Research article

mRNA:guanine-*N7* cap methyltransferases: identification of novel members of the family, evolutionary analysis, homology modeling, and analysis of sequence-structure-function relationships

Janusz M Bujnicki*[1], Marcin Feder[1,2], Monika Radlinska[2] and Leszek Rychlewski[1,3]

Address: [1]Bioinformatics Laboratory, International Institute of Cell and Molecular Biology, ul. ks. Trojdena 4, 02-109 Warsaw, Poland, [2]Institute of Microbiology, Warsaw University, ul. Miecznikowa 1, 02-093 Warsaw, Poland and [3]BioInfoBank, ul. Limanowskiego 24A, 60-744 Poznan, Poland

E-mail: Janusz M Bujnicki* - iamb@bioinfo.pl; Marcin Feder - marcin@bioinfo.pl; Monika Radlinska - monika@bioinfo.pl; Leszek Rychlewski - leszek@bioinfo.pl

*Corresponding author

Published: 22 June 2001

BMC Bioinformatics 2001, **2**:2

This article is available from: http://www.biomedcentral.com/1471-2105/2/2

Received: 21 May 2001
Accepted: 22 June 2001

Abstract

Background: The 5'-terminal cap structure plays an important role in many aspects of mRNA metabolism. Capping enzymes encoded by viruses and pathogenic fungi are attractive targets for specific inhibitors. There is a large body of experimental data on viral and cellular methyltransferases (MTases) that carry out guanine-N7 (cap 0) methylation, including results of extensive mutagenesis. However, a crystal structure is not available and cap 0 MTases are too diverged from other MTases of known structure to allow straightforward homology-based interpretation of these data.

Results: We report a 3D model of cap 0 MTase, developed using sequence-to-structure threading and comparative modeling based on coordinates of the glycine *N*-methyltransferase. Analysis of the predicted structural features in the phylogenetic context of the cap 0 MTase family allows us to rationalize most of the experimental data available and to propose potential binding sites. We identified a case of correlated mutations in the cofactor-binding site of viral MTases that may be important for the rational drug design. Furthermore, database searches and phylogenetic analysis revealed a novel subfamily of hypothetical MTases from plants, distinct from "orthodox" cap 0 MTases.

Conclusions: Computational methods were used to infer the evolutionary relationships and predict the structure of Eukaryotic cap MTase. Identification of novel cap MTase homologs suggests candidates for cloning and biochemical characterization, while the structural model will be useful in designing new experiments to better understand the molecular function of cap MTases.

Background

Transcripts produced by RNA polymerase II are modified at their 5' end by the addition of a methylated 5'-terminal cap structure $m^7G(5')ppp(5')N$, which directs pre-mRNA to the processing and transport pathways in the cell nucleus and regulates both mRNA turnover and the initiation of translation [1,2]. Cap is formed by a series of three enzymatic reactions as follows: an RNA triphosphatase (TPase) removes the γ-phosphate at the 5' end of the transcript, a GTP:RNA guanylyltransferase (GTase) adds a GMP residue to the 5' diphosphate end in a 5'-to-5' orientation, and an RNA:guanine-N7 (m^7G) methyltransferase (cap 0 MTase, for simplicity referred to hereafter as cap MTase) adds the methyl group to the guanine [3]. Mutations in the TPase, GTase, or cap MTase of the yeast capping apparatus that inhibit any of these activities are lethal *in vivo* [4,5,6]. The capping apparatus differs significantly in fungi, metazoans, protozoa and viruses in respect to the evolutionary origin and structure of individual subunits and the subunit composition of the proteins that carry the three activities [3]. Hence, the capping enzymes encoded by viral, fungal and protozoal pathogens are attractive targets for specific inhibitors that would exert limited effect on the host enzyme.

The mechanisms and structures of cellular and viral capping enzymes have been extensively studied. The crystal structures of the GTase from Chlorella virus PBVCV-1 [7] and the TPase from yeast [8] have been solved and used to guide extensive site-directed mutagenesis experiments [9,10,11]. However, there are a few important gaps in our understanding of capping enzymes. For instance, there is a large body of mutagenesis data on cap MTase [5,12,13,14,15,16]; however, its structure remains unknown. Therefore, many important details of the cap binding and m^7G methyltransfer reaction mechanism remain unexplained.

Cap MTase belongs to the AdoMet-dependent MTase superfamily [13], which contains numerous remotely related families of DNA, RNA, protein, and small molecule-modifying enzymes [17]. To date, three-dimensional structures have been determined for more than a dozen MTases. The common fold of the catalytic domain, which bears the AdoMet binding site and the active site, has been identified (reviewed in [18]). Despite low sequence similarity, the catalytic domains of typical MTases display a common tertiary architecture, similar to the Rossmann-fold, but with a unique peripheral β-hairpin structure instead of a typical right-handed β-α turn [19]. Another characteristic feature of many MTase families is the presence of an additional "variable" domain, which is primarily responsible for substrate recognition and binding. This domain has been initially characterized in DNA:cytosine-C5 (m^5C) MTases and dubbed TRD (for target recognition domain). More recently, it was determined that the majority of TRDs of individual MTase families are unrelated. They occur in different locations in the primary structure of the protein and fold into different structures, suggesting that they have originated from independent gene fusions ([18]. Nevertheless, it has been shown that the TRDs of m^5C MTases are structurally similar, even though only several common residues could be delineated in their sequences that are critical for stability of the hydrophobic core and interactions of the TRD with the substrate. Moreover, based on the sequence-to-structure threading, it has been predicted that the TRDs of type I DNA MTases (a subclass of enzymes that modify adenine in DNA) share the common fold with the TRD of m^5C MTases [20]. This prediction has been later supported by mutagenesis studies [21]. Therefore, aside from the structural and evolutionary diversity among TRDs, some MTase families may share conserved homology in the catalytic and substrate binding domains, even though their sequences seem dissimilar.

The prolonged unavailability of the atomic structure of cap MTase prompted us to predict its structure and construct a three-dimensional model, which is accompanied by an evolutionary study. The results form this report should aid in the interpretation and design of mutagenesis experiments and provide a framework for comparative sequence-structure-function analysis of members of the MTase family. Cap MTases exhibited limited similarities to other MTases in the common AdoMet-binding region, and the substrate-binding site could not be unambiguously identified, based on sequence analysis and mutagenesis results [13]. Therefore, we resorted to the sequence-to-structure threading method to find a structural template for homology modeling. We report here that cap MTases are related in structure to the glycine *N*-MTase. In addition, we carried out extensive database searches to identify novel genes that exhibit homology to known cap MTases, which may encode yet unidentified RNA modification enzymes.

