Calpains in Cyanobacteria and the Origin of Calpains

Dominika Vešelényiová (dominika.veselenyiova@fpvucm.sk)
University of Ss. Cyril and Methodius in Trnava https://orcid.org/0000-0002-1965-160X
Lenka Raabová
University of Ss. Cyril and Methodius in Trnava
Mária Schneiderová
University of Ss. Cyril and Methodius in Trnava
Matej Vesteg
Matej Bel University
Juraj Krajčovič
University of Ss. Cyril and Methodius in Trnava

Research Article

Keywords: alphaproteobacteria, Archaea, bacteria, Chamaesiphon, CysPC, cysteine protease, Fischerella

Posted Date: August 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-751618/v1

Abstract

Calpains are cysteine proteases involved in many cellular processes. They are an ancient and large superfamily of enzymes responsible for the cleavage and irreversible modification of a large variety of substrates. They have been intensively studied in humans and other mammals, but information about calpains in bacteria is scarce. Calpains have not been found among Archaea to date. In this study, we have investigated the presence of calpains in selected cyanobacterial species using *in silico* analyses. We show that calpains defined by possessing CysPC core domain are present in cyanobacterial genera *Anabaena*, *Aphanizomenon*, *Calothrix*, *Chamaesiphon*, *Fischerella*, *Microcystis*, *Scytonema* and *Trichormus*. Based on *in silico* protein interaction analysis, we have predicted putative interaction partners for identified cyanobacterial calpains. The phylogenetic analysis including cyanobacterial, other bacterial and eukaryotic calpains divided bacterial and eukaryotic calpains into two separate monophyletic clusters. We propose two possible evolutionary scenarios to explain this tree topology: 1) the eukaryotic ancestor or an archaeal ancestor of eukaryotes obtained calpain gene from an unknown bacterial donor, or alternatively 2) calpain gene had been already present in the last common universal ancestor and subsequently lost by the ancestor of Archaea, but retained by the ancestor of Bacteria and by the ancestor of Eukarya. Both scenarios would require multiple independent losses of calpain genes in various bacteria and eukaryotes.

Introduction

Calpains (EC 3.4.22.17) are an ancient superfamily of cysteine proteases activated by Ca^{2+} ions. They are cytosolic, non-lysosomal proteins with cysteine-rich domains, and they are evolutionary well-conserved. The first calpain was discovered by Guroff (1964) when he purified a calcium-activated enzyme present in a soluble fraction of the rat brain. Other calpains have been later identified in various organisms including mammals, invertebrates, plants and fungi as well as in some bacteria (Goll et al. 2003), but they have not been found in Archaea (Rawlings 2015).

Calpains are divided into two groups based on their structure: classical and non-classical ones (Fig. 1). Classical calpains are composed of large and small subunits that are assembled to a functional heterodimer after activation by calcium ions. The large subunit of classical calpains is composed of N-terminal domain, catalytic CysPC domain composed of protease core domains 1 and 2 (PC1 and PC2), C2-like domain and a penta-EF hand domain of the large subunit (PEFl). Small subunit contains only two conserved domains: penta-EF hand domain of the small subunit (PEFs) and a glycine-rich region. Calpains found in bacteria belong to the group of non-classical calpains, they are monomers lacking the small subunit, and they have in common with other calpains only the catalytic calpain domain CysPC (Rawlings 2015).

Although the number of calpain studies has dramatically increased in recent years, they have focused mainly on mammalian calpains, because of their biomedical and clinical importance (for a review see Ono et al. 2016). Calpains are involved in the development of various diseases including limb-girdle

muscular dystrophy (Richard et al. 1995; Wang et al. 2015), type II diabetes (Buraczynska et al. 2013), neurodegenerative disorders (Nixon 2003; Vosler et al. 2008) and cancer (Storr et al. 2011). The information about calpains in unicellular eukaryotes as well as bacteria remains scarce.

Cyanobacteria are one of the most ancient and major groups of Gram-negative bacteria. They obtain energy by photosynthesis. Cyanobacteria are the only diazotrophs producing oxygen as a by-product of the photosynthetic process (Berman-Frank et al. 2003). Their invention of oxygenic photosynthesis ultimately led to the Great Oxidation Event (GOE) cca 2.3 billion years ago (Luo et al. 2016). The chloroplasts of eukaryotic supergroup Archaeplastida comprising glaucophytes, and red and green algae including land plants (Adl et al. 2019) have originated from cyanobacteria in the process termed primary endosymbiosis (Raven and Allen 2003).

Cyanobacteria are an enormously diverse group with high adaptive capacity and many species have the ability to tolerate extreme conditions (Gaysina et al. 2018). Therefore, they have colonized almost all habitats on the Earth with the access to sunlight and they play a significant role in biochemical processes in nature (Kulasooriya 2012). In the past, the taxonomy of cyanobacteria was based mainly on morphological and only rarely also on ecological criteria. In modern taxonomy, more focus is on polyphasic approach, which combines morphological, biochemical, phylogenetic, genomic and ecological data (Komárek 2016). Cyanobacteria currently comprise at least 4,780 described species (Guiry and Guiry 2021).

Due to the advances in genomics, transcriptomics and proteomics, the genetic makeup of cyanobacteria has been studied more intensively and their significance in biotechnological applications has increased. Cyanobacteria can be a source of bioactive compounds including pharmaceuticals and toxins (Maurya et al. 2018).

Several calcium binding proteins (CaBPs) have been discovered in cyanobacteria. These proteins play a significant role in bacterial cells, mainly in processes such as cell division and development, motility, homeostasis, stress response, secretion, molecular transport, cellular signalling, and host-pathogen interactions (Domínguez et al. 2015). Nevertheless, the information about cysteine proteases from the calpain superfamily in cyanobacteria remains limited. In this study, we have conducted bioinformatic search for calpain homologs in proteomes of various selected cyanobacterial species, mainly colonizing extreme environments and species with biotechnological significance. The putative interacting partners of cyanobacterial calpains have also been identified *in silico*. We have also performed the phylogenetic analysis of calpain core CysPC domain to infer the phylogenetic position of the identified cyanobacterial calpains.

Material And Methods

We have searched for calpains *in silico* in proteomes of 50 selected cyanobacterial species (Supplementary Table 1). Our selection was focused on cyanobacteria from extreme biotopes as well as those with biotechnological potency. The proteomic data are available online in NCBI GenBank

(https://www.ncbi.nlm.nih.gov/genbank/) and Uniprot Proteomes (https://www.uniprot.org/help/proteomes_manual) databases. Since calpain superfamily is relatively divergent and it has many members, to identify potential calpain sequences, we created Hidden Markov Model (HMM) of calpain catalytic core domain (CysPC) – the most common and well conserved calpain motif. For HMM creation, annotated calpain sequences from various organisms were obtained from the UniProt database (https://www.uniprot.org/). HMM was built using HMMER 3.2.1 (Potter et al. 2018). This model was then applied to cyanobacterial proteomes and putative calpain sequences were identified.

