# Supporting Information for "The evolutionary history of Cys-tRNA$^{\text{Cys}}$ formation"

Patrick O'Donoghue, Anurag Sethi, Carl R. Woese and Zaida Luthey-Schulten

## 1 Methods

**Sequence information.** The Institute for Genomic Research (TIGR), Joint Genome Institute (JGI), and NCBI genomic databases were searched for pseudogenes of SepRS and SepCysS using the `tblastn` program with a profiles composed of the known members of these protein families. These searches were carried out in the genomes of the euryarchaeal phyla that do not have proteins coding for the indirect pathway of cysteine aminoacylation and also on all known microbial genomes. No candidate pseudogenes were identified. Sequence searches were carried out using BLAST [1] with sequences or profile-based methods, as in ref. [2], to retrieve homologs belonging to a particular protein family.

**Alignments.** Due to the high sequence similarity among the SepRSs (sequence identity range of 53%-94%) and separately among the SepCysSs (sequence identity range of 65%-92%), alignments for the two groups were generated using CLUSTAL W [3]. Aside from small manual adjustments and the addition of some new sequences, the individual alignments of the PheRS $\alpha$-subunits, PheRS $\beta$-subunits, CysRSs and GluRSs are the same as those used by Woese and colleagues in ref. [4].

Of the sequences included in our study, the average sequence identity between the SepRSs and the PheRS $\alpha$-subunits is 23%, while the SepRSs share only 15% sequence identity on average with the PheRS $\beta$-subunits. The PheRS $\beta$-subunits and $\alpha$-subunits also share about 15% sequence identity on average. The higher similarity between the SepRSs and the PheRS $\alpha$-subunits, along with blocks of high conservation allow a fairly accurate alignment to be established between the SepRSs and the PheRS $\alpha$-subunits. These groups were initially aligned using CLUSTAL W with subsequent manual adjustments. Because sequence similarity relationships between the SepRS/PheRS $\alpha$-subunit group and the PheRS $\beta$-subunits is too low to generate a composite alignment with sequence methods alone, a structural alignment between the crystallographic structures of the $\alpha$ and $\beta$ chain PheRSs from *Thermus thermophilus* (PDB code 1b7y) was required. The SepRS/PheRS $\alpha$-subunit group sequences were then aligned to the PheRS $\beta$-subunits according to the structural alignment. The next most closely related aaRS, AlaRS (1riq), was used as an outgroup to root the SepRS/PheRS $\alpha$-subunit/PheRS $\beta$-subunit phylogeny. A structural alignment between AlaRS and the

PheRS structures was used to align AlaRS to the larger group. A similar procedure, making use of the crystallographic structures of the *E. coli* CysRS (1li5), *T. thermophilus* GluRS (1gln), *E. coli* GlnRS (1gtr), and the outgroup protein, *P. horikoshii* LysRS (1irx), was used to align the CysRSs to the GluRSs.

Both the group I and group II desulfurases were aligned individually using CLUSTAL W. Sequence-based QR factorization [2] was applied with 45% sequence identity threshold to these alignments to remove redundancy within these groups. Sequences belonging to phylogenetically relevant groupings were added if there were no representatives present from these groupings in the reduced sets. The group I and the group II desulfurases were aligned using the profile-profile alignment algorithm in CLUSTAL W. Based on the structural alignment of 1ecx, 1p3w, and 1kmj, minor adjustments were made to improve the alignment of the group I desulfurase to the group II desulfurases. A representative set of the SepCysS proteins were also made by applying sequence-based QR with a threshold of 55% sequence identity. The representative set of desulfurases were aligned to the evolutionary profile (EP) of SepCysS proteins using profile-profile alignment in CLUSTAL W with subsequent manual adjustments.

**Phylogenetic analysis.** Distance-based trees were constructed using PHYLIP v.3.6 [5]. Phylogenetic analysis, performed according to current protocols [4,6], involved a combination Maximum Parsimony (MP), to search the tree topology space, and Maximum likelihood (ML), which was used to rank the tree topologies and assign branch lengths. Since the MP/ML protocol is computationally intensive, especially with large numbers of sequences, representative sequence sets were used. Sequence-based QR factorization [2] was used on the bacterial sequences with a threshold such that one or more members of each of the major bacterial phyla were retained. Only two or three representative eukaryotes were retained, as well as three *Crenarchaea*. In order to determine if there is a congruence with the SepRS and SepCysS trees, all the *Euryarchaea* that contain SepRS and SepCysS are explicitly represented in the other trees. Some euryarchaeal classes, such as the *Halobacteriales*, *Thermococcales* and *Thermoplasmatales*, were left out or only minimally represented.

For a given alignment, PAUP4b10 [7] was used to generate the thousand most parsimonious tree topologies. Gaps were included as a character and the BLOSUM45 amino acid substitution matrix was used [8,9]. All default options were used except for the following options: none of the branches were collapsed during parsimony (collapse = no), sequences were added using the simple method (addseq=simple), and the trees which are saved during a search for optimal trees are used as input during branch swapping procedure (allswap=yes).

The thousand most parsimonious trees were analyzed using maximum likelihood analysis as implemented in PHYML v.2.4.4 [10] with the options `phyml alignment.phy 1 i 1 0 JTT 0 8 e treesFromPAUP.tre` `n n` to determine the tree topology that best represents the sequence alignment. At this step, the branch lengths and branch order are not optimized for each of the thousand trees. For the phylogenetic tree from SepRS, PheRS and AlaRS, the tree topology, branch lengths, and rate parameters were optimized with PHYML using the command `phyml alignment.phy 1 i 1 0 JTT 0 8 e bestTreefromPHYML.tre y y`. The maximum likelihood tree in PHYML was constructed using the JTT model [11] without invariable sites, a gamma distribution with eight rate categories and the shape parameter estimated from the data. PROTML, from the Molphy 3.2 package [12], was used to compute the local bootstrap proportions based on the re-estimation of log likelihoods (RELL) method for all the phylogenetic trees shown. PROTML was used with the following command : `protml -j -uRX -I alignment.phy PHYML.optimizedTree.usertre`. For the SepCysS tree (twelve SepCysS proteins and a cysteine desulfurase as an outgroup), it was found that the branch lengths and bootstrap values provided by PROTML were comparable to the values obtained using the maximum likelihood bootstrap procedure in PHYML, with tree topology and branch lengths optimized for each bootstrap replicate. Since the approximate RELL method employed in PROTML is 60 times faster than the PHYML method, PROTML was used to obtain the branch lengths and local bootstrap proportions for the alignments with a larger number of sequences containing all the SepCysS proteins and representatives of different phyla of cysteine desulfurase proteins. TreeViewX [13] and MATLAB Release 14 (Mathworks, Natick, MA) phylogenetic tree tool were used for displaying and drawing trees.