Results and Discussion

Sequence analysis

Following a thorough search of the sequence databases, sequences of genuine cap MTases displayed homology with members of the cellular and viral cap MTase families (15 and 8 sequences, respectively, with the pairwise amino acid identity < 90%; Figure 1). No significant sequence similarity to the putative cap MTase domains from alphaviruses, dsRNA viruses or yeast killer plasmids could be detected. Nevertheless, we noticed similarities between the cellular cap MTases (BLAST e-value $< 10^{-4}$) and a putative protein F4F15.320 from *Arabidopsis thaliana*, which lacked several conserved regions. Reciprocal BLAST searches queried with At_F4F15.320 revealed that it is indeed closely related to cap MTases (e-value $< 10^{-5}$) and may represent a paralog of another cap MTase-like protein from *A. thaliana* (At_F3H11.3), which has been identified at earlier stages of the search. Using the sequence of At_F4F15.320 as a query, we retrieved its close homologs from plants *Glycine max* (combination of cDNA clones sd21c10.y2, sn79e11.y1 and

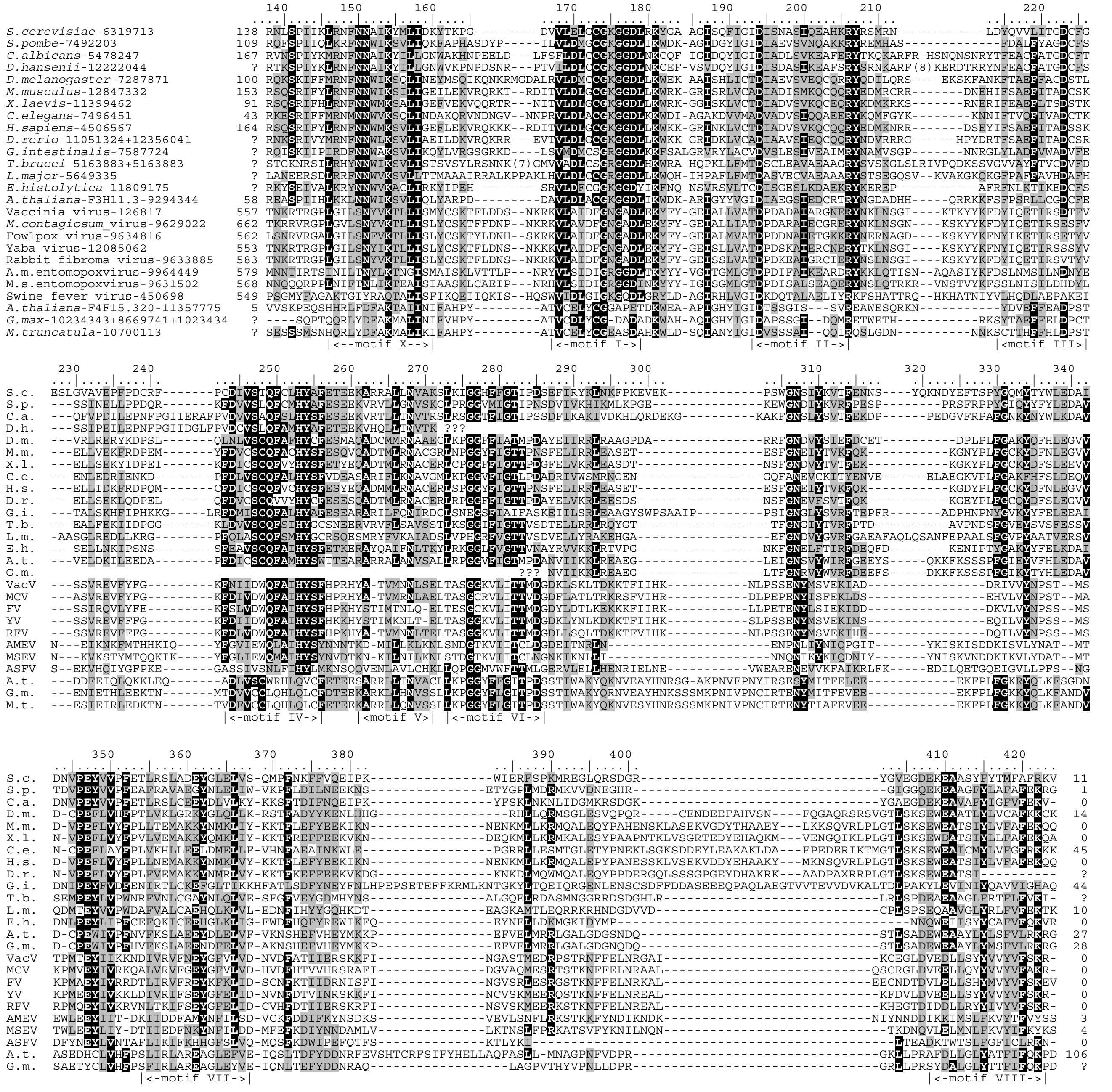

Figure 1

Multiple alignment of the cap MTase family. Only the sequences, whose amino acid identity was lower than 90% were included. Conserved motifs are labeled according to the nomenclature described for the AdoMet-dependent MTase superfamily [18]. Invariant residues are highlighted in black, conserved residues are highlighted in gray. Numbers show the size of termini (unknown in preliminary sequences) omitted for clarity

sr53f01.y1) and *Medicago truncatula* (the sequence predicted from the cDNA clone pGVSN-24P11 was truncated at the C-terminus). These new sequences possessed most features that were most common to "orthodox" cap MTases, but missing from At_F4F15.320. We assumed that the cDNA sequences from *G. max* and *M. truncatula* were more likely to correspond to native proteins, while the sequence of At_F4F15.320, deduced from the genomic data, might contain frameshifts and/or incorrectly predicted intron/exon junctions. We corrected the prediction of splice sites in At_F4F15.320 by comparisons of its DNA and protein sequences with those of its newly identified homologs. All conserved elements could be restored and the resulting sequence scored significantly higher in BLAST searches against genuine cap MTases (e-values < 10^{-15}).

The sequences of viral, cellular cap MTases and the newly identified subfamily of putative MTases from higher plants exhibited relatively high similarity in their N-terminus. However, after different substitution matrices, gap opening and extension penalties were used in both PSI-BLAST and CLUSTALX, their C-termini were found to align poorly, and we observed substantial variation in the multiple sequence alignment in this region. Therefore, even the construction of a global alignment, including known viral and cellular cap MTases, presented considerable challenges. Since the sequences within the subfamilies showed high similarity and could be aligned over their entire length, we resorted to profile-to-profile alignment using FFAS, which was proven superior to pairwise or sequence-to-profile alignments [22]. The results revealed that all three subfamilies shared homology not only at the N-terminus, but also at several moderately conserved regions at the C-terminus, separated by regions of high variability. Remarkably, the pattern of the secondary structures predicted for individual subfamilies using JPRED agreed very well with the alignment reported by FFAS. For example, not only were the number and order of predicted helices and strands the same, but also aligned well, showing strong correlation with the aforementioned blocks of moderately conserved sequence (the final alignment is shown in Figure 1).

To improve the multiple sequence alignment and to provide a structural framework for the interpretation of experimental studies and phylogenetic analysis, we attempted to predict the tertiary structure of cap MTase using sequence-to-structure threading and homology modeling. The rationale behind this approach is that most of the alignment errors that are undetectable at the level of primary and secondary structure would manifest themselves in the model. They could be identified and corrected by computer software for the evaluation of tertiary structures, followed by the analysis of graphic representations with a trained eye. We submitted sequences of several members of each subfamily, as well as artificial "consensus" sequences that represented the individual subfamilies or the entire family to the MetaServer ([http://bioinfo.pl/meta]), which combines several programs for prediction of secondary structure, solvent accessibility and fold recognition (i.e. detection of the known structure) that are most compatible with the query sequence (see Materials and Methods). A similar strategy based on only one fold recognition algorithm was recently applied for analysis of heterotrimeric PCNA family members [23].

