METHIONINE ADENOSYLTRANSFERASE AS A USEFUL MOLECULAR SYSTEMATICS TOOL REVEALED BY PHYLOGENETIC AND STRUCTURAL ANALYSES

Gabino F. Sánchez-Pérez[1], José M. Bautista [2], María A. Pajares*[1]

[1]Instituto de Investigaciones Biomédicas "Alberto Sols" (CSIC-UAM), Arturo Duperier 4, 28029 Madrid (Spain).

[2]Departamento de Bioquímica y Biología Molecular IV, Facultad de Veterinaria, Universidad Complutense de Madrid, 28040 Madrid (Spain).

*Corresponding author: María A. Pajares, Instituto de Investigaciones Biomédicas "Alberto Sols" (CSIC-UAM), Arturo Duperier 4, 28029 Madrid, Spain. Phone number: 34-91-5854414, FAX number: 34-91-5854401, e-mail:mapajares@iib.uam.es

SUMMARY

Structural and phylogenetic relationships among *Bacteria* and *Eukaryota* were analyzed by examining 292 methionine adenosyltransferase (MAT) amino acid sequences with respect to the crystal structure of this enzyme established for *Escherichia coli* and rat liver. Approximately 30% of MAT residues were found to be identical in all species. Five highly conserved amino acid sequence blocks did not vary in the MAT family. We detected specific structural features that correlated with sequence signatures for several clades, allowing taxonomical identification by sequence analysis. In addition, the number of amino acid residues in the loop connecting β-strands A2 and A3 served to clearly distinguish sequences between eukaryotes and eubacteria. The molecular phylogeny of MAT genes in eukaryotes can be explained in terms of functional diversification coupled to gene duplication or alternative splicing and adaptation through strong structural constraints. Sequence analyses and intron/exon junction positions among nematodes, arthropods and vertebrates support the traditional Coelomata hypothesis. In vertebrates, the liver MAT I isoenzyme has gradually adapted its sequence towards one providing a more specific liver function. MAT phylogeny also served to cluster the major bacterial groups, demonstrating the superior phylogenetic performance of this ubiquitous, housekeeping gene in reconstructing the evolutionary history of distant relatives.

Keywords: Methionine adenosyltransferase, S-adenosylmethionine, evolution, methionine metabolism.

S-adenosylmethionine (SAM) is the main methyl donor in the transmethylation of numerous essential cell constituents (DNA, neurotransmitters, phospholipids, and many small molecules) [1]. After decarboxylation, SAM acts as a propylamine group donor in the biosynthesis of some polyamines (spermine and spermidine) [2]. Its importance is reflected by the fact that this molecule participates in as many reactions as ATP. However, contrary to ATP which is produced in a large number of reactions, SAM synthesis occurs in only one reaction catalyzed by methionine adenosyltransferase (MAT, EC 2.5.1.6). MAT is generally a homotetrameric enzyme that uses methionine and ATP in a reaction dependent on the presence of K^+ and Mg^{2+} ions to render SAM, pyrophosphate and inorganic phosphate [3; 4].

To date, many structure/function relationship studies have used either the *Escherichia coli* (c-MAT) or rat liver (rl-MAT) enzyme. These studies have provided a relevant amount of information on key residues of the protein, including cysteines and active-site amino acids [5; 6; 7; 8; 9]. Since the description of the first MAT gene [10], a substantial number of genes encoding MATs of different origins have been cloned and characterized [11; 12; 13; 14; 15; 16]. The data obtained indicate exceptional conservation of the gene sequence among highly divergent species. At the amino acid level, c- and rl-MATs have been estimated to show 52% identity [17]. The crystal structures of these MATs are the only ones presently available that indicate that conservation also occurs at the structural level, and this is reflected by the essentially identical organization of the domains in the monomer [7; 18].

The development of molecular phylogenetics has been generally based on small-subunit (SSU) and large-subunit (LSU) ribosomal RNA analysis [19]. However, several recent concerns have challenged the validity of rRNA as a unique phylogenetic marker. These concerns are related to biases in base composition, disparities in evolutionary rates among lineages, position-

dependent substitution patterns, alignment ambiguities among very distant species, etc. Thus, recent efforts have focused on assessing the use of other genes [20; 21; 22; 23] and large combined protein sequence data sets [24; 25] to reconstruct evolutionary relationships among organisms. It has even been suggested that it will be possible to reconstruct a robust universal phylogeny only if a core of conserved markers, not affected by lateral gene transfer, is identified [26].

Despite a high degree of sequence and structural conservation in MAT, a large number of representative species for which there are available sequences, and vast knowledge on MAT structure and functionality in highly divergent species, this enzyme has not yet been considered as a possible phylogenetic marker. Although attempts have been made to reconstruct partial phylogenies [15; 27; 28] and MAT has been included in studies examining sets of several proteins [29; 30; 31], no universal phylogenetic evaluation has used MAT as the marker. Hence, the aim of the present study was to assess the performance of MAT in phylogenetic reconstructions using the 292 sequences available to date, and to demonstrate its utility in molecular systematic studies.

Results and Discussion

MAT sequence identification and characterization

MAT sequences retrieved by database mining led to the identification of 303 candidate sequences in almost every eukaryote and bacterium, but none in archaea, for which non homologous replacement by a new type of MAT has been recently described [32]. The absence of the MAT gene from the complete genomes of members of the genus *Chlamydia* and the microsporidia *Encephalitozoon cuniculi*, both intracellular parasites [33; 34], is remarkable. Sequences for *Rickettsia prowazekii* and *R. typhi*, also obligate intracellular parasites, were excluded from the analysis because of the recent detection of stop codons in their MAT gene sequences. This may be interpreted as indicative of a certain degree of genome degeneration [35]. The lack or degeneration of the MAT gene in these species may be explained by functional redundancy (obsolescence) or by the existence of another methyl donating pathway (the host produces the methyl donor to be used by the parasite or another compound replaces SAM).

Giardia lamblia, one ancient eukaryote, showed highly divergent sequence and was thus only included in preliminary studies, in which it appeared as the earliest eukaryote with unique structural features. *G. lamblia* has a 41 amino acid insertion in the loop that connects helix 4 with β-strand A4, the significance of which is unknown .

Sequence conservation vs. 3D structure constraints

A final alignment of 392 positions was obtained from the 292 MAT amino acid sequences considered. Positional identity was difficult to establish in areas located at the N- and C-terminals and at loops connecting secondary structure elements. These ambiguous positions were therefore excluded, leaving 330 parsimony-informative positions for the final analysis.

MAT protein alignment revealed the presence of 57 amino acids located in identical positions in 100% of MATs analyzed and 61 additional residues that were conserved in 90% of the species studied. This indicates that approximately 30% of MAT residues are identical in all species. Moreover, 49 amino acid positions were identical in 75-90% of the species, 45 in 60-75%, and 39 in 50-60%. Residue conservation was inhomogeneous along the MAT sequence. Further, regions such as the N-terminal, C-terminal and some intermediate regions showed a lower degree of identity, probably due the capacity of these areas to absorb a high variety of substitutions without affecting the overall conformation of the molecule and its function.

Core areas of greatest amino acid conservation were observed along 5 stretches which we denoted blocks I, II, III, IV and V (Figure 1). Block I comprises residues 20-47, including the β-strand A1 and the α-helix 1, and one of the methionine binding motifs, ^{29}GHPDK^{33} which is preserved in all MAT sequences [7]. Block II is defined by two separate areas bearing residues 114-122 and 132-143, flanking the flexible loop at the active site of the enzyme. This loop has been recently shown to be involved in controlling the catalytic efficiency of the enzyme [36]. Conservation of these two areas could be due to the need to preserve the correct orientation for the loop, or to the fact that it contains the ATP binding motif ^{132}GAGDQG^{137} [17]. The consensus sequence for ATP binding sites has been defined as GxGDxG plus a lysine located 16-28 residues upstream [37]. However, MAT seems to show no variation in this sequence, which is always GAGDQG, highlighting the significant role of alanine and glutamine residues in this enzyme's ATP binding motif. Block III comprises residues 177-189 that form part of β-strand A2, including D180 and K182, two of the amino acids involved in catalysis [7; 18]. Block IV is the largest and most conserved including 54 amino acids (246-300). This block contains the central loop connecting the N-terminal and central domains, as well as many residues directly involved

in substrate and cation binding [7; 18]. In addition, this block contains a high-glycine stretch (254-281 with 10 Gly), which is fully preserved in all MATs. Finally, block V corresponds to the C-terminal and includes α-helix 9 (372-389). Besides these blocks, two further reasonably conserved regions were detected: the first comprises β-strands B1 and B2 with two conserved motifs, residues 55-59 and 70-75, and residue E58, involved in a saline bond with the central loop [7]; and the second area includes residues 303-337, which form part of β-strands C2, C3, and the end of helix 5. However, it must be emphasized that all the amino acids involved in substrate binding and catalysis occur in the blocks described above, and are fully conserved in the MAT family.

To asses the relationship between evolutionary conservation and surface accessibility, the amino acids were also classified according to the degree of identity among MAT sequences as four categories: fully (100%), highly (75-99%), moderately (50-75%) and poorly conserved (<50%) (Figure 2). Buried residues were often observed among the amino acids within the first and second categories (>75%), whereas exposed amino acids were poorly represented in these categories. This over-representation of buried residues among the most conserved residues could reflect their involvement in catalytic activity, correct folding and the stability of the final structure. The fact that they establish the highest number of interactions among residues in the protein structure and their hydrophobic character, may be indicative of their role in the folding nuclei of the monomeric intermediate, according to the overall folding mechanism that has been established [38]. This area is later involved in the association process that leads to the dimer, the minimum active unit of MAT enzymes [2].