**Modeling and refinement by molecular dynamics** As described previously [2], structural templates for homology modeling were found using profiles of all the available sequences of SepRS (target) in a BLAST search [1] of the sequences in the ASTRAL database. The initial alignment of the template to target was generated using CLUSTAL's profile-profile alignment method, with a profile of the SepRSs and a profile of the template (PDB code 1b7y) and its homologs. This alignment was manually modified based on the agreement of secondary structure prediction from PSIPRED [14] and the secondary structure elements in the structural alignment. Homology modeling was performed using MODELLER v. 6.2 [15] with loop optimization. The procedure was repeated for SepCysS and the three well-resolved structures of the pyridoxal 5′-phosphate (PLP)-dependent transferase cysteine desulfurases as structural templates (PDB code 1kmj, 1ecx, and 1p3w). All three structures have E-value $< 10^{-8}$ in the BLAST search with an EP of SepCysS proteins. The active site of proteins belonging to PLP-dependent transferase family is formed at the interface of two monomers in a dimeric structure. The dimeric structure of MJ1678 was modeled using the dimeric structure 1p3w as a template.

In order to generate a more realistic model of the active site of *M. jannaschii* SepRS, a 25,000 fs successive minimization and a 1 ns equilibration simulation were performed using molecular dynamics (MD) package NAMD [16] and the classical force field, CHARMM [17]. For the solvent, explicit water molecules and ions were included. Force field parameters for the *O*-phosphoseryl-adenylate (Sep-AMP) substrate were generated by analogy with molecules already parameterized in the CHARMM forcefield, including serine and adenosine monophosphate. The topology and parameter files for Sep-AMP are available on the web [1].

We modeled the active site of MJ1678 with PLP in the aldimine form and *O*-phosphoserine attached to the trinucleotide CCA, representing the acceptor stem of $\mathrm{tRNA}^{\mathrm{Cys}}$. The parameters for PLP were found in ref. [18]. The protonation state of PLP was determined, assuming neutral pH [19], and the position of PLP in the active site was modeled from the structure 1kmj. MD was used to minimize and equilibrate the modeled structure in a box of explicit solvent. The modeled structure was first minimized and then equilibrated for 2 ns with the residue K234 bound to PLP in the aldimine form. To determine the probable docking site of the aminoacyl-like tRNA in the active site, the electrostatic potential was calculated with particle mesh Ewald (PME) [20] approximation averaged over 100 frames in a 2 ns of the equilibration. The substrate *O*-phosphoserine bound to CCA was then added. The original placement of *O*-phosphoserine was based on the position of substrate cysteine in 1kmj. The initial conformation for *O*-phosphoserine attached to the CCA end of the tRNA was taken from the conformation of cysteine attached to $\mathrm{tRNA}^{\mathrm{Cys}}$ bound to the elongation factor in the structure 1b23 [21]. The modeled structure with PLP and the Sep-CCA substrate was equilibrated for 3 ns. The parameters, available on the web[1], for the ester bond linkage in *O*-phosphoserine attached to the tRNA were determined by analogy with the ester bond linkage in fatty acids.

## 2 Further results

**Cysteine Metabolism and aminoacylation.** There are two principle pathways for cysteine metabolism, e.g., [22]. In the bacterial and plant cysteine metabolism pathway *O*-acetyl serine is formed from the amino acid serine which is catalyzed by the enzyme serine acetyltransferase (SAT or CysE). *O*-acetylserine is then converted to cysteine by the enzyme *O*-acetylserine sulfhydralase (OASS or CysM). In the mammalian pathway, methionine is first converted to homoserine by a series of reactions. Homoserine is then converted to cystathionine which is catalyzed by the enzyme cystathionine $\beta$-synthase (CBS or Cys4). Cystathionine is finally converted to cysteine by the enzyme cystathionine $\gamma$-lyase (CGL or Cys3). Recently, it was found that the homolog of CysE in some archaeal organisms phosphorylates rather than acetylates serine. The CysM

[1] http://www.scs.uiuc.edu/∼schulten/publications.html

homolog in these organisms then converts *O*-phosphoserine to cysteine, with sulfide as the sulfur source [23]. Table 1 shows the presence of proteins involved in cysteine metabolism and Cys-tRNA$^{\mathrm{Cys}}$ formation in the sequenced or partially sequenced archaeal organisms.

Table 1, which is based in part on Table 1 in ref. [22], was expanded and verified by constructing an EP for each enzyme, and finding additional homologs via a profile BLAST search over the genomes and draft genomes of the organisms listed. Each enzyme belongs to a larger family of proteins, which are involved in aminoacylation or amino acid metabolism. The protein found in the above database search was then used as a query in a reverse BLAST search over the NCBI-NR database to confirm its function. In addition, by isolating monophyletic groups containing proteins of the same function, phylogenetic analysis was performed to confirm the annotation of the protein. Using a simple BLAST search, as in [22], some enzymes (enclosed in brackets in Table 1) were found to be related to a family of proteins involved in cysteine metabolism but could not be confirmed to have the function to which they have been assigned.