All fold recognition algorithms systematically reported the AdoMet-dependent methyltransferase fold and particularly the glycine *N*-methyltransferase (GNMT) structure [24] as the best modeling template for members of the cap MTase family, regardless of which sequence was used as a query (data not shown). Importantly, while most of the MTase structures showed compatibility with the cap MTase sequence only in their N-terminal domains, which span MTase motifs X and I-VI, the alignments with the GNMT structure were frequently reported to span the full length of both the target and the template. Moreover, the models based on these alignments formed compact, mutually superimposable structures. The protein arginine MTase Hmt1 [25] also produced long alignments; however, these were inconsistent with one another and resulted in models with largely misfolded C-termini (data not shown). Therefore, the GNMT structure was chosen as the main modeling template along with other MTase structures serving as guides in modeling of the variable regions, such as those with relatively large insertions and deletions that result in truncation or extension of secondary structural elements. The *S. cerevisiae* cap MTase (ScABD1 protein) was chosen as the representative modeling target since it is well-characterized experimentally [5,13,16], and it lacks certain non-conserved insertions that are present in the cap MTase from *C. albicans* or higher Eukaryota (Figure 1). The modeling posed a formidable challenge because of the variability between alternative target-template alignments that were reported by different servers. Hence, our modeling strategy followed a modified version of "multiple models" approach [26] and employed a variety of alternative models based on alignments with several templates evaluated by several different criteria (see Materials and Methods). Refinement of the model was done in parallel with refinement of the multiple sequence alignment and involved additional threading and modeling for updated consensus sequences. This procedure was utilized until the final model could not be further improved. The final target-template alignment is shown in Figure 2.

Sequence conservation and evolutionary relationships in the cap MTase family

Analysis of the multiple sequence alignment revealed amino acid residues that are conserved among all or most of the individual family members as well as some differences between the three subfamilies (Figure 1). Together with the unrooted phylogenetic tree, which was calculated from this alignment (Figure 3), these results demonstrate that viral and cellular cap MTases and the newly identified proteins from green plants originate from three phylogenetically distinct lineages. It should be noted that sequences obtained from preliminary data and cDNA clones may contain errors that can influence

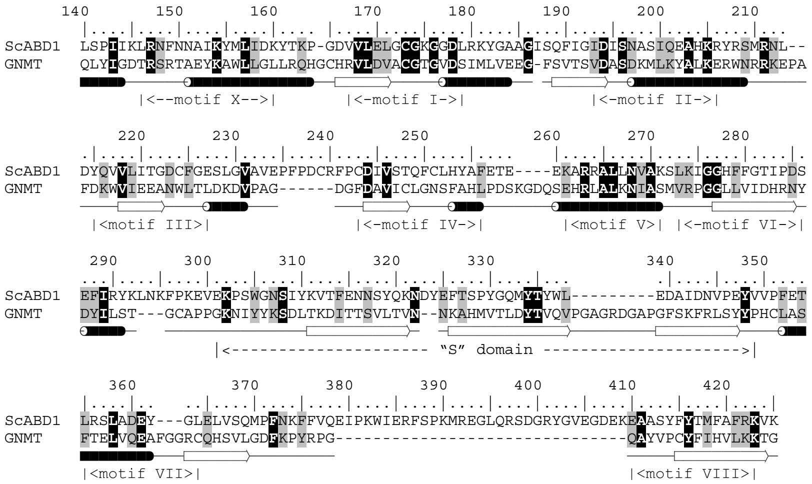
140 150 160 170 180 190 200 210
ScABD1 LSPIIKLRNFNNAIKYMLIDKYTKP-GDVVLELGCGKGGDLRKYGAAGISQFIGIDISNASIQEAHKRYRSMRNL--
GNMT QLYIGDTRSRTAEYKAWLLGLLRQHGCHRVLDVACGTGVDSIMLVEEG-FSVTSVDASDKMLKYALKERWNRRKEPA
|<--motif X-->|
|<-motif I->|
|<-motif II->|
220 230 240 250 260 270 280
ScABD1 DYQVVLITGDCFGESLGVAVEPFPDCRFPCDIVSTQFCLHYAFETE----EKARRALLNVAKSLKIGGHFFGTIPDS
GNMT FDKWVIEEANWLTLDKDVPAG------DGFDAVICLGNSFAHLPDSKGDQSEHRLALKNIASMVRPGGLLVIDHRNY
|<motif III>|
|<-motif IV->|
|<motif V>|
|<-motif VI->|
290 300 310 320 330 340 350
ScABD1 EFIRYKLNKFPKEVEKPSWGNSIYKVTFENNSYQKNDYEFTSPYGQMYTYWL---------EDAIDNVPEYVVPFET
GNMT DYILST---GCAPPGKNIYYKSDLTKDITTSVLTVN--NKAHMVTLDYTVQVPGAGRDGAPGFSKFRLSYYPHCLAS
|<--------------------- "S" domain --------------------->|
360 370 380 390 400 410 420
ScABD1 LRSLADEY---GLELVSQMPFNKFFVQEIPKWIERFSPKMREGLQRSDGRYGVEGDEKEAASYFYTMFAFRKVK
GNMT FTELVQEAFGGRCQHSVLGDFKPYRPG------------------------------QAYVPCYFIHVLKKTG
|<motif VII>|
|<motif VIII>|

Figure 2

Alignment of the structural template GNMT and the ScABD1 sequence. Identical residues are highlighted. Secondary structure elements detected in GNMT (virtually identical to those in the ScABD1 model) are shown as cylinders (α-helices) and arrows (β-strands).

the outcome of the comparative analysis. The fact that several sequences are truncated could have affected the calculated phylogenetic distances (i.e. the length of the branches). For instance, an elongated C-terminus of putative MTase from *Giardia intestinalis* (low-pass HTG sequence MJ2197) may be an artifact, for it shows little similarity to yeast MTases and could not be unambiguously threaded onto the GNMT structure. However, we believe that the predicted topology of the branching pattern is correct because it remained unchanged after the incomplete or preliminary sequences were removed. The topology is also strongly supported by bootstrap analysis, and the phylogenetic groups correlate well with the presence of sequence signatures that can be regarded as synapomorphies (shared features derived from a common ancestor). For instance, the MTases from yeast possess a unique insertion close to the C-terminus.

Another peculiarity emerging from the alignment is the mosaic character of similarity of the new plant MTase family to viral and "orthodox" cellular enzymes. For instance, the "ENYM" patch (corresponding to aa 306-309 in ScABD1) is conserved in viral MTases and in the newly identified proteins, but absent from "orthodox" cellular enzymes, and the "PLFGXKY" patch (corresponding to aa 328-334 in ScABD1) is common to all cellular enzymes, but absent from viral MTases (Figure 1). It should be noted however that viral and "orthodox" cellular enzymes are mutually similar in many regions that are otherwise considerably diverged in the plant MTases, which suggest that the latter are more ancient. We could not predict with confidence a potential function for the subfamily of plant MTases. It is possible that they represent genuine cap MTases from the yet unidentified plant viruses or small nuclear RNA capping enzymes. Their function remains to be determined experimentally.