Surface mapping of the level of evolutionary conservation for each amino acid in the protein structure may help to identify functionally and/or structurally significant regions [39]. For

this purpose, we used available crystallographic data for rl-MAT I/III and color-coded the surface according to the previously defined residue conservation categories (Figure 3A-D). The conservation pattern shows that the preserved blocks defined above occur in the inner channel, where the active site is located, and in the conserved area exposed at both the entrance and walls of this channel. High conservation among residues at the subunit interface is, therefore, consistent with a role for these amino acids in the structure and function of MAT, as reported for other oligomeric enzymes [40]. The study of individual enzyme families reveals how binding and catalysis are optimized in nature through the inclusion of mutations that improve efficiency in cases in which no new function has been acquired [41]. However, the perfect conservation of active site residues in MATs is an exception to this rule, since it indicates the preservation of the catalytic mechanism during evolution with no modification. This would suggest that the special features of the reaction catalyzed by MAT cannot be easily improved.

It is also of interest that several residues outside the subunit interface are also highly conserved. These residues were found to be mainly located in loops connecting secondary structure elements, thus suggesting a key role in preserving the correct orientation between them. The relevance of preserving this orientation probably reflects their essential contribution to final protein folding.

Sequence alignment using conventional algorithms revealed 60% identity between eukaryotic and bacterial MATs. The high degree of sequence conservation indicates severe restrictions for the substitution of certain amino acids. Such restrictions may be determined by two factors. First, the location of certain residues in the active site is important because of their role in catalysis. Mutations affecting these or adjacent residues modify their relative orientation leading to considerably reduced enzyme activity [7]. The second factor is that correct orientation

and positioning of certain secondary structure elements seems to depend strongly on the presence of some amino acids in the connecting loops, thus their substitution may lead to wrongly folded structures with no activity. To date, no such modifications have been identified in any MAT sequence though these loops have not yet been mapped by site-directed mutational analysis. Evolutionary studies could, nevertheless, take advantage of these restrictions in the sequence. Residue changes in areas of low variation that could be functionally absorbed in a certain evolutionary setting, are very unlikely to be repeated or to revert back at a different evolutionary time. Thus, the mutational study of certain areas of the MAT sequence may serve to clarify certain evolutionary relationships due to the low probability of mutational saturation at the protein level.

MAT as a phylogenetic marker

To establish whether MAT could be a useful tool for reconstructing phylogenetic relationships among clades, we used a final data set of 330 positions for distance and parsimony phylogenetic analyses (Figure 4). A general view of the unrooted MAT tree shows the separate grouping of *Eukaryota* and *Eubacteria* with high bootstrap support indicating a common evolutionary origin. Protein structure comparisons showed that sequences belonging to each group can be clearly distinguished according to the number of residues involved in the loop connecting β-strands A2 and A3. Specifically, this loop was normally 3-4 residues shorter in bacteria, except for *Campylobacter jejuni* and *Deinococcus radiodurans*. This difference has structural implications as shown in Figure 5. A longer loop allows the establishment of 18 favorable interactions between N- and C-terminal domains, including the formation of a salt bridge between D192 and R313. However, in eubacteria, only 2 can be formed among such

interactions including a salt bridge and hence both domains remain more distant. Close inspection of the structural models constructed from the crystallographic data available, shows higher rigidity for the monomer in the eukaryota, due to contact of N- and C-terminal domains [7; 18].

Members of the MAT family can be grouped into several clades, mostly corresponding to the main taxonomy arrangements. Analysis of MAT phylogeny allows the identification of consensus sequences and specific structural features for the groups (table 2). Additional specific characteristics for each major taxonomic group are detailed below:

1.Eukaryotes

A general view of the eukaryotic MAT phylogeny supports an animal-fungal clade excluding green plants. This is consistent with results obtained using SSU rRNA [42] and protein reconstructions [43; 44; 45].

1.1 Viridiplantae

Three to four MAT gene copies were identified in the plants. These show differential expression among tissues and during development [27; 28; 46; 47; 48; 49]. Our analysis also supports (bootstrap value >99%) previous plant MAT gene classification as two types (types I and II) [27]. This topology suggests ancestral duplication, but the absence of MAT genes in conifers prevents us knowing whether duplication occurred before or after the divergence of *Magnoliophyta* from *Coniferophyta*. MAT type I duplications indicate low sequence divergence, thus duplication events are likely to have occurred in a narrow time frame. Consequently, phylogenetic relationships in plants cannot be confidently resolved by sequence similarity analysis. The functional requirements of other MAT isoenzymes may stem from their differential regulation

and/or from new SAM functions observed in plants (e.g., production of the phytohormone ethylene) [50].

It is generally accepted that the ancestors of vascular plants were similar to green algae (*Chlorophyta*). In our phylogenetic analysis, the green alga *Chlamydomonas reinhardtii* branches earliest from the plant clade with high bootstrap support (100%). Some genes may be non-functional because of extensive divergence (e.g., *Cicer arieticum*) or loss of the C-terminal conservation shown by the remaining MATs (e.g., *Gossypium hirsutum 3*). Some of these MAT copies may no longer be needed as in other gene duplications and thus degenerate to pseudogenes.

The MAT genes corresponding to conifers appeared apart from the remaining plants, indicating earlier divergence for this group. The inclusion of one of the two genes isolated from *Pinus contorta* among the *Leuconostocaceae*, which were initially considered an outgroup for the plants, deserves special mention [28]. Based on the high degree of confidence for this node, the most parsimonious explanation for this would be a horizontal gene transfer event.

1.2 Fungi

In this phylogenetic analysis, fungi clustered as a monophyletic group forming the second branch point from that leading to the animals. This topology is well-supported by other studies [51]. Moreover, the MAT tree shows the early separation of the phylum *Basidiomycota* from *Ascomycota*, with high bootstrap support (100%). Only one MAT gene could be identified in most of the fungal species. The exception was *Saccharomyces cerevisiae* which has recently duplicated its genome [52] and has two similar copies of the MAT gene.

1.3 Nematoda

Interestingly, *Caenorhabditis elegans* presents 5 MAT genes. One of them, C49F5.1, was separated by an earlier gene duplication explaining its large phylogenetic distance and least similarity to the other *C. elegans* MATs. Genome analysis of the MAT genes yielded the following observations: a) four genes on autosomic chromosome IV, whereas C49F5.1 appears X-linked; b) this last gene presents 4 exons compared to the others showing 6; and c) three MAT genes were identified in *C. briggsae*, one of which showed similar characteristics to C49F5.1. Thus, MAT genes in nematodes probably arose from one single copy that underwent duplication, leading to genes on chromosomes X and IV. Further duplications led to the four copies identified on chromosome IV. Divergences among the five gene sequences could be attributed to a higher mutational rate of the genes on chromosome X, or to the fact that the duplications in chromosome IV have recently taken place. Just as MAT type II in plants, C49F5.1 seems to be the result of a duplication event required for different regulation or new functions of the final reaction product, SAM.

1.4 Arthropoda

Arthropods differed from the other organisms containing specialized tissues in that only one MAT gene was identified. However, genomic analysis revealed the possibility that this gene may suffer alternative splicing of exon 4 which codifies the flexible loop. Differences among both exons occur in the highly variable region within the loop, whereas conserved regions remain unchanged. These results suggest a role for the flexible loop in the regulation of arthropod MAT that eliminates the need for duplication and tissue specific distribution for each gene, i.e., the most common mechanism seen in MAT from higher organisms.

1.5 Vertebrata

There are three isoforms of MAT in mammalian tissues that are encoded by two genes. MAT I and MAT III are tetrameric and dimeric forms, respectively, of the same gene product (MAT1A), which is mainly expressed in adult liver [53]. The ubiquitously expressed MAT II isoform is a heterotetramer ($\alpha_2\beta_2$), whose catalytic subunit (α) is encoded by a different gene (MAT2A)[2]. The protein isoforms also differ in their regulation and catalytic properties (affinities for methionine) [2; 53]. In every vertebrate reported here, both genes presented around 85% identity, differences always corresponding to the same regions. The MAT tree clearly separates both gene types and locates the duplication event after the divergence from Urochordata, but before the divergence from Teleostomi. This is not surprising, if we consider that many early chordate gene families were formed or expanded by large-scale DNA duplications [54]. Interestingly, liver ontogeny in vertebrates is concomitant with this MAT gene duplication suggesting that tissue-specialized enzyme forms were required for adaptation to the functions of this new organ. MAT I/III has been suggested as a marker for liver development, due to the differential expression of its isoforms in fetal and adult tissue [55].

A further significant issue is identifying specific isozyme residues to which to ascribe functional roles for these conserved amino acids, such as has been done for other homotetrameric enzymes [56; 57]. This identification uses information from the sequence alignment of vertebrate MAT isozymes and is based on the criterion that the residue is conserved in orthologs and is distinct among paralogs. Surprisingly, not all the residues considered liver-specific occurred in all our vertebrate clades suggesting that the specialization of the liver MAT gene took place gradually. Thus, cysteine 121 is found at the flexible loop of the liver enzyme and, being the target for regulation by nitrosylation of the protein [58], it mediates the response to oxidative stress. This residue is present in mammalian liver MAT, but not in that of *Gallus gallus*. By comparing

MAT I/III and MAT II among mammals, birds, amphibian and fishes we were able to identify further analogous positions (Table 2). Along with functional studies, the characterization of new MAT sequences from species intermediate between mammals and ancient chordates, may clarify the evolutionary adaptation related to functional diversification of the MAT liver gene.

