Acta Scientific MICROBIOLOGY (ISSN: 2581-3226)

Volume 6 Issue 10 October 2023

Research Article

The Lack of Concordance in Evolutionary Pattern of Carboxysome Proteins – Repercussions of HGT or Diverse Evolutionary Potential?

Gurpreet Kaur Sidhu[1], Panchsheela Nogia[1], Vandana Tomar[1], Rajesh Mehrotra[2] and Sandhya Mehrotra[2]*

[1]Plant Molecular Biology and Biochemistry Laboratory, Birla Institute of Technology and Science, Department of Biological Sciences, Pilani, India

[2]Department of Biological Sciences, Birla Institute of Technology and Science, KK Birla Goa Campus, Zuarinagar, Goa, India

***Corresponding Author:** Sandhya Mehrotra, Department of Biological Sciences, Birla Institute of Technology and Science, KK Birla Goa Campus, Zuarinagar, Goa, India.

DOI: 10.31080/ASMI.2023.06.1305

Received: August 10, 2023

Published: September 23, 2023

Abstract

Carboxysomes are microcompartments enclosing the primary photosynthetic enzyme Ribulose 1, 5 Bisphosphate Carboxylase/ Oxygenase (RuBisCO), an adaptation to help overcome the loose specificity of the latter for carbon dioxide. These carboxysomes, which exist in cyanobacteria and a few other eubacteria are composed of a protein shell wherein a well organized multi-protein assembly acts as the carbon concentrating mechanism (CCM). The present study was conducted to find out the presence/absence of the carboxysome forming proteins across various phyla of eubacteria in order to trace their evolutionary path. The analysis was conducted using the CCM proteins of *Gloeobacter* violaceus PCC 7421, an early diverging cyanobacterium.

While α carboxysome proteins are also found in other phyla of eubacteria such as proteobacteria, complete set of β carboxysome constituting proteins are found only in β cyanobacteria. The study supports the fact that shell proteins of carboxysomes are evolutionarily linked to shell proteins of microcompartments involved in ethanolamine utilization and propanediol utilization pathways. Moreover, the CcmM and CcmN proteins have possibly originated by domain shuffling or gene fusion like mechanisms. The CcmM, CcmN and CcmO, the multidomain proteins were found to have an evolutionary pattern different from that of CcmK and CcmL leading to cumulative effect on phylogeny of complete operon which was found to be only moderately similar to most conserved regions of genome. The latter (CcmK and CcmL) also being more conserved suggest less robustness to mistranslation possibly due to tight selection of the protein structure evidently responsible for creating an environment suitable for microcompartment pathway it encloses.

Keywords: Carbon Concentrating Mechanism; Carboxysome; BMC Domain; Domain Shuffling; Microcompartment; Evolution

Introduction

Cell compartmentalization was considered an asset of eukaryotes alone until the discovery of carboxysomes in cyanobacteria in 1961 [1]. The discovery of such compartments was delayed largely because of the icosahedral shape and nano-scale size which caused these to be mistaken as viral capsids. Bacterial microcompartments (BMCs) are polyhedral protein complexes (40-200 nm in diameter), encasing metabolic enzymes encapsulated in a protein shell and are used by bacteria to optimize metabolic processes such as carbon fixation, ethanolamine utilization, 1,2 propanediol utilization etc. [2]. The BMCs serve to prevent toxic metabolites formed by the enclosed enzymes from entering the cytoplasm, pre-

vent the loss of volatile intermediates and prevent the interference of competing substrates [3,4]. Confinement of metabolic pathways further, is assumed to provide better efficiency and enhanced protein stability due to exclusion of oxidative damage [5]. The major kinds of BMCs discovered till date include carboxysomes (carbon fixation) [1], Pdu BMC (1, 2 propanediol utilization) [6] and Eut BMC (Ethanolamine utilization) [7].

The most extensively studied BMC is carboxysomes. Carboxysomes are a part of a carbon concentrating mechanism of cyanobacteria and some proteobacteria. The carboxysomes evolved in order to overcome the inefficiency of enzyme Ribulose-1, 5-Bisphosphate Carboxylase/Oxygenase (RuBisCO), which it suffers at the hands of photorespiration because of its fickle specificity for carbon dioxide and oxygen [8]. The CCM in β cyanobacteria (β cyanobacteria is a lineage of cyanobacteria possessing RuBisCO IB form in comparison to α cyanobacteria which contain the Form IA RuBisCO; the two groups form distinct clades in phylogeny based on 16S rRNA as well) constitute structural proteins CcmK, CcmL and CcmO which form the carboxysome shell and CcmM, CcmN which are enclosed by the carboxysome shell along with RuBisCO [9-11]. The CcmK and CcmO possess the BMC domain while the CcmL protein constitutes the Pfam03319 domain. The BMC domain proteins form the 20 flat facets of the shell while the Pfam03319 domain proteins form the pentamers that introduce curvature to the carboxysome shell by forming the 12 vertices [12]. The hexamers and the pentamers formed by the shell proteins also have a central pore which acts as a passage for metabolites, bicarbonate ions, RuBP (Ribulose 1,5 Bisphosphate) and 3-Phosphoglycerate (3-PGA) [12,13]. The CcmM protein has an N-terminal γ carbonic anhydrase like domain [14-17] and a C-terminal of RuBisCO small subunit (SSU) repeats. CcmM and RuBisCO form protein complexes within the carboxysomes such that the C-terminal interacts with the RuBisCO while the N-terminal is towards the outer shell of the carboxysomes [18]. The carboxysomal protein CcmN contains bacterial transferase hexapeptide repeat domains. The CcmN protein has been reported to interact with the carboxysome shell proteins as well as the core proteins, hence playing an important role in the formation of shell around the core proteins [19]. Kinney., *et al.* (2012) [19] also state the importance of ~18 C-terminal residues of CcmN in carrying out this interaction, as CcmNΔ18 mutant strains did not form carboxysome structures.