Putative calpains found by HMM were further analysed. Conserved domains were visualized using Conserved Domain Database (CDD) (https://www.ncbi.nlm.nih.gov/cdd) and Pfam (https://pfam.xfam.org/), and catalytic sites were identified. The sequences, which did not contain full-length CysPC domain, were excluded from analyses. The identified cyanobacterial sequences were aligned using MAFFT (https://www.ebi.ac.uk/Tools/msa/mafft/) and the sequence logo was generated using WebLogo (https://weblogo.berkeley.edu/logo.cgi) TMHMM v. 2.0 server (http://www.cbs.dtu.dk/services/TMHMM/) was used to identify putative transmembrane regions.

SmartBLAST (https://blast.ncbi.nlm.nih.gov/smartblast/) was used to search for homologs of cyanobacterial calpains in other bacteria as well as in eukaryotes. 3D structure of putative calpains was predicted by Phyre2 (Kelley et al. 2015) to verify that the identified sequences are really calpains. String DB was used for the prediction of putative interactions with other proteins (Szklarczyk et al. 2015) to elucidate possible function of cyanobacterial calpains. String DB classifies protein interaction partners into three categories: (I) Known interactions, which have been experimentally determined and/or the information is based on curated databases; (II) Predicted interactions, where the prediction can be based on gene neighbourhood, gene fusions or gene co-occurrence, and (III) Other, with the interaction based on text mining, co-expression or protein homology (Szklarczyk et al. 2015).

We gathered annotated calpain sequences from various organisms from the UniProt database and we included them in phylogenetic analysis together with the identified cyanobacterial calpains (Supplementary Table 2). Only the regions corresponding to catalytic CysPC domain were used for the phylogenetic analysis. All CysPC sequences were aligned in MAFFT (https://www.ebi.ac.uk/Tools/msa/mafft/) with automatic settings for amino acid sequences. IQ-Tree (Minh et al., 2020) was used for the construction of phylogenetic trees. Out of 168 models, WAG + F + I + G4 (Whelan and Goldman 2001) was selected as the best suiting model for our dataset. Bootstrap was set to 1 000. Phylogenetic tree was visualized using ITOL (https://itol.embl.de/).

Results

Since information about calpains in bacteria is still limited, we decided to search for these cysteine proteases in 50 selected cyanobacterial species (Supplementary Table 1). Using HMM of calpain catalytic domain (CysPC) and subsequent conserved domain prediction by CDD and Pfam, we identified

13 putative calpain sequences (Table 1) in 10 of 50 cyanobacterial species (10/50; 20%). Calpains were found in 7 species (*Anabaena minutissima, Aphanizomenon flosaquae, Calothrix parasitica, Fischerella thermalis*, *Fischerella muscicola*, *Scytonema hofmannii* and *Trichormus variabilis*) belonging to order Nostocales, one species (*Microcystis aeruginosa*) belonging to Chroococcales and two species (*Chamaesiphon minutus* and C*hamaesiphon polymorphus*)belonging to Synechococcales. Three putative calpains were identified in *S. hofmanii*, two in *F. thermalis*, while only one calpain was identified in other cyanobacterial species.

The Supplementary table 3 includes the sequences of 13 cyanobacterial calpains identified in this study. Their sequence length ranges from 382 amino acid residues in *F. thermalis* to significantly longer sequences in *C. minutus* and *C. polymorphus* (1145 and 1160 amino acid residues, respectively). The domain structure of all 13 putative calpains was analysed (Table 1). All identified calpains contain conserved CysPC domain at the C-terminus and the most of them contain also bacterial pre-peptidase C-terminal domain (PPC) (or more PPCs) at the N-terminus (Table1, Fig. 2). PPC is, however, usually present at the C-terminus of bacterial secreted peptides (Yeats et al. 2003).

The alignment of CysPC domains from 13 cyanobacterial calpains and the sequence logo generated from this alignment is presented in Fig. 3. All CysPC domains should share a catalytic triad of amino acids typical for calpains – Cys (C), His (H) and Asn (N). C and H residues were correctly aligned for all 13 CysPC domains, while N residue for 12 of them (Fig. 3). For *Scytonema hofmanni 2* (one of three putative calpains identified in this species), D (Asp) residue can be seen in place of N residue in the alignment. Nevertheless, there are present two N residues in the C-terminal part of the *Scytonema hofmanni 2* CysPC domain suggesting that it should be functional but N was incorrectly aligned.

Although most calpains are cytosolic, few of eukaryotic calpains can be also found in organelles such as mitochondria – human calpain 10 (Goll et al. 2003; Ni et al. 2016) or they are anchored in plasma membrane as in the case of plant calpain DEK1 (Galletti et al.2015; Lid et al. 2002). Thus, we also performed prediction of transmembrane regions in cyanobacterial calpains. We did not identify transmembrane regions in any of cyanobacterial calpains suggesting their cytosolic localization.

Smart BLAST was used to confirm that the identified sequences belong to the calpain superfamily. All 13 sequences show similarity with members of this superfamily. However, the level of sequence identity is relatively low (~30%). This might be due to the lack of annotated bacterial calpain sequences in public databases and only a limited number of well-studied calpains from unicellular eukaryotes and bacteria.

Using String DB, a tool for functional annotation and prediction of interaction partners of proteins (Szklarczyk et al. 2015), we predicted the putative interactions of cyanobacterial calpains with other proteins. Since the majority of predicted interaction partners for cyanobacterial calpains have not been yet annotated, we used protein BLAST to search for their homologs. We found ten putative interaction partners for calpains in *A. minutissima*, *C. parasitica*, *S. hofmannii* (*S. hofmanii 1*, one of three identified calpains), *F. thermalis* (*Fischerella thermalis1*, one of two identified calpains),*F. muscicola*as well as in both *Chamaesiphon* spp.neighbourhood, although most interaction partners are different for different

cyanobacterial species. Four calpain interaction partners were found *M. aeuruginosa*, while one of them (Chromosome segregation protein) is known as experimentally determined interaction partner. We were unable to predict interaction partners for two studied cyanobacteria, namely *Aphanizomenon flosaquae* and *Trichormus variabilis*. For one additional identified calpain from *F. thermalis* (*F. thermalis 2*)and two additional calpains from *S. hofmannii* (*S. hofmanii 2* and *3*),no potential interaction partners were identified using StringDB. Potential interaction partners of cyanobacterial calpains are shown in Fig. 4 and their potential functions is discussed in the Discussion section.