**Specific relationship between *M. burtonii* 2 and *M. hungatei* SepCysSs.** The relationship between the *M. burtonii* 2 and *M. hungatei* SepCysSs (see Fig. 1) is supported by sequence signature analysis, which shows that these two proteins contain 28 uniquely conserved residues, as opposed to 12 and 13 signature residues that define the *Methanosarcinaceae* and the *Methanococcales*, respectively, and more dramatically by the fact that only these proteins contain a homologous N-terminal extension of 40 residues. In genomic position, the *M. burtonii* 2 SepCysS is adjacent to two genes, encoding *O*-acetylhomoserine sulfhydrylase and homoserine *O*-acetyltransferase, that are both involved in homocysteine biosynthesis, which is linked to the cysteine metabolic pathway [24]. The closest relatives of these two genes are their homologs in *M. hungatei*, which are in close genomic proximity to SepCysS. As the genomic context similarities show [25], SepCysS was horizontally transfered along with (at least) these two other genes.

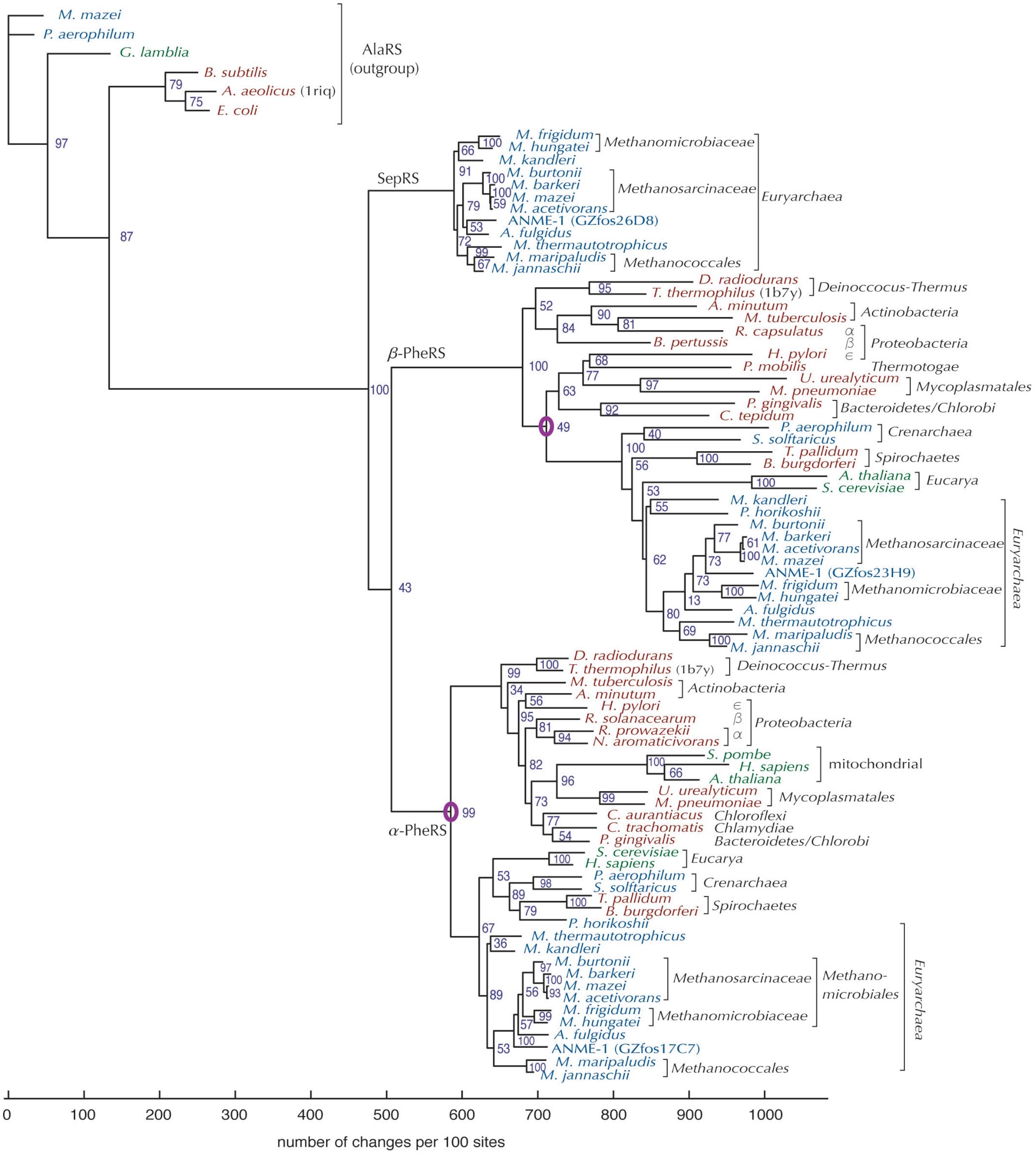


Figure 4: Phylogenetic tree is shown for homologous groups including SepRS, and PheRS. Organism names are color coded as *Bacteria* (red), *Archaea* (blue) and *Eucarya* (green). All local bootstrap probabilities are shown to the right of each branching. AlaRS sequences used to root the tree are shown explicitly. PDB codes for crystallographic structures used in the structural alignment are in parentheses.