The Pdu microcompartments sequester an intermediate of 1,2 propanediol degradation (propionaldehyde) so as to utilize it as carbon source under anaerobic or micro aerobic environments. The Pdu BMCs are formed of 14 different polypeptides viz. PduABB'CDEGHJKOPTU [20]. The shell is composed of BMC domain proteins PduABB'JKTU and the enzymes of the metabolic pathway constitute Pdu CDE (B_{12} dependent dehydratase), PduGH (diol dehydratase), PduO (adenosyl transferase) and PduP (Propionaldehyde dehydrogenase) [20]. Eut BMC is found in bacteria utilizing ethanolamine as carbon, nitrogen and energy source, formed as a result of phosphatidyl ethanolamine degradation in mammalian gastrointestinal tract [2]. The exact protein composition of Eut BMCs is unknown till date because the purification of the Eut BMCs is still not reported. The versatility in the role of BMCs suggests that they contribute to metabolic innovation in bacteria in a broad range of environments [2].

The evolutionary relationships between different types of microcompartments are not very clear but evidently they all contain a conserved outer shell structure, enclosing certain enzymatic reactions. As discussed by [21] Rae., *et al.* (2013), the formation of two types of carboxysomes (α and β) is by convergent evolution. The β carboxysomes are lumen centric and possess higher intrinsic order with the lumen proteins capable of self-assembly, while the α carboxysomes are shell centric being capable of assembling the shell even in the absence of the core proteins [22]. Such observations suggest that possibly the α carboxysomes came into existence by recruitment of carboxysome core proteins to pre-existing BMC shell in primeval cell [21]. Moreover, as reported by [23] Fan., *et al.* (2012) α carboxysomes and Pdu microcompartment are capable of incorporating enzyme complexes through connection with the BMC shell. Alternatively, as conceived by [21] Rae., *et al.* (2013) RuBisCO IA could have incorporated into a pre-existing microcompartment shell by the targeting method, provided there was co-existence of another type of BMC in the same cell. There are reports of organisms possessing more than one type of BMC viz. *Salmonella enterica serovar typhimurium* contains both Pdu and Eut microcompartments [24].

The mechanism by which the bacterial microcompartments control the passage of various metabolic molecules (Bicarbonate, RuBP and 3-PGA in carboxysomes; 1,2-Propanediol, Cobala-

min, 1-propanol and Propionyl-CoA in Pdu microcompartments; Ethanolamine, acetaldehyde in Eut microcompartments) is not yet known, but analyzing these pathways may bring forth novel properties of biological protein shells such as carboxysomes. The study of the carboxysome proteins is timely as the reconstitution of these shells into higher plants forms an important target in order to enhance plant productivity [25]. Further, the tapping of these proteins is also essential to explore their possible use as containers for drug delivery in living systems [26]. The present study involves the screening of various phyla of Eubacteria in order to identify homologs of carboxysome proteins and to add insight into the evolutionary path of the CCM proteins. The analysis was designed upon the belief that the computation of evolutionary distances between various organisms with respect to the CCM proteins will yield the closest relatives of the early diverging cyanobacteria for the carboxysome proteins. In contrast to the expectation, there exists a difference in the evolution trajectory of the CCM proteins despite the fact that they co-exist and are involved in same biochemical pathway.

Materials and Methods

Identification of early diverging cyanobacteria

The 16srRNA, *rbcL* and *pyrH* sequences and ccm operon sequences were retrieved from 37 cyanobacterial strains available at Kazusa Genome Resource (http://genome.microbedb.jp/) and the former were concatenated by sequence matrix software [27]. The sequences retrieved were aligned using ClustalW [28] in GUIDANCE2 server [29,30]. The GUIDANCE2 server provides MSA of the sequences as well as MSA with unreliable columns/sequences removed. The trimmed alignment file was then used for evaluation of best fit substitution model in MEGA 6.0 [31]. The best model on basis of BIC score was then used to carry out phylogenetic analysis with 1000 replicates of bootstrapping. Sequences of *Escherichia coli str* K-12, *Staphylococcus aureus*, *Rhodopseudomonas palustris* CGA009 and *Chlorobium tepidum* TLS (where appropriate) were used for outgrouping the phylogenetic trees. The best fit models for datasets evaluated was General Time Reversible model [32] for concatenated sequences and operon sequences.

Identification of CCM homologs across all Eubacterial phyla

The amino acid sequences of carbon concentrating mechanism proteins of *Gloeobacter violaceus* PCC 7421 were retrieved from Kazusa Genome Resource. The *Gloeobacter violaceus* PCC 7421 proteins were used as the query since according to *16s rRNA-rbcL-pyrH* analysis, it is an early diverging cyanobacteria. The accession numbers for each CCM proteins of *G. violaceus* used in the study are mentioned in table 1.

Table 1: The Accession numbers and the conserved domain profile of the proteins of CCM from *G. violaceus* as mentioned in NCBI database.

Sr. No	Protein Name	Accession number	Protein length	Domain present
1	CcmK	BAC90037.1	102aa	BMC_ccmK domain
2	CcmL	WP_011142092.1	100aa	EutN_ccmL domain
3	CcmM	WP_011142091.1	668aa	Lbh-gamma CA; Rubisco SSU domain
4	CcmN	NP_925038.1	201aa	LbetaH superfamily domain
5	CcmO	NP_925037.1	245aa	BMC_ccmK domain

The protein sequences were blasted (BLASTp) against phyla of Eubacteria (http://blast.ncbi.nlm.nih.gov/Blast.cgi) [33,34]. The exercise brought forth several proteins as hits which either perform the same function as the query sequence or perform a different function but still possess appreciable sequence homology. The BLASTp hits obtained were surveyed such that the matching sequences with an expect value equal to or lower than 10^{-5} to at least one species of each of the phyla of Eubacteria were selected. To further validate the BLASTp search results, the query sequences were used to carry out tBLASTn search to make conclusive analysis on absence or wrong annotation of a gene. The protein hits obtained were verified by CDART [35] for the domain present and hence classified as involved in carboxysome function/formation or otherwise (role in Pdu/Eut microcompartments or other functions).