Finally, sequence comparison and the evolution of intron/exon junction positions among bilateria (data not shown) suggest the presence of a common ancestor for arthropods and vertebrates that branched from nematodes. This observation is in line with previous studies based on 18S rRNA that assign nematodes and arthropods to a common clade (Ecdysozoa) separate from vertebrates [59]. Conversely, our results are more in agreement with a recent study based on more than 100 proteins that supports the traditional Coelomata hypothesis [31] grouping arthropods with vertebrates apart from nematodes. This reinforces the idea that MAT is a better marker than rRNA for resolving the branching order of the main animal lineages.

2. Bacteria

Five clades of bacteria were clearly distinguished in our molecular phylogeny of MAT based on the presence of a single gene. These clades are: *Proteobacteria*, *Cyanobacteria*, *CFB group*, *Actinobacteria* and *Firmicutes*. Single representative species of several phyla appear as independent branches of the tree: *Aquificae* (*A. aeolicum*), *Thermotogae* (*T. maritima*), *Deinococcus-Thermus* (*D. radiodurans*), *Chloroflexi* (*C. aurianticus*), *Planctomycetes* (*G. obscuriglobis*), *Fibrobacteres* (*F. succinogenes*), *Spirochaetes* (*B. burgdorferi*) and *Fusobacteria* (*F. nucleatum*). The anaerobic detoxifying bacterium *Dehalococcoides ethenogenes* also appeared as an independent branch confirming its position in a unique phylogenetic group as described by the use of 16S rRNA [60]. There is little resolution in the MAT tree reflecting the

position of these phyla, and hence no consistent phylogenetic relationships with other groups could be clearly established.

2.1.Proteobacteria

Proteobacteria have been previously divided into five phylogenetically distinct groups (α, β, γ, δand ε)[19]. In the present phylogenetic analysis, proteobacterial MAT sequences did not consistently form a clade. However, the α-, β-, γ- and δ- groups were found to cluster individually.

2.1.1 α-Proteobacteria

All the α-proteobacteria MAT considered grouped with near 100% confidence but branched apart from the rest of proteobacteria because of the extension of several loops only detected in this class (Table 2). The phylogenetic tree presents an early branching point bearing several members of the order *Rickettsiales*, indicating the early separation of this order from the remaining α-proteobacteria. The order *Rhizobiales* was also clearly separated from the rest of the α-proteobacteria. The most distinctive feature of species of this order was the presence of a 7 amino acid insertion between helix 1 and β-strand B1. Other characteristics typical of rhizobial MATs were: a) a flexible loop at least 2 residues longer than in the other α-proteobacteria; and b) an insertion between β-strand B2 and helix 2. The exception to this rule was shown by *Methylobacterium extorquens* with none of these peculiarities, indicating the very early separation of this family from the remaining rhizobiales.

2.1.2 β*-Proteobacteria*

Based on this phylogenetic tree, MAT sequences from β- and γ-proteobacteria show a common origin, whereas relationships with the rest of the proteobacteria remain obscure, given the <50% bootstrap value for the nodes separating these bacterial classes. Differences between

both groups rely, for example, on the presence of two extra residues in the flexible loop of the β-proteobacteria. Of interest is the presence of *Acidithiobacillus ferrooxidans* among the β-proteobacteria, despite its taxonomical classification as belonging to the order *Chromatiales* of the γ-proteobacteria. However, the high statistical support for this branch in our analysis and the absence of the typical characteristics of a γ-proteobacteria suggest its reclassification within β-proteobacteria.

2.1.3 γ-Proteobacteria

Most of the γ-proteobacteria clustered in the tree, showing relationships among the different orders in this class. All the members of the *Pasteurellales* included in the tree grouped together, as occurred for the *Enterobacteria*. MAT sequences in members of the order *Legionellales* slightly diverged from those of the remaining γ-proteobacteria. This leads to the early branching of *Legionellales* from the γ-proteobacteria group.

2.2 Cyanobacteria

Based on MAT phylogeny, the *Cyanobacteria* appeared as an independent taxonomic group that was well differentiated by specific insertions (Table 2). Two of the four members of subsection I included in this study (*Prochlorococcus marinus* and *Synechococcus sp*.) carried a 7-residue insertion between helix 4 and β-strand A4. In contrast, remaining cyanobacteria only have 6 amino acids at this position. In addition, these two species have an extra residue inserted in the loop between helix 2 and β-strand B3. This could explain why *P. marinus* and *Synechococcus sp.* occur apart from the other members of subsections I, III and IV.

2.3 CFB group

The CFB group was included in a homogeneous cluster due to the specific characteristics of this phylum (Table 2). It should be highlighted that phylogenetic reconstruction places

Chlorobium tepidum on an early branch from the CFB group, although it presents distinctive features such as a 4-residue insertion in the flexible loop, and the absence of insertions in the loop between helix 4 and β-strand A4. This position in the tree indicates the existence of a common ancestor for both phyla and their subsequent divergence. The class *Sphingobacteria* reflects this group's peculiarities: the loop between β-strand A3 and helix 4 does not include the 10-12 residue insertion detected in members of the class *Bacteroidetes*. Moreover, in *Chlorobium* this insertion bears 7 residues.

2.4 Actinobacteria

Members of the group *Actinomycetales* clustered perfectly together, with the order *Bifidobacteriales* remaining apart. The following modifications to the general characteristics were, nevertheless, observed: a) longer insertions in the flexible loop in *B. longum* and *S. fradiae*; b) *Mycobacterium avium*, *M. leprae* and *M. tuberculosis* have a 3-residue insertion in the loop connecting β-strand B2 and helix 2; c) the insertion in the loop between helix 4 and β-strand A4 is absent in *Thermobifida* and differs in length between *Streptomyces* (4 residues) and *Mycobacterium*, *Corynebacterium* and *Bifidobacterium* (3-2 residues).

2.5 Firmicutes

2.5.1 Bacilli/Clostridia. All *Clostridia* and *Bacilli* clustered together but separately defined each class. Moreover, within *Bacilli*, the orders *Bacillales* and *Lactobacillales* were also separately arranged. The topology observed for the family *Leuconostaceae* is an exception, which can be explained by the early divergence from a common ancestor giving rise to a branch for *Leuconostocaceae* and a later divergence rendering the *Bacilli* and *Clostridia* families. Alternatively, the MAT sequence in *Leuconostocaceae* diverged rapidly.

2.5.2 Mollicutes

This class clustered apart from the other *Firmicutes* (*Clostridia*, *Bacilli*). This might be explained by the presence of a 7-amino acid deletion in the loop connecting helix 2 and β-strand B3. *Mycoplasma pneumoniae* and *M. genitalium* clustered apart from *M. pulmoni* and *U. urealyticum*. These last species presented a 2-amino acid insertion in the loop between helices 8 and 9. In contrast, the separation of *M. pneumoniae* and *M. genitalium* may be attributed to differential structural features: a) a flexible loop two amino acids shorter; b) a residue inserted in the loop between helix 4 and β-strand A4; c) another residue inserted in the loop connecting β-strand C3 and helix 6; and d) the insertion of a residue in the loop between helices 7 and 8. In addition, the positioning of *Mycoplasma* among this group is congruent with results derived from protein fusion trees, in contrast to rRNA trees [26]. This finding is yet another confirmation of the idea that protein sequences more accurately reflect the phylogenetic position of *Mycoplasma* [61].

Conclusions

The present study is a first attempt at using the housekeeping MAT gene as a marker in eukarya and bacterial systematics, as an alternative to rRNA and other protein reference markers. This work is the result of integrating data yielded by intensive data mining, robust aligning of MAT sequences and structural-functional analyses. Through the detection of fully conserved regions in the MAT protein of all species, regions varying even among close relatives and characteristic structural features missing from certain taxonomic groups, this new tool enabled us to resolve phylogenetic relationships between close and distant relatives with high bootstrap support.

Materials and Methods

MAT sequences

MAT amino acid sequences were deduced from DNA sequence data available from complete or nearly complete publicly available genomes by conducting a TBLASTN search using rl-MAT as probe. For further references to amino acid positions rl-MAT is used as the consensus sequence. Candidate sequences were identified as MAT when they met the following criteria: a) a length of 370-414 amino acids; b) an N-terminal sequence containing the motif ^{21}FTSESVxEGHPDK^{33}; and c) a C-terminal including the motif ^{379}GHFGxxxxxWE^{389}. MAT sequences were obtained from GenBank, The Institute for Genomic Research (TIGR), DOE Joint Genome Institute, The Sanger Institute, University of Oklahoma, Baylor College of Medicine, Genome Sequencing Center at Washington University Medical School, Columbia Genome Center, Whitehead/MIT Genome Center, Genoscope, DNA Data Bank of Japan, Université Catholique de Louvain and European Bioinformatics Institute. A complete list of the sequences and their sources is provided in table 1.

