* and *Pyrococcus*

furiosus, each contain two putative RNA m^5C MTases in this subfamily. The two eubacterial homologs listed in Table 1 are represented by incomplete sequences; provisional assignment to subfamily IV is based primarily on homology in the core region.

Subfamily V contains seven archaebacterial members of unknown function. This subfamily contains only the core sequence. Three Archaebacteria, *Archaeoglobus fulgidus*, *P.horikoshii* and *P.furiosus*, each contain two putative RNA m^5C MTases in this subfamily. There is also one eubacterial RNA m^5C MTase homolog that appears to possess only the core sequence. It is provisionally assigned to this subfamily.

Subfamily VI contains two archaebacterial members of unknown function that are uniquely characterized by an internal insert between two portions of the conserved core sequence.

The proteins encoded by the eukaryotic ORFs in subfamilies VII and VIII are more diverged from Fmu than the members of groups I–VI. A yeast ORF (*NCL1*) from subfamily VII has recently been identified as encoding a non-essential nuclear protein (23). Disruption of *NCL1* leads to increased sensitivity to paramomycin, an aminoglycoside antibiotic that affects translational fidelity. The core sequences of subfamily VII contain the sequence motifs (discussed below) which clearly identify them as members of the family of probable m^5C MTases. Motif N1, whose function is unknown, is more diverged or absent. Members of the family have N- and C-terminal extensions with conserved regions whose lengths are uncertain due to incomplete sequence data. Subfamily VIII possesses a variation in one of the prime signature motifs for RNA m^5C MTases (motif IV discussed below; ProSerCys rather than ProCys) and assignment of the 5 position of C as the target of methyl transfer is therefore less certain for this subfamily. Members of subfamily VIII have at most a minimal C-terminal extension; there is an N-terminal extension with a conserved region whose length is uncertain.

The phylogenetic distribution of the Fmu subfamily (subfamily I) (Table 1) suggests that this enzyme originated in a common ancestor of many eubacterial phyla, in particular the Firmicutes (the Gram-positives) and the Proteobacteria phylum (which includes *E.coli*). Also the wide distribution suggests continuing and strong selection pressure to retain this enzyme. Analysis of small subunit rRNA sequences suggests that secondary loss of the activity is correlated with loss of the substrate C in specialized genera such as Mycoplasma (in the Gram-positives), in divisions such as the epsilon division of the Proteobacteria (which includes *H.pylori and C.jejuni*), and perhaps in entire secondarily reduced phyla such as the Spirochaetes (*T.pallidum, B.burgdorferi)*. Representatives of subfamilies II, VII and VIII show a distribution suggestive of an ancient origin in a common ancestor of Fungi and Animalia, with secondary loss of VII and VIII in the nematode *Caenorhabditis elegans*. The eubacterial homologs in subfamilies III–V show widely scattered phylogenetic distributions more consistent with lateral transmission (Table 1).

Sequence motifs and structural homologies in the AdoMet-dependent DNA and RNA MTases

Alignments of known DNA and RNA AdoMet MTases exhibit an ordered set of up to 10 motifs, designated I–X (9,10) for which functions have been assigned from structural studies (24–27). Motifs I–V contain binding elements for AdoMet. Motifs IV and VI contain signature sequences that target specific bases. Motif IV contains a ProCys consensus for the DNA m^5C MTases or a Pro(Tyr/Phe) consensus for the exocyclic amino m^6A and m^4C MTases. Motif VI contains an acidic residue for recognition of the target C of m^5C MTases or hydrophobic residues in exocyclic amino MTases (6,7,11). Two other motifs found in AdoMet-dependent MTases, VIII and X, are well conserved within but not between separate families of DNA MTases. The linear arrangements of motifs I–X are found as circular permutations characteristic of different AdoMet-dependent MTase families (8). Most DNA m^5C MTases have the arrangement, I–VI…VIII….X; one unusual DNA m^5C MTase is circularly permuted to have motif X at the N-terminus (28), a pattern also seen in the DNA exocyclic-amino MTase family gamma (29). Regardless of their linear arrangement, motifs VIII and X occupy similar positions in the three-dimensional structures, and hence probably serve similar functions. Recent work suggests that for polynucleotide substrates, a likely function for motif VIII is to help stabilize the target base in a position flipped out from its normal position in the secondary structure (30).

Crystal structures show that the DNA m^5C MTases are folded into two domains. The larger 'catalytic' domain is composed of the N-terminal region and usually has a small contribution from the C-terminus. This domain consists of a typical MTase fold containing the motifs that provide the AdoMet binding and catalytic sites. The core region (excluding motif N1) of the RNA m^5C MTase family corresponds to this domain of the DNA m^5C MTases. The smaller variable domain provides sequence-specific recognition of DNA substrates. The RNA super-family does not contain this smaller domain, but either the extensions and inserts that characterize the subfamilies or separate subunits may serve as RNA binding domains.