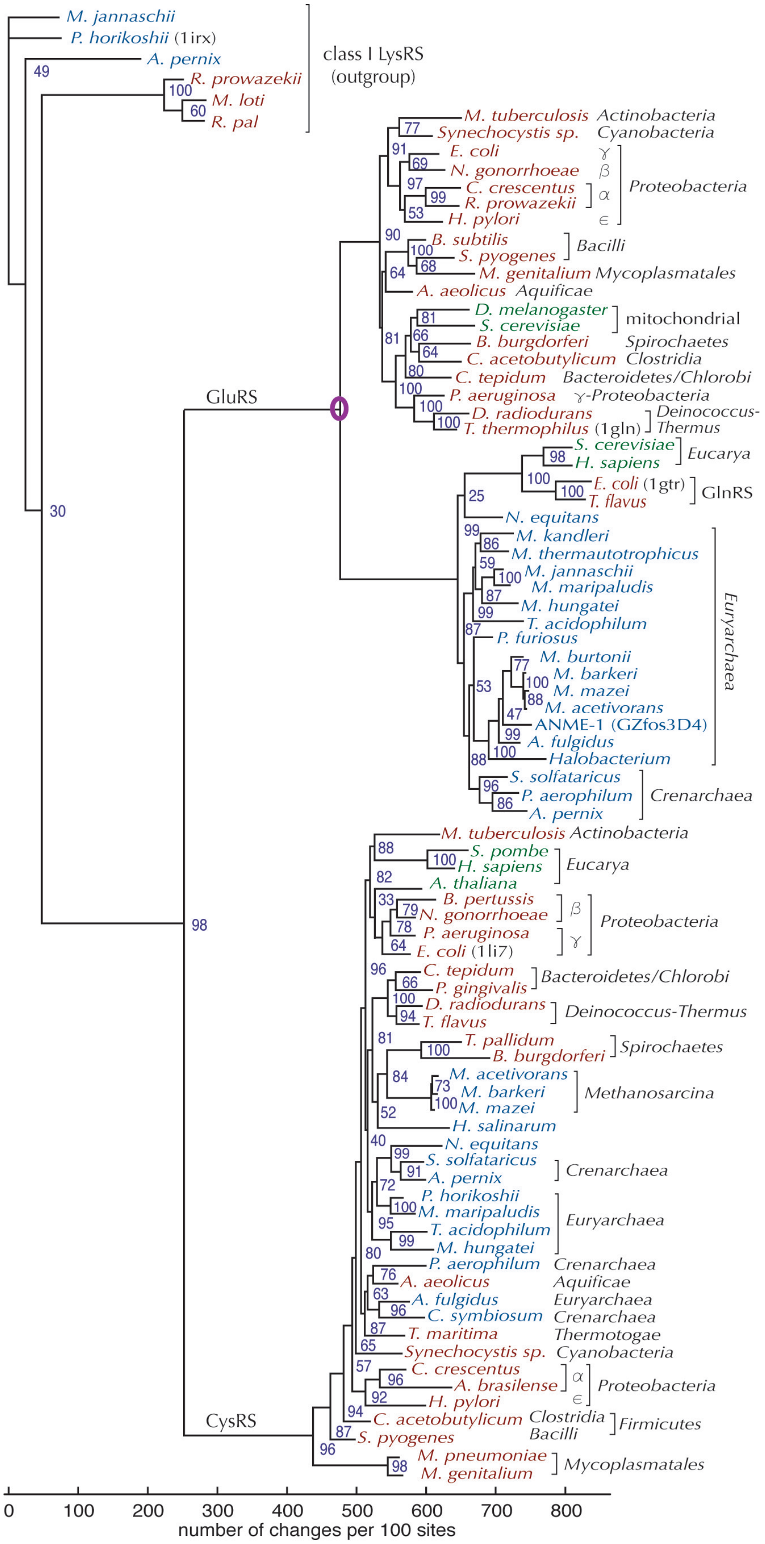


Figure 5: Phylogenetic tree is shown for homologous groups including GluRS and CysRS. LysRS sequences used to root the tree are shown explicitly. Other details as in Fig. 4.