To determine evolutionary relationships between cyanobacterial calpains and calpains present in other prokaryotes and eukaryotes, we performed phylogenetic analysis of the CysPC domain. In contrast to other parts of calpain sequences, CysPC domain is highly conserved in all calpains. It consists of approximately 350 amino acid residues. The result of phylogenetic analysis are shown in Fig.5. Bacterial and eukaryotic CysPC domains are clearly separated into two monophyletic clusters. All cyanobacterial calpains, except for *S. hofmanii 2*, form a monophyletic cluster within bacteria.

Discussion

Proteases are a large and divergent group of enzymes with various biological functions including developmental processes, acquisition of nutrients and stress responses (Pérez-Lloréns et al. 2003). Calpains are well conserved cysteine proteases that have been found in a wide range of eukaryotic organisms and some bacteria, but they have not been found in Archaea. Most of information about calpains comes from the study of humans, animals and plants, and only little is known about their distribution, structure and function in bacteria. In this study, we have investigated 50 species of cyanobacteria to determine, whether calpains are present in this taxonomic group.

We have identified calpains in 10 cyanobacterial species based on Hidden Markov Model of the catalytic CysPC domain typical for calpain proteins. The number of identified cyanobacterial species possessing calpains is relatively low, but as it has been shown previously, cyanobacteria are a highly diverse group and their genome content varies significantly even at the species and strain levels (Mohanta et al. 2017). CysPC domain is in cyanobacteria often associated with PPC domain (Table 1), which is typically present in bacterial secreted proteins and it is found at their C-terminus (Yeats et al., 2003), while in cyanobacterial calpains, it is found at the N-terminus. The transmembrane helical regions are absent from all putative cyanobacterial calpains suggesting their cytosolic localization. These findings are consistent with the study of calpains in other bacteria that also do not possess any predictable transmembrane regions (Rawlings 2015).

Calpains are known to be involved in many cellular processes in eukaryotes such as aleurone bilayer development and positional cell division in plants (Olsen et al. 2015), and brain function, memory formation and the development of many pathological processes in mammals (Ono et al. 2016). Calpains cleave a wide range of substrates, among which are e.g. protein kinases, receptor molecules and proteins involved in signal transduction. It has been proposed that calpains play main role in regulation of cell

signalling rather than in protein digestion (Wang et al. 1989; Moriyasu and Wayne 2004). However, their function in bacteria remains unknown.

Our *in silico* interaction analysis revealed that cyanobacterial calpains from *C. parasitica*, *F. muscicola*, *F. thermalis* and *S. hofmannii* (*S. hofmannii 1*) interact with methionine synthase (Fig. 4). Methionine synthase catalyses the transfer of a methyl group from methyl-cobalamin to homocysteine (Deobald et al. 2020). S8 peptidase is putatively interacting with calpains from *A. minutissima*, *C. minutus* and *C. polymorphus*. S8 endopeptidases are thermostable secreted proteins involved in nutrition that are present in almost all taxa except for fungi (Li et al. 2016).

SecA, TamB and collagen triple helix repeat protein were identified as putative calpain interacting partners in *C. minutus* and C. *polymorphus*. SecA protein has a role in coupling the hydrolysis of ATP to the transfer of proteins into and across the bacterial plasma membrane (Fröderberg et al. 2004). Cyanobacterial SecA is mainly found as soluble homodimer in the cytosol, but a small fraction of this protein is associated with cytoplasmic as well as thylakoid membrane suggesting that SecA is involved in protein translocation across these membranes in cyanobacteria (Nakai et al. 1994). TamB is a component of the translocation and assembly module autotransportercomplex. It functions in translocation of autotransporters across the outer membrane of Gram-negative bacteria. Collagen triple helix repeat protein predominantly consists of the Glycine (G) -X- Tyrosine (Y) repeats and the polypeptide chains form a triple helix. Collagens are generally extracellular structural proteins involved in formation of connective tissues (Mayne and Brewton 1993), mostly known in eukaryotes, but it has been shown that some bacterial species can possess collagen-like proteins as well. Among cyanobacteria, collagen-like protein was identified for the first time in the filamentous cyanobacterium *Trichodesmium erythraeum* (Layton et al. 2008; Price and Anandan 2013). This protein can bind antibody against human collagen. It is localized between adjacent cells of filaments and it was shown to play an important role in the structural development of filaments (Layton et al. 2008; Price and Anandan, 2013). *Chamaesiphon* spp. are currently not known to form filaments, however, according to phylogenetic analysis based on 16S rDNA, the genus is closely related to Gomontiellaceae, which phylogenetically clusters with other groups of filamentous cyanobacteria that belong to Oscillatoriales (Kurmayer et al., 2018).

Three annotated putative calpain interaction partners were identified in *A. minutissima* - BPSL0067 family protein, cadherin repeat-containing protein and S8 peptidase. Cadherins are adhesion molecules responsible for cell-cell interactions and the maintenance of appropriate intercellular spacing (Guan et al. 2014). They also play key role in stress responses (Ivanov et al. 2001; Tripathi and Sowdhamini 2008).

In *C. parasitica*, six annotated putative interaction partners were predicted - glycotransferase, glycoside hydrolase family 3 protein, methionine synthase, tandem 95 repeat protein and transposase ISAzo13. Proteins that belong to glycoside hydrolase family 3 are widely distributed in bacteria, fungi, and plants. They often cleave a broad range of substrates and function in a variety of cellular processes, including cellulosic biomass degradation, remodelling of cell wall in both bacteria and plants, energy metabolism and pathogen defence (Harvey et al. 2000; Lee et al. 2003).

Although *F. thermalis* and *F. muscicola* are closely related filamentous cyanobacteria, the proteins predicted to interact with their calpains are different, except for methionine synthase. In *F. thermalis*, calpain could also interact with beta-N-acetylhexosaminidase, glycosyltransferase and radical SAM ((Sterile α Motif) protein (Fig. 4). Beta-N-acetylhexosaminidase might also interact with calpain found in *S. hofmannii*. In bacteria, it is known to play a role in peptidoglycan recycling pathway important for cell wall development (Litzinger et al. 2010). N-acetylhexosaminidase is classified in glycoside hydrolase family 3 (McDonald et al. 2015) which potentially interacts also with calpain from *C. parasitica*.

WD40-like protein was identified as aa putative calpain interaction partner in *F. muscicola* (Fig. 4). WD40-repeat proteins belong to a large protein family and they are implicated in a variety of functions ranging from signal transduction and transcription regulation to cell cycle control, autophagy and apoptosis in eukaryotes. All WD40-repeat proteins play a role in coordinating multi-protein complex assemblies. The repeating units serve as a rigid scaffold for protein interactions (Mishra et al. 2014; Stirnimann et al. 2010). In bacteria, over 4000 WD40 proteins have been identified which is more than 6% of all known WD40 proteins. Although only a small number of bacterial genomes encode WD40 proteins, they are most abundant in cyanobacteria and Planctomycetes (Hu et al. 2017). WD40 proteins function in bacteria as signal transductors and they are also involved in nutrient synthesis (Hu et al. 2017).