Alignment and sequence analysis

Sequences were aligned using the program BIOEDIT [62]. The alignment was manually refined to correct for large inserts and ambiguities. Final alignments contained 292 sequences 370-414 amino acids long from which 392 positions were selected for phylogenetic reconstruction. A total number of 330 informative positions were taken into account. Identity percentages were calculated using MULTALIN [63]. For additional information visit our database home page at http://www.iib.uam.es/MAT.

Evolutionary analysis

MAT's evolutionary tree was established from the 292 sequence alignment of the 303 sequences available using neighbor-joining and parsimony methods. MAT sequences for the following microorganisms were discarded in the final analysis due to ambiguity: *Amoeba proteus*, *Campylobacter jejuni*, *Escherichia coli II*, *Giardia lamblia*, *Mesorhizobium loti II*, *Nostoc sp. PCC 7120. II*, *Rickettsia prowazekii*, *Rickettsia typhi*, *Treponema pallidum*, *Leptospira interrogans* and *Tropherma whipplei*. Ambiguity may arise when there is a loss in the phylogenetic signal, obsolescence, insufficient taxon sampling or high sequence divergence. The phylogenetic analysis was performed using the PHYLIP package v.3.5 [64]. Distances were calculated using Dayhoff in PROTDIST followed by the use of NEIGHBOR or PROTPARS for tree reconstruction. Phylogenetic trees were drawn using the MEGA2 [65] and ATV programs [66]. Statistical support for the groups in the tree was evaluated by bootstrap analysis of 500 iterations using SEQBOOT.

Amino acid solvent accessibility

Coordinates for rl-MAT crystal structure [7] were used to calculate the solvent accessibility (SA) of each individual residue in the dimeric structure using the SwissPdbViewer [67]. Amino acids with a SA <10% are regarded as buried, whereas residues with SA >10% are considered exposed [68].

Footnotes

Acknowledgments

This work has been supported by grants of Fondo de Investigación Sanitaria of the Instituto de Salud Carlos III (01/1077 and RCMN C03/08) and MCYT (BMC-2002-00243) (to M.A.P.), and MCYT (PM99-0049-C02-01) (to J.M.B.).

Abbreviations: MAT, ATP:L-methionine adenosyltransferase; c-MAT, *E. coli* methionine adenosyltransferase; rl-MAT, rat liver methionine adenosyltransferase; SAM, S-adenosylmethionine; L-*cis*AMB, L-2-amino-4-methoxy-*cis*-but-3-enoic acid.

.

References

1. Cantoni, G. L. (1975). Biological methylation: selected aspects. *Annu Rev Biochem* 44, 435-51.
2. Mato, J. M., Alvarez, L., Ortiz, P. & Pajares, M. A. (1997). S-adenosylmethionine synthesis: molecular mechanisms and clinical implications. *Pharmacol Ther* 73, 265-80.
3. Cantoni, G. L. (1953). S-adenosylmethionine: a new intermediate formed enzymatically from L-methionine and adenosinetriphosphate. *J Biol Chem* 204, 403-416.
4. McQueney, M. S. & Markham, G. D. (1995). Investigation of monovalent cation activation of S-adenosylmethionine synthetase using mutagenesis and uranyl inhibition. *J Biol Chem* 270, 18277-84.
5. Mingorance, J., Alvarez, L., Sanchez-Gongora, E., Mato, J. M. & Pajares, M. A. (1996). Site-directed mutagenesis of rat liver S-adenosylmethionine synthetase. Identification of a cysteine residue critical for the oligomeric state. *Biochem J* 315 (Pt 3), 761-6.
6. Taylor, J. C. & Markham, G. D. (1999). The bifunctional active site of s-adenosylmethionine synthetase. Roles of the active site aspartates. *J Biol Chem* 274, 32909-14.
7. Gonzalez, B., Pajares, M. A., Hermoso, J. A., Alvarez, L., Garrido, F., Sufrin, J. R. & Sanz-Aparicio, J. (2000). The crystal structure of tetrameric methionine adenosyltransferase from rat liver reveals the methionine-binding site. *J Mol Biol* 300, 363-75.
8. Taylor, J. C. & Markham, G. D. (2000). The bifunctional active site of S-adenosylmethionine synthetase. Roles of the basic residues. *J Biol Chem* 275, 4060-5.
9. Sanchez-Perez, G. F., Gasset, M., Calvete, J. J. & Pajares, M. A. (2003). Role of an Intrasubunit Disulfide in the Association State of the Cytosolic Homo-oligomer Methionine Adenosyltransferase. *J Biol Chem* 278, 7285-7293.
10. Markham, G. D., DeParasis, J. & Gatmaitan, J. (1984). The sequence of metK, the structural gene for S-adenosylmethionine synthetase in Escherichia coli. *J Biol Chem* 259, 14505-7.
11. Thomas, D. & Surdin-Kerjan, Y. (1987). SAM1, the structural gene for one of the S-adenosylmethionine synthetases in Saccharomyces cerevisiae. Sequence and expression. *J Biol Chem* 262, 16704-9.
12. Larsson, J. & Rasmuson-Lestander, A. (1994). Molecular cloning of the S-adenosylmethionine synthetase gene in Drosophila melanogaster. *FEBS Lett* 342, 329-33.
13. Mautino, M. R., Barra, J. L. & Rosa, A. L. (1996). eth-1, the Neurospora crassa locus encoding S-adenosylmethionine synthetase: molecular cloning, sequence analysis and in vivo overexpression. *Genetics* 142, 789-800.
14. Yocum, R. R., Perkins, J. B., Howitt, C. L. & Pero, J. (1996). Cloning and characterization of the metE gene encoding S-adenosylmethionine synthetase from Bacillus subtilis. *J Bacteriol* 178, 4604-10.
15. Chiang, P. K., Chamberlin, M. E., Nicholson, D., Soubes, S., Su, X., Subramanian, G., Lanar, D. E., Prigge, S. T., Scovill, J. P., Miller, L. H. & Chou, J. Y. (1999). Molecular characterization of Plasmodium falciparum S-adenosylmethionine synthetase. *Biochem J* 344 Pt 2, 571-6.
16. Reguera, R. M., Balana-Fouce, R., Perez-Pertejo, Y., Fernandez, F. J., Garcia-Estrada, C., Cubria, J. C., Ordonez, C. & Ordonez, D. (2002). Cloning expression and

characterization of methionine adenosyltransferase in Leishmania infantum promastigotes. *J Biol Chem* 277, 3158-67.
17. Horikawa, S., Ishikawa, M., Ozasa, H. & Tsukada, K. (1989). Isolation of a cDNA encoding the rat liver S-adenosylmethionine synthetase. *Eur J Biochem* 184, 497-501.
18. Takusagawa, F., Kamitori, S. & Markham, G. D. (1996). Structure and function of S-adenosylmethionine synthetase: crystal structures of S-adenosylmethionine synthetase with ADP, BrADP, and PPi at 28 angstroms resolution. *Biochemistry* 35, 2586-96.
19. Woese, C. R. (1987). Bacterial evolution. *Microbiol Rev* 51, 221-71.
20. Gupta, R. S. & Golding, G. B. (1993). Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria, and eukaryotes. *J Mol Evol* 37, 573-82.
21. Gupta, R. S., Aitken, K., Falah, M. & Singh, B. (1994). Cloning of Giardia lamblia heat shock protein HSP70 homologs: implications regarding origin of eukaryotic cells and of endoplasmic reticulum. *Proc Natl Acad Sci U S A* 91, 2895-9.
22. Eisen, J. A. (1995). The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. *J Mol Evol* 41, 1105-23.
23. Baldauf, S. L., Palmer, J. D. & Doolittle, W. F. (1996). The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc Natl Acad Sci U S A* 93, 7749-54.
24. Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. *Nat Genet* 28, 281-5.
25. Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. & Doolittle, W. F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* 290, 972-7.
26. Brochier, C., Bapteste, E., Moreira, D. & Philippe, H. (2002). Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet* 18, 1-5.
27. Schroder, G., Eichel, J., Breinig, S. & Schroder, J. (1997). Three differentially expressed S-adenosylmethionine synthetases from Catharanthus roseus: molecular and functional characterization. *Plant Mol Biol* 33, 211-22.
28. Lindroth, A. M., Saarikoski, P., Flygh, G., Clapham, D., Gronroos, R., Thelander, M., Ronne, H. & von Arnold, S. (2001). Two S-adenosylmethionine synthetase-encoding genes differentially expressed during adventitious root development in Pinus contorta. *Plant Mol Biol* 46, 335-46.
29. Feng, D. F., Cho, G. & Doolittle, R. F. (1997). Determining divergence times with a protein clock: update and reevaluation. *Proc Natl Acad Sci U S A* 94, 13028-33.
30. Mushegian, A. R., Garey, J. R., Martin, J. & Liu, L. X. (1998). Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. *Genome Res* 8, 590-8.
31. Blair, J. E., Ikeo, K., Gojobori, T. & Hedges, S. B. (2002). The evolutionary position of nematodes. *BMC Evol Biol* 2, 7.
32. Graham, D. E., Bock, C. L., Schalk-Hihi, C., Lu, Z. J. & Markham, G. D. (2000). Identification of a highly diverged class of S-adenosylmethionine synthetases in the archaea. *J Biol Chem* 275, 4055-9.
33. Wicher, V., Baughn, R. E., Fuentealba, C., Shadduck, J. A., Abbruscato, F. & Wicher, K. (1991). Enteric infection with an obligate intracellular parasite, Encephalitozoon cuniculi, in an experimental model. *Infect Immun* 59, 2225-31.

