Protein Engineering vol.15 no.2 pp.101–108, 2002

In silico identification, structure prediction and phylogenetic analysis of the 2′-*O*-ribose (cap 1) methyltransferase domain in the large structural protein of ssRNA negative-strand viruses

Janusz M.Bujnicki[1,2,3] and Leszek Rychlewski[1,2]

[1]Bioinformatics Laboratory, International Institute of Cell and Molecular Biology, ul. ks. Trojdena 4, 02–109 Warsaw and
[2]BioInfoBank, ul. Limanowskiego 24A, 60–744 Poznan, Poland
[3]To whom correspondence should be addressed, at the Warsaw address.
E-mail: iamb@bioinfo.pl

The *Escherichia coli RrmJ* gene product has recently been shown to be the 23S rRNA:U2552 specific 2′-*O*-ribose methyltransferase (MTase) (RrmJ). Its structure has been solved and refined to 1.5 Å resolution, demonstrating conservation of the three-dimensional fold and key catalytic side chains with the vaccinia virus VP39 protein, which functions as an mRNA 5′m^{7}G-cap-N-specific 2′-*O*-ribose MTase. Using the amino acid sequence of RrmJ as an initial probe in an iterative search of sequence databases, we identified a homologous domain in the sequence of the L protein of non-segmented, negative-sense, single-stranded RNA viruses. The plausibility of the prediction was confirmed by homology modeling and checking whether important residues at substrate/ligand-binding sites were conserved. The predicted structural compatibility and the conservation of the active site between the novel putative MTase domain and genuine 2′-*O*-ribose MTases, together with the available results of biochemical studies, strongly suggest that this domain is a 5′m^{7}G-cap-N-specific 2′-*O*-ribose MTase (i.e. the cap 1 MTase). Evolutionary relationships between these proteins are also discussed.

Keywords: bioinformatics/homology modeling/ molecular evolution/mRNA capping/protein structure/ RNA methyltransferase

Introduction

Methylated 5′-terminal cap structures have been described in most eukaryotic and many viral mRNAs. In all cap structures, including the 'minimal' cap 0 [m^{7}G(5′)ppp(5′)N)] an N^7-methylguanosine (m^{7}G) is attached through a 5′–5′ triphosphate bridge to the penultimate nucleoside. In some molecules additional 2′-*O*-ribose methylations are found at the penultimate and the antepenultimate nucleosides, forming the cap 1 [m^{7}G(5′)ppp(5′)Nm] and cap 2 [m^{7}G (5′)ppp(5′)NmpNm] structures, respectively (Banerjee, 1980; Varani, 1997; Furuichi and Shatkin, 2000). Cap 0 is usually synthesized in the nucleus by the sequential action of three enzymatic activities: mRNA triphosphatase, guanylyltransferase and m^{7}G-methyltransferase (MTase). In HeLa cells, the cap 1 and cap 2 structures are generated sequentially by two distinct enzymes, localized in the nucleus and the cytoplasm, respectively (Langberg and Moss, 1981). The monomethylated cap structure is essential for efficient initiation of translation and mRNA stability and transport from the nucleus to the cytoplasm. In contrast, the cap 2 structure (as in small nuclear RNAs) facilitates import of RNA from the cytoplasm to the nucleus (Fischer and Luhrmann, 1990).

While the 'minimal' capping apparatus from yeast and vaccinia has been extensively characterized, little is known about the cap 1 and cap 2 MTases except for the vaccinia virus protein VP39, which acts as a cap 1 MTase and a non-catalytic, smaller subunit of the heterodimeric poly(A) polymerase. Structural and functional studies on VP39 have highlighted many aspects of this protein's 2′-*O*-MTase function (Hodel *et al.*, 1996, 1998, 1999). All the cap MTases characterized so far utilize *S*-adenosylmethionine (AdoMet) as the methyl group donor and belong to a family of proteins with a common, albeit weakly conserved, signature of the cofactor-binding region (Kagan and Clarke, 1994; Cheng and Blumenthal, 1999). Therefore, one may envisage identification of other classes of cap MTases and phylogenetic classification based on their degree of similarity to the VP39 protein. However, sequence similarity searches failed to identify any homologs of VP39 outside pox viruses, presumably due to the extreme divergence of subfamilies of various viral and cellular cap 1 MTases. On the other hand, extensive sequence analysis of proteins from positive-strand RNA viruses carried out by Koonin and co-workers allowed them to delineate a putative MTase domain, tentatively associated with the activity required for cap formation (Koonin *et al.*, 1992; Rozanov *et al.*, 1992; Koonin, 1993). Nevertheless, it was unclear which of the cap methylations are carried out by the predicted domain, since its sequence did not show pronounced similarity to any other cap MTase subfamilies.

Recently, the structure of another 2′-*O*-MTase, namely the 23S rRNA:U2552 specific RrmJ MTase from *Escherichia coli*, has been solved at 1.5 Å resolution, revealing the three-dimensional fold and architecture of the active site common with VP39, despite the fact that these proteins lack overall sequence similarity (Bugl *et al.*, 2000). Also, in the mammalian reovirus λ2 protein structure solved at 3.6 Å resolution, two AdoMet-dependent MTase-like domains have been delineated (Reinisch *et al.*, 2000). Remarkably, owing to the lack of similarity to other proteins in the sequence databases, the authors could not unambiguously determine which of these domains is the cap 0 and which is the cap 1 MTase and, according to our analysis, their tentative assignment should be reversed (Bujnicki and Rychlewski, 2001). Problems with identification of the sequence elements specific for 2′-*O*-ribose MTases are a good illustration of the degree of divergence in this protein family.