Identification of the closest homologs of CCM proteins of *G. violaceus*

In order to carry out the evolutionary distance analysis, the amino acid sequences of the homologs of the proteins under study were retrieved from NCBI and a dataset created after removing redundant sequences. These individual datasets were then uploaded into GUIDANCE2 server to carry out MSA by ClustalW and remove

the poorly aligned regions of the sequence. The output file from GUIDANCE2 server after removal of unreliable columns was then used for estimating evolutionary distances in MEGA6.0. The distance of each of the proteins from the corresponding protein of *G. violaceus PCC 7421* was used to identify closest relatives of the early diverging cyanobacteria. The distances are basically number of amino acid/nucleotide substitutions per site between sequences, estimated by equal input model [36] for amino acid sequences and Jukes Cantor model [37] for nucleotide sequences in MEGA 6.0.

Results and Discussions

Gloeobacter violaceus PCC 7421 is an early diverging cyanobacterium

The 16s rRNA, *pyrH,* and *rbcL* sequence based cluster analysis shows two distinct groups of cyanobacteria i.e., the α cyanobacteria and the β cyanobacteria (Figure 1). Two major clades are formed: Clade A constituting *Synechococcus sp* JA-2-3B'a(2-13), *Synechococcus sp* JA-3-3Ab and *Gloeobacter violaceus* PCC 7421 and another clade constituting the rest of the cynaobacteria. The second clade further branches into two sub clades, one of α cyanobacteria (clade C) and the other of β cyanobacteria (clade B). It is to be noted that *Synechococcus elongatus* PCC 6301 and *Synechococcus elongatus* PCC 7942 although are β cyanobacteria, are found to be clustered with α cyanobacteria. Similar results were also obtained by 16s rRNA analysis of cyanobacteria done by [38-43] Nelissen., *et al.* (1995), Memon., *et al.* (2013), Gupta and Mathews, (2010), Dvorak., *et al.* (2014), Soo., *et al.* (2014) and De Rienzi., *et al.* (2013). *Gloeobacter violaceus* PCC 7421, as also reported by previous studies was found to be early diverging. *Gloeobacter violaceus* PCC 7421, a unicellular cyanobacterium dwells on the calcareous rocks in mountainous regions of Switzerland [44]. *G. violaceus* possesses several unique characteristics: it lacks thylakoid membrane and the photosynthetic machinery is situated in the cytoplasmic membrane [44], it shows presence of morphologically distinct phycobilisomes [45] and absence of SQDG (Sulfoquinovosyl diacylglycerol) which has an important role in photosystem stabilization [46]. These characteristics also suggest that this organism has retained ancestral cyanobacterial properties.

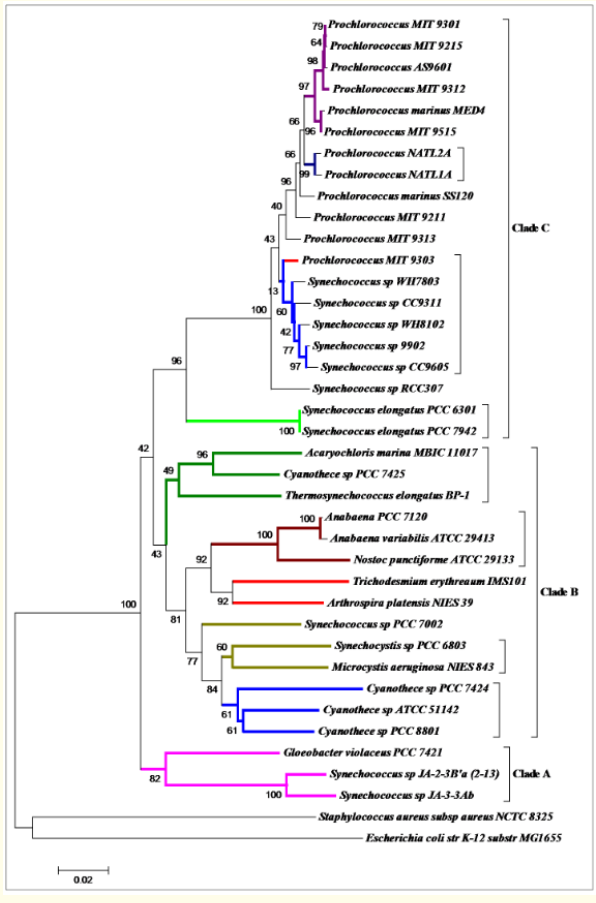

Figure 1: Maximum Likelihood phylogenetic representation of cyanobacteria based on 16s rRNA-pyrH-rbcL sequence. The evolutionary history was estimated on basis of General Time Reversible model with 1000 bootstrapping replicates. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories {+G, parameter = 0.3531}). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 0.0000% sites). The outgrouping of the tree was done using the 16s rRNA sequences of *Staphylococcus aureus*, Escherichia coli, *Rhodopseudominas palustris* CGA009 and *Chlorobium tepidum* TLS.

The complete set of β carboxysome proteins found in β cyanobacteria only

The BLASTp results for CcmK have been listed in table SI. The cyanobacteria, as expected were found to possess the CcmK protein. Further, in the Actinobacteria and Proteobacteria phyla, several significant hits were noted, which include both CcmK protein as well as other proteins (proteins with function/s other than formation of carboxysomes but with similar domain). The homologs of CcmK obtained in the analysis, apart from CcmK include, Ethanolamine utilization protein EutM, Ethanolamine utilization protein EutK, Propanediol utilization protein PduA, Propanediol utilization

protein PduJ and Propanediol utilization protein PduT. Seven out of 23 reported genes for Pdu microcompartment are homologous to the carboxysome shell proteins [6]. The proteins found to have structural and functional analogy to the CcmK and CcmO proteins of carboxysomes are CsoS1A-D and CsoS2 in α carboxysomes, PduB, PduA, PduJ, PduK, PduU and PduT in Pdu microcompartments and EutL, EutM and EutS in Eut microcompartments [2,47].