34. Zomorodipour, A. & Andersson, S. G. (1999). Obligate intracellular parasites: Rickettsia prowazekii and Chlamydia trachomatis. *FEBS Lett* 452, 11-5.
35. Andersson, J. O. & Andersson, S. G. (1999). Genome degradation is an ongoing process in Rickettsia. *Mol Biol Evol* 16, 1178-91.
36. Taylor, J. C., Takusagawa, F. & Markham, G. D. (2002). The active site loop of S-adenosylmethionine synthetase modulates catalytic efficiency. *Biochemistry* 41, 9358-69.
37. Wierenga, R. K., Terpstra, P. & Hol, W. G. (1986). Prediction of the occurrence of the ADP-binding beta alpha beta-fold in proteins, using an amino acid sequence fingerprint. *J Mol Biol* 187, 101-7.
38. Gasset, M., Alfonso, C., Neira, J. L., Rivas, G. & Pajares, M. A. (2002). Equilibrium unfolding studies of the rat liver methionine adenosyltransferase III, a dimeric enzyme with intersubunit active sites. *Biochem J* 361, 307-15.
39. Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E. & Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 19, 163-4.
40. Notaro, R., Afolayan, A. & Luzzatto, L. (2000). Human mutations in glucose 6-phosphate dehydrogenase reflect evolutionary history. *Faseb J* 14, 485-94.
41. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2002). Plasticity of enzyme active sites. *Trends Biochem Sci* 27, 419-26.
42. Van de Peer, Y., Baldauf, S. L., Doolittle, W. F. & Meyer, A. (2000). An updated and comprehensive rRNA phylogeny of (crown) eukaryotes based on rate-calibrated evolutionary distances. *J Mol Evol* 51, 565-76.
43. Stechmann, A. & Cavalier-Smith, T. (2002). Rooting the eukaryote tree by using a derived gene fusion. *Science* 297, 89-91.
44. Roger, A. J., Sandblom, O., Doolittle, W. F. & Philippe, H. (1999). An evaluation of elongation factor 1 alpha as a phylogenetic marker for eukaryotes. *Mol Biol Evol* 16, 218-33.
45. Baldauf, S. L. (1999). A Search for the Origins of Animals and Fungi: Comparing and Combining Molecular Data. *Am Nat* 154, S178-S188.
46. Espartero, J., Pintor-Toro, J. A. & Pardo, J. M. (1994). Differential accumulation of S-adenosylmethionine synthetase transcripts in response to salt stress. *Plant Mol Biol* 25, 217-27.
47. Vander Mijnsbrugge, K., Van Montagu, M., Inze, D. & Boerjan, W. (1996). Tissue-specific expression conferred by the S-adenosyl-L-methionine synthetase promoter of Arabidopsis thaliana in transgenic poplar. *Plant Cell Physiol* 37, 1108-15.
48. Lee, J. H., Chae, H. S., Hwang, B., Hahn, K. W., Kang, B. G. & Kim, W. T. (1997). Structure and expression of two cDNAs encoding S-adenosyl-L-methionine synthetase of rice (Oryza sativa L.). *Biochim Biophys Acta* 1354, 13-8.
49. Gomez-Gomez, L. & Carrasco, P. (1998). Differential expression of the S-adenosyl-L-methionine synthase genes during pea development. *Plant Physiol* 117, 397-405.
50. Kende, H. (1993). Ethylene biosynthesis. *Annu Rev Plant Physiol Plant Mol Biol* 44, 283-307.
51. Baldauf, S. L. & Palmer, J. D. (1993). Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc Natl Acad Sci U S A* 90, 11558-62.

52. Wolfe, K. H. & Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708-13.
53. Lu, S. C., Gukovsky, I., Lugea, A., Reyes, C. N., Huang, Z. Z., Chen, L., Mato, J. M., Bottiglieri, T. & Pandol, S. J. (2003). Role of S-adenosylmethionine in two experimental models of pancreatitis. *Faseb J* 17, 56-8.
54. McLysaght, A., Hokamp, K. & Wolfe, K. H. (2002). Extensive genomic duplication during early chordate evolution. *Nat Genet* 31, 200-4.
55. Gil, B., Casado, M., Pajares, M. A., Bosca, L., Mato, J. M., Martin-Sanz, P. & Alvarez, L. (1996). Differential expression pattern of S-adenosylmethionine synthetase isoenzymes during rat liver development. *Hepatology* 24, 876-81.
56. Chang, S. H. & Kemp, R. G. (2002). Role of Ser530, Arg292, and His662 in the allosteric behavior of rabbit muscle phosphofructokinase. *Biochem Biophys Res Commun* 290, 670-5.
57. Pezza, J. A., Choi, K. H., Berardini, T. Z., Beernink, P. T., Allen, K. N. & Tolan, D. R. (2003). Spatial Clustering of Isozyme-specific Residues Reveals Unlikely Determinants of Isozyme Specificity in Fructose-1,6-bisphosphate Aldolase. *J Biol Chem* 278, 17307-13.
58. Avila, M. A., Mingorance, J., Martinez-Chantar, M. L., Casado, M., Martin-Sanz, P., Bosca, L. & Mato, J. M. (1997). Regulation of rat liver S-adenosylmethionine synthetase during septic shock: role of nitric oxide. *Hepatology* 25, 391-6.
59. Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A. & Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387, 489-93.
60. Hendrickson, E. R., Payne, J. A., Young, R. M., Starr, M. G., Perry, M. P., Fahnestock, S., Ellis, D. E. & Ebersole, R. C. (2002). Molecular analysis of Dehalococcoides 16S ribosomal DNA from chloroethene-contaminated sites throughout North America and Europe. *Appl Environ Microbiol* 68, 485-95.
61. Kamla, V., Henrich, B. & Hadding, U. (1996). Phylogeny based on elongation factor Tu reflects the phenotypic features of mycoplasmas better than that based on 16S rRNA. *Gene* 171, 83-7.
62. Hall, T. A. (1999). BioEdit: a user-friendly biological alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Sex* 41, 95-98.
63. Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* 16, 10881-90.
64. Felsenstein, J. (1993). PHYLIP (phylogeny inference package) version 3.5c. *University of Washington, Seattle*.
65. Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* 17, 1244-5.
66. Zmasek, C. M. & Eddy, S. R. (2001). ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics* 17, 383-4.
67. Guex, N. & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18, 2714-23.
68. Goldman, N., Thorne, J. L. & Jones, D. T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149, 445-58.

Figure legends

Figure 1. Evolution of the MAT protein. This diagram shows the analysis of 292 different MAT sequences. Rat liver MAT (lower row) was used as the reference sequence. The lettering for rat liver MAT indicates amino acid solvent accessibility (SA): blue, buried (SA<10%); black, exposed (SA>10%). Immediately above the rat liver MAT sequence, we show the identity consensus sequence for the 292 sequences analyzed. At each position, the most frequent amino acid is coded according to its frequency (blue: 100%; green: 90-99%; red: 75-90%; yellow: 60-75%). The conserved blocks described in the text are boxed and identified with Roman numerals. The locations of α-helices and β-strands as found in the rat liver MAT 3D structure are provided above the identity consensus sequence. The structure of the first 16 residues (lowercase) in rat liver MAT is not known.

Figure 2. Conservation and 3-dimensional structure. Each bar represents the distribution of MAT amino acids among the different conservation categories for exposed, total, and buried residues, respectively.

Figure 3. Diagram showing MAT sequence conservation in 3-dimensional structure. The conservation pattern is color-coded on the molecular surface of rat liver MAT: dark violet indicates maximal conservation (100-90% identity), white indicates an average conservation level (90-50%) and dark turquoise indicates maximal variability (<50% identity). (**A**) Dimer viewed from the active site entrance. (**B**) Side view of the dimer after a 90º rotation to the right showing the exposed surface. (**C**) A view of the monomer from the monomer-monomer interface. (**D**) Side view of the monomer after a 180º rotation showing the exposed surface.

Figure 4. Phylogenetic tree derived from MAT sequences. Numbers on the branches show the percentage occurrence of nodes in 500 bootstrap replicates in the neighbor-joining and maximum parsimony analyses. Bootstrap values are indicated only when greater than 70%. For simplicity, the branches within each major group were collapsed.

Figure 5. Differences between rat liver MAT (**A**) (PDB ID code: 1QM4) and *E. coli* MAT (**B**) (PDB ID code: 1fug) for the atomic interactions between loops connecting A2 and A3 strands in the N-terminal domain (yellow ribbons and red backbone and side chains) and loops connecting helix 5, C2 and C3 strands in the C-terminal domain (blue ribbons and cyan backbone and side chains).

Tables

Table 1. List of MAT sequences used in the present analysis.