The non-segmented, negative-sense, single-stranded RNA viruses [order *Mononegavirales* (MNV)] comprise many human and animal pathogens of significant epidemiological importance, including respiratory syncytial virus, measles, mumps, rabies, parainfluenza, vesicular stomatitis and Marburg and Ebola viruses, and several plant pathogens (abbreviations of the names of viruses analyzed in this work are given in Table I). Complete nucleotide sequences have been determined for more than 20 MNV, revealing five common genes main-

Table I. The protein sequences used in the phylogenetic analysis, with the corresponding non-redundant database accession numbers and positions of the MTase domain in the sequence

Name	GI	Host	Position
EBOM	10313999	Ebola virus strain Mayinga	1808–2009
EBOSM	8477362	Ebola virus strain Sudan Maleo-79	1805–2006
MABVP	464697	Marburg virus (strain Popp)	1924–2123
MABVM	9627506	Marburg virus strain Musoke	1924–2123
BEFV	10086573	Bovine ephemeral fever virus	1690–1890
VSVSJ	1173173	Vesicular stomatitis virus (strain San Juan).	1643–1843
VSVJH	133616	Vesicular stomatitis virus (ser. N.Jersey / strain Hazelhurst)	1643–1843
VSVJO	133617	Vesicular stomatitis virus (ser. N.Jersey / strain Ogden)	1643–1843
RABVR	8648086	Rabies virus strain RC-HL	1677–1877
RABVS	133609	Rabies virus (strain SAD B19)	1677–1877
APVC	1688090	Avian pneumovirus strain CVL 14/1	1664–1857
HRSVA	133602	Human respiratory syncytial virus (strain A2)	1823–2014
BRSVA	3643021	Bovine respiratory syncytial virus (strain A51908)	1819–2010
NDVB	133604	Newcastle disease virus (strain Beaudette C/45)	1748–1964
SV5	3914878	Simian virus 5 (strain W3)	1778–1994
SV41	548838	Simian parainfluenza virus 41.	1788–2004
PI2HT	133605	Human parainfluenza virus 2 (strain Toshiba)	1783–1999
MVJL	7861769	Mumps virus strain Jeryl Lynn	1784–2000
PRV	2121316	Porcine rubulavirus	1778–1994
HV	9630565	Hendra virus	1813–2023
TPMV	9634976	Tupaia paramyxovirus	1838–2048
SVE	133612	Sendai virus (strain Enders)	1774–1984
PI1HT	4566775	Human parainfluenza virus 1 strain C35	1774–1984
PI3BS	6760241	Bovine parainfluenza virus 3 strain Shipping Fever	1778–1988
PI3H4	133606	Human parainfluenza 3 virus (strain NIH 47885)	1778–1988
RINDR	730620	Rinderpest virus (strain RBOK)	1758–1964
MEASA	548836	Measles virus (strain AIK-C)	1758–1964
CDVA	5733648	Canine distemper virus strain A75/17	1758–1964
PDVU	1707654	Phocine distemper virus strain Ulster 88	1758–1964
TlyA	7340781	*M.ulcerans*	61–248
YgdE	418436	*E.coli* K12	183–357
RrmJ	120571	*E.coli* K12	30–209
GCRV	9971835	Grass carp reovirus	481–664
HRV	67152	Human reovirus (1ej6 in PDB)	477–660

tained in highly similar order (N–P–M–G–L), varied in some cases by insertions (Pringle and Easton, 1997; Conzelmann, 1998). One of the *bona fide* homologous components of the viral ribonucleoprotein core is the large (L) protein, which functions as the RNA-dependent RNA polymerase (Tordo *et al.*, 1992). Other activities attributed to the L protein include mRNA capping, cap 0 and cap 1 methylation, poly(A) polymerase and protein kinase.

The capping and poly(A) polymerase reactions are all tightly coupled to RNA polymerization and fail to respond to exogenous substrates, hence the sequence–structure–function relationships of individual enzymatic activities have been difficult to establish. Nevertheless, several different temperature-sensitive mutants in the L protein exhibited complementation, suggesting that the particular activities are linked with distinct domains (Flamand and Bishop, 1973). Amino acid sequence alignments revealed several short, closely spaced segments, which are well conserved among all L proteins. Some of these motifs were recognized as parts of the RNA-dependent RNA polymerase and two putative protein kinase domains (McClure and Perrault, 1989). However, no systematic analysis has hitherto been performed that would allow the precise delineation of the exact boundaries and assignment of function to the individual domains. The lack of established domain structure in the L protein of MNV might have been one of the causes of problems of Zanotto *et al.* in phylogenetic analysis of the presumed polymerase domain (Zanotto *et al.*, 1996).

In vesicular stomatitis virus (VSV), the monomethylated G(5′)ppp(5′)Am structure, generated by the cap 1 MTase, is a preferred substrate for the cap 0 (m7G) MTase (Testa and Banerjee, 1977; Hammond and Lesnaw, 1987). In this respect, the VSV system is distinct from the vaccinia and reovirus systems, in which the cap 0 structure is necessary for cap 1 methylation to occur (Furuichi and Shatkin, 2000). It has been shown that the VSV mRNAs lacking the cap 0 structures are poor templates for protein synthesis *in vitro* (Testa and Banerjee, 1977; Horikami and Moyer, 1982). Therefore, inhibitors of either cap 0 or cap 1 MTase activities of the L protein may be used as antiviral drugs, but to date the regions involved in catalysis of the methyl transfer have not been mapped on to the primary structure of the L protein, which hampers knowledge-based drug design. Hence, *in silico* structure prediction of the domains responsible for the L protein activities is of obvious importance.

In this paper, we report the identification of a putative MTase domain in the L protein. Based on extensive bioinformatics analysis, including iterative database searches, structure prediction and molecular modeling, we demonstrate that this domain shares key features with known 2′-*O*-ribose MTases. We analyze sequence and structural similarities to other 2′-*O*-ribose MTases, propose possible roles for conserved residues and discuss evolutionary relationships among the representative members of the family of the putative cap 1 MTases of MNV.