The BLASTp results for CcmL are mentioned in table SII. The search from the cyanobacterial protein database revealed several significant hits which include both CcmL protein as well as proteins such as Ethanolamine utilization protein EutN. Similarly, α proteobacteria and β proteobacteria gave equal number of significant hits for CcmL and other proteins i.e., EutN. The proteins in Pdu and Eut microcompartments responsible for forming the vertices are PduN and EutN, respectively and hence share appreciable sequence homology. The remaining phyla of Eubacteria were not found to possess CcmL protein, instead they have a hypothetical protein with significant homology to CcmL. The hypothetical protein could either be CcmL protein or a protein involved in formation of microcompartment for some pathway other than Calvin cycle of photosynthesis as the CDART analysis revealed conserved domain.

The domains present in CcmM are Lbh_gamma CA and RuBisCO small subunit (SSU) domain. The BLASTp results for CcmM are mentioned in table SIII. The BLAST hits from cyanobacteria phylum were CcmM proteins from other cyanobacteria as well as carbonic anhydrase. No complete homologs of CcmM were found in proteobacteria. This implies that the homolog identified as hypothetical protein is not complete CcmM protein, but the proteins have carbonic anhydrase/RuBisCO SSU domains in common. The several of the hits obtained are carbonic anhydrase as the N-terminal of CcmM has γ carbonic anhydrase domain. The *Synechococcus elongatus* PCC 7942 CcmM N-terminal bears 60% amino acid similarity to γ carbonic anhydrase of *Methanosarcina thermophila* and possibly forms a similar structural arrangement. Apart from cyanobacteria, all phyla under study were observed to possess homologs other than CcmM protein. The various homologs obtained include carbonic anhydrase, acetyltransferase, ferripyochelin binding protein (*Methanothermobacter* sp CaT2, Archaea, CA domain), siderophore binding protein (Organism: *Halobacteriovorax marinus*, Delta proteobacteria, has CA domain), phenylacetic acid degradation protein PaaY, 2,3,4,5-tetrahydropyridine-2,6-dicarboxylate-N-acetyltransferase, hexapeptide repeat containing protein, UDP-3-O-[3 hydroxymyristoyl] glucosamine –N-acyltransferase, protein YrdA and hypothetical protein.

Very few CcmN homologs were found in the various phyla under investigation (table SIV). Only Cyanobacteria and Firmicutes gave hits of CcmN protein and other proteins which include hypothetical protein, hexapeptide repeat containing transferase, transferase and carbonic anhydrase (all of these proteins contain the LbetaH domain). The various homologs of CA/LbetaH domain of CcmM and CcmN, could be the possible candidates for evolutionary ancestors of these CCM lumen proteins. The most common molecular mechanisms responsible for formation of multidomain proteins include non-homologous recombination (also called domain shuffling) [48], fusion of genes [49,50] and fission of genes [51,52]. At the time of emergence of CCM in primitive cyanobacteria, the pre-existing domains possibly underwent certain evolutionary mechanism viz. domain shuffling, gene fusion or gene fission in order to develop proteins such as CcmM and CcmN, which are crucial for a functional low environmental Ci phenotype. This theory could be supported by the fact that CcmM and CcmN do not have any complete homologs in any of the reported phyla available at NCBI, apart from the domain centric homologs.

The CcmO protein contains two repeats of the BMC domain. In cyanobacteria phyla, hits were obtained for both CcmO and other proteins which include hypothetical protein (table SV). The phyla which gave hits other than CcmO, include Bacteroidetes/Chlorobi, Chlamydiae, Chloroflexi, Gemmatimonadetes, Spirochaetes and Tenericutes. The hits obtained in these phyla include EutM, carboxysomes shell protein, hypothetical protein, PduT, microcompartment protein and propanediol utilization protein. Very few CcmO homologs were obtained in Actinobacteria, Fibrobacteres/Acidobacter, Firmicutes, Fusobacteria, Planctomycetes and γ proteobacteria.

The phyla found to possess homologs for CcmK, CcmL and CcmO include Actinobacteria, Fibrobacteres, Fusobacteria, Planctomycetes, α proteobacteria, β proteobacteria, Gamma proteobacteria, Cyanobacteria and Firmicutes. The genome of the organisms other than cyanobacteria, possessing CcmK/L homologs were screened for the presence of other CCM proteins, but were not found to possess other CCM proteins (*Acidimicrobium ferrooxidans* (Actinobac-

teria), *Bradyrhizobium sp* ORS 278 and *Bradyrhizobium sp* BTAi1 (α proteobacteria)). These bacteria have been reported to have α carboxysomes which have a different set of proteins apart from the shell proteins for carboxysome structure formation (α carboxysomes). While as also observed in the present analysis there are no reports for presence of the β carboxysomes or its gene repertoire in any other group of organisms except the β cyanobacteria (table 2).

The analysis reveals that the carbon concentrating mechanism microcompartment proteins, the ethanolamine utilization pathway shell proteins and the propanediol utilization mechanism shell proteins are evolutionarily linked to each other. The phylogenetic analysis of BMC domain proteins from carboxysomes, Pdu BMC, Eut BMC and grp-type BMC shows segregation of the carboxysomal shell proteins from the rest and moreover, the occurrence of BMCs other than carboxysomes is not uniform among organisms of a particular phyla or even genera [53].

Table 2: The distribution of the β carboxysome proteins across various phyla of Eubacteria.