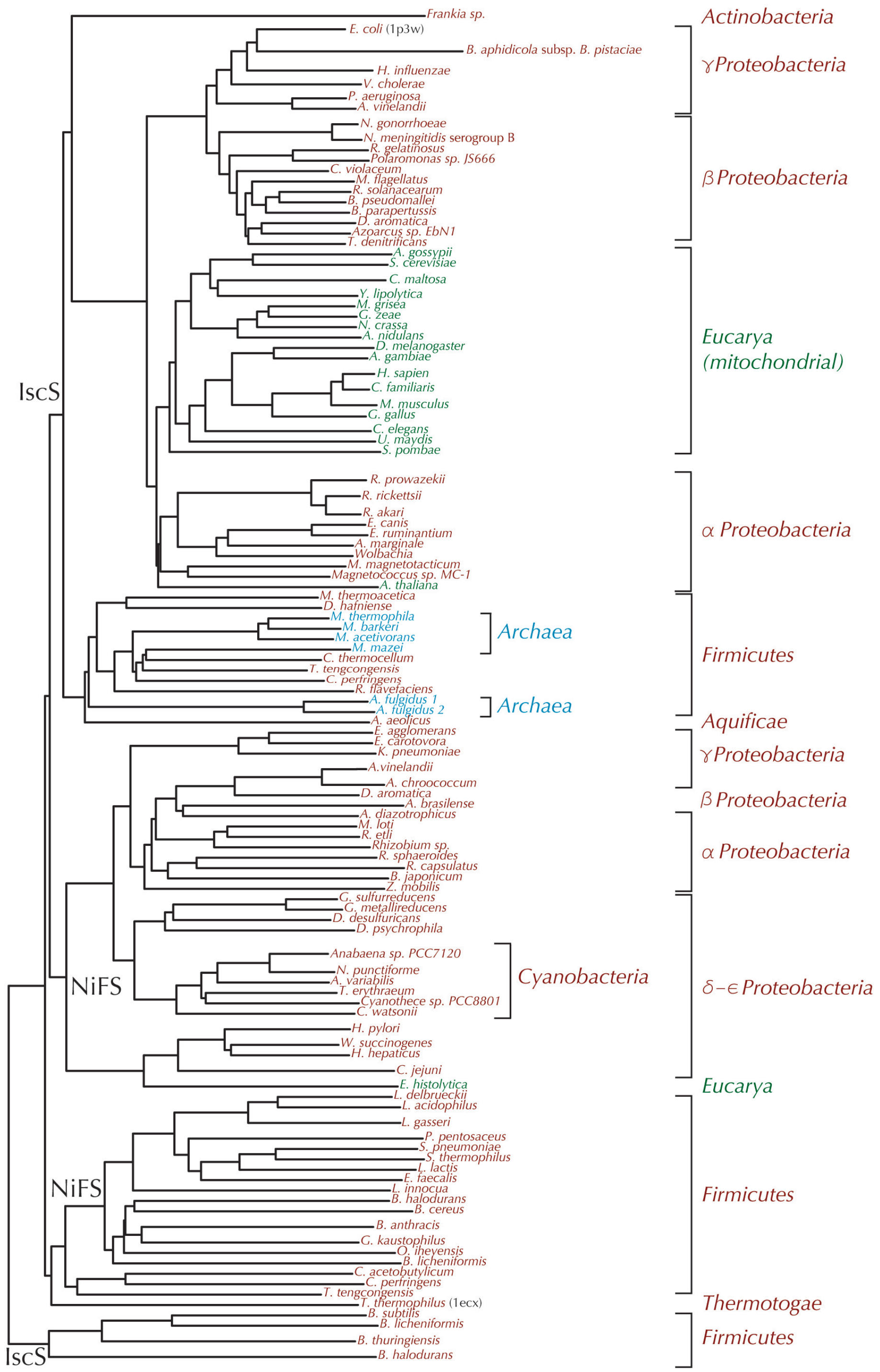


Figure 6: Phylogenetic tree is shown for group 1 cysteine desulfurase. The tree was made using the distance based neighbor joining method in PHYLIP. Other details as in Fig. 4.

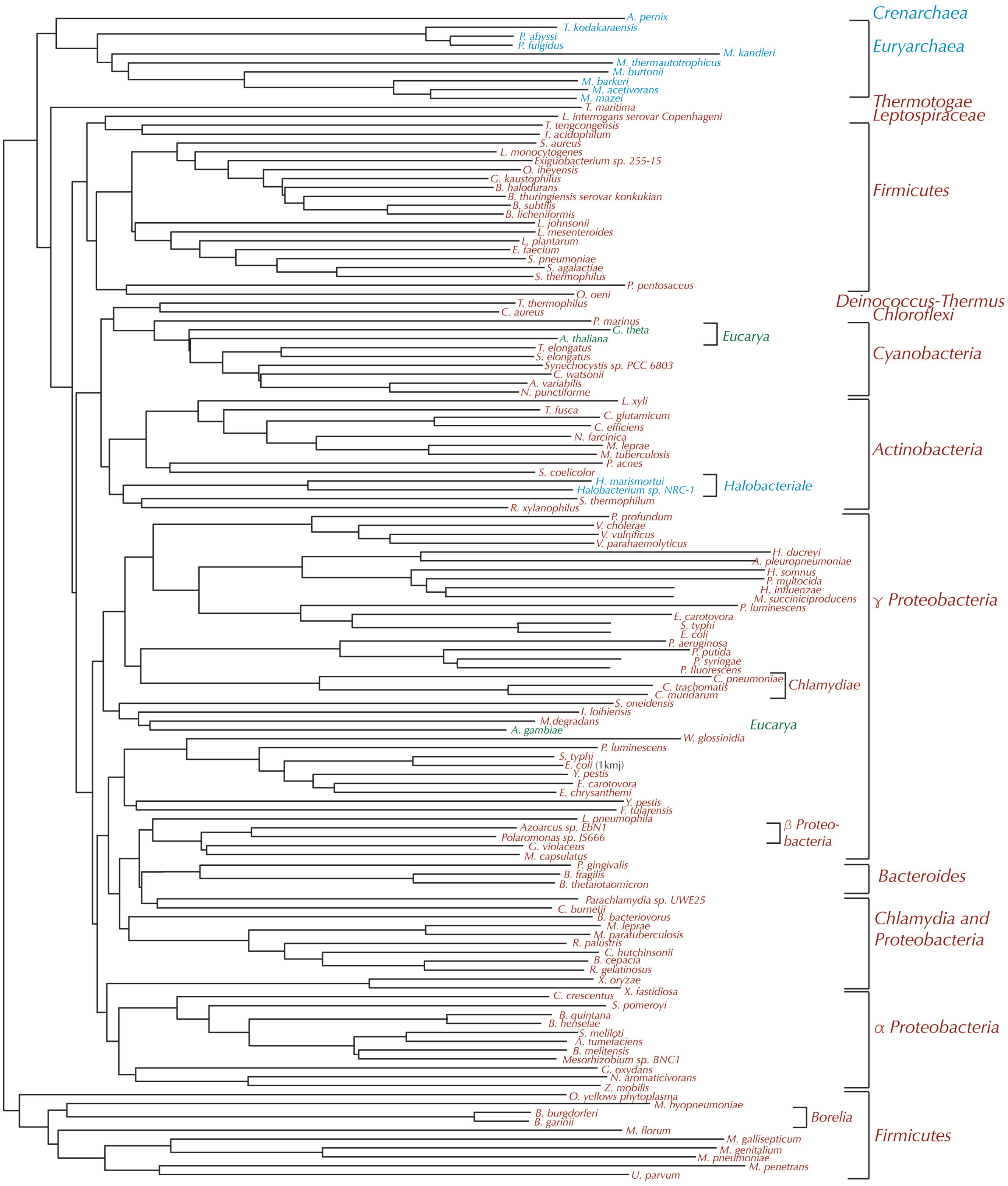


Figure 7: Phylogenetic tree is shown for group 2 cysteine desulfurase. The tree was made using the distance based neighbor joining method in PHYLIP. The *Euryarchaea* show a phylogenetic pattern similar to the canonical 16S rRNA tree. Other details as in Fig. 4.