The putative interaction partners of calpain 1 identified in *S. hofmanii*, play a role in stress responses including expression of specific genes, activation of specific proteins, transport of proteins between membranes and altering DNA replication progress (Matelska et al. 2017; Imamura and Asaima 2009; Paget and Helmann 2003).

Calpain from *F. muscicola* was also predicted to interact with SAM-dependent methyltransferase which are responsible for nitrogen sequestration (Sofia 2001). Another interaction partner is Can B-type protein – a calcium-dependent, calmodulin-stimulated protein phosphatase containing a regulatory subunit of calcineurin with calcium sensitivity (Li et al. 2016). Our result show that predicted interaction partners of identified cyanobacterial calpains differ significantly among studied cyanobacterial species.

We also conducted phylogenetic analysis of calpain core CysPC domain to infer the phylogenetic position of cyanobacterial calpains. Phylogenetic analysis revealed the monophyly of bacterial as well as of eukaryotic CysPCs with bootstrap support 98 and 97, respectively (Fig. 5). No horizontal gene transfers of CysPC domain from bacteria to eukaryotes or *vice versa* were detected using our taxon sampling. The tree is, however, indicative of the real evolutionary relationships within neither bacteria nor eukaryotes. CysPC is thus unlikely to be a suitable marker for inferring the evolutionary relationships between organisms and it is also possible that several horizontal transfers of calpains have occurred within bacteria as well as within eukaryotes.

With the exception of *S. hofmannii 2*, all cyanobacterial CysPC domains are a monophyletic group within bacterial CysPC domains (Fig. 5). The alignment of cyanobacterial CysPC domains also confirms that CysPC domain 2 from *S. hofmannii* is the most divergent in comparison to other cyanobacterial CysPC

domains (Fig. 3). The tree topology also disproves the hypothesis that cyanobacteria, from which chloroplasts of Archaeplastida evolved, were the endosymbiotic donors of archaeplastidial calpains.

The explanation of the origin of eukaryotic calpains depends on the opinion about the origin of eukaryotes themselves. The most popular hypothesis for the origin of eukaryotes suggests that eukaryotes evolved by the endosymbiosis of an alphaproteobacterial ancestor of mitochondria in an archeal host (Martin and Müller 1998), probably from the group Asgard archaea (Spang et al. 2019; Liu et al. 2021). Since archaea do not possess calpains, while some alphaproteobacteria do, under this scenario, the host archaeal cell could have obtained calpain gene from an alphaproteobacterial endosymbiont. This scenario would be supported if alphaproteobacterial CysPC domains would be placed at the base of eukaryotic CysPCs in the phylogenetic tree with high bootstrap support. Since this is not the case (Fig. 5), our tree does not support alphaproteobacterial origin of eukaryotic calpains. Nevertheless, the hypothesis, that an archaeal ancestor of eukaryotes or the last common ancestor of eukaryotes obtained the calpain gene from an unknown bacterial donor, e.g. *via* an ancient horizontal gene transfer, cannot be rejected.

Alternative, currently less popular but still plausible, hypothesis suggests that Archaea and Eukarya are sister groups *sensu* Carl Woese et al. (1990). Under this scenario, Archaea and Eukarya had a common undefined ancestor. This ancestor might have been even more complex than all contemporary archaea, Archaea domain might have arisen *via* reductive evolution of this archaeo-eukaryotic ancestor and the differences between genome contents of contemporary archaeal lineages could be explained by differential gene losses (Vesteg and Krajčovič, 2011; Vesteg et al. 2012; Forterre 2015). Considering this scenario, the calpain gene could have been already present in the last universal common ancestor, lost in the ancestor of Archaea, while retained in the ancestor of Bacteria and in the ancestor of Eukarya. Since calpain genes are universally distributed in neither bacteria nor eukaryotes, all mentioned scenarios would require multiple independent losses of calpain genes in various bacterial and eukaryotic lineages.

Declarations

Acknowledgements

This work was supported by the Scientific Grant Agency of the Slovak Ministry of Education and the Academy of Sciences (grant VEGA 1/0694/21) and by project ITMS 26210120024 supported by the Research & Development Operational Program funded by the ERDF.

Author Contributions

D. Vešelényiová performed most bioinformatic analyses and prepared the first draft of the manuscript; M. Schneiderová was involved in the identification of cyanobacterial calpains; L. Raabová contributed to the choice of cyanobacteria for analyses and interpretation of results; M. Vesteg contributed to the design

of phylogenetic analysis and its interpretation. J. Krajčovič conceived the study. All authors edited the manuscript and approved its final form.

Fundings: This work was supported by the Scientific Grant Agency of the Slovak Ministry of Education and the Academy of Sciences (grant VEGA 1/0694/21) and by project ITMS 26210120024 supported by the Research & Development Operational Program funded by the ERDF.

Conflicts of interest/Competing interests: Not applicable

References

1. Adl SM, Bass D, Lane CE at al (2019) Revisions to the classification, nomenclature, and diversity of Eukaryotes. J Eukaryot Microbiol 66: 4-119
2. Berman-Frank I, Lundgren P, Falkowski P (2003) Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria.Res Microbiol 154: 157-164
3. Buraczynska M, Wacinski P, Stec A, Kuczmaszewska A (2013) Calpain-10 gene polymorphisms in type 2 diabetes and its micro- and macrovascular complications. J Diabetes Complications 27: 54-58
4. Deobald D, Hanna R, Shahryari S, Layer G, Adrian L. (2020) Identification and characterization of a bacterial core methionine synthase. Sci Rep 10: 1-13
5. Domínguez DC, Guragain M, Patrauchan M (2015) Calcium binding proteins and calcium signaling in prokaryotes. Cell Calcium, 10: 1-13
6. Forterre P (2015) The universal tree of life: an update. Front Microbiol 6: 717
7. Fröderberg L, Houben ENG, Baars L, Luirink J, De Gier JW (2004) Targeting and translocation of two lipoproteins in Escherichia coli via the SRP/Sec/YidC pathway. J Biol Chem 279: 31026-31032
8. Galletti R, Johnson KL, Scofield S, San-Bento R, Watt AM, Murray JAH, Ingram GC (2015) Defective Kernel 1 promotes and maintains plant epidermal differentiation. Development 142: 1978-1983
9. Gaysina LA, Saraf A, Singh P (2018) Cyanobacteria in diverse habitats. In: Mishra AK, Tiwari DN, Rai AN (ed.) Cyanobacteria: From Basic Science to Applications. Academic Press, London, United Kingdom, 1-28
10. Goll DE, Thompson VF, Li H, Wei W, Gong J. (2003) The calpain system. Physiol Rev 83: 731–801
11. Guiry MD, Guiry GM (2020) AlgaeBase. World-wide electronic publication. https://www.algaebase.org/ Accessed 28 January 2021
12. Guroff G (1964) A neural, calciu-activated proteinase from the soluble fraction of rat brain. J Biol Chem 239: 149-155
13. Hu XJ, Li T, Wang Y, Xiong Y, Wu X H, Zhang DL, Ye ZQ, Wu YD (2017) Prokaryotic and highly-repetitive WD40 proteins: A systematic study. Sci Rep 7: 1-13
14. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE (2015) The Phyre2 web portal for protein modeling, prediction and analysis.Nat Prot 10: 845-858