The crystal structure of the RNA MTase ErmC' and the solution structure of the related ErmAM have recently been determined (31,32). These enzymes are members of a family of enzymes which confer erythromycin resistance to microbes by methylating N6 of a highly conserved A residue in large subunit rRNAs (A2058 in *E.coli*) (33). The *erm*-related gene products show conserved core sequence motifs homologous to the AdoMet binding motifs and base-specific motifs of the exocyclic amino DNA m^6A and m^4C MTase families. Analysis of these motifs and their location in the structures shows them to be organized in the same fold found in the DNA m^5C and m^6A/m^4C MTase families. The sequence and structural homologies of these proteins strongly suggest evolutionary relationships connecting AdoMet-dependent MTases, including the RNA and DNA m^5C and m^6A/m^4C MTases.

The larger region of the conserved core of the putative RNA m^5C MTases contains eight conserved motifs ('X'–I–II–III–IV–V–VI–'VIII') that as a group have strong homologies to the signature motifs described above. Figure 4 shows the core region of two representatives of the putative RNA m^5C MTase superfamily aligned with conserved motifs from representatives of three other nucleic acid MTase families. Alignments of motif regions, N1, 'X' I, IV, VI and 'VIII', for all eight subfamilies of the putative RNA m^5C MTases are shown in Figure 2. 'X' and 'VIII' are two conserved motifs that flank motifs I–VI and thus are in the positions of motifs X and VIII, respectively, in the unusual DNA m^5C MTase described above. The conserved motifs provide evidence that all of the

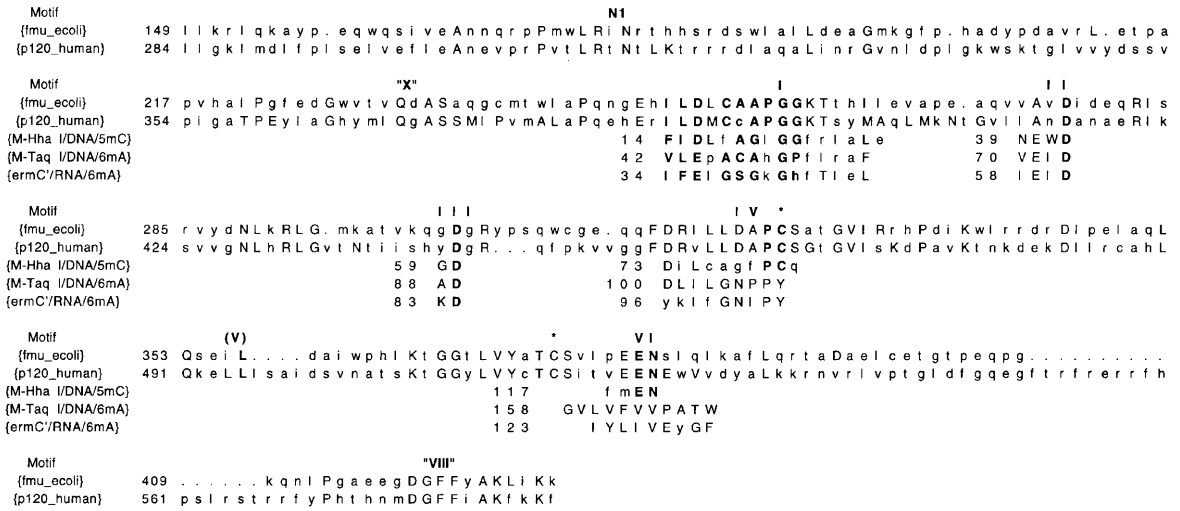


Figure 4. RNA m^{5}C MTases aligned with the signature motifs of representatives of other AdoMet-dependent MTases. Two RNA m^{5}C MTases, Fmu and P120, are shown. Capital letters indicate residues that are moderately to highly conserved in subfamilies I–VI. The two Cys residues that are totally conserved throughout the superfamily are indicated with *. R M-*Hha* I, M-*Taq* I and ErmC' are representatives of different superfamilies of AdoMet MTases. The alignment of the conserved regions shown is based on published crystal structures (31). Analogous conserved regions from different families have been given standardized motif names (I–VIII and X) (8). Bold letters in motifs I–VI indicate residues likely to be involved in substrate and AdoMet binding.

Fmu homologs are MTases that specifically target the 5 position of C.