Materials and methods

Sequence analysis

The non-redundant (nr) database at NCBI was extensively searched with the PSI-BLAST algorithm (Altschul *et al.*, 1997), using the sequence of RrmJ, VP39 and subsequently some newly retrieved representative sequences as queries. The expectation (e)-value cutoff value was varied in the range 10^{-6}–10^{-12}, depending on the visual inspection of the alignments reported by the program for different queries. Low-complexity sequence regions were left unmasked. Full-length protein sequence alignments were reconstructed using ClustalX (Thompson *et al.*, 1997) based of the degapped PSI-BLAST output processed using BIB-VIEW (http://bioinfo.pl/bibview.pl). Secondary and tertiary structure predictions were carried out via the protein structure prediction MetaServer-Pcons interface (http://bioinfo.pl/meta/) (Bujnicki *et al.*, 2001; Lundstrom *et al.*, 2001) using 150–400 amino acid fragments of the L protein as queries.

Molecular modeling

Homology modeling was carried out following a modified version of the 'multiple models' approach (Pawlowski *et al.*, 1997) using MODELLER (Sali and Blundell, 1993) to generate several alternative preliminary models based on threading-derived pairwise target–template alignments and PROMODII (Guex and Peitsch, 1997) to merge the best-scored fragments of preliminary models into the final structure. The preliminary models were obtained using unrefined pairwise alignments reported by PSI-BLAST (Altschul *et al.*, 1997), FFAS (Rychlewski *et al.*, 2000), 3DPSSM (Kelley *et al.*, 2000), BIOINBGU (Fischer, 2000) and GenThreader (Jones, 1999) and merged using secondary structure prediction results and energy evaluation to resolve ambiguities. The structure of an insertion 20 aa long was predicted using the *ab initio* protein folding server I-SITES/ROSETTA (Simons *et al.*, 1997) and manually inserted into the homology-modeled core. Energy minimization was carried out using GROMOS96 (Scott *et al.*, 1999) until all inconsistencies in geometry were rectified and all the short contacts were relieved. The stereochemical and energetic properties of modeling intermediates and of the final model were evaluated using WHATCHECK (Hooft *et al.*, 1996) and VERIFY3D (Eisenberg *et al.*, 1997). Semi-automated and manual manipulations with protein structures and sequence–structure alignments were conducted using SWISS-PDB VIEWER (Guex and Peitsch, 1997).

Phylogenetic analysis

Phylogenetic inference was carried out using the conserved regions of the refined sequence–structure alignment of catalytic domains of the predicted viral cap 1 MTase based on the neighbor-joining method (Saitou and Nei, 1987). A corrected distance matrix was calculated from sequences according to the JTT model (Jones *et al.*, 1992). Bootstrapping analysis was performed, generating 100 replicates of the sequence alignment. The majority-rule consensus tree was visualized using TREEVIEW (Page, 1996).

Results and discussion

Database searches and multiple sequence alignment

The evolutionary relationships among various nucleic acid MTase families have been studied (Gustafsson *et al.*, 1996), but for a long time there was no indication of any structural or evolutionary relatedness between various families of 2′-*O*-ribose MTases. As a part of a larger project, aiming at identification and classification of novel RNA MTases among the uncharacterized or putative proteins in sequence databases, we conducted exhaustive database searches using sequences and structures of *bona fide* 2′-*O*-ribose MTases as queries. VP39 and its close homologs turned out to be a poor query, since even with the relaxed cutoff (e-value <0.1) all PSI-BLAST (Altschul *et al.*, 1997) searches converged after several iterations yielding no sequences except the 2′-*O*-ribose MTases from poxviruses. Conversely, when the sequence of *E.coli* RrmJ was used as an initial probe with a default cutoff (e-value = 10^{-3}), the search 'exploded', ultimately reporting similarities to many AdoMet-dependent MTase families, including enzymes known to modify proteins, various small molecules and bases in RNA and DNA (data not shown). Therefore, we tested alternative cutoff values, monitoring the addition of MTases exhibiting conservation of both the AdoMet-binding region and the ribose-binding residues common to VP39 and RrmJ (Bugl *et al.*, 2000).

The stringent cutoff of 5×10^{-6} turned out to be optimal for the search for putative 2′-*O*-ribose MTases initiated with the RrmJ sequence, since all hits reported retained the residues implicated in ribose binding in structurally characterized members of the family. The second iteration revealed significant similarity (e-value = 9×10^{-7}) of RrmJ and its homologs to the putative methyltransferase domain of the large non-structural protein of ssRNA positive-strand viruses (Rozanov *et al.*, 1992; Koonin, 1993; Koonin and Dolja, 1993). Remarkably, whereas Koonin and co-workers could not predict the function of cap 0 or cap 1 MTase (or both) based on their analysis of amino acid conservation patterns, our study of recently solved structures strongly suggests that the MTase domain that they identified shares the key features of the 2′-*O*-ribose MTase and therefore most likely functions as the cap 1 MTase. Iterating the PSI-BLAST search resulted in the accumulation of numerous closely related sequences of the putative cap 1 MTase of positive-strand viruses and members of the TlyA family (Aravind and Koonin, 1999). Further, in the third iteration a similarity of 2′-*O*-MTases to a fragment of the RNA polymerase (L protein) sequence from the mumps virus, which belongs to the genus *Rubulavirus* of ssRNA negative-strand viruses, has been reported with a score of 4×10^{-6} and in further iterations more related sequences from rubulaviruses and paramyxoviruses were retrieved with e-values as high as 10^{-13}. We have also detected a new putative 2′-*O*-MTase YgdE in *E.coli* and related bacteria (Bujnicki and Rychlewski, 2000).