15. Komárek J (2016) A polyphasic approach for the taxonomy of cyanobacteria: principles and applications. Eur J Phycol51: 346-353
16. Kurmayer R, Christiansen G, Holzinger A, Rott E. (2018) Single colony genetic analysis of epilithic stream algae of the genus *Chamaesiphon* spp. Hydrobiologia 811: 61-75
17. Kulasooriya S. (2012) Cyanobacteria: pioneers of planet Earth.Ceylon J Sci 40: 71-88
18. Layton BE, D'Souza AJ, Dampier W, Zeiger A, Sabur A, Jean-Charles J (2008). Collagen's triglycine repeat number and phylogeny suggest an interdomain transfer event from a Devonian or Silurian organism into *Trichodesmium erythraeum.* J Mol Evol 66: 539
19. Le SQ, Gascuel O. (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25: 1307-1320
20. Li HJ, Tang BL, Shao X, Liu BX, Zheng XY, Han XX, Li1 PY, Zhang XY, Song, XY, Chen XL (2016) Characterization of a new S8 serine protease from marine sedimentary *Photobacterium* sp. A5-7 and the function of its protease-associated domain. Front Microbiol 7: 1-17
21. Lid SE, Gruis D, Jung R, Lorentzen JA, Ananiev E, Chamberlin M, Niu X, Meeley R, Nichols S, Olsen OA (2002) The defective kernel 1 (dek1) gene required for aleurone cell development in the endosperm of maize grains encodes a membrane protein of the calpain gene superfamily. Proc Natl Acad Sci USA99: 5460-5465
22. Liu Y, Makarova KS, Huang WC, Wolf YI, Nikolskaya AN, Zhang X, Cai M, Zhang CJ, Xu W, Luo Z, Cheng L, Koonin EV, Li M (2021) Expanded diversity of Asgard archaea and their relationships with eukaryotes. Nature 593: 553-557
23. Luo G, Ono S, Beukes N. J, Wang DT, Xie S, Summons RE (2016) Rapid oxygenation of Earth's atmosphere 2.33 billion years ago. Sci Adv 2: e1600134
24. Martin W, Müller M (1998) The hydrogen hypothesis for the first eukaryote. Nature 392: 37-41
25. Maurya SK, Niveshika, Mishra R (2018) Importance of bioinformatics in genome mining of cyanobacteria for production of bioactive compounds. In: Mishra AK, Tiwari DN, Rai AN (ed.) Cyanobacteria: From Basic Science to Applications. Academic Press, London, United Kingdom, 477-506
26. Mayne R, Brewton RG (1993) New members of the collagen superfamily. Curr Opin Cell Biol 5: 883-890
27. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol, 37: 1530-1534
28. Mishra AK, Muthamilarasan M, Khan Y, Parida SK, Prasad M (2014) Genome-wide investigation and expression analyses of WD40 protein family in the model plant foxtail millet (*Setaria italica* L.). *PLoS ONE* 9: e86852
29. Mohanta TK, Pudake RN, Bae H (2017) Genome-wide identification of major protein families of cyanobacteria and genomic insight into the circadian rhythm. Eur J Phycol 52: 149-165

30. Moriyasu Y, Wayne R (2004) A novel calcium-activated protease in *Chara corallina*. Eur J Phycol 39: 57-66
31. Muñoz-Gómez SA, Hess S, Burger G, Franz Lang B, Susko E, Slamovits CH, Roger AJ (2019) An updated phylogeny of the alphaproteobacteria reveals that the parasitic rickettsiales and holosporales have independent origins. eLife 8: e42535
32. Nakai M, Goto A, Nohara T, Sugita D, Endo T. (1994) Identification of the SecA protein homolog in pea chloroplasts and its possible involvement in thylakoidal protein transport.J Biol Chem 269: 31338-31341.
33. Ni R, Zheng D, Xiong S, Hill DJ, Sun T, Gardiner RB, Fan GC, Lu Y, Abel ED, Greer PA, Peng T (2016) Mitochondrial calpain-1 disrupts ATP synthase and induces superoxide generation in type 1 diabetic hearts: A novel mechanism contributing to diabetic cardiomyopathy. Diabetes 65: 255-268
34. Nixon RA (2003) The calpains in aging and aging-related diseases. Ageing Res Rev 2: 407-418
35. Olsen OA, Perroud PF, Johansen W, Demko V (2015) DEK1; missing piece in puzzle of plant development. Trend Plant Sci 20: 70–71
36. Ono Y, Saido TC, Sorimachi H (2016) Calpain research for drug discovery: Challenges and potential. Nat Rev Drug Discov 15: 854–876
37. Pérez-Lloréns J, Benítez E, Vergara JJ, Berges JA (2003) Characterization of proteolytic enzyme activities in macroalgae. Eur J Phycol 38: 55-64
38. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD (2018) HMMER web server: 2018 update. Nucleic Acid Res 46: W200-W204. https://www.ebi.ac.uk/Tools/hmmer/. Accessed 30 January 2021
39. Price S, Anandan S. (2013) Characterization of a novel collagen-like protein TrpA in the cyanobacterium *Trichodesmium erythraeum* IMS101. J Phycol 49: 758-764
40. Raven JA, Allen, JF (2003) Genomics and chloroplast evolution: What did cyanobacteria do for plants? Genome Biol 4: 209
41. Rawlings ND (2015) Bacterial calpains and the evolution of the calpain (C2) family of peptidases. Biol Direct 10: 1–12
42. Richard I, Broux O, Allamand V et al. (1995) Mutations in the proteolytic enzyme calpain 3 cause limb-girdle muscular dystrophy type 2A. Cell 81**:** 27-40
43. Sofia HJ (2001) Radical SAM, a novel protein superfamily linking unresolved steps in familiar biosynthetic pathways with radical mechanisms: functional characterization using new analysis and information visualization methods. Nucleic Acid Res 29: 1097-1106
44. Spang A, Stairs CW, Dombrowski N, Eme L, Lombard J, Caceres EF, Greening C, Baker BJ, Ettema TJG (2019) Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. Nat Microbiol 4: 1138-1148
45. Stirnimann CU, Petsalaki E, Russell RB, Müller CW (2010) WD40 proteins propel cellular networks. Trend Biochem Sci 35: 565-574