Relationship of the Fmu homologs to the DNA m^{5}C MTases

Conserved motifs I–VI of the putative RNA m^{5}C MTases bear a striking resemblance to the six AdoMet and cytosine-binding motifs (I–VI) of the large catalytic domain of DNA-m^{5}C MTases. Following is a comparison of conserved motifs of the putative RNA m^{5}C MTases to the motifs of DNA m^{5}C MTases using Fmu and M-*Hha*I as examples of the two families (Fig. 4).

Motif I is one of the most conserved motifs in the DNA m^{5}C MTases (residues 18–22, FAGLGG, in M-*Hha*I). It is rich in Gly and other small side-chain amino acids that allow a tight turn that interacts with AdoMet. In Fmu the analog of motif I appears at residues 254–259 (CAAPGG), and it is also one of the most conserved motifs in the RNA m^{5}C MTase family alignment.

Motifs II and III are among the least conserved motifs in the DNA-m^{5}C MT family. Each contains an acidic residue which provides part of the AdoMet binding site (E40 and D60 in M-*Hha* I). Motif II is usually spaced 10–20 residues after motif I, and motif III occurs about 20 residues thereafter. In Fmu, the candidates for motifs II and III are those which contain D267 and D303.

In the DNA m^{5}C MTase family, conserved motif IV provides binding residues for both AdoMet and cytosine and includes the active site Cys that forms a covalent adduct with the 6 carbon of the cytosine ring during catalysis (34). The catalytic Cys follows about 20 residues after the acidic residue of motif III, and it is invariably preceded by a Pro, the side-chain of which interacts with AdoMet (Pro80Cys81 in M-*Hha* I). As expected, the ProCys of motif IV is not found in the MTases that methylate the exocyclic amino groups of A or C, since covalent adduct formation with the target base is not an essential feature of these reactions. In Fmu, motif IV contains Pro324Cys325 and is one of the three most highly conserved motifs of the RNA m^{5}C MTase family. Although there is a second completely conserved Cys in the RNA m^{5}C MTase family, Cys325 is likely to be the nucleophilic catalyst. A recent report of a mutagenic analysis of the yeast subfamily II (p120) homolog shows that the Cys residue in the ProCys motif is essential for function, while the second conserved Cys is not (35).

Motif V is not always clearly identifiable in members of the DNA m^{5}C MTase family (Leu100 in M-*Hha* I). The suggested function is to provide a hydrophobic residue, usually Leu, as part of the AdoMet binding site. In Fmu, the candidate for motif V is that which contains the highly conserved L357.

Motif VI, a third highly conserved motif of the DNA m^{5}C MTase family, provides a glutamate in the sequence GluAsn as an important contributor to cytosine-binding (E119 in M-*Hha*I). The Glu side chain carboxyl forms H-bonds to N3 and N4 of the target cytosine, serving a role in determining cytosine specificity. In Fmu, motif VI contains E381N382 and is highly conserved throughout the RNA m^{5}C MTase family. This motif region also contains the second Cys residue which is totally conserved in this family.

Summary

Using Fmu as a probe to search the available sequence databases, we uncovered more than 50 proteins that are likely to be RNA m^{5}C methyltransferases. Only one Fmu homolog, YebU, was found in the *E.coli* genome database, and we propose that this enzyme is responsible for producing one or both of the remaining m^{5}C residues in rRNA (C1407 of 16S RNA or C1962 of 23S RNA). Of significance is the finding that human P120, a proliferation-associated nucleolar antigen and important tumor marker, is also likely to be an RNA m^{5}C MTase. Comparison

of sequences of the putative RNA m^5C MTases to those of DNA m^5C MTases revealed that at least six conserved signature motifs in the DNA MTases were also found in the RNA m^5C MTases. From this comparison, we identified Cys325 of Fmu, and the corresponding Cys residue in each of the Fmu homologs, as the probable candidate for the catalytic nucleophile in the enzymatic reaction.

ACKNOWLEDGEMENT

This work was supported by USPHS Grant GM51232 (D.V.S.) from the National Institutes of Health. Complete information on the funding source and institute supplying sequence though the TIGR website can be found at http://www.tigr.org/tdb/mdb/mdb.html