General features of the three-dimensional model of S. cerevisiae cap MTase

As expected, the model of *S. cerevisiae* cap MTase (ScABD1) is similar to the GNMT template and displays the essential features of the typical AdoMet-dependent MTase fold [18,19] (Figure 4). All insertions and deletions (indels) localize primarily to surface loop structures of GNMT and do not disrupt the core elements. Modeling of indels present in other family members results also in reasonably folded structures (data not shown), indicating that the target-template alignment is well optimized. Only several insertions longer than 20 amino acid residues (including the region E379-K409 in ScABD1) that mapped to solvent-exposed regions were not modeled because the present methodology does not allow for conformational prediction of such large fragments of the polypeptide chain or of their packing against the common core.

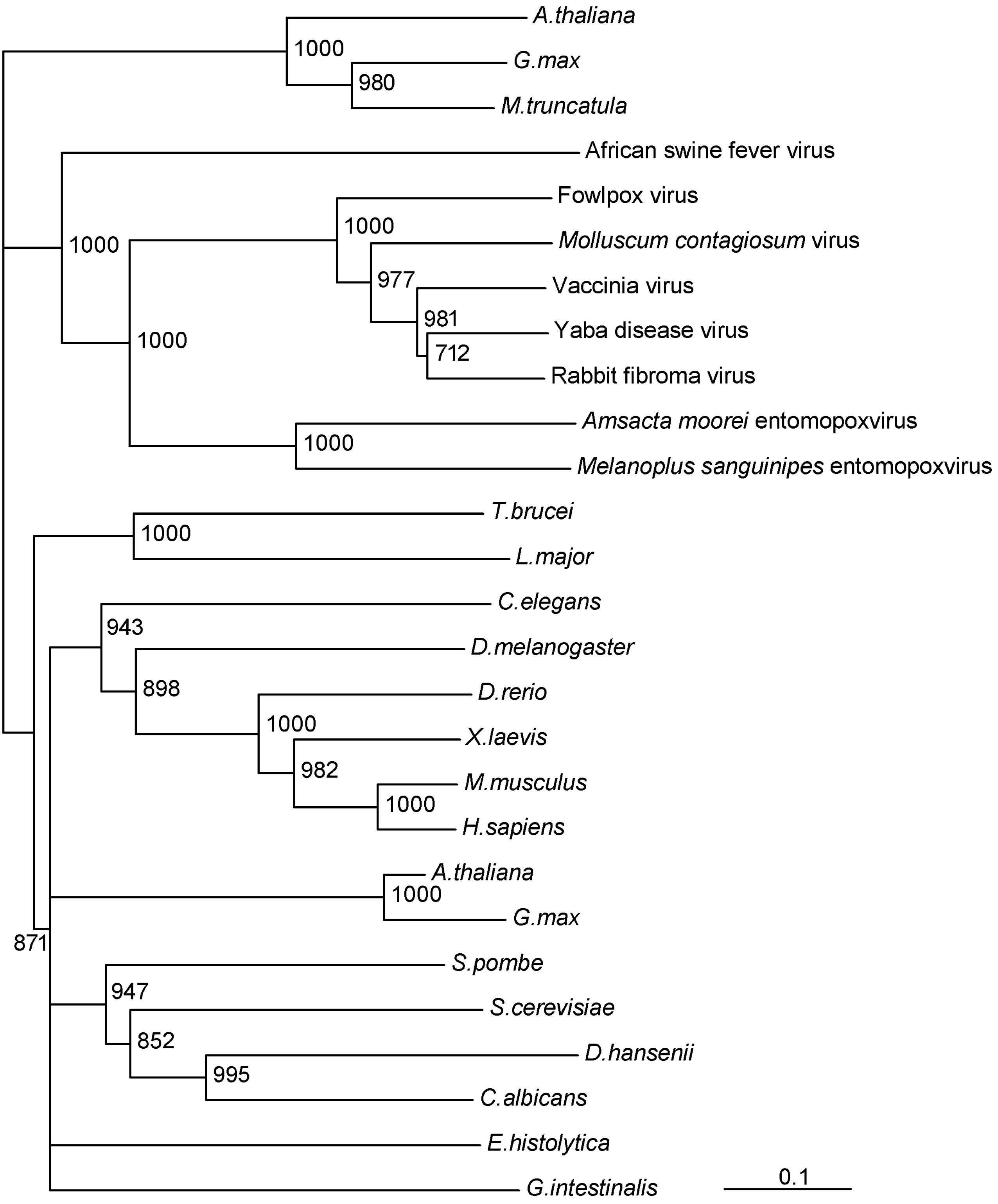

Figure 3

The phylogenetic tree of the cap MTase family. The numbers at the nodes indicate the statistical support of the branching order by the bootstrap criterion. The nodes with bootstrap support < 50% are shown as unresolved. The bar at the bottom of the phylogram indicates the evolutionary distance, to which the branch lengths are scaled based on the estimated divergence.

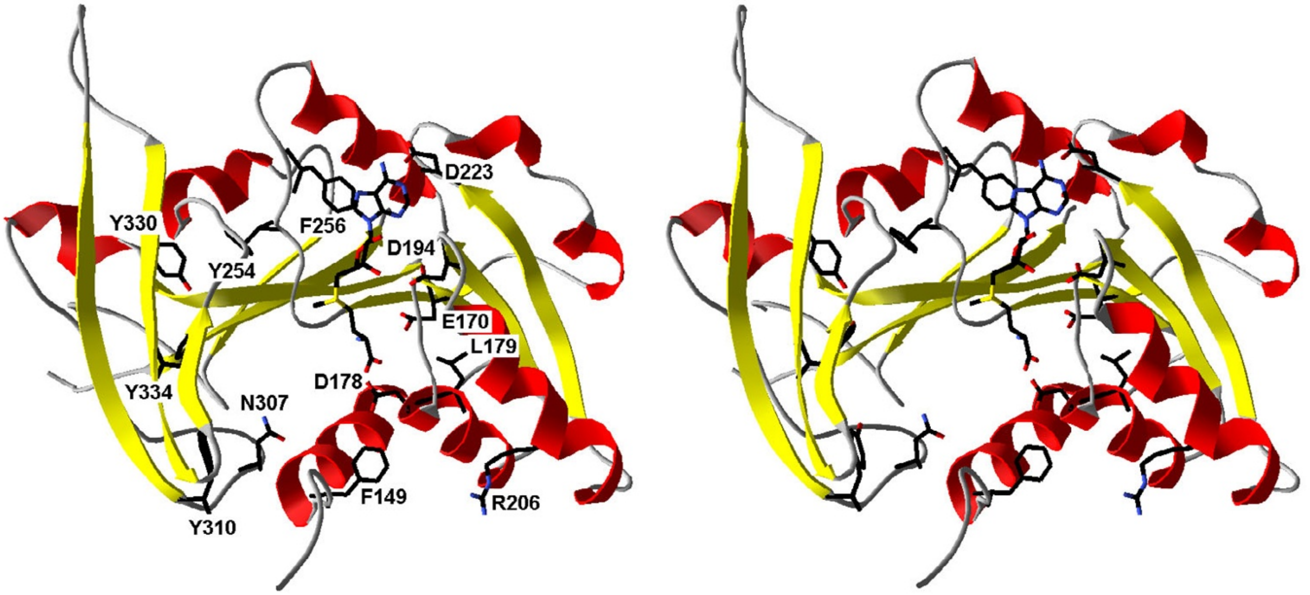

Figure 4
A cartoon diagram of the modeled ScABD1 protein. The selected functionally important residues discussed in the text are shown in wireframe representation and labeled.