46. Storr SJ, Carragher NO, Frame MC, Parr T, Martin SG (2011) The calpain system and cancer. Nat Rev Cancer 11: 364-374
47. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, Von Mering C (2015) STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acid Res 43: D447-D452
48. Vosler PS, Brennan CS, Chen J (2008) Calpain-mediated signaling mechanisms in neuronal injury and neurodegeneration. Mol Neurobiol 38: 78-100
49. Wang CH, Liang WC, Minami N, Nishino I, Jong YJ (2015) Limb-girdle muscular dystrophy type 2A with mutation in CAPN3: The first report in Taiwan. Pediatr Neonatol 56: 62-65
50. Wang KKW, Villalobo A, Roufogalis BD (1989) Calmodulin-binding proteins as calpain substrates. Biochem J 262: 693-706
51. Vesteg M, Krajčovič J (2011) The falsifiability of the models for the origin of eukaryotes. Curr Genet 57: 367-390
52. Vesteg M, Šándorová Z, Krajčovič J (2012) Selective forces for the origin of spliceosomes. J Mol Evol 74: 226-231
53. Woese CR, Kandler O, Wheelis ML (1990) Toward a natural system of organisms: proposal for the domainsArchaea, Bacteria, and Eukarya. Proc Natl Acad Sci USA 87: 4576-4579
54. Yeats C, Bentley S, & Bateman A (2003) New knowledge from old: *In silico* discovery of novel protein domains in Streptomyces coelicolor. BMC Microbiology 3: 3

Tables

Table 1: Information about calpains found in cyanobacteria. All sequences possess a single catalytic core domain of calpains (CysPC). Most sequences also possess one, two or five bacterial pre-peptidase C-terminal domains (PPC). All identified CysPC domains contain three catalytic sites (CS) typical for calpains.

Organisms	Sequence length	CS	Conserved domain(s)	Pfam ID	Number of domains
Anabaena minutissima	608		CysPC	cl00051	1
Aphanizomenon flosaquae	964		CysPC	cl00051	1
			PPC	pfam04151	5
Calothrix parasitica	998		CysPC	cl00051	1
			PPC	pfam04151	1
Chamaesiphon minutus	1145		CysPC	cl00051	1
			PPC	pfam04151	1
Chamaesiphon polymorphus	1160		CysPC	cl00051	1
			PPC	pfam04151	1
Fischerella muscicola	585		CysPC	cl00051	1
			PPC	pfam04151	1
Fischerella thermalis 1	585		CysPC	cl00051	1
			PPC	pfam04151	1
Fischerella thermalis 2	392		CysPC	cl00051	1
Microcystis aeruginosa	677		CysPC	cl00051	1
			PPC	pfam04151	2
Scytonema hofmannii 1	578		CysPC	cl00051	1
Scytonema_hofmannii 2	509		CysPC	cl00051	1
Scytonema hofmannii 3	442		CysPC	cl00051	1
Trichormus variabilis	827		CysPC	cl00051	1
			PCC	pfam04151	2

Figures

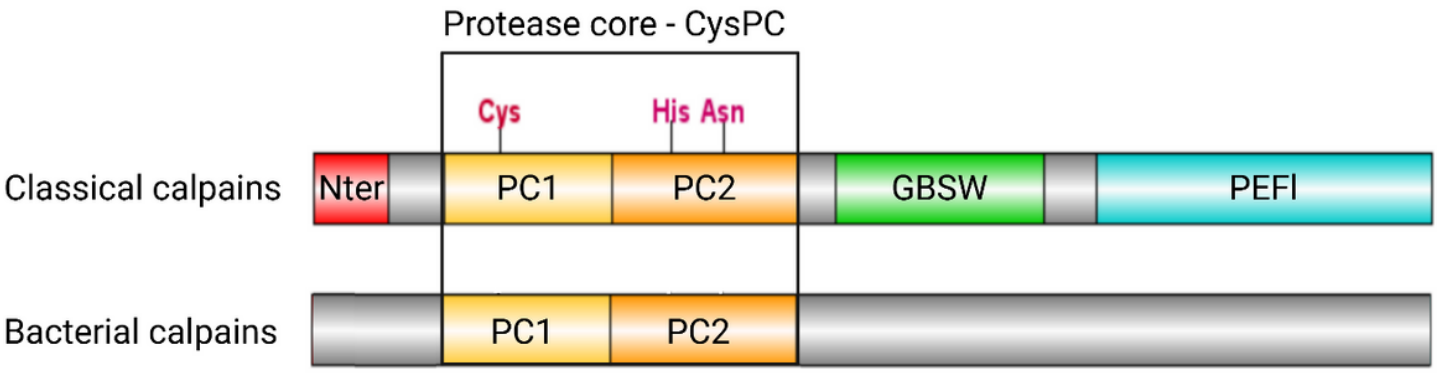

Figure 1

Structural comparison of classical and non-classical calpains. Classical eukaryotic calpains are heterodimers composed of large and small subunits. Each subunit is composed of conserved domains. Large subunit of classical calpains is composed of N-terminal domain, CysPC domain composed of protease core domains 1 and 2 (PC1 and PC2), calpain-type β sandwich (GBSW) domain and a penta-EF

hand domain of the large subunit (PEFl). Small subunit contains two conserved domains: penta-EF hand domain (PEFs) and a glycine-rich region (GR). Classical calpains are absent from bacteria. Non-classical calpains present in some bacteria (but also in some eukaryotes) are monomers, typically with only a single conserved domain – calpain catalytic domain CysPc composed of PC1 and PC2.