Specie	**Accesion Nº / Genome center**
Acanthamoeba castellanii	**927487**
Acidithiobacillus ferroxidans	TIGR
Acinetobacter ADP-1	Genoscope
Actinidia chinensis 1	**726030**
Actinidia chinensis 2	**726028**
Actinobacillus actinomycetemcomitans	Univ. Oklahoma
Agrobacterium tumefaciens	**17934278**
Ajellomyces capsulatus	Univ. Washington
Anaplasma phagocytophila	TIGR
Anopheles gambiae	**21291484**
Aquifex aeolicus	**7387875**
Arabidopsis thaliana 1	**15217781**
Arabidopsis thaliana 2	**15234354**
Arabidopsis thaliana 3	**15229033**
Arabidopsis thaliana 4	**15228048**
Ascaris lumbricoides	EBI Parasites
Ascobolus immersus	**836960**
Aspergillus fumigatus	TIGR
Aspergillus nidulans	Whitehead/MIT
Azotobacter vinelandii	**23102724**
Bacillus anthracis	**21402812**
Bacillus cereus 1	TIGR
Bacillus cereus 2	**29898393**
Bacillus halodurans	**20138752**
Bacillus stearothermophilus	Univ. Oklahoma
Bacillus subtilis	**7434008**
Bacteroides forsythus	TIGR
Bacteroides fragilis	Sanger Institute
Bacteroides thetaiotaomicron	**29340533**
Bifidobacterium longum	**23327085**
Bombyx mori	EST NCBI
Bordetella avium	Sanger Institute
Bordetella bronchiseptica	Sanger Institute
Bordetella pertussis	Sanger Institute
Borrelia burgdorferi	**7434004**
Botrytis cinerea	Genoscope
Bradyrhizobium japonicum	**27354222**
Brassica juncea 1	**10443981**
Brassica juncea 2	**14600070**
Brassica juncea 3	**14600072**
Brucella melitensis	**17988253**

Brucella suis	**23349048**
Buchnera aphidicola	**11386917**
Burkholderia cepacia	Sanger Institute
Burkholderia fungorum	**22985056**
Burkholderia mallei	TIGR
Burkholderia pseudomallei	Sanger Institute
Caenorhabditis briggsae 1	Sanger Institute
Caenorhabditis briggsae 2	Sanger Institute
Caenorhabditis briggsae 3	Sanger Institute
Caenorhabditis elegans 1	**17538494**
Caenorhabditis elegans 2	**1753849**
Caenorhabditis elegans 3	**21106027**
Caenorhabditis elegans 4	**7509275**
Caenorhabditis elegans 5	**17551082**
Camellia sinensis	**7594741**
Campylobacter jejuni	**11258525**
Candida albicans	**7271000**
Carboxydothermus hydrogenoformans	TIGR
Carica papaya	**22774026**
Catharanthus roseus 1	**1655576**
Catharanthus roseus 2	**1655578**
Catharanthus roseus 3	**1655580**
Caulobacter crescentus	**16124306**
Chlamydomonas reinhardtii	DOE Joint Institute
Chlorobium tepidum	**21646663**
Chloroflexus aurantiacus	**22971187**
Cicer arietinum	**1808591**
Ciona intestinalis	**23586111**
Clavibacter michiganensis	Sanger Institute
Clostridium acetobutylicum	**15896110**
Clostridium botulinum	Sanger Institute
Clostridium perfringens	**18311159**
Clostridium tetani	**28202517**
Coccidioides posadasii	TIGR
Colwellia psychroerythraea	TIGR
Corynebacterium diphtheriae	TIGR
Corynebacterium efficiens	**23493563**
Corynebacterium glutamicum	**19552815**
Coxiella burnetii	TIGR
Cryptococcus neoformans	TIGR
Cryptosporidium parvum	TIGR
Cytophaga hutchinsonii	**23137174**
Danio rerio 1	**28278852**
Danio rerio 2	TIGR
Dechloromonas aromatica	DOE Joint Institute
Dehalococcoides ethenogenes	TIGR
Deinococcus radiodurans	**15805667**
Dendrobium crumenatum	**16226059**

Desulfovibrio desulfuricans	**23475965**
Desulfovibrio vulgaris	TIGR
Dianthus caryophyllus	**7434012**
Dichelobacter nodosus	TIGR
Dictyostelium discoideum	Sanger Institute
Drosophila melanogaster	**7296263**
Drosophila pseudoobscura	Baylor
Ehrlichia chaffensis	TIGR
Ehrlichia ruminantium	Sanger Institute
Elaeagnus umbellata 1	**13540316**
Elaeagnus umbellata 2	**13540318**
Enterococcus faecalis	**29342835**
Escherichia coli (metK)	**1708999**
Escherichia coli (metX)	**26250367**
Fibrobacter succinogenes	TIGR
Fusarium sporotrichioides	Univ. Oklahoma
Fusobacterium nucleatum	**19703697**
Gallus gallus	TIGR
Gemmata obscuriglobus	TIGR
Giardia lamblia	**29247850**
Glycine max 1	TIGR
Glycine max 2	TIGR
Glycine max 3	TIGR
Gossypium hirsutum 1	TIGR
Gossypium hirsutum 2	TIGR
Gossypium hirsutum 3	TIGR
Haemophilus influenzae	**1170942**
Haemophilus somnus	**23467639**
Helicobacter pylori	**3024119**
Helicobacter pylori J99	**6685665**
Heliobacillus mobilis	**27262362**
Homo sapiens 1	**4557737**
Homo sapiens 2	**284394**
Hordeum vulgare 1	**7434000**
Hordeum vulgare 2	TIGR
Hordeum vulgare 3	TIGR
Ictalurus punctatus	TIGR
Klebsiella pneumoniae	Univ. Washington
Lactobacillus gasseri	DOE Joint Institute
Lactobacillus plantarum	**28378057**
Lactococcus lactis	**13878576**
Lactuca sativa 1	TIGR
Lactuca sativa 2	TIGR
Legionella pneumophila	Columbia
Leishmania infantum	**20387266**
Leptospira interrogans	**24215333**
Leuconostoc mesenteroides	**23023832**
Listeria innocua	**1680041**

Listeria monocytogenes	**16411100**
Litchi chinensis	**30142157**
Lotus japonicus 1	**21907982**
Lotus japonicus 2	TIGR
Lycopersicon esculentum 1	**1084406**
Lycopersicon esculentum 2	**481566**
Lycopersicon esculentum 3	**1084408**
Lycopersicon esculentum 4	TIGR
Magnaporthe grisea	TIGR
Magnetococcus sp. MC-1	**23000957**
Magnetospirillum magnetotacticum	**23015399**
Mannheimia haemolytica	Baylor
Medicago truncatula 1	TIGR
Medicago truncatula 2	TIGR
Medicago truncatula 3	TIGR
Medicago truncatula 4	TIGR
Mesembryanthemum crystallinum	**1724104**
Mesorhizobium loti 1	**20803994**
Mesorhizobium loti 2	**13475107**
Methylobacterium extorquens	Univ. Washington
Methylococcus capsulatus	TIGR
Microbulbifer degradans	**23027148**
Moraxella catarrhalis	DDJB Japan
Mus musculus 1	**476917**
Mus musculus 2	**13097429**
Musa acuminata	**2305014**
Mycobacterium avium	TIGR
Mycobacterium bovis	Sanger Institute
Mycobacterium leprae	**15214074**
Mycobacterium marinum	Sanger Institute
Mycobacterium smegmatis	TIGR
Mycobacterium tuberculosis	**3915763**
Mycoplasma genitalium	**1346527**
Mycoplasma penetrans	**26553547**
Mycoplasma pneumoniae	**2500686**
Mycoplasma pulmonis	**15829173**
Myxococcus xanthus	**27804841**
Neisseria gonorrhoeae	Univ. Oklahoma
Neisseria meningitidis	**11258516**
Neorickettsia sennetsu	TIGR
Neurospora crassa	**2133316**
Nicotiana tabacum	**7230379**
Nitrosomonas europaea	**22954639**
Nostoc sp. PCC 7120 1	**17231616**
Nostoc sp. PCC 7120 2	**17132339**
Novosphingobium aromaticivorans	**23110658**
Oceanobacillus iheyensis	**22777999**
Oenococcus oeni	**23038264**