The N-terminal 139 residues and the C-terminal 12 residues of ScABD1 were not included in the model. The present C-terminus maps to the "back" of the cap MTase, while the N-terminus protrudes outside its AdoMet-binding/catalytic face. Shuman and coworkers carried out deletion mutagenesis of ScABD1 and reported that mutants lacking the 130 N-terminal residues or the 10 C-terminal residues were fully active, whereas the activity of a mutant lacking 143 N-terminal residues was reduced to ~ 5% of that of the wild type ScABD1 [13]. The GNMT template lacks the counterpart of the C-terminal 12 residues of ScABD1 [24]. This region is not conserved among other AdoMet-dependent MTases [18], suggesting it is dispensable for the stability of the MTase core. A C-terminal deletion of 55 amino acids (residues 381-436) was lethal [5]. According to our model this region encompasses the antiparallel β-strand that takes part in formation of the hydrophobic core and its loss would destabilize the MTase structure.

In GNMT, which unlike most MTases is a tetrameric enzyme, the N-terminus is swapped between the subunits and acts as a cork to the active site entrance [24]. In the enzymatically active "open" conformation of GNMT, the residues 1-40 are disordered. It has been suggested that this region takes part in auto-inhibitory and forced product release mechanisms [27]. As discussed by Takusagawa and coworkers [27], the flexible loops of many enzymes can stabilize the active site when ligands bind. We hypothesize that the corresponding region to the N-terminus of N150 in ScABD1 may form a conformationally variable structure that affects the binding of the ligands. Analysis of the molecular surface of the model (Figure 5) reveals two large pockets that are predicted to serve as the cofactor- and guanine-binding sites and a neighboring positively charged area that may bind the phosphate groups of the mRNA chain. The predicted flexible N-terminus is located at the border of the putative mRNA-binding patch, "below" the two pockets. Thus, participation of this region in regulation of substrate binding and catalysis seems plausible.

Structure-based analysis of site-directed mutagenesis results

Alanine scanning mutagenesis of the cofactor-binding region in *S. cerevisiae* and human cap MTases revealed that of all residues of the sequence patch 168-VLELGCG-KGGDLRKY-182 (motif I) only E170, G174, and D178 are essential [14]. The data for *C. albicans* suggest that the counterparts of L169 and L179 in ScABD1 (L204 and L214 in CaABD1) are also indispensable for both *in vitro* and *in vivo* activity of the enzyme. In ScABD1, L179 packs against the side chain of E170, while in CaABD1 L214 packs against the shorter side chain of D205, suggesting that the L214A substitution creates a larger cavity in the hydrophobic core of CaABD1 that may perturb the overall architecture of the AdoMet-binding site. The E170D mutant of ScABD1 was viable [16], suggesting that a carboxylate is essential at this position. Indeed, an acidic residue is present at this position in most AdoMet-

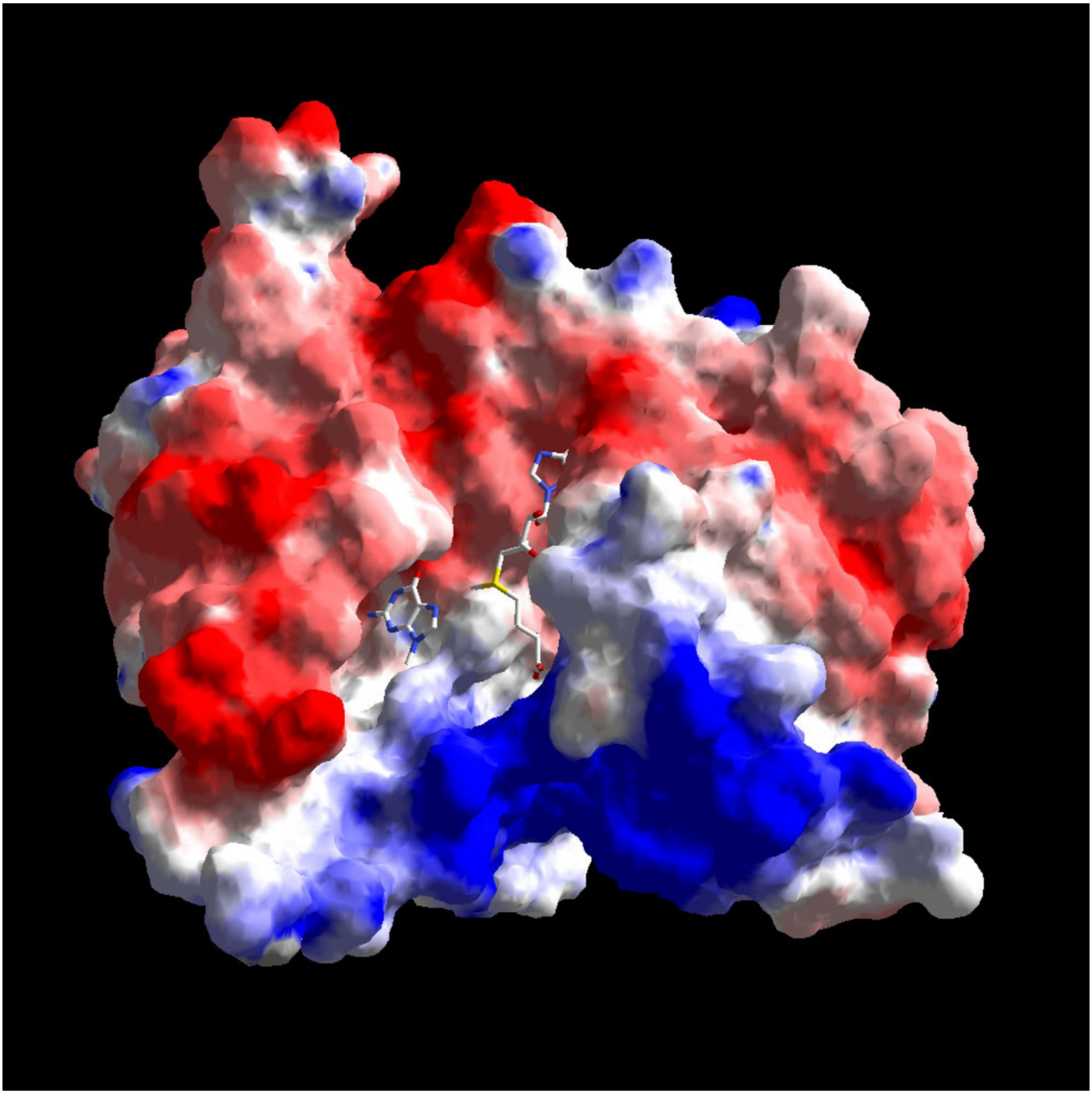

Figure 5

Crossed stereoview of the electrostatic potential mapped onto the molecular surface of ScABD1. The values of surface potentials are expressed as a spectrum ranging from -10 kT/e (deep red) to +10 kT/e (deep blue). The guanine moiety (docked by hand only to help visualize the size of the "molecular basket") and AdoMet are shown in wireframe representation.

dependent MTases [18]. In the recently solved high-resolution structure of rRNA:2'-*O*-ribose MTase RrmJ, a D57 residue present at this position coordinates the nitrogen atom of AdoMet *via* an ordered water molecule [28]. However, this carboxylate is replaced by Ala or Ser in cap MTases from all viruses except for ASFV (Figure 1). Notably, all viral MTases (again, except for ASFV) possess an Asp residue in the position corresponding to G172 in ScABD1. Modeling suggests that in viral MTases and in the ScABD1 double mutants E170A G172D, the carboxylate would make an equivalent, direct contact to the AdoMet moiety (data not shown), which suggests a textbook case of correlated mutations. This important difference between cellular and viral cap MTases may be utilized in rational drug design.