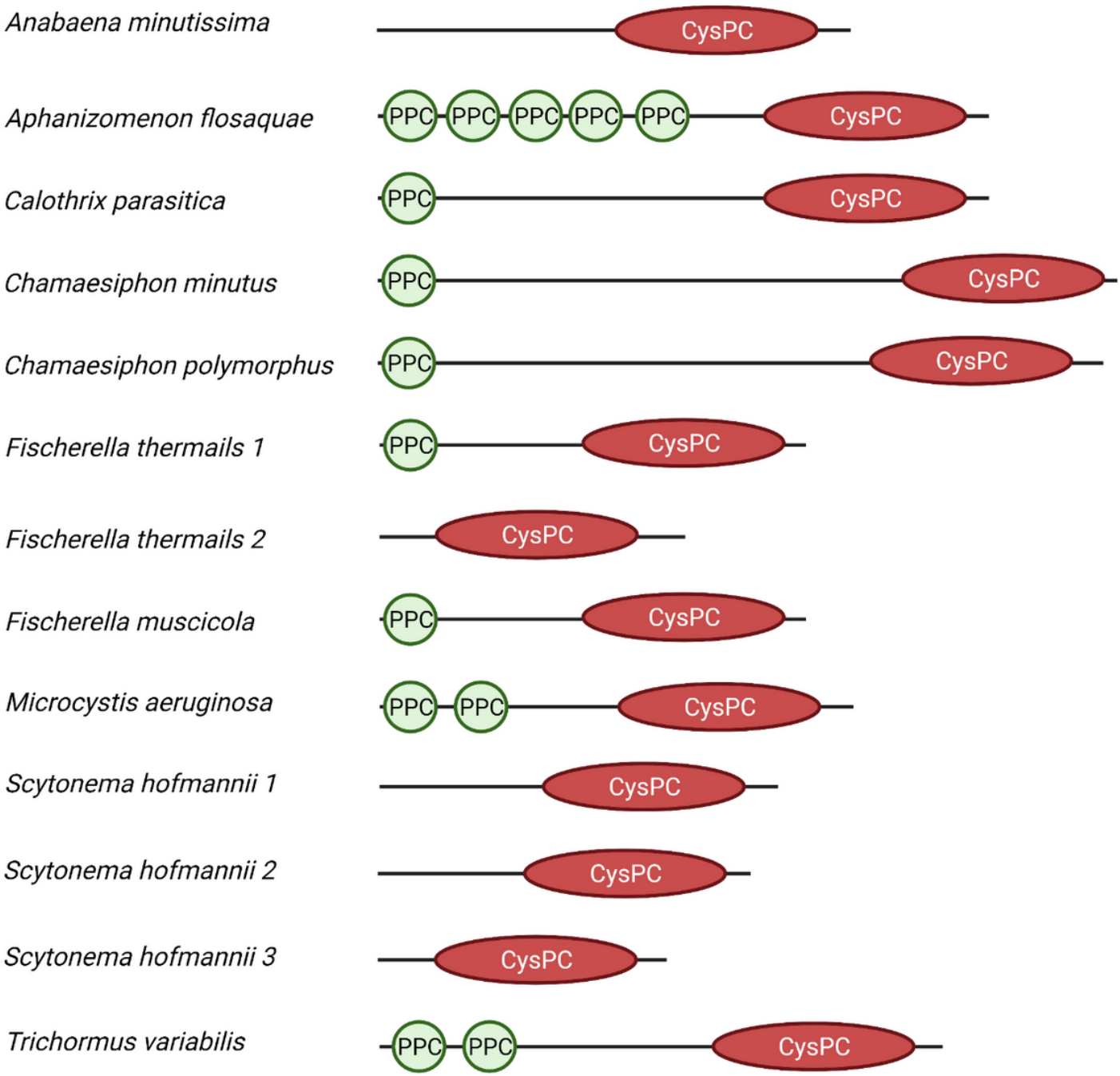

Figure 2

Characteristics of identified cyanobacterial calpains. All calpains contain CysPc conserved domain (ellipse) with a catalytic triad C, H, N typical for calpains. Some calpains possess the second conserved

domain, PPC (circle), which can be present in several copies.

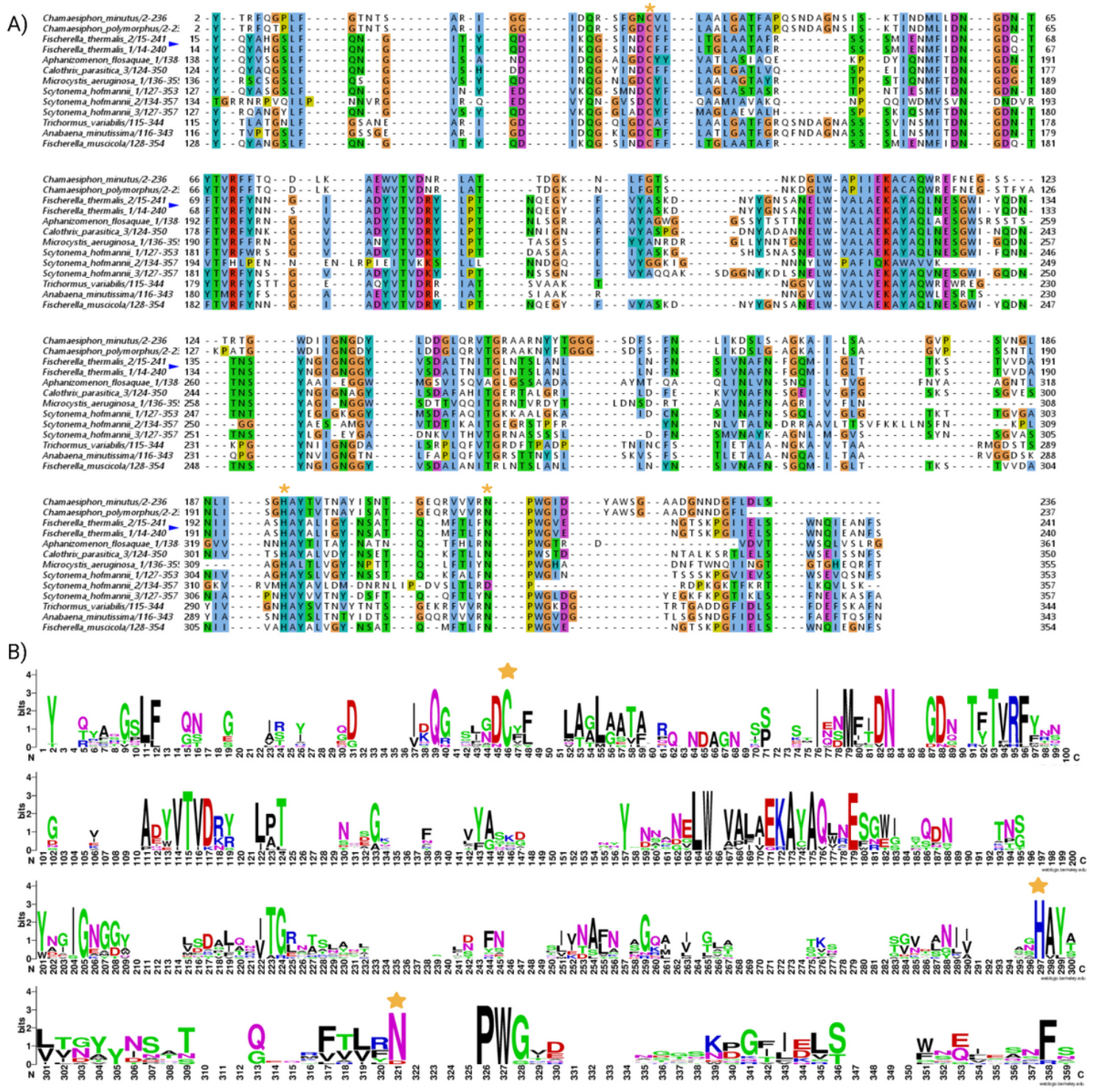

Figure 3

Multiple sequence alignment of the CysPC region of identified cyanobacterial calpains (A) and the generated sequence logo. The presence of conserved C, H and N residues characteristic for CysPC domain is marked by yellow asterisks.