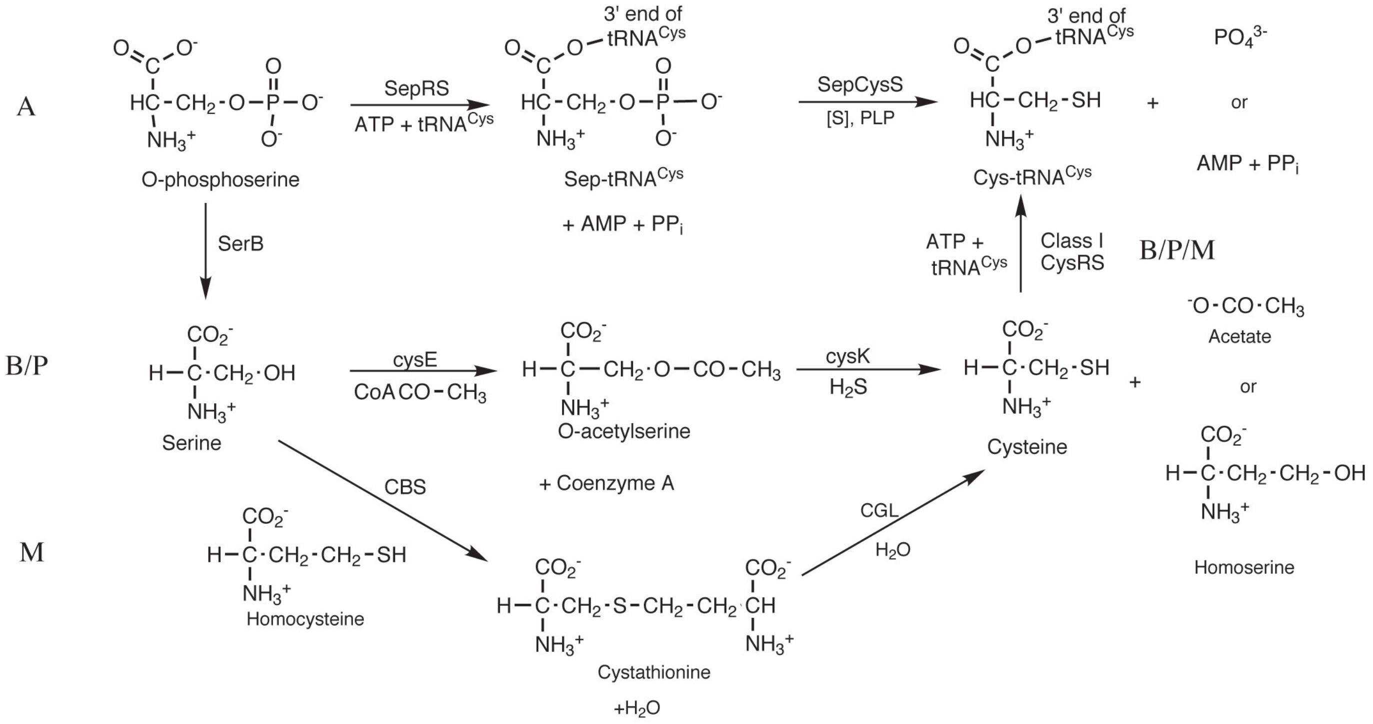


Figure 8: Pathways for biosynthesis and aminoacylation of cysteine are shown. A denotes the archaeal pathway, while B/P denotes the bacterial and plant pathway, and M denotes the mammalian pathway.

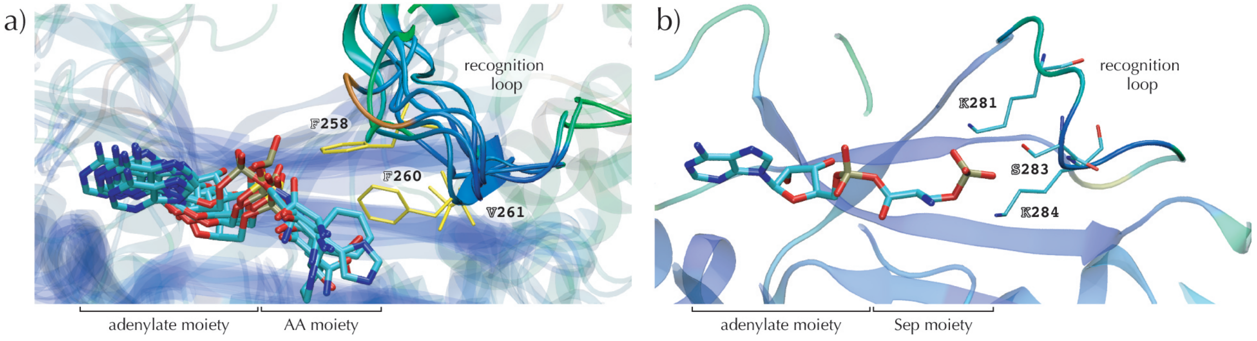


Figure 9: Structural conservation in the class II aaRS active site. a) Structural superposition of class II aaRS catalytic domains (background, transparent), co-crystallized with their cognate aminoacyl-adenylate substrate or substrate analog (sticks). In PheRS, one Val and two Phe residues are in the recognition loop (solid) and make direct contact with the cognate amino acid. b) A view of the equilibrated model of SepRS with the *O*-phosphoseryl-adenylate substrate bound and some of the putative recognition residues marked.