Alanine substitution of the other two residues in the AdoMet-binding region, i.e. D194 (in motif II) and R206 (in motif III) abrogates the enzyme's activity; however, conservative substitutions at these positions that allow retention of the charged functional groups are tolerated [13]. According to our model, the carboxylate D194 directly coordinates the ribose oxygens while the role of R206 is less clear. It might participate in the binding of phosphate groups of the nascent RNA chain, for it forms part of a large basic patch at the protein surface (Figure 5). However, this hypothesis remains to be tested experimentally. Other polar residues that may be a part of this predicted binding patch include R147, N150, N151, K154, and Y155. All of these residues have been individually mutated to Ala without an observed loss of activity *in vivo* [16]. It is conceivable that non-specific binding of mRNA depends on electrostatic interactions that are not dependent on the presence of the individual functional groups in the protein. It would be interesting to test if such simultaneous mutagenesis of the above residues can significantly lower the affinity of the protein to the mRNA substrate. Another group of non-essential residues includes E202, Y207, R208, Y215, and D223. They are not universally conserved in the cap MTase family and are localized in the variable edge region of the Rossmann-fold-like AdoMet-binding subdomain. In our model, only D223 is predicted to form a contact with the adenine ring of the AdoMet moiety. It would be interesting to test if the *in vitro* AdoMet-binding activity of cap MTase is influenced by the D223A mutation.

Alanine scanning mutagenesis identified only three residues in the central and C-terminal part of ScABD1 that are essential for catalysis [5,16]. In our model Y254 is buried in a hydrophobic core and F256 stacks with the adenine ring of AdoMet in a manner similar to W117 in the GNMT structure [27], which explains why hydrophobic residues at these positions are required for the cap MTase activity. According to our model, Y330 is located in the lid subdomain where it forms the external wall of the "molecular basket". A bulky hydrophobic residue is present at this position in all cap MTases (Figure 1).

Among the residues from the central and C-terminal part of ScABD1 that were found to be nonessential [5,13,16], D244, G276, G277, E287, E361, Y362, G363, L366, V367, and K423 are not located at the protein surface near any of the predicted binding sites, and their substitution should not influence the structure or function of cap MTase. Residues T282, P284, W305, and Y416 form a cluster at the surface of the catalytic domain near the entrance to the "molecular basket" where the substrate might bind. However, it is not apparent from the modeled structure alone if they can take part in catalysis or binding; therefore, their localization in the model is not inconsistent with the experimental data. W383, E385, E408, and E410 map to the loop between the C-terminal β-strands, which has not been included in the present model because its structure could not be predicted with confidence. On the other hand, F279, F419, and F421 are in the hydrophobic core of the catalytic domain. Y348, V349, and V350 form part of an interface between the "lid" subdomain and the catalytic domain, and F314 is located in the lid subdomain on the inside of the "molecular basket". No straightforward explanation is offered by the current model to rationalize why substitutions of conserved residues at these important locations do not have any influence on ScABD1 activity *in vivo*. It is possible that introducing cavities at these positions in the protein core may slightly destabilize the structure but not disrupt the overall fold, which allows cap MTase to retain its activity. We suggest that introducing polar or charged side chains at these positions (for instance Arg) should disrupt the protein core and render the enzyme inactive.

Structure-based prediction of guanine-binding residues

In GNMT, the additional "S" domain, composed of a three-stranded β-sheet, forms the wall of a large "molecular basket" structure, which may accommodate a variety of small molecules, including AdoMet, tetrahydrofolate and polycyclic aromatic hydrocarbon molecules, such as benzopyrene (reviewed in [29]). According to the secondary structure prediction and threading algorithms, cap MTases also possess a similar structure that is a good candidate for a target-binding site. Therefore, we looked for conserved residues that could correspond to the guanine-binding site at the inner walls of the "molecular basket" of the cap MTase model. To our knowledge, all structurally characterized MTases, which modify bases in nucleic acids and do not employ covalent bond formation with the target, use aromatic or aliphatic side chains to bind the base to be methylated via hydrophobic interactions with the heterocycle for stabilization in the active site. Examples of such MTases, for which the structure of the active site was determined experimentally or predicted from sequence analysis, include enzymes generating N6-methyladenine in DNA and RNA (reviewed in [30]), N4-methylcytosine in DNA [31], and N2-guanine in RNA [32]. On the other hand, various polar residues could be implicated in specific contacts to the hydrophilic edge of the base as it has been proposed for DNA amino-MTases [31].

We have identified several conserved residues that localize to the inner surface of the "basket", which have not yet been tested whether they are important for the catalytic activity of ScABD1. They include the invariant Asn residue (N307 in ScABD1), which may hydrogen bond to the O6 atom, the N2 amino group of the target guanine and the aromatic residue (Y310 in ScABD1), which could be involved in stacking interactions with the aromatic heterocycle. The aromatic residue is present at this position only in proteins from entomopoxviruses and in "orthodox" cellular enzymes (Figure 1), except for the protein from *C. elegans*, which is substituted by a Cys residue that is quite big and hydrophobic and may be involved in van der Waals interactions with the target guanine. Interestingly, in all sequences from viruses (except for ASFV) and in putative proteins from higher plants, a Tyr residue is present at the position corresponding to S308 in ScABD1. Modeling of viral MTases and *in silico* mutagenesis of ScABD1 suggests that aromatic residues located in these two alternative positions could have their side chains oriented in a similar manner (data not shown), arguing for a similar case of correlated mutations as described previously for a carboxylate residue involved in AdoMet binding (see above).

We considered several candidates for the possible second aromatic residue localized in the vicinity of the conserved N307, and the methyl group of AdoMet. F250A, W305A, and Y416A mutants were shown to be functional *in vivo* [16]. However, it would be interesting to test whether mutations at these positions influence binding of the substrate *in vitro*. F149 in ScABD1, which has not been analyzed by mutagenesis, is located further away from the predicted substrate-binding site (7.5 Å from N307) and is conserved only in cellular proteins. Nevertheless, the conformation of the N-terminus of the template GNMT structure changes between the closed and open forms of the protein [27]. As we proposed (see above), the corresponding region may be mobile also in the ScABD1 structure, thus F149 could be relocated upon target binding. It cannot be ruled out that the mechanism of cap binding by cap 0 MTases is different from that of cap 1 MTase from vaccinia virus [33], or that target base binding by nucleic acid amino-MTases does not necessarily employ stacking of the guanine base between two aromatic sidechains. Another possibility remains that the side chains involved in guanine binding are poorly conserved between the subfamilies, or the corresponding alanine mutants retain their function *in vivo* [16] and a more sensitive approach is required to analyze the influence of mutations on the efficiency of catalysis.