Visual analysis of the PSI-BLAST output saved as multiple sequence alignment revealed that all sequence fragments of the L protein of negative-strand viruses retained the presumptive ribose-binding C-terminal subdomain with three conserved side chains, but lacked the N-terminal AdoMet-binding subdomain with the characteristic glycine-rich motif I. Consequently, the fourth residue typical of 2′-*O*-MTases, namely invariant Lys from motif X, which in RrmJ is most proximal to the N-terminus, could not be observed in the initial alignment. To test the possibility that the L protein of viruses from the family Paramyxoviridae and possibly from other MNV includes an intact domain similar to 2′-*O*-MTases, we followed two distinct strategies. Firstly, we carried out a series of reciprocal PSI-BLAST searches with default cutoff values, using the reported fragments of the L protein with varying extensions at both termini to localize the previously

missed conserved Lys residue and motif I. Secondly, we used the same sequence fragments to carry out sequence-to-structure threading in order to determine if the domain under consideration is indeed similar to the structurally characterized MTases. Both the sequence- and structure-based strategy turned out to be successful. Querying PSI-BLAST with the fragment of the mumps virus L protein spanning residues 1764–2010 resulted in a multiple sequence alignment encompassing both of the previously identified C-termini of the catalytic domain, and also two additional sequence blocks, one of which contained an invariant Lys residue and the other closely resembled the GxGxG pattern typical for the conserved AdoMet-binding loop (Kagan and Clarke, 1994; Fauman *et al.*, 1999). On the other hand, threading algorithms evaluated the MTase fold as the most likely candidate for the structure of the query sequence. For instance, 3D-PSSM reported similarity of the above-mentioned region to the RrmJ structure with a highly significant score of 0.00271 (the detailed results of the search reported by all servers are available online at http://bioinfo.pl/meta/target.pl?id=2357). The results of the threading analysis carried out for other sequences in the data set of potential MTase domains extracted from the L proteins confirmed their high propensity to assume the MTase fold. Further, carrying out PSI-BLAST searches for the same sequence fragments resulted in the retrieval of a bulk of sequences from MNV in the earliest iterations, followed by eukaryotic and prokaryotic homologs of the RrmJ MTase and the putative cap 1 MTases from positive-strand viruses (data not shown).

Interestingly, in the sixth iteration of the search initiated with the PI2HT putative MTase domain a hit to the grass carp reovirus (GCRV) VP1 protein has been reported. This region in GCRV VP1 corresponds to a part of the MTase I domain in the recently solved crystal structure of the human reovirus λ2 protein. Extending the profile-to-profile alignment of sequences from the MNV to the pair of reoviral proteins using FFAS (Rychlewski *et al.*, 2000) resulted in a perfect match of the four invariant residues (Figure 1). No significant overall sequence similarity or generally conserved residues was observed when we attempted to align the 2′-*O*-MTase family to the MTase II domain. This result supports our independent prediction, based on analysis of protein structures alone, that the reoviral MTase I domain is more similar to 2′-*O*-ribose

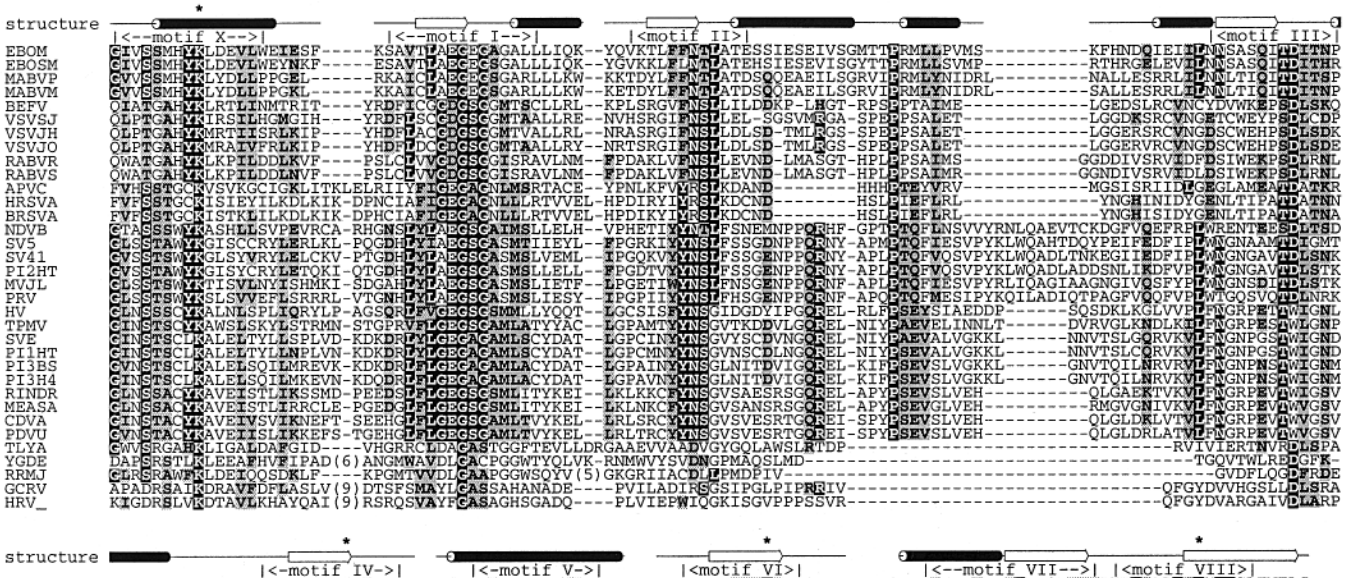

Fig. 1. Multiple alignment of the predicted cap 1 MTases with representative sequences of several genuine and putative 2′-*O*-ribose MTase subfamilies. For clarity of presentation, only representative sequences were chosen from closely related species. Conserved motifs are labeled according to the nomenclature described for the 'orthodox' AdoMet-dependent MTase superfamily (Fauman *et al.*, 1999). Identical residues are highlighted in black and conservatively substituted residues are highlighted in gray.

cap 1 MTases, while the MTase II domain most likely functions as the m^7G-specific cap 2 MTase (Bujnicki and Rychlewski, 2001).