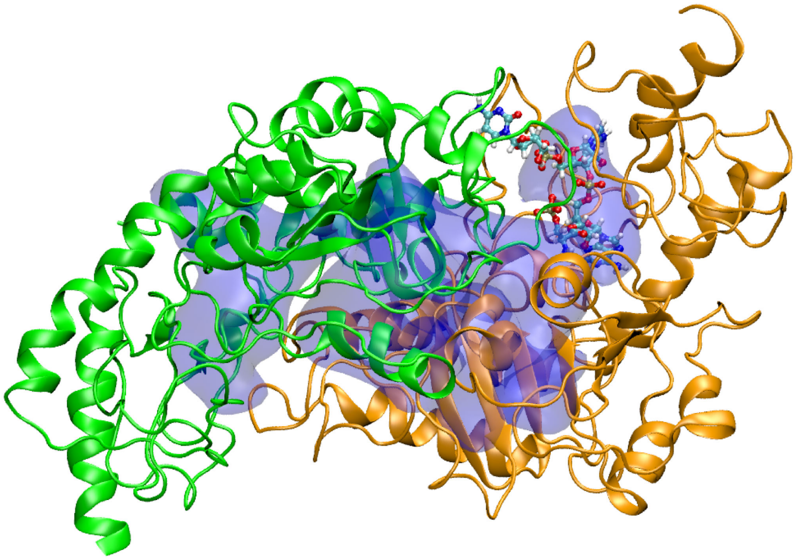
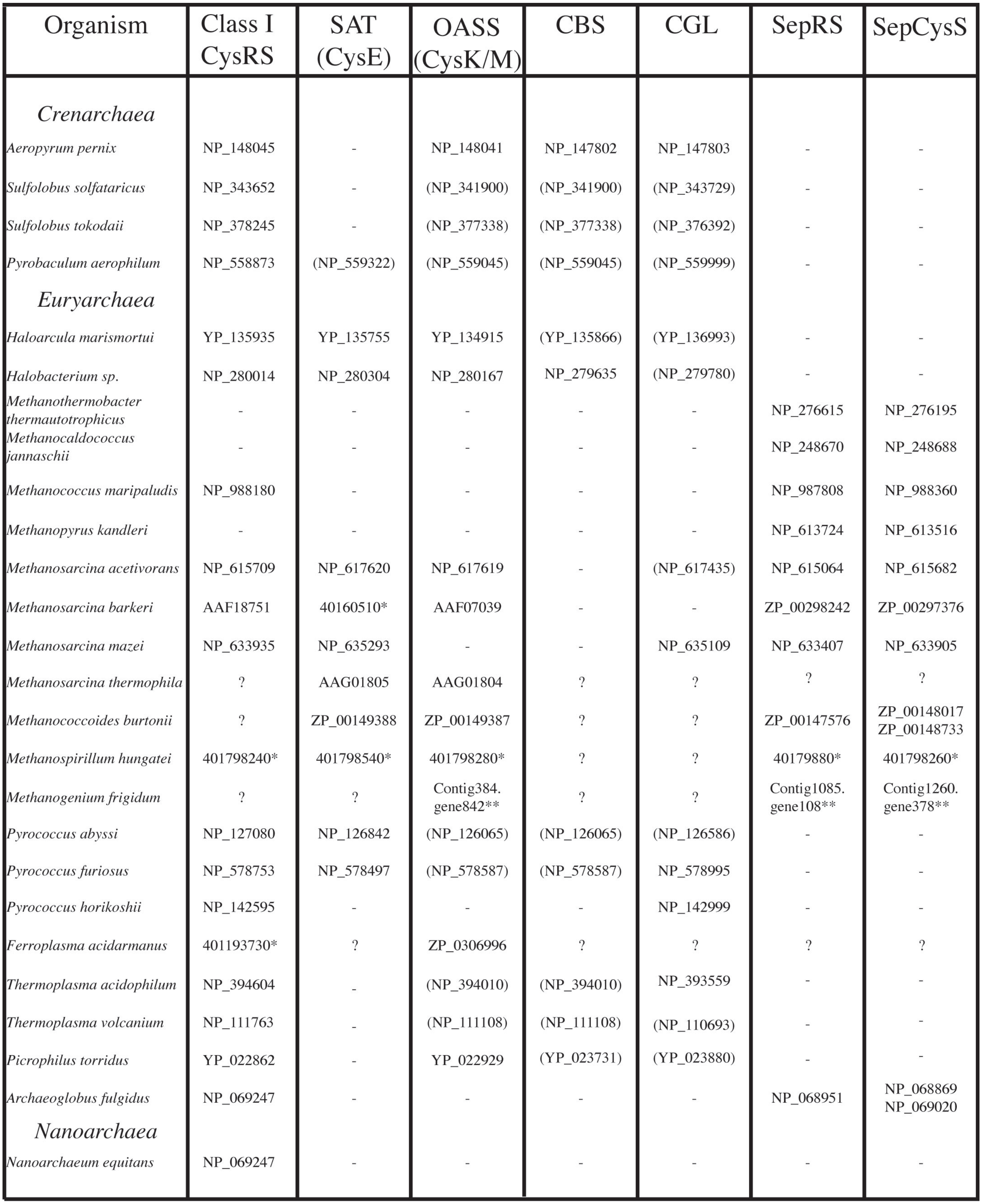


Figure 10: Positive electrostatic surface of protein bound to PLP calculated using PME. The largest positive electrostatic potential was found to be approximately 54 $k_BT/e$ at room temperature. The positive potential surface shown in the figure corresponds to 20 $k_BT/e$ and overlaps with the substrate recognition site of MJ1678, indicating that electrostatics plays a major role in SepCysS-tRNA docking.