Phyla with no CCM protein homologs	Phyla with homologs for CcmK, CcmL and CcmO	Phyla with domain homologs for CcmM and CcmN domain	Phyla with homologs for all CCM (β carboxysome) proteins
Armatimonadetes	Actinobacteria	Archaea	Cyanobacteria
ε Proteobacteria	Fibrobacteres	Aquificae	
Thermodesulfobacteria	Fusobacteria	Caldiserica	
	Planctomycetes	Chlamydiae	
	α Proteobacteria	Chrysiogenetes	
	β Proteobacteria	Deferribacteres	
	γ Proteobacteria	Deinococcus	
	Firmicutes	Dictyoglomi	
		Elusimicrobia	
		Gemmatimonadetes	
		Nitrospinae	
		Nitrospirae	
		Spirochaetes	
		Tenericutes	

Was the CCM operon passed on by vertical succession or horizontal gene transfer?

The phylogenetic tree of the carboxysome operon (Figure 2) shows two major clades constituting α cyanobacteria and β cyanobacteria, respectively. This result is in accordance with the expectation, since both α carboxysomes and β carboxysomes are formed of different set of proteins with the exception of BMC proteins, which are found in both. The subclades formed for α cyanobacteria are in accordance with the 16s rRNA phylogeny, validating the fact that the α carboxysome operon was formed *in-situ* and passed in vertical succession. The topological arrangement of β cyanobacteria demonstrates some variance with respect to the 16s rRNA based phylogeny. An important observation that can be made from the topology of carboxysome operon based tree is that although the arrangement of the various organisms is at some variance from that in case of 16s rRNA tree but the clustering of the organisms in the sub-clades formed is conserved. Further, although the early diverging forms cluster together, but it would not be correct to consider them to be the first forms to have acquired CCM.

Roughly 50% of prokaryotic genome is reported to be operonic [54] evolved under selection pressure [39]. Memon., *et al.* (2013) hypothesize that the most conserved and moderately conserved operons are formed *in-situ* and inherited vertically, on the basis of comparative phylogenetic analysis between the operon under study and the 16s rRNA sequence. An appreciable similarity between the phylogenetics based on 16s rRNA sequence and any other genetic sequence, if observed, is possible only if the latter is passed on by vertical inheritance [39].

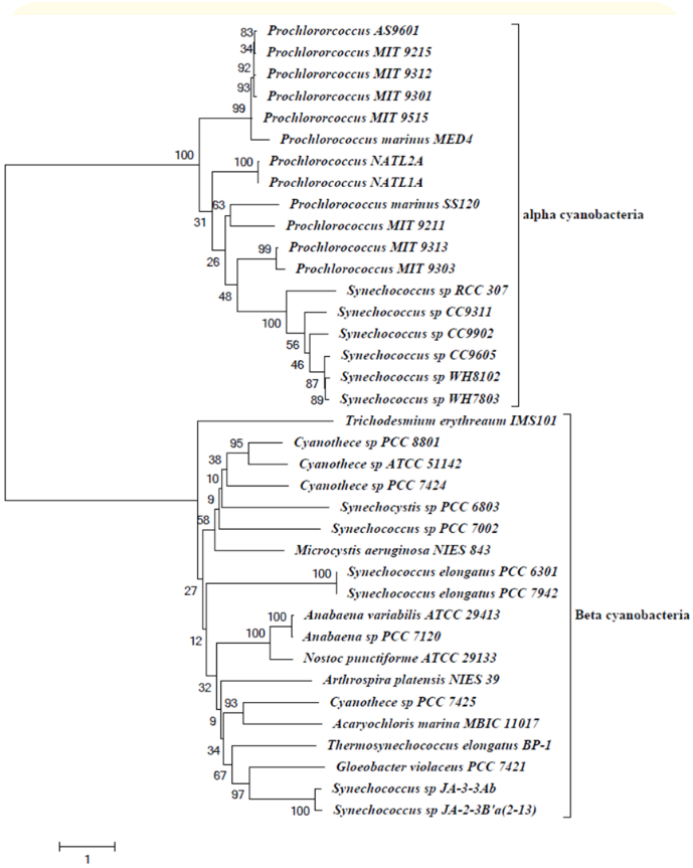

Figure 2: Maximum Likelihood phylogenetic representation of cyanobacteria based on carboxysome operon sequence. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches. The evolutionary distances were computed using the General Time Reversible model with 1000 bootstrapping replicates. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories {+G, parameter = 1.8049}). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 0.0000% sites).

During the study it was noted that the *cso* operon constitutes the carboxysome encoding genes as well as the RuBisCO encoding genes, while most of the β cyanobacterial carboxysomes constitute only the CCM genes. Only five of the β cyanobacteria, available in Kazusa Genome Resource namely, *Synechococcus elongates* PCC 6301, *Synechococcus elongatus* PCC 7942, *Synechococcus* sp PCC 7002, *Microcystis aeruginosa* NIES 843 and *Cyanothece* sp PCC 7424 were found to have rubisco genes in tandem with the carboxysome encoding genes. Further, while most of the β cyanobacteria were found to have CcmO as a part of the operon, several had CcmO encoding gene distant from the other CCM encoding genes (in *Synechocystis* sp PCC 6803, *T. elongatus* BP-1, *M. aeruginosa* NIES 843, *Synechococcus* sp PCC 7002, *Cyanothece* sp ATCC 51142, *Cyanothece* sp PCC 7424 and *Cyanothece* sp PCC 8801). The occurrence of CcmO encoding gene away from the rest of the CCM encoding genes suggests lack of resistance to maintaining the operonic arrangement. It would be interesting to identify the regulatory mechanisms of these proteins for carboxysome formation. Moreover, while *ccmk/ccmk1/ccmk2* were found to be located adjacent to the rest of CCM genes, *ccmk3/ccmk4* if present were always found to be located at different loci of the genome (table SII). It must be noted that the CCM genes, present on direct or complementary strand, were found to have a conserved arrangement, i.e., the order of the genes is *ccmO*, *ccmN*, *ccmM, ccmL* and *ccmK*. Amongst the organisms considered for the study, in one organism only, i.e., *Trichodesmium erythraeum* IMS101 there is an insertion of a hypothetical protein in-between *ccmM* and *ccmN*, and hence it forms a separate branch in the phylogenetic tree of CCM operon of β cyanobacteria. On the basis of the conservation of the complete set of CCM encoding genes in an operon, *G. violaceus* PCC 7421, *Synechococcus* sp JA-2-3Ba'(2-13) and *Synechococcus* sp JA-3-3Ab, would be most suitable to study the carboxysome. In other cyanobacteria analyzed in the present study, some parts of the carboxysome encoding genes are lying separately in the genome, hence requiring a separate set of genetic engineering steps to get the complete picture of CCM assembly and function.