The alignment presented in Figure 1 includes 30 sequences from MNV, whose pairwise amino acid identity was <90%, and representatives of other genuine and putative 2′-*O*-ribose MTases. Several blocks of conserved residues can be delineated, which correspond to nine motifs typical for the 'orthodox' MTases. To the best of our knowledge, functional roles for most of the highly conserved residues in the sequence of the domain analyzed in this study have not yet been reported. By analogy with the known MTase structures, we suggest that motifs I–III are involved in binding of the methyl group donor, while the four invariant acidic and basic residues from motifs X, IV, VI and VIII form the core of the active site.

In addition to the conserved motifs common for the majority of MTases, the putative cap 1 MTase domain in the L protein contains a variable region between motifs II and III. This region is absent from other 2′-*O*-ribose MTases. Interestingly, proteins from the α subfamily of DNA:m^6A MTases possess a variable region between motifs II and III, which forms an autonomous domain (the so-called 'target recognition domain', TRD) implicated in recognition of the specific sequence in the DNA (Tran *et al.*, 1998). The chemotaxis receptor methyltransferase CheR also possesses an additional small domain in the same region; this domain is dissimilar to the TRD of α-m^6A MTases and is involved in specific interactions with the methylated receptor (Djordjevic and Stock, 1998). However, the 'variable' domains of α-m^6A and CheR MTases are of relatively constant length and exhibit conservation of the key hydrophobic residues, suggesting that a common fold is retained within each family, while the length of the insertion present in the predicted cap 1 MTase domain is extremely variable. Moreover, we could detect only one Pro residue that may be conserved in this subfamily, which suggests that this part of the protein is structurally variable.

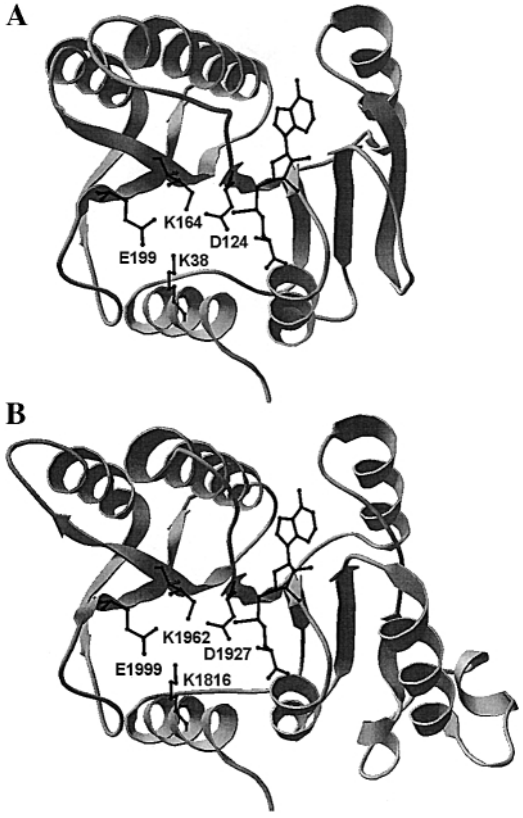

Fig. 2. Comparison of cartoon diagrams of catalytic domain structures of (**A**) RrmJ (Bugl *et al.*, 2000) and (**B**) modeled cap 1 MTase of Ebola virus. The functionally important residues are shown in wireframe representation.

Molecular modeling

It is known that detection of distant homologs either by sequence searches or by threading does not necessarily translate into correct alignments, from which sequence–structure–function relationships could be accurately inferred (Smith *et al.*, 1997). The independent verification of threading, refinement of sequence alignments with known structures and estimation of reliability of certain regions in the alignments can be addressed by modeling. For that reason and to gain insight into the molecular basis of intriguing similarities in the active site and into considerable differences in the sequence of the 'variable' regions, the structure of the putative 2′-*O*-ribose MTase of Ebola virus was predicted by homology modeling, using the coordinates of RrmJ as the template. We did not use the structures of the vaccinia virus and reovirus cap 1 MTases, since their sequences exhibited much lower similarity to the whole range of potential target sequences from MNV and we observed large discrepancies between alternative threading-based alignments including these structures. Besides, we could not identify a common cap-binding motif in these two MTases or a region at the N- or C-terminus of the predicted MTase domain in the L protein that would exhibit similarity to the cap-binding site of either cap 1 MTase of known structure. Therefore, only the conserved core spanning the AdoMet- and ribose-binding sites could be predicted with confidence. Still, the modeling was not trivial, since RrmJ and the viral proteins shared low sequence identity and therefore we resorted to a modified version of the 'multiple models approach' of Pawlowski *et al.* (Pawlowski *et al.*, 1997) (see Materials and methods) to ensure that the number of possibly misaligned sequence segments between the target and the template was reduced to the minimum. The preliminary models showed very good agreement in the predicted core elements; there were only few relative shifts of sequence segments, almost exclusively in the peripheral structures, which very strongly supports the presented fold-recognition results (data not shown).