| Organism | Class I CysRS | SAT (CysE) | OASS (CysK/M) | CBS | CGL | SepRS | SepCysS |
|---|---|---|---|---|---|---|---|
| *Crenarchaea* | | | | | | | |
| *Aeropyrum pernix* | NP_148045 | - | NP_148041 | NP_147802 | NP_147803 | - | - |
| *Sulfolobus solfataricus* | NP_343652 | - | (NP_341900) | (NP_341900) | (NP_343729) | - | - |
| *Sulfolobus tokodaii* | NP_378245 | - | (NP_377338) | (NP_377338) | (NP_376392) | - | - |
| *Pyrobaculum aerophilum* | NP_558873 | (NP_559322) | (NP_559045) | (NP_559045) | (NP_559999) | - | - |
| *Euryarchaea* | | | | | | | |
| *Haloarcula marismortui* | YP_135935 | YP_135755 | YP_134915 | (YP_135866) | (YP_136993) | - | - |
| *Halobacterium sp.* | NP_280014 | NP_280304 | NP_280167 | NP_279635 | (NP_279780) | - | - |
| *Methanothermobacter thermautotrophicus* | - | - | - | - | - | NP_276615 | NP_276195 |
| *Methanocaldococcus jannaschii* | - | - | - | - | - | NP_248670 | NP_248688 |
| *Methanococcus maripaludis* | NP_988180 | - | - | - | - | NP_987808 | NP_988360 |
| *Methanopyrus kandleri* | - | - | - | - | - | NP_613724 | NP_613516 |
| *Methanosarcina acetivorans* | NP_615709 | NP_617620 | NP_617619 | - | (NP_617435) | NP_615064 | NP_615682 |
| *Methanosarcina barkeri* | AAF18751 | 40160510* | AAF07039 | - | - | ZP_00298242 | ZP_00297376 |
| *Methanosarcina mazei* | NP_633935 | NP_635293 | - | - | NP_635109 | NP_633407 | NP_633905 |
| *Methanosarcina thermophila* | ? | AAG01805 | AAG01804 | ? | ? | ? | ? |
| *Methanococcoides burtonii* | ? | ZP_00149388 | ZP_00149387 | ? | ? | ZP_00147576 | ZP_00148017<br>ZP_00148733 |
| *Methanospirillum hungatei* | 401798240* | 401798540* | 401798280* | ? | ? | 40179880* | 401798260* |
| *Methanogenium frigidum* | ? | ? | Contig384. gene842** | ? | ? | Contig1085. gene108** | Contig1260. gene378** |
| *Pyrococcus abyssi* | NP_127080 | NP_126842 | (NP_126065) | (NP_126065) | (NP_126586) | - | - |
| *Pyrococcus furiosus* | NP_578753 | NP_578497 | (NP_578587) | (NP_578587) | NP_578995 | - | - |
| *Pyrococcus horikoshii* | NP_142595 | - | - | - | NP_142999 | - | - |
| *Ferroplasma acidarmanus* | 401193730* | ? | ZP_0306996 | ? | ? | ? | ? |
| *Thermoplasma acidophilum* | NP_394604 | - | (NP_394010) | (NP_394010) | NP_393559 | - | - |
| *Thermoplasma volcanium* | NP_111763 | - | (NP_111108) | (NP_111108) | (NP_110693) | - | - |
| *Picrophilus torridus* | YP_022862 | - | YP_022929 | (YP_023731) | (YP_023880) | - | - |
| *Archaeoglobus fulgidus* | NP_069247 | - | - | - | - | NP_068951 | NP_068869<br>NP_069020 |
| *Nanoarchaea* | | | | | | | |
| *Nanoarchaeum equitans* | NP_069247 | - | - | - | - | - | - |

Table 1: The genes which encode protein involved in the cysteine metabolism and direct cysteinylation of tRNA$^{\mathrm{Cys}}$. Asterisks indicate gene object identifiers from the Integrated Microbial Genomes database at JGI [25] and double asterisks indicate the contig number and gene number from the draft genome sequence of *M. frigidum* [26]. Dashes indicate the absence of a gene, and question marks signify that the enzyme could not be found in the partially sequenced genome. Other codes are the NCBI-NR database gene identifiers.

## References

[1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W, & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.

[2] Sethi, A., O'Donoghue, P, & Luthey-Schulten, Z. (2005) *Proc. Natl. Acad. Sci. USA* **102**, 4045–4050.

[3] Thompson, J. D., Higgins, D. G, & Gibson, T. J. (1994) *Nucl. Acids Res.* **22**, 4673–4680.

[4] Woese, C. R., Olsen, G., Ibba, M, & Söll, D. (2000) *Microbiol. Mol. Bio. Rev.* **64**, 202–236.

[5] Felsenstein, J. (1989) *Cladistics* **5**, 164–166.

[6] Brochier, C., Forterre, P, & Gribaldo, S. (2004) *Genome Biol.* **5**, R17.

[7] Swofford, D. (2003) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.* (Sinauer Associates, Sunderland, MA).

[8] Henikoff, S & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919.

[9] Marsh, T. L., Reich, C. I., Whitelock, R. B, & Olsen, G. J. (1994) *Proc. Natl. Acad. Sci. USA* **81**, 4180–4184.

[10] Guindon, S & Gascuel, O. (2003) *Syst. Biol.* **52**, 696–704.

[11] Jones, D. T., Taylor, W. R, & Thornton, J. M. (1992) *CABIOS* **8**, 275–282.

[12] Adachi, J & Hasegawa, M. (1996) *Comp. Sci. Monogr.* **28**, 1–150.

[13] Page, R. D. M. (1996) *Comp. Appl. Biosci.* **12**, 357–358.

[14] Jones, D. T. (1999) *J. Mol. Biol.* **292**, 195–202.

[15] Marti-Renom, M. A., Stuart, A., Fiser, A., Sanchez, F., Melo, F, & Sali, A. (2000) *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325.

[16] Kale, L., Skeel, R., Bhandarkar, M., Brunner, R., A. Gursoy, N. K., Phillips, J., Shinozaki, A., Varadarajan, K, & Schulten., K. (1999) *J. Comp. Phys.* **151**, 283–312.

[17] MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S, & *et. al.* (1998) *J. Phys. Chem. B* **102**, 3586–3616.

[18] Mustata, G. I., Soares, T. A, & Briggs, J. M. (2003) *Biopolymers* **70**, 186–200.

[19] Kossekova, G., Miteva, M, & Atanasov, B. (1996) *J. Photochem. Photobiol. B* **32**, 71–79.

[20] Ewald, P. (1921) *Ann. Phys.* **64**, 253–287.

[21] Nissen, P., Thirup, S., Kjeldgaard, M, & Nyborg, J. (1999) *Structure Fold. Des.* **7**, 143–156.

[22] Ambrogelly, A., Kamtekar, S., Sauerwald, A., Ruan, B., Tumbula-Hansen, D., Kennedy, D., Ahel, I, & Söll, D. (2004) *Cell. Mol. Life Sci.* **61**, 2437–2445.

[23] Oda, Y., Mino, K., Ishikawa, K, & Ataka, M. (2005) *J. Mol. Biol.* **351**, 334–344.

[24] White, R. H. (2003) *Biochim. Biophys. Acta* **1624**, 46–53.

[25] Markowitz, V., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I, & *et. al.* (2006) The integrated microbial genomes (IMG) system. Special database issue, in press.

[26] Saunders, N. F. W., Thomas, T., Curmi, P. M. G., Mattick, J. S., Kuczek, E., Slade, R., Davis, J., Franzmann, P. D., Boone, D., Rusterholtz, K, & *et. al.* (2003) *Gen. Res.* **13**, 1580–1588.