REFERENCES

1. Smith,J.E., Cooperman,B.S. and Mitchell,P. (1992) *Biochemistry*, **31**, 10825–10834.
2. Gu,X.R., Gustafsson,C., Ku,J., Yu,M. and Santi,D.V. (1999) *Biochemistry*, **38**, 4053–4057.
3. Tscherne,J.S., Nurse,K., Popienick,P., Michel,H., Sochacki,M. and Ofengand,J. (1999) *Biochemistry*, **38**, 1884–1892.
4. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
5. Genetics Computer Group (1997) *Wisconsin Package v.9.1*. Genetics Computer Group, Inc.
6. Cheng,X. (1995) *Annu. Rev. Biophys. Biomol. Struct.*, **24**, 293–318.
7. Cheng,X.D. (1995) *Curr. Opin. Struct. Biol.*, **5**, 4–10.
8. Malone,T., Blumenthal,R.M. and Cheng,X.D. (1995) *J. Mol. Biol.*, **253**, 618–632.
9. Pósfai,J., Bhagwat,A.S. and Roberts,R.J. (1988) *Gene*, **74**, 261–265.
10. Pósfai,J., Bhagwat,A.S., Pósfai,G. and Roberts,R.J. (1989) *Nucleic Acids Res.*, **17**, 2421–2435.
11. Kumar,S., Cheng,X.D., Klimasauskas,S., Mi,S., Posfai,J., Roberts,R.J. and Wilson,G.G. (1994) *Nucleic Acids Res.*, **22**, 1–10.
12. Schluckebier,G., O'Gara,M., Saenger,W. and Cheng,X.D. (1995) *J. Mol. Biol.*, **247**, 16–20.
13. Brimacombe,R. (1995) *Eur. J. Biochem.*, **230**, 365–383.
14. Bakin,A. and Ofengand,J. (1995) *Nucleic Acids Res.*, **23**, 3290–3294.
15. Maden,B.E. (1990) *Prog. Nucleic Acid Res. Mol. Biol.*, **39**, 241–303.
16. Valdez,B.C., Perlaky,L., Henning,D., Saijo,Y., Chan,P.K. and Busch,H. (1994) *J. Biol. Chem.*, **269**, 23776–23783.
17. Koonin,E.V. (1994) *Nucleic Acids Res.*, **22**, 2476–2478.
18. Hong,B., Brockenbrough,J.S., Wu,P. and Aris,J.P. (1997) *Mol. Cell. Biol.*, **17**, 378–388.
19. Freeman,J.W., McGrath,P., Bondada,V., Selliah,N., Ownby,H., Maloney,T., Busch,R.K. and Busch,H. (1991) *Cancer Res.*, **51**, 1973–1978.
20. Fonagy,A., Swiderski,C., Wilson,A., Bolton,W., Kenyon,N. and Freeman,J.W. (1993) *J. Cell. Physiol.*, **154**, 16–27.
21. Fonagy,A., Swiderski,C., Dunn,M. and Freeman,J.W. (1992) *Cancer Res.*, **52**, 5250–5256.
22. Busch,H., Busch,R.K., Freeman,J.W. and Perlaky,L. (1991) *Boll. Soc. Ital. Biol. Sper.*, **67**, 739–750.
23. Wu,P., Brockenbrough,J.S., Paddy,M.R. and Aris,J.P. (1998) *Gene*, **220**, 109–117.
24. Cheng,X.D., Kumar,S., Posfai,J., Pflugrath,J.W. and Roberts,R.J. (1993) *Cell*, **74**, 299–307.
25. Klimasauskas,S., Kumar,S., Roberts,R.J. and Cheng,X.D. (1994) *Cell*, **76**, 357–369.
26. O'Gara,M., Roberts,R.J. and Cheng,X.D. (1996) *J. Mol. Biol.*, **263**, 597–606.
27. O'Gara,M., Klimasauskas,S., Roberts,R.J. and Cheng,X.D. (1996) *J. Mol. Biol.*, **261**, 634–645.
28. Xu,S.Y., Xiao,J.P., Posfai,J., Maunus,R. and Benner,J. (1997) *Nucleic Acids Res.*, **25**, 3991–3994.
29. Gong,W.M., Ogara,M., Blumenthal,R.M. and Cheng,X.D. (1997) *Nucleic Acids Res.*, **25**, 2702–2715.
30. Pues,H., Bleimling,N., Holz,B., Wolcke,J. and Weinhold,E. (1999) *Biochemistry*, **38**, 1426–1434.
31. Bussiere,D.E., Muchmore,S.W., Dealwis,C.G., Schluckebier,G., Nienaber,V.L., Edalji,R.P., Walter,K.A., Ladror,U.S., Holzman,T.F. and AbadZapatero,C. (1998) *Biochemistry*, **37**, 7103–7112.
32. Yu,L., Petros,A.M., Schnuchel,A., Zhong,P., Severin,J.M., Walter,K., Holzman,T.F. and Fesik,S.W. (1997) *Nature Struct. Biol.*, **4**, 483–489.
33. Eady,E.A., Ross,J.I. and Cove,J.H. (1990) *J. Antimicrobial Chemother.*, **26**, 461–465.
34. Wu,J.C. and Santi,D.V. (1987) *J. Biol. Chem.*, **262**, 4778–4786.
35. King,M., Ton,D. and Redman,K.L. (1999) *Biochem. J.*, **337**, 29–35.