The final averaged and optimized model passed all the tests implemented in the stereochemistry-evaluating WHATCHECK suite (Hooft *et al.*, 1996) and in the VERIFY3D program, which uses contact potentials to assess whether the modeled amino acid residues occur in the environment typical for globular proteins with hydrophobic core and solvent-exposed surface (Eisenberg *et al.*, 1997). It is worth emphasizing that the stereochemistry of even a plain wrong model can be refined to an acceptable degree; however, the calculation of energy based on observed contacts would still indicate that the polypeptide chain is misfolded. Moreover, the reasonable energies are rarely observed for misfolded structures. Thus, the scores reported for our model by WHATCHECK (*Z*-score –4.1) and VERIFY3D (average score 0.3, no regions scored lower than 0) suggest that both its three-dimensional fold and the conformation of individual residues are reasonable.

Sequence–structure–function relationships

The modeled putative cap 1 MTase domain of the Ebola virus L protein, (aa 1808–2009) resembles its modeling template, structure of the 2′-O-ribose RrmJ MTase (Figure 2). All

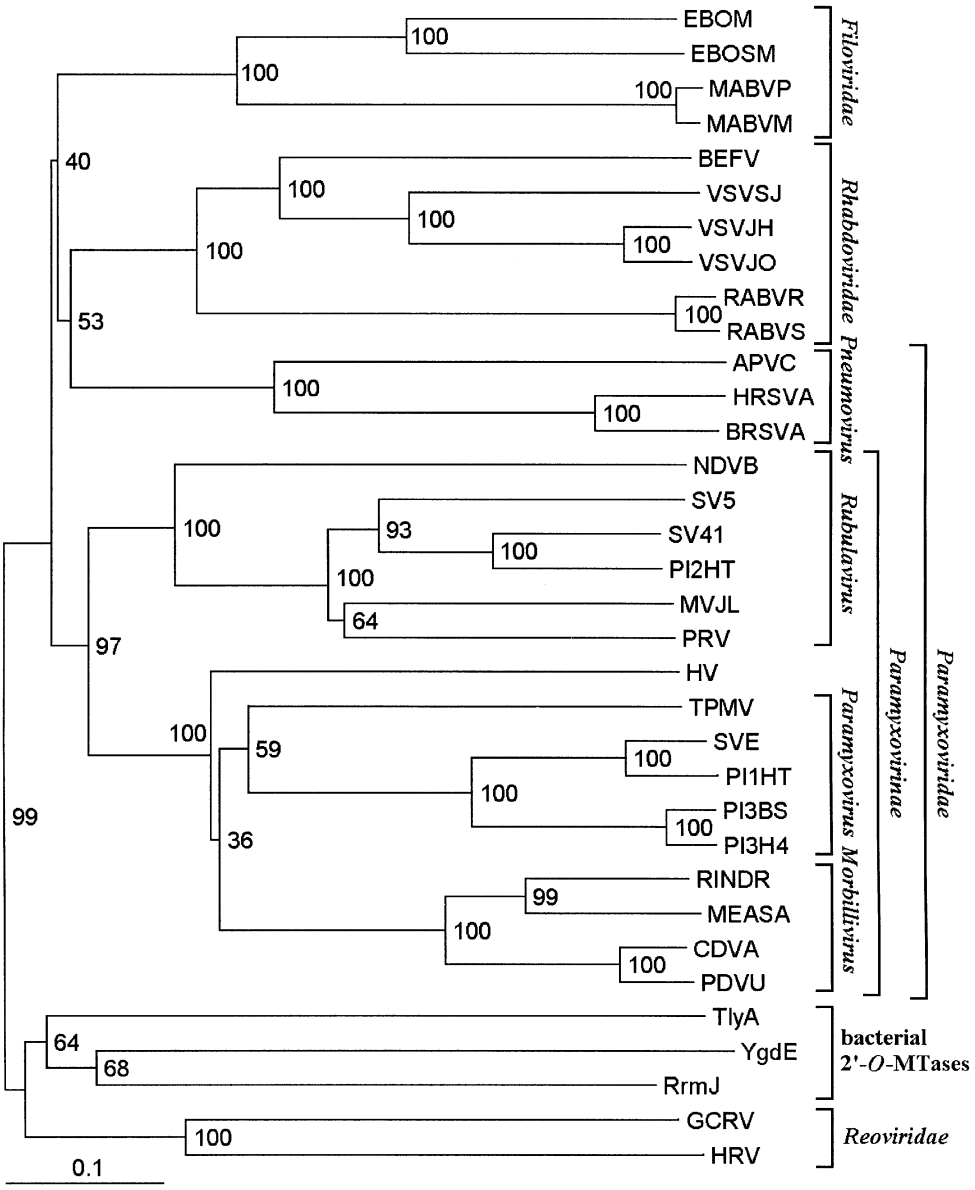

Fig. 3. The phylogenetic tree of the predicted cap 1 MTase domain of MNV. Names were taken from Table I. The names of the taxons are outlined. The numbers at the nodes indicate the statistical support of the branching order by the bootstrap criterion. The bar at the bottom of the phylogram indicates the evolutionary distance, to which the branch lengths are scaled based on the estimated divergence.

insertions in the sequence of the viral MTase are localized in the connectors between the secondary structure elements of RrmJ and are modeled as loops; these modifications do not interfere with the active site of the MTase. The only major difference between the two structures is the insertion in the viral protein that forms an elaboration of the edge of the common β-sheet, distant from the cofactor- and ribose-binding sites. It is rather unlikely that it is involved in recognition of the substrate; however, in the absence of experimental data it is not possible to propose any function for this region other than interactions with other proteins, possibly other domains of the L protein.