Diversity of CCM proteins

All CCM proteins were analyzed for pairwise distances to find out the relation of one protein with respect to the others on the

basis of sequence homology. Further, the distance of various cyanobacteria was analyzed with respect to *G. violaceus* PCC 7421, the most primitive cyanobacteria, using the number of substitutions per site of the sequences under study. The analysis brought forth the closest relatives of *G. violaceus* PCC 7421 on the basis of 16s rRNA and CCM protein sequence homology. The distance obtained for individual sequence depicts the distance of each cyanobacteria from *G. violaceus* PCC 7421, for each protein respectively and hence the probable time of emergence of CCM in that particular cyanobacteria. This analysis by itself fails to give the exact closest relatives of *G. violaceus* PCC 7421 due to difference in scale of distance for each protein; hence a box and whiskers plot was used to analyze the data.

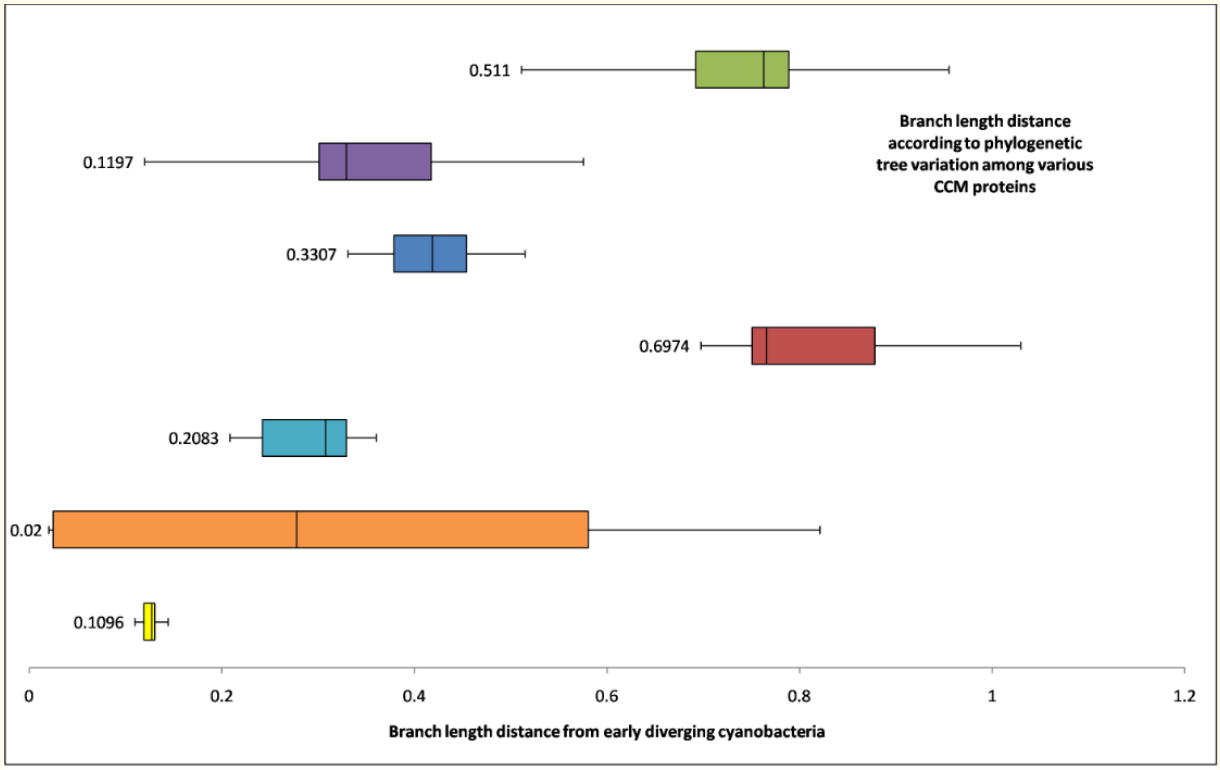

Figure 3: Box and Whiskers plot for the distance of the 16s rRNA (yellow), CCM operon (orange), CcmL (light blue), CcmN (red), CcmO (dark blue), CcmM (purple) and CcmK (green) sequences (on Y axis) with respect to *Gloeobacter violaceus* PCC 7421 sequences, depicting the degree of evolutionary distance (X axis) existing among the sequences under study.

The thickness of the bars shows degree of variation in the distance of the various β cyanobacteria from *G. violaceus* PCC 7421 (Figure 3). The distance of various cyanobacteria from *G. violaceus* PCC 7421 on basis of 16S rRNA shows very little variation, as expected, since it is the most conserved region of the genome. The proteins CcmN and CcmM show greater degree of variation. CcmK, although expected to be very diverse due to several duplicates of the protein found in each genome, the bar on the graph shows it to be comparatively conserved. This result was obtained, because only CcmK/CcmK1/CcmK2 was used for this particular analysis (i.e., the CcmK protein closest to *G. violaceus* PCC 7421 CcmK from ML tree).

Further, information that can be extracted from the plot is the identification of close relatives of *G. violaceus* PCC 7421 with respect to the CCM proteins. The cyanobacteria corresponding to distance values lying between the minimum value and the first quartile of the set of values for a particular protein were considered as close relatives (mentioned in Figure 4).