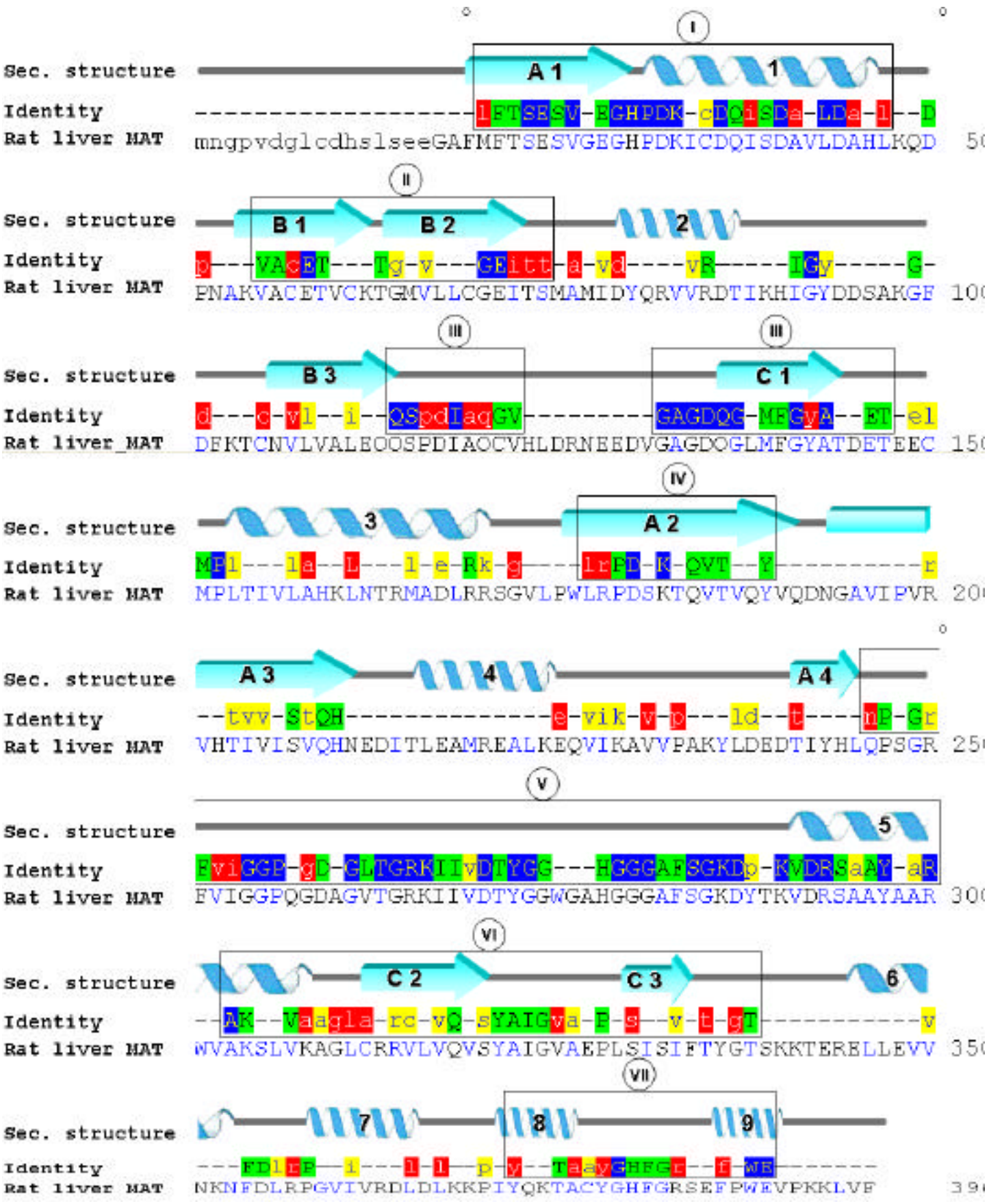
Oryza sativa 1	**450549**
Oryza sativa 2	**1778821**
Oryza sativa 3	**8468037**
Pasteurella multocida	**13431697**
Pectobacterium carotovorum	Sanger Institute
Pectobacterium chrysanthemi	TIGR
Petunia x hybrida 1	**1084428**
Petunia x hybrida 2	**5726594**
Phanerochaete chrysosporium	DOE Joint Institute
Phaseolus lunatus	**18157331**
Photorhabdus asymbiotica	Sanger Institute
Phytophthora infestans	**23394401**
Pinus banksiana	**1033190**
Pinus contorta 1	**10441429**
Pinus contorta 2	**10441431**
Pinus contorta 3	TIGR
Pinus contorta 4	TIGR
Pinus contorta 5	TIGR
Pisum sativum	**2129889**
Plasmodium berghei	Sanger Institute
Plasmodium chabaudi	Sanger Institute
Plasmodium falciparum	**10129955**
Plasmodium knowlesi	Sanger Institute
Plasmodium yoelii	**23482440**
Populus deltoides	**497900**
Porphyromonas gingivalis	TIGR
Prevotella intermedia	TIGR
Prochlorococcus marinus 1	**23132136**
Prochlorococcus marinus 2	**23122176**
Pseudomonas aeruginosa	**15595743**
Pseudomonas fluorescens	**23060543**
Pseudomonas putida	**26991645**
Pseudomonas syringae	**23471420**
Psychrobacter sp.273-4	DOE Joint Institute
Ralstonia eutropha	DOE Joint Institute
Ralstonia metallidurans	**22979451**
Ralstonia solanacearum	**17544853**
Rattus norvegicus 1	**92483**
Rattus norvegicus 2	**19705457**
Rhizobium leguminosarum	Sanger Institute
Rhodobacter sphaeroides	**22957691**
Rhodopseudomonas palustris	**22961217**
Rhodospirillum rubrum	**22965557**
Rickettsia prowazekii	**7387879**
Rickettsia typhi	**11133588**
Saccharomyces cerevisiae 1	**1346525**
Saccharomyces cerevisiae 2	**83324**
Salmonella typhi	**16766391**

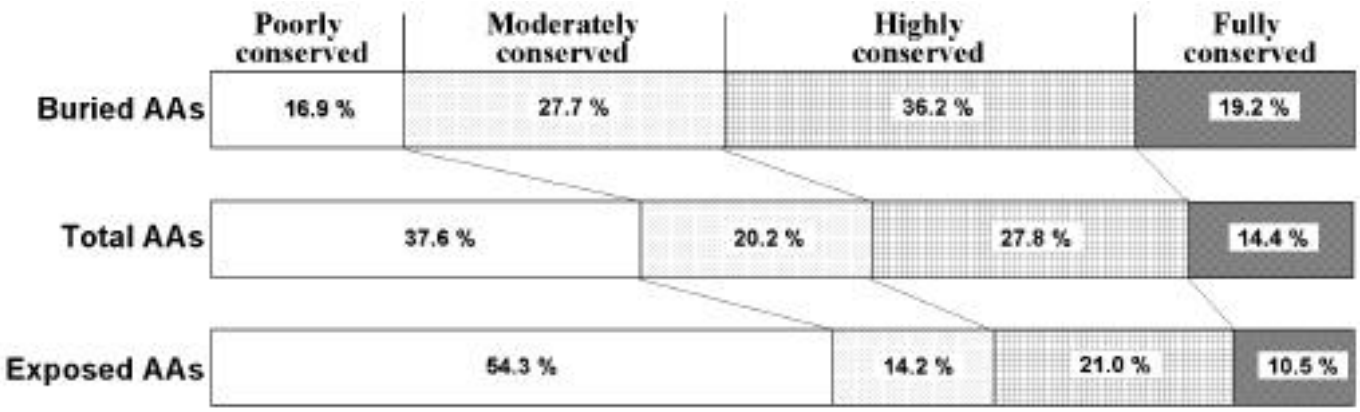
Schistosoma japonicum	EBI Parasites
Schizosaccharomyces pombe	**7493352**
Secale cereale	TIGR
Serratia marcescens	Sanger Institute
Shewanella oneidensis	**24372516**
Shigella flexneri	**24114197**
Silicibacter pomeroyi	TIGR
Sinorhizobium meliloti	**15964164**
Solanum tuberosum 1	TIGR
Solanum tuberosum 2	TIGR
Solanum tuberosum 3	TIGR
Sorghum bicolor	TIGR
Spiroplasma kunkelii	Univ. Oklahoma
Staphylococcus aureus	**1709003**
Staphylococcus epiderminis	**9624212**
Streptococcus agalactiae	**29611810**
Streptococcus equi	Sanger Institute
Streptococcus gordonii	TIGR
Streptococcus mitis	TIGR
Streptococcus mutans	**29611808**
Streptococcus pneumoniae	**15902715**
Streptococcus pyogenes	**19746332**
Streptococcus sobrinus	TIGR
Streptococcus suis	Sanger Institute
Streptococcus thermophilus	Univ. C. Louvain
Streptococcus uberis	Sanger Institute
Streptomyces avermitilis	**29833416**
Streptomyces coelicolor	**21219978**
Streptomyces fradiae	**15554326**
Streptomyces pristinaespiralis	**2294502**
Streptomyces spectabilis	**7387884**
Suaeda maritima	**11992267**
Synechococcus sp.	**23134392**
Synechocystis sp.	**7434007**
Takifugu rubripes 1	DOE Joint Institute
Takifugu rubripes 2	DOE Joint Institute
Tetraodon nigroviridis	Genoscope
Thalassiosira pseudonana	DOE Joint Institute
Theileria annulata	Sanger Institute
Theileria parva	TIGR
Thermobifida fusca	**23017995**
Thermosynechococcus elongatus	**29611806**
Thermotoga maritima	**7387883**
Treponema pallidum	**15639781**
Trichodesmium erythraeum	**23040652**
Triticum aestivum 1	TIGR
Triticum aestivum 2	TIGR
Triticum aestivum 3	TIGR

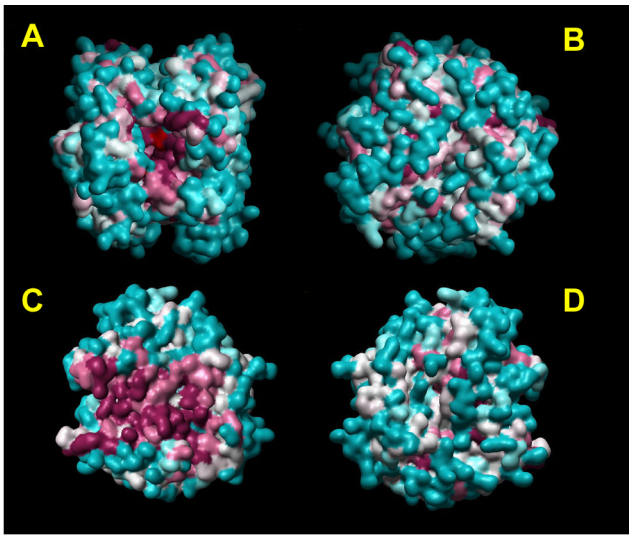
Triticum aestivum 4	TIGR
Tropheryma whipplei	**28572562**
Trypanosoma brucei	TIGR
Trypanosoma cruzi	Sanger Institute
Ureaplasma urealyticum	**13357974**
Vibrio cholerae	**9654898**
Vibrio parahaemolyticus	**28899380**
Vibrio vulnificus	**27364907**
Wigglesworthia brevipalpis	**24324056**
Wolbachia sp.	TIGR
X-bacteria	Direct submission
Xanthomonas axonopodis	**21241583**
Xanthomonas campestris	**21230235**
Xenopus laevis 1	TIGR
Xenopus laevis 2	**27882050**
Xenopus tropicalis	Sanger Institute
Xylella fastidiosa	**15836994**
Yersinia enterocolitica	Sanger Institute
Yersinia pestis	**16121235**
Zea mays 1	TIGR
Zea mays 2	TIGR