The availability of the model allows the proposal of residues that may be involved in interactions with the cofactor and the target RNA molecule. The predicted AdoMet-binding residues include 1836–GEGAGAL–1842 (motif I), 1856–NTL–1858 (motif II), 1902–TDIT–1905 (motif III) and two invariant residues that participate in the methyl transfer reaction, i.e. K1816 (motif X) and D1927 (motif IV). The other two invariant catalytic residues are K1962 (motif VI) and E1999 (motif VIII). Other side chains that may interact with the substrate include E1929, S1994, R1996, S1997 and Y2001. Not surprisingly, most of these residues are conserved in the viral proteins, although this conservation does not necessarily extend to all 2′-*O*-ribose MTases, presumably because of the differences in substrate specificity. The role of other conserved residues in the loop comprising aa 1989–1999 is most likely to stabilize the structure of that region. These predictions can be tested by site-directed mutagenesis experiments.

Phylogenetic analysis

In spite of the low degree of sequence similarity among putative cap 1 MTases from individual families of MNV, the present study shows that they all originate from a common ancestral enzyme. To evaluate the evolutionary relationships between the MTase domains and verify if they are consistent

with relationships between the polymerase domain and the established taxonomy of MNV (http://www.ncbi.nlm.nih.gov/ICTV/overview/negssrna.html) (van Regenmortel *et al.*, 2000), phylogenetic trees were inferred from the alignment as described in Materials and methods. A consensus tree inferred using the neighbor-joining method with default parameters is shown in Figure 3.

The aligned sequences of MNV can be divided into six lineages, based on both evolutionary branching order topology and the observed distribution of shared sequence signatures, such as specific insertions/deletions or characteristic amino acid residue patterns. These six lineages can be grouped in two major clades: the genera *Rubulavirus*, *Paramyxovirus* and *Morbillivirus*, i.e. the subfamily Paramyxovirinae and the families Rhabdoviridae and Filoviridae that cluster together with the genus *Pneumovirus*. This topology closely resembles the order of taxa recognized in the universal system of virus taxonomy (van Regenmortel *et al.*, 2000), albeit with two exceptions. Among Paramyxovirinae, the Hendra virus (HV, classified as a species in the genus *Morbillivirus*) forms an outgroup to both the genus *Paramyxovirus*, with the L protein of Tupaia paramyxovirus reported as the first BLAST hit (score = 160 bits, e-value = 1×10^{-38}) and the genus *Morbillivirus*, with the measles virus reported at a lower position in the BLAST ranking (score = 136 bits, e-value = 3×10^{-31}). The bootstrap confidence of this topology is low (36%); nevertheless, this observation suggests that from the 'evolutionary point of view' the predicted cap 1 MTase domain of HV is as old as the two other lineages. Interestingly, our analysis suggest that in respect to the phylogeny of cap 1 MTases, the family Paramyxoviridae is split: the subfamily Pneumovirinae seems more closely related to the families Rhabdoviridae and Filoviridae than to the subfamily Paramyxovirinae. This is supported both by high bootstrap values (99%) and by the result of BLAST searches initiated with the profile based on the sequences of cap 1 MTase domain from APVC, HRSVA and BRSVA (for rabies virus the score = 46.5 bits, e-value = 3×10^{-4}, while the members of the subfamily Paramyxovirinae are reported with e-values >100). Interestingly, in the phylogenetic tree based on the polymerase domains reported by Stec *et al.* (Stec *et al.*, 1991), HRSVA branches out earlier than all other genera within the subfamily Paramyxovirinae, while members of the family Rhabdoviridae (rabies virus and vesicular stomatitis virus) form an outgroup. To clarify further the issue of possible mosaic origin of the L protein of pneumoviruses and its relationship to other MNV, a more extensive analysis involving phylogenomic analysis of all individual domains is necessary; however, such study is beyond the scope of this paper.

Conclusions

Using iterative searches of sequence databases and sequence-to-structure threading, we have detected the presence of a putative 2′-*O*-ribose (cap 1) MTase domain in the L protein of MNV and predicted the key residues involved in binding and the methyl transfer reaction. The homology of the putative cap 1 MTase domain of MNV to other 2′-*O*-ribose MTases, which served as a basis for comparative molecular modeling and phylogenetic analysis, is not *prima facie* evident and is also not recognizable by standard, non-iterated algorithms for pairwise sequence comparison. Therefore, the alignment presented in this work will be a good starting point for the creation of specific sequence profiles used for identification of other 2′-*O*-ribose MTases in genomic data and the structural model may be useful in development of antiviral drugs. Our results will help in understanding better the biology of MNV and their relationship to other viruses. Further biochemical and structural studies of the L proteins from MNV, including site-directed mutagenesis and X-ray crystallography, may be used in clarifying the structural constraints imposed by the function of these enzymes and their divergence from the common ancestor shared with many ribose MTases and other AdoMet-dependent enzymes. In addition, our phylogenetic analysis revealed the unusual relationship of the cap 1 MTase domain of pneumoviruses to other paramyxoviruses, which suggests that the evolution of the L proteins should be studied based on comparison of a series of phylogenies inferred from alignments of all individual domains.

Acknowledgements

This work was supported by KBN (grant 8T11F01019 to J.M.B.) and BioInfoBank.