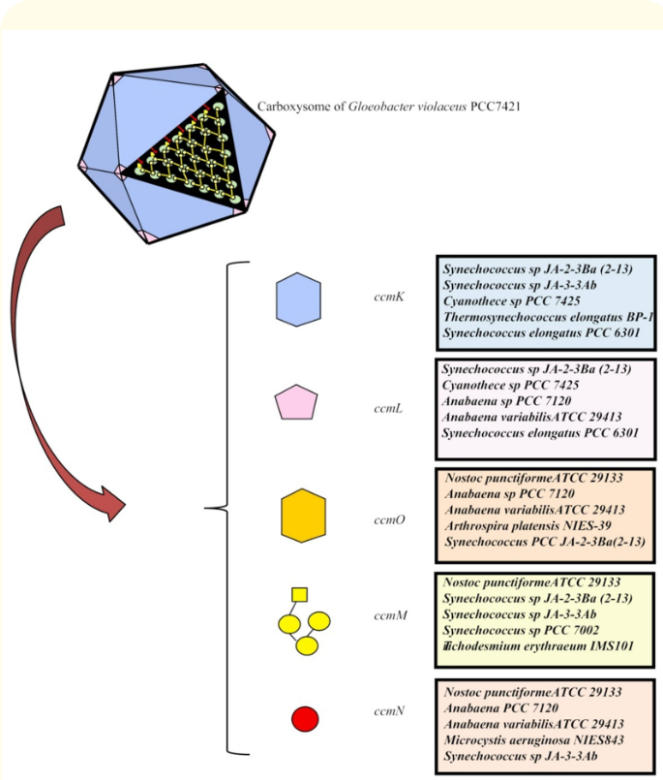

Figure 4: The closest homologs of CCM proteins of *Gloeobacter violaceus* PCC 7421, among the β-cyanobacteria available at Cyanobase found by phylogenetic analysis.

According to 16s rRNA-rbcL-pyrH sequence, *T. elongatus* BP-1, *Synechococcus* sp JA-3-3Ab, *Cyanothece* sp PCC 7425, *Synechococcus* sp JA-2-3B'a(2-13) and *Acaryochloris marina* MBIC11017 are closest to *G. violaceus PCC 7421*. Further according to ccm operon sequence *Synechococcus* sp JA-2-3Ba(2-13), *Synechococcus* sp JA-3-3Ab, *Cyanothece* sp PCC 7425, *T. elongatus* BP-1 and *Acaryochloris marina* MBIC11017 are closest relatives according to distance. Similarly, for CcmK and CcmL, the closest homologs are *Synechococcus* sp JA-2-3B'a(2-13), *Synechococcus* sp JA-3-3Ab, *Cyanothece* sp PCC 7425, *T. elongatus* BP-1 and *S. elongatus* PCC 6301 and *Synechococcus* sp JA-2-3B'a(2-13), *Cyanothece* sp PCC 7425, *Anabaena sp* PCC 7120, *A. variabilis* ATCC 29413 and *S. elongatus* PCC 6301, respectively. For CcmM, CcmN and CcmO are *Nostoc punctiforme* ATCC29133, *Synechococcus* sp JA-2-3B'a(2-13), *Synechococcus* sp JA-3-3Ab,*Synechococcus* sp PCC 7002, *T. erythraeum* IMS101; *Nostoc punctiforme* ATCC29133, *Anabaena* sp PCC 7120, *A. variabilis* ATCC 29413 *M aeruginosa* NIES843, *Synechococcus* sp JA-3-3Ab; *Nostoc punctiforme* ATCC29133, *Anabaena* sp PCC 7120, *A. variabilis* ATCC 29413, *A. platensis* NIES39, *Synechococcus* sp JA-2-3B'a(2-13), respectively.

The analysis also shows lack of concordance in evolution of the CCM proteins. The CcmK and CcmL proteins appear to have co-evolved, separately from CcmM, CcmN and CcmO which appear to have evolved together. The multiple sequence alignment of CcmM and CcmN proteins shows a comparatively conserved C terminal and a more diverse N terminal region. To consider the possibility that N terminal of both the proteins had undergone recent evolutionary changes in accordance to the proteins they interact with in BMC formation and hence resulted in this disparity, the evolutionary distance of the C and N terminals (separately) of the proteins was computed against *G. violaceus* proteins. Despite the variability, the analysis brought forth the same set of organisms as the close relatives of respective proteins in *G. violaceus*, viz. *N. punctiforme* for CcmM, CcmN and CcmO.

The occurrence of *Synechococcus sp* sequences as close relative of *G. violaceus* CcmK and CcmL and *N. punctiforme* sequences as close relative of *G. violaceus* CcmM, CcmN and CcmO suggests the possibility of acquiring the shell proteins from other BMC containing early ancestors and evolution of lumen proteins at a later stage. However, this revelation does not reflect in the phylogeny based on complete operon which is moderately concordant with the 16S rRNA evolution, signaling organism specific evolutionary patterns which cannot be confirmed by current analysis. The variation in protein evolution across species could be explained by the differences in the underlying mutation rates which in turn are governed by differences in DNA methylation, fidelity of DNA-repair mechanisms or production of DNA-damaging agents [55].

Conclusions

The present analysis reveals that α and β carboxysomes are encoded by distinct set of proteins apart from the shell proteins that share the BMC domain, and hence form distinct phylogenetic clades. Further, the comparison of the carboxysome genes based phylogeny with that of the 16s rRNA shows moderate congruency between the two and hence suggests *in-situ* formation of the carboxysome and its subsequent vertical succession. As pointed out by several researchers, the carboxysomes are believed to have come into emergence after the divergence of the α and β cyanobac-