Conclusions

In this report, we used computational methods to infer the evolutionary relationships and predict the structure of cap MTase. A tertiary model has been built for the Eukaryotic enzyme and used to interpret the available mutation data and guide the comparative sequence analysis. We propose that cap MTases share the catalytic domain and the "S" domain with glycine N-MTases, which raises the possibility that these two families of N-MTases are relatively closely related. Moreover, we have identified a novel family of putative MTases that are specific to green plants and share structure and mechanism with cap MTases. Therefore, the alignment presented in this work will be a good starting point for further analysis of other N-MTase subfamilies that may share the "molecular basket" structure. Our analysis of the AdoMet-binding site in cap MTases, combined with evolutionary considerations, highlighted a case of correlated mutation in viral enzymes, which may be important for design of specific antivirals. We also used the model to predict the guanine binding site and identify conserved residues that may serve catalytic or structural function, which can be tested by site-directed mutagenesis. A putative non-specific mRNA binding patch was also proposed. Prior to the experimental solution of the structure of cap MTase, our model will be useful in designing new experiments to better understand the molecular function of cap MTases, whereas the identification of a novel family of genes will aid in identifying candidates for cloning and biochemical characterization. We hope that the prediction of numerous structural and functional features presented in this paper will advance these studies.

Materials and Methods

Sequence analysis

PSI-BLAST [34] and FFAS [22] algorithms were used to search the non-redundant version of current sequence databases (nr) and the publicly available complete and incomplete genome sequences via the Gene Relational DataBase (GRDB) ([http://grdb.bioinfo.pl]). The EST (expressed sequence tag), STS (sequence-tagged site), HTG (high throughput genomic) and GSS (genome survey sequence) divisions of the GenBank database [35] were searched at NCBI ([http://www.ncbi.nlm.nih.gov/]) using TBLASTN [36]. Fragments of sequences were assembled into partial ORFs using the sequences of genuine cap MTases as guides; the predicted splicing sites were verified in reciprocal BLAST searches against the database comprising sequences of cap MTase homologs. All sequences were subsequently realigned using the CLUSTALX program [37] to the degapped profiles obtained from the multiple sequence alignments reported by BLAST. Manual adjustments were introduced based on the BLAST pairwise comparison, secondary structure prediction, threading results, and finally, superposition of modeled structures (see below).

Phylogenetic analysis

The number of amino acid replacements per sequence position in the alignment was estimated using the JTT

[38] model. The sampling variance of the distance values was estimated from 1000 bootstrap resamplings of the alignment columns. The evolutionary inference was performed according to the neighbor-joining method of Saitou and Nei [39]. Multiple runs were conducted with randomized sequence input order to avoid the tree being caught in a local statistical minimum.

Structure prediction

In search for structurally characterized homologs of cap MTase we used the MetaServer available at http://bioinfo.pl/meta/ [40], which uses fold recognition methods such as FFAS [22], 3DPSSM [41], BIOINBGU [42], GenThreader [43], SAM-T99 [44], FUGUE ([http://www-cryst.bioc.cam.ac.uk/~fugue/]), and 123D+ [45]. These methods "thread" the query sequence (the target) onto every fold in libraries of structures (templates) and return 10 alignments that scored best according to the criterion of compatibility, which is specific for a given algorithm. The results are collected by the MetaServer and submitted to the Pcons neural network (J.Lundstrom, L.Rychlewski, J.M.Bujnicki, and A.Eloffson, manuscript submitted), which compares the models and the associated scores and produces a ranking of potentially best predictions. Pcons differs from other "consensus" methods since it predicts the quality of a model and not simply if a correct fold is recognized or not. This is especially advantageous in cases where several alternative folds are reported or if the correct fold is reported by most servers, but the alignments differ, or if one needs to choose the best template from several similar structures. In addition to prediction of the three-dimensional fold, the MetaServer displays the independently predicted secondary structure according to PSIPRED [46], SAM - T99 [44], and JPRED [47], with the latter server reporting also the predicted solvent accessibility profile. These predictions were compared with secondary structures and solvent accessibility calculated directly from the threading-based models of cap MTases.

Modeling

Homology modeling was carried out following a modified version of the "multiple models" approach [26]. Using the SWISS-MODEL/PROMOD II server [48] and the GROMOS forcefield for energy minimization [49] we generated a set of preliminary models based on threading-derived pairwise target-template alignments obtained from the MetaServer. The preliminary models were then superimposed using SWISS-PDB VIEWER [50] and the best fragments were merged into the final structure. The choice of fragments was based on the evaluation of their stereochemical and energetic parameters by WHATCHECK [51] and PROSA II software embedded within PROMOD II [52], consensus between the individual methods, agreement with the independently predicted pattern of secondary structures and solvent accessibility (see above).

Acknowledgements

We would like to thank Frank King for critical reading of the manuscript. This work was supported by KBN (grant 8T11F01019 to J.M.B.) and BioInfoBank. Coordinates of the model have been deposited in the Protein Data Bank with entry code 1IC3.

References

1. Banerjee AK: **5'-terminal cap structure in eucaryotic messenger ribonucleic acids.** *Microbiol. Rev.* 1980, **44**:175-205
2. Lewis JD, Izaurralde E: **The role of the cap structure in RNA processing and nuclear export.** *Eur. J Biochem* 1997, **247**:461-469
3. Shuman S: **Structure, mechanism, and evolution of the mRNA capping apparatus.** *Prog. Nucleic Acid. Res. Mol. Biol* 2000, **66**:1-40
4. Schwer B, Shuman S: **Mutational analysis of yeast mRNA capping enzyme.** *Proc. Natl. Acad. Sci U.S.A.* 1994, **91**:4328-4332
5. Mao X, Schwer B, Shuman S: **Mutational analysis of the Saccharomyces cerevisiae ABD1 gene: cap methyltransferase activity is essential for cell growth.** *Mol. Cell Biol* 1996, **16**:475-480
6. Tsukamoto T, Shibagaki Y, Imajoh-Ohmi S, Murakoshi T, Suzuki M, Nakamura A, Gotoh H, Mizumoto K: **Isolation and characterization of the yeast mRNA capping enzyme beta subunit gene encoding RNA 5'-triphosphatase, which is essential for cell viability.** *Biochem Biophys. Res. Commun.* 1997, **239**:116-122
7. Hakansson K, Doherty AJ, Shuman S, Wigley DB: **X-ray crystallography reveals a large conformational change during guanyl transfer by mRNA capping enzymes.** *Cell* 1997, **89**:545-553
8. Lima CD, Wang LK, Shuman S: **Structure and mechanism of yeast RNA triphosphatase: an essential component of the mRNA capping apparatus.** *Cell* 1999, **99**:533-543
9. Wang SP, Deng L, Ho CK, Shuman S: **Phylogeny of mRNA capping enzymes.** *Proc. Natl. Acad. Sci U.S.A.* 1997, **94**:9573-9578
10. Pei Y, Lehman K, Tian L, Shuman S: **Characterization of Candida albicans RNA triphosphatase and mutational analysis of its active site.** *Nucleic Acids Res.* 2000, **28**:1885-1892
11. Pei Y, Schwer B, Hausmann S, Shuman S: **Characterization of Schizosaccharomyces pombe RNA triphosphatase.** *Nucleic Acids Res.* 2001, **29**:387-396
12. Mao X, Shuman S: **Vaccinia virus mRNA (guanine-7-)methyltransferase: mutational effects on cap methylation and AdoHcy-dependent photo-cross-linking of the cap to the methyl acceptor site.** *Biochemistry* 1996, **35**:6900-6910
13. Wang SP, Shuman S: **Structure-function analysis of the mRNA cap methyltransferase of Saccharomyces cerevisiae.** *J Biol Chem.* 1997, **272**:14683-14689
14. Saha N, Schwer B, Shuman S: **Characterization of human, Schizosaccharomyces pombe, and Candida albicans mRNA cap methyltransferases and complete replacement of the yeast capping apparatus by mammalian enzymes.** *J Biol Chem.* 1999, **274**:16553-16562
15. Yamada-Okabe T, Mio T, Kashima Y, Matsui M, Arisawa M, Yamada-Okabe H: **The Candida albicans gene for mRNA 5-cap methyltransferase: identification of additional residues essential for catalysis.** *Microbiology.* 1999, **145 (Pt 11)**:3023-3033
16. Schwer B, Saha N, Mao X, Chen HW, Shuman S: **Structure-function analysis of yeast mRNA cap methyltransferase and high-copy suppression of conditional mutants by AdoMet synthase and the ubiquitin conjugating enzyme Cdc34p.** *Genetics* 2000, **155**:1561-1576
17. Cheng X, Blumenthal RM: **S-adenosylmethionine-dependent methyltransferases: structures and functions.** *Singapore, World Scientific Inc.* 1999
18. Fauman EB, Blumenthal RM, Cheng X: **Structure and evolution of AdoMet-dependent MTases.** *In S-Adenosylmethionine-dependent methyltransferases: structures and functions. (Edited by Cheng X, Blumenthal RM) Singapore, World Scientific Inc.* 1999, :1-38
19. Bujnicki JM: **Comparison of protein structures reveals monophyletic origin of the AdoMet-dependent methyltransferase family and mechanistic convergence rather than recent differentiation of N4-cytosine and N6-adenine DNA methylation.** *In Silico Biol.* 1999, **1**:1-8