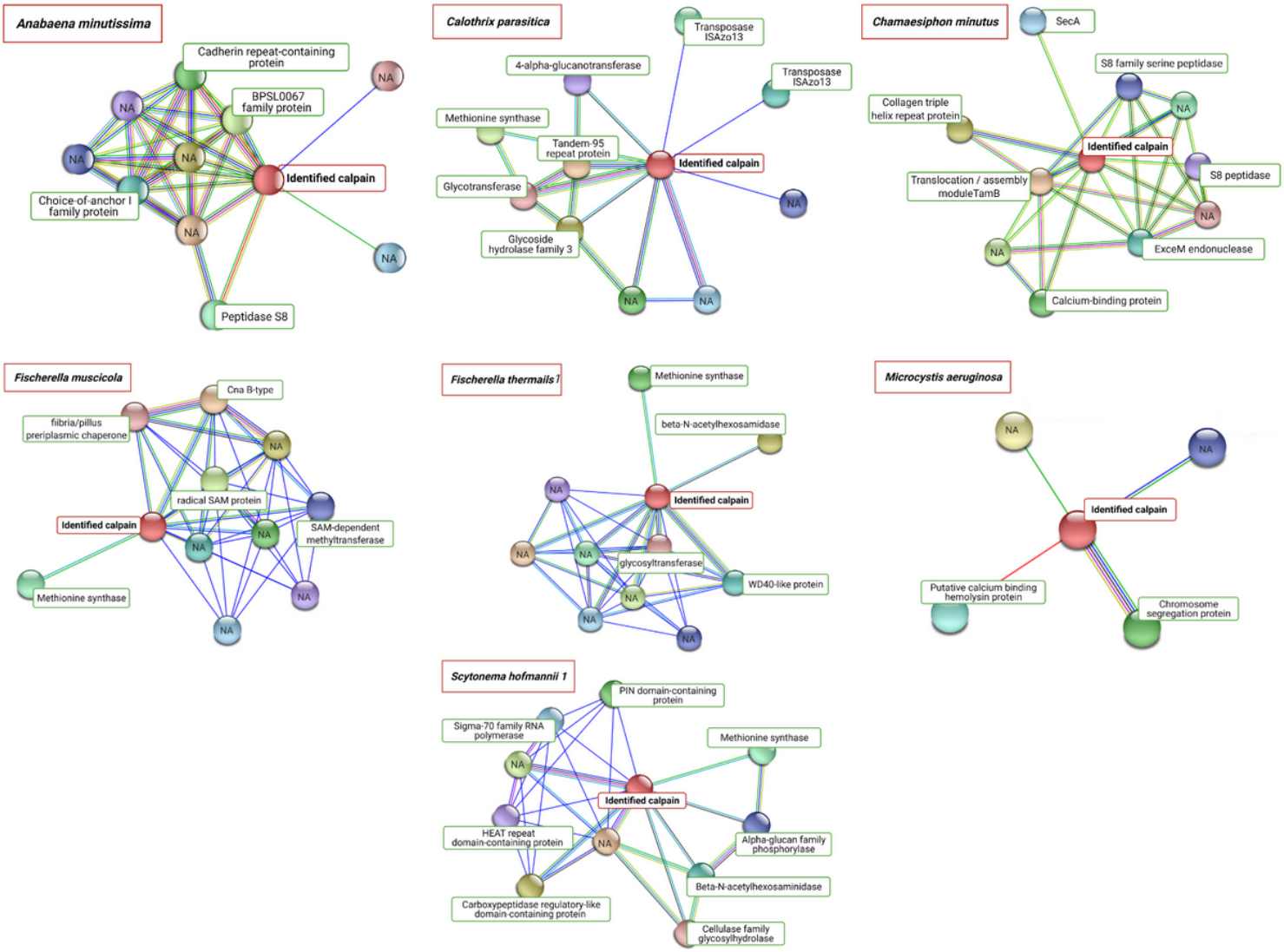

Figure 4

Prediction of interaction partners for cyanobacterial calpains by StringDB. Some predicted interaction partners are not annotated (NA). Identified calpains are shown in red and annotated proteins in green frames, respectively. The evidence for each interaction is determined by the color of connecting lines: (I) Known interactions: experimentally determined (pink), information from curated databases (light blue); (II) Predicted interactions: gene neighbourhood (green), gene fusions (red), gene co-occurrence (dark blue), and (III) Other evidence: text mining (yellow), co-expression (black), protein homology (purple).

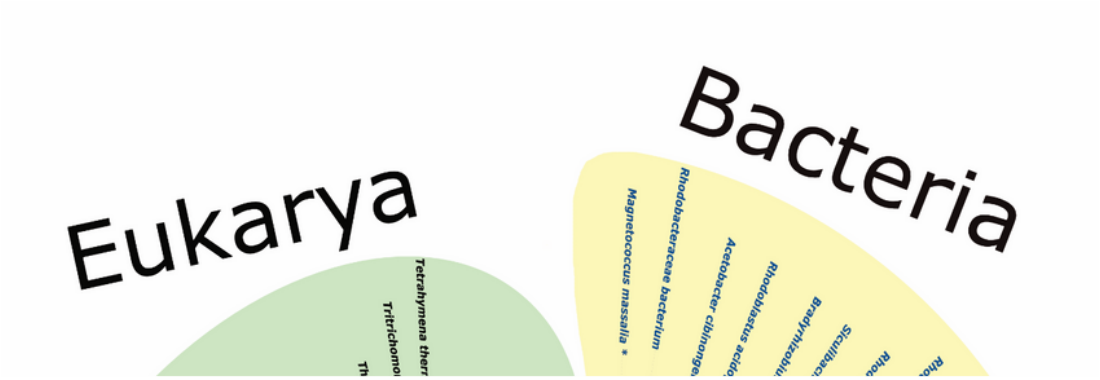

Figure 5

Unrooted phylogenetic tree of calpain catalytic CysPC domain. Cyanobateria are in red and alphaproteobacteria in blue. * Recent phylogenetic studies of alphaproteobacteria suggested that magnetotactic bacteria including Magnetococcus spp. should be excluded from alphaproteobacteria and placed into the separate class Magnetococcia (Muñoz-Gómez et al. 2019).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytables.docx