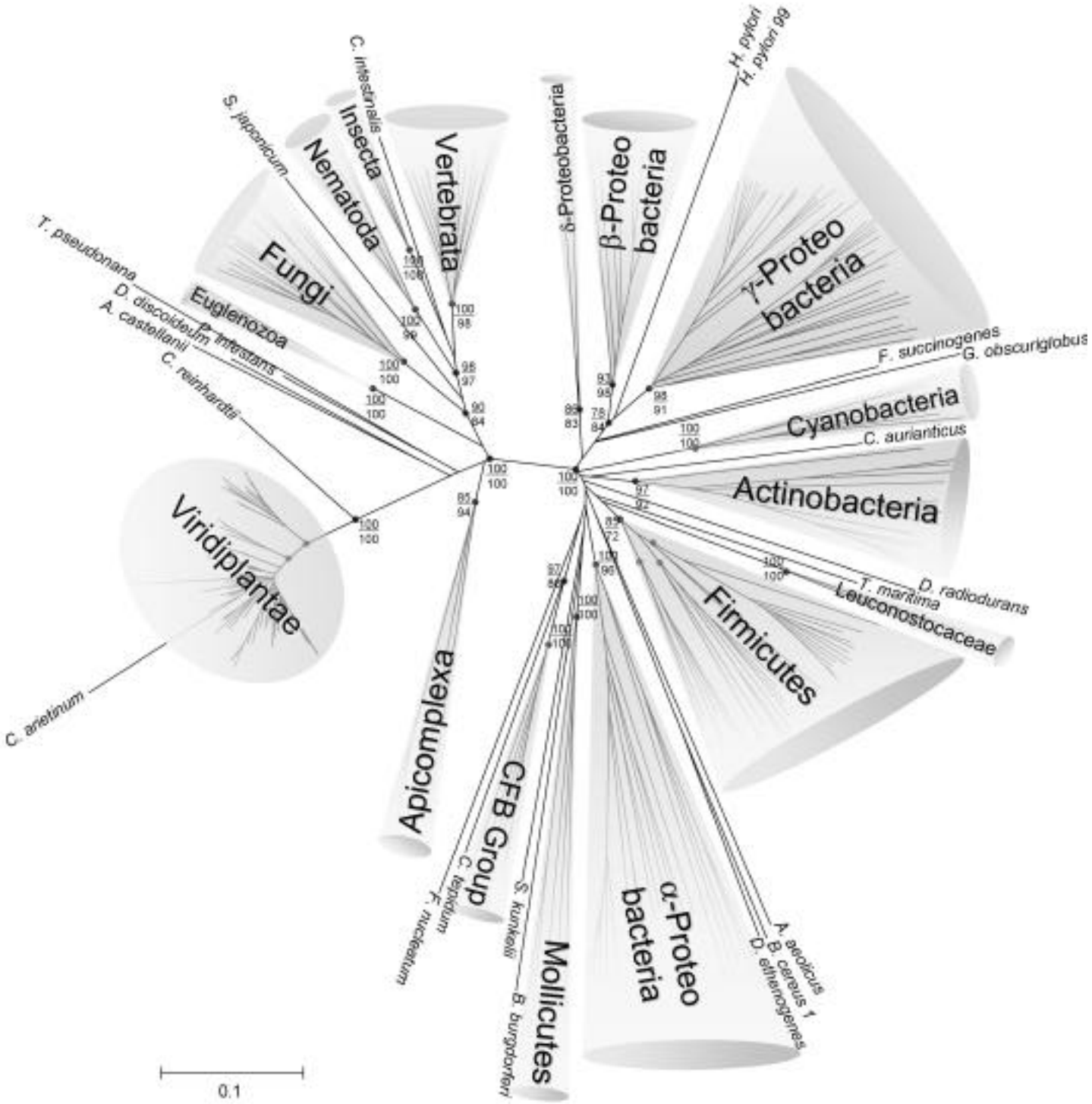
Table 2. Specific structural features shown by each group on the MAT phylogenetic tree.

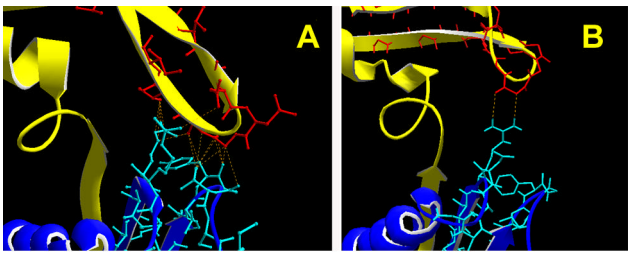
Eukaryota	The loop connecting β-strands A2-A3 has 10 residues		
	Plants	Presence of the sequences ^{93}FXSXDVXLXAD^{103}, ^{216}NDEIA^{220}, and ^{331}VFVD^{334}	
		Insertion of 3-4 residues in the loop connecting helixes 7 and 8, between positions 369-370	
	Vertebrates	**Fishes**	Present specific residues for the MAT I enzyme, C69, C377 and V262
		Birds/reptiles	Present additional specific residues such as R82 and D167, besides those shown in fishes
		Mammals	Present a typical cysteine in the MAT I flexible loop, C121 besides the residues specific for fishes, birds and reptiles
Eubacteria	Deletion of 3-4 residues in the loop connecting β-strands A2-A3		
	Proteobacteria	Deletion of 2 residues in the loop connecting helix 2 and β-strand B3	
		Insertion of 4 residues between helix 4 and β-strand A4	
		Isertion of 4-5 residues between helixes 8-9	
	Cyanobacteria	Insertion of 4 residues in the loop between β-strand B3 and helix 4	
		Insertion of 6-7 residues in the loop connecting helix 4 and β-strand A4	
		Insertion of 7 residues in the loop between helixes 7 and 8	
		Insertion of 2 residues in the loop connecting helixes 8 and 9	
	CFB group	Insertion of 1-2 residues in the loop connecting helix 3 and β-strand A2	
		Insertion of 2 residues in the loop connecting β-strand C3 and helix 6	
		Insertion of 9-11 residues in the loop connecting helix 4 and β-strand A4	
		Insertion of 21-22 residues in the loop connecting helixes 8 and 9	
	Actinobacteria	Insertion of 8 residues in the flexible loop	
		Insertion of 2-4 residues in the loop connecting helix 4 and β-strand A4	
	Firmicutes	**Mollicutes**	Deletion of 7 residues in the loop connecting helix 2 and β-strand B3


Sec. structure
Identity
Rat liver MAT
mngpvdglcdhslseeGAFMFTSESVGEGHPDKICDQISDAVLDAHLKQD
PNAKVACETVCKTGMVLLCGEITSMAMIDYQRVVRDTIKHIGYDDSAKGF
DFKTCNVLVALEQQSPDIAQCVHLDRNEEDVGAGDQGLMFGYATDETEEC
MPLTIVLAHKLNTRMADLRRSGVLPWLRPDSKTQVTVQYVQDNGAVIPVR
VHTIVISVQHNEDITLEAMREALKEQVIKAVVPAKYLDEDTIYHLQPSGR
FVIGGPQGDAGVTGRKIIVDTYGGWGAHGGGAFSGKDYTKVDRSAAYAAR
WVAKSLVKAGLCRRVLVQVSYAIGVAEPLSISIFTYGTSKKTERELLEVV
NKNFDLRPGVIVRDLDLKKPIYQKTACYGHFGRSEFPWEVPKKLVF


Poorly conserved
Moderately conserved
Highly conserved
Fully conserved
Buried AAs
16.9 %
27.7 %
36.2 %
19.2 %
Total AAs
37.6 %
20.2 %
27.8 %
14.4 %
Exposed AAs
54.3 %
14.2 %
21.0 %
10.5 %


A
B
C
D


Vertebrata
Insecta
Nematoda
Fungi
Euglenozoa
Viridiplantae
Apicomplexa
CFB Group
Mollicutes
α-Proteo bacteria
β-Proteo bacteria
γ-Proteo bacteria
δ-Proteobacteria
Cyanobacteria
Actinobacteria
Firmicutes
Leuconostocaceae
C. intestinalis
S. japonicum
T. pseudonana
D. discoideum
A. castellanii
P. tetrasans
C. reinhardtii
C. arietinum
H. pylori
H. pylori 99
F. succinogenes
G. obscuriglobus
C. aurantiacus
T. maritima
D. radiodurans
A. aeolicus
B. cereus 1
D. ethanogenes
S. kunkeii
B. burgdorferi
C. tepidum
F. nucleatum
0.1


A
B