20. Sturrock SS, Dryden DT: **A prediction of the amino acids and structures involved in DNA recognition by type I DNA restriction and modification enzymes.** *Nucleic Acids Res.* 1997, **25**:3408-3414
21. O'Neill M, Dryden DT, Murray NE: **Localization of a protein-DNA interface by random mutagenesis.** *EMBO J.* 1998, **17**:7118-7127
22. Rychlewski L, Jaroszewski L, Li W, Godzik A: **Comparison of sequence profiles. Strategies for structural predictions using sequence information.** *Protein Sci* 2000, **9**:232-241
23. Venclovas C, Thelen MP: **Structure-based predictions of Rad1, Rad9, Hus1 and Rad17 participation in sliding clamp and clamp-loading complexes.** *Nucleic Acids Res.* 2000, **28**:2481-2493
24. Fu Z, Hu Y, Konishi K, Takata Y, Ogawa H, Gomi T, Fujioka M, Tak U: **Crystal structure of glycine N-methyltransferase from rat liver.** *Biochemistry* 1996, **35**:11985-11993
25. Weiss VH, McBride AE, Soriano MA, Filman DJ, Silver PA, Hogle JM: **The structure and oligomerization of the yeast arginine methyltransferase, HMT1.** *Nat. Struct. Biol.* 2000, **7**:1165-1171
26. Pawlowski K, Jaroszewski L, Bierzynski A, Godzik A: **Multiple model approach - dealing with alignment ambiguities in protein modeling.** *Pac. Symp. Biocomput.* 1997, :328-339
27. Huang Y, Komoto J, Konishi K, Takata Y, Ogawa H, Gomi T, Fujioka M, Takusagawa F: **Mechanisms for auto-inhibition and forced product release in glycine N-methyltransferase: crystal structures of wild-type, mutant R175K and S-adenosylhomocysteine-bound R175K enzymes.** *J Mol. Biol* 2000, **298**:149-162
28. Bugl H, Fauman EB, Staker BL, Zheng F, Kushner SR, Saper MA, Bardwell JC, Jakob U: **RNA methylation under heat shock control.** *Mol. Cell* 2000, **6**:349-360
29. Takusagawa F, Ogawa H, Fujioka M: **Glycine N-methyltransferase, a tetrameric enzyme.** *In S-Adenosylmethionine-dependent methyltransferases: structures and functions. (Edited by Cheng X, Blumenthal RM) Singapore, World Scientific Inc.* 1999, :93-122
30. Schluckebier G, Labahn J, Granzin J, Saenger W: **M.TaqI: possible catalysis via cation-pi interactions in N-specific DNA methyltransferases.** *Biol. Chem.* 1998, **379**:389-400
31. Gong W, O'Gara M, Blumenthal RM, Cheng X: **Structure of PvuII DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment.** *Nucleic Acids Res.* 1997, **25**:2702-2715
32. Bujnicki JM: **Phylogenomic analysis of 16S rRNA:(guanine-N2) methyltransferases suggests new family members and reveals highly conserved motifs and a domain structure similar to other nucleic acid amino-methyltransferases.** *FASEB J* 2000, **14**:2365-2368
33. Hu G, Gershon PD, Hodel AE, Quiocho FA: **mRNA cap recognition: dominant role of enhanced stacking interactions between methylated bases and protein aromatic side chains.** *Proc. Natl. Acad. Sci U.S.A.* 1999, **96**:7149-7154
34. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res.* 1997, **25**:3389-3402
35. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, *et al*: **Database resources of the national center for biotechnology information.** *Nucleic Acids Res.* 2001, **29**:11-16
36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J. Mol. Biol.* 1990, **215**:403-410
37. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res.* 1997, **25**:4876-4882
38. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput. Appl. Biosci.* 1992, **8**:275-282
39. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol. Biol. Evol.* 1987, **4**:406-425
40. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L: **LiveBench-1: continuous benchmarking of protein structure prediction servers.** *Protein Sci* 2001, **10**:352-361
41. Kelley LA, McCallum CM, Sternberg MJ: **Enhanced genome annotation using structural profiles in the program 3D-PSSM.** *J. Mol. Biol* 2000, **299**:501-522
42. Fischer D: **Hybrid fold recognition: combining sequence derived properties with evolutionary information.** *Pac. Symp. Biocomput.* 2000119-130
43. Jones DT: **GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences.** *J Mol. Biol* 1999, **287**:797-815
44. Karplus K, Barrett C, Hughey R: **Hidden Markov models for detecting remote protein homologies.** *Bioinformatics* 1998, **14**:846-856
45. Alexandrov NN, Nussinov R, Zimmer RM: **Fast protein fold recognition via sequence to structure alignment and contact capacity potentials.** *Pac. Symp. Biocomput.* 1996, :53-72
46. McGuffin LJ, Bryson K, Jones DT: **The PSIPRED protein structure prediction server.** *Bioinformatics* 2000, **16**:404-405
47. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ: **JPred: a consensus secondary structure prediction server.** *Bioinformatics* 1999, **14**:892-893
48. Peitsch MC: **ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling.** *Biochem. Soc. Trans.* 1996, **24**:274-279
49. Scott WRP, Hunenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Kruger P, van Gunsteren WF: **The GROMOS biomolecular simulation program package.** *J. Phys. Chem.* 1999, **103**:3596-3607
50. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.** *Electrophoresis* 1997, **18**:2714-2723
51. Hooft RW, Vriend G, Sander C, Abola EE: **Errors in protein structures.** *Nature* 1996, **381**:272
52. Sippl MJ: **Recognition of errors in three-dimensional structures of proteins.** *Proteins* 1993, **17**:355-362