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
Aravind,L. and Koonin,E.V. (1999) *J. Mol. Evol.*, **48**, 291–302.
Banerjee,A.K. (1980) *Microbiol. Rev.*, **44**, 175–205.
Bugl,H., Fauman,E.B., Staker,B.L., Zheng,F., Kushner,S.R., Saper,M.A., Bardwell,J.C. and Jakob,U. (2000) *Mol. Cell*, **6**, 349–360.
Bujnicki,J.M. and Rychlewski,L. (2000) *Acta Microbiol. Pol.*, **49**, 253–260.
Bujnicki,J.M. and Rychlewski,L. (2001) *Genome Biol.*, **2**, 38.1–38.6
Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) *Bioinformatics*, **17**, 750–751.
Cheng,X. and Blumenthal,R.M. 1999. *S-Adenosylmethionine-dependent Methyltransferases: Structures and Functions*. World Scientific, Singapore.
Conzelmann,K.K. (1998) *Annu. Rev. Genet.*, **32**, 123–162.
Djordjevic,S. and Stock,A.M. (1998) *Nature Struct. Biol.*, **5**, 446–450.
Eisenberg,D., Luthy,R. and Bowie,J.U. (1997) *Methods Enzymol.*, **277**, 396–404.
Fauman,E.B., Blumenthal,R.M. and Cheng,X. (1999) In Cheng,X. and Blumenthal,R.M. (eds), *S-Adenosylmethionine-dependent Methyltransferases: Structures and Functions*. World Scientific, Singapore, pp. 1–38.
Fischer,D. (2000) *Pac. Symp. Biocomput.*, 119–130.
Fischer,U. and Luhrmann,R. (1990) *Science*, **249**, 786–790.
Flamand,A. and Bishop,D.H. (1973) *J. Virol.*, **12**, 1238–1252.
Furuichi,Y. and Shatkin,A.J. (2000) *Adv. Virus Res.*, **55**, 135–184.
Guex,N. and Peitsch,M.C. (1997) *Electrophoresis*, **18**, 2714–2723.
Gustafsson,C., Reid,R., Greene,P.J. and Santi,D.V. (1996) *Nucleic Acids Res.*, **24**, 3756–3762.
Hammond,D.C. and Lesnaw,J.A. (1987) *Virology*, **159**, 229–236.
Hodel,A.E., Gershon,P.D., Shi,X. and Quiocho,F.A. (1996) *Cell* **85**, 247–256.
Hodel,A.E., Gershon,P.D. and Quiocho,F.A. (1998) *Mol. Cell*, **1**, 443–447.
Hodel,A.E., Quiocho,F.A. and Gershon,P.D. (1999) In Cheng,X. and Blumenthal,R.M. (eds), *S-Adenosylmethionine-dependent Methyltransferases: Structures and Functions*. World Scientific, Singapore, pp. 255–282.
Hooft,R.W., Vriend,G., Sander,C. and Abola,E.E. (1996) *Nature*, **381**, 272
Horikami,S.R. and Moyer,S.A. (1982) *Proc. Natl Acad. Sci. USA*, **79**, 7694–7698.
Jones,D.T. (1999) *J. Mol. Biol.*, **287**, 797–815.
Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) *Comput. Appl. Biosci.*, **8**, 275–282.
Kagan,R.M. and Clarke,S. (1994) *Arch. Biochem. Biophys.*, **310**, 417–427.
Kelley,L.A., McCallum,C.M. and Sternberg,M.J. (2000) *J. Mol. Biol.*, **299**, 501–522.
Koonin,E.V. (1993) *J. Gen. Virol.*, **74**, 733–740.
Koonin,E.V. and Dolja,V.V. (1993) *Crit. Rev. Biochem. Mol. Biol.*, **28**, 375–430.
Koonin,E.V., Gorbalenya,A.E., Purdy,M.A., Rozanov,M.N., Reyes,G.R. and Bradley,D.W. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 8259–8263.
Langberg,S.R. and Moss,B. (1981) *J Biol. Chem.*, **256**, 10054–10060.
Lundstrom,J., Rychlewski,L., Bujnicki,J.M. and Elofsson,A. (2001) *Protein Sci.*, **10**, 2354–2362.
McClure,M.A. and Perrault,J. (1989) *Virology*, **172**, 391–397.
Page,R.D. (1996) *Comput. Appl. Biosci.*, **12**, 357–358.

Pawlowski,K., Jaroszewski,L., Bierzynski,A. and Godzik,A. (1997) *Pac. Symp. Biocomput.*, 328–339.
Pringle,C.R. and Easton,A.J. (1997) *Semin. Virol.*, **8**, 49–57.
Reinisch,K.M., Nibert,M.L. and Harrison,S.C. (2000) *Nature*, **404**, 960–967.
Rozanov,M.N., Koonin,E.V. and Gorbalenya,A.E. (1992) *J. Gen. Virol.*, **73**, 2129–2134.
Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) *Protein Sci.*, **9**, 232–241.
Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.*, **4**, 406–425.
Sali,A. and Blundell,T.L. (1993) *J. Mol. Biol.*, **234**, 779–815.
Scott,W.R.P. *et al.* (1999) *J. Phys. Chem.*, **103**, 3596–3607.
Simons,K.T., Kooperberg,C., Huang,E. and Baker,D. (1997) *J. Mol. Biol.*, **268**, 209–225.
Smith,T.F., Lo,C.L., Bienkowska,J., Gaitatzes,C., Rogers,R.G.J. and Lathrop,R. (1997) *J. Comput. Biol.*, **4**, 217–225.
Stec,D.S., Hill,M.G. and Collins,P.L. (1991) *Virology*, **183**, 273–287.
Testa,D. and Banerjee,A.K. (1977) *J. Virol.*, **24**, 786–793.
Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) *Nucleic Acids Res.*, **25**, 4876–4882.
Tordo,N., de Haan,P., Goldbach,R. and Poch,O. (1992) *Virology*, **3**, 341–357.
Tran,P.H., Korszun,Z.R., Cerritelli,S., Springhorn,S.S. and Lacks,S.A. (1998) *Structure*, **6**, 1563–1575.
van Regenmortel,M.H.V. *et al.* (2000) *Virus Taxonomy: the Classification and Nomenclature of Viruses. The Seventh Report of the International Committee on Taxonomy of Viruses.* Academic Press, San Diego.
Varani,G. (1997) *Structure*, **5**, 855–858.
Zanotto,P.M., Gibbs,M.J., Gould,E.A. and Holmes,E.C. (1996) *J. Virol.*, **70**, 6083–6096.

Received March 27, 2001; revised October 5, 2001; accepted November 11, 2001