teria, considering the difference in the set of genes involved in the carboxysomes of both. The phylogenetic analysis, the conservation of genetic arrangement and sequence of the β carboxysome proteins adds to the evidence of *in-situ* formation of the ccm operon. Further, the structural and functional homology to the BMC shell proteins of other microcompartments (pdu/eut) signifies the evolutionary relation between the two. The CcmM and CcmN proteins were found to have no protein homolog/s amongst the proteins available at NCBI database apart from the homologs as result of common domains, suggesting the evolution of these by mechanisms like domain shuffling, gene fusion or gene fission. The CcmK proteins of the β cyanobacteria when analyzed were found to be present in several copies, emphasizing the importance of the protein in shell formation. The present analysis was conducted to identify closest relatives of early diverging cyanobacteria with respect to the CCM proteins, with the objective of tracing their evolutionary paths. While the phylogenies as well as the evolutionary distances suggest a relationship among the cyanobacteria in accordance with that obtained from the most conserved regions of the genome, a similar analysis on the basis of the individual CCM proteins shows some disparity. As suggested by [56] Csaba., *et al.* (2006), the proteins evolve at different rates depending on several factors including selection on protein structure and function and robustness to mistranslation. The CcmK and CcmL proteins have been found to be more conserved in comparison to CcmM, CcmN and CcmO, signaling the former to be less robust to variation in sequence as compared to the latter and hence the lack in concordance of the evolution of the proteins involved in the same pathway.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Dr. Sandhya Mehrotra conceived and designed the experiments. GKS analysed the data and wrote the first draft of the manuscript. PN, VT and RM contributed to the writing of the manuscript. All authors agree with manuscript results and conclusions. GKS and SM jointly developed the structure and arguments for the paper and made critical revisions and approved final version. All authors reviewed and approved of the final manuscript.

Funding

GKS is thankful to the UGC-BSR for her fellowship. PN is thankful to CSIR for Senior Research fellowship. VT is thankful to DST-inspire fellowship program of DST, India. This work was supported by SERB project EMR/2016/002470 sanctioned by the government of India to SM.

Acknowledgments

The authors are grateful to Birla Institute of technology and science, Pilani, Rajasthan, India for providing infrastructural and logistic support.

Table SI: pBLAST hits for amino acid sequence of CcmK from Gloeobacter violaceus PCC 7421 from various phyla of Eubacteria and Archaeabacteria

S.No	Phylum	Total Hits	$E<10^{-5}$	ccmK and $E<10^{-5}$	Other proteins and $E<10^{-5}$	Names of other proteins
1	Archaeabacteria	9	0	0	0	-
2	Actinobacteria	>100	>100	31	69	PduA,EutM, BMC domain protein
3	Aquificae	11	0	0	0	-
4	Armatimonadetes					
5	Bacteroidetes/Chlorobi	100	11	-	11	Carboxysome shell protein, EutM, pduT, hypothetical protein
6	Caldiserica	42	0	0	0	-
7	Chlamydiae	100	0	0	0	-
8	Chloroflexi	14	3	0	3	EutM, carboxysome shell protein
9	Chrysiogenetes	13	0	0	0	-
10	Cyanobacteria	100	100	100	0	-

11	Deferribacteres	2	0	0	0	-
12	Deinococcus-thermus	2	0	0	0	-
13	Dictyoglomi	4	0	0	0	-
14	Elusimicrobia	6	0	0	0	-
15	Fibrobacteres/Acidobacter	21	12	1	11	EutM, Carboxysome shell protein, microcompartment protein
16	Firmicutes	100	100	8	92	EutM, pduA, carboxysome shell protein
17	Fusobacteria	100	79	2	77	Carboxysome shell protein, EutM, pduT, hypothetical protein
18	Gemmatimonadetes	11	5	0	5	EutM
19	Nitrospinae	5	0	0	0	-
20	Nitrospirae	101	0	0	0	-
21	Planctomycetes	46	36	1	35	Carboxysome shell protein, EutM, microcompartment protein
22	Alpha Proteobacteria	52	33	22	10	EutM, pduA
23	Beta Proteobacteria	54	46	34	12	EutM, pduA, EutK, pduT
24	Gamma Proteobacteria	100	100	41	59	EutM, pduJ, pduA
25	Delta Proteobacteria	101	70	40	30	EutM, pduA, EutK, Biotin ligase
26	Epsilon Proteobacteria	2	0	0	0	-
27	Spirochaetes	24	20	0	20	Carboxysome shell protein, EutM, hypothetical protein
28	Synergistetes	71	36	3	33	Carboxysome shell protein, EutM, pduT
29	Tenericutes	12	4	0	4	Propanediol utilization protein, EutM
30	Thermodesulfobacteria	12	0	0	0	-
31	Thermotogae	13	0	0	0	-

Table SII: pBLAST hits for amino acid sequence of CcmL from *Gloeobacter violaceus PCC 7421* from various phyla of Eubacteria and Archaeabacteria

S. No	Phylum	Total Hits	$E<10^{-5}$	CcmL and $E<10^{-5}$	Other proteins and $E<10^{-5}$	Names of other proteins
1	Archaeabacteria	2	0	0	0	-
2	Actinobacteria	78	22	1	21	EutN, PduN, hypothetical protein
3	Aquificae	1	0	0	0	-
4	Armatimonadetes	9	0	0	0	-
5	Bacteroidetes/Chlorobi	100	13	0	13	EutN, hypothetical protein
6	Caldiserica	3	0	0	0	-
7	Chlamydiae	3	0	0	0	-
8	Chloroflexi	10	1	0	1	EutN
9	Chrysiogenetes	8	0	0	0	-
10	Cyanobacteria	100	100	86	14	EutN
11	Deferribacteres	2	0	0	0	-

12	Deinococcus-thermus					
13	Dictyoglomi	8	0	0	0	-
14	Elusimicrobia	50	0	0	0	-
15	Fibrobacteres/Acidobacter	18	11	0	11	EutN, hypothetical protein
16	Firmicutes	100	100	10	90	EutN, propanediol utilization protein, hypothetical protein
17	Fusobacteria	38	33	2	31	EutN, hypothetical protein
18	Gemmatimonadetes	13	3	0	3	EutN, hypothetical protein
19	Nitrospinae	9	0	0	0	-
20	Nitrospirae	9	0	0	0	-
21	Planctomycetes	52	31	8	23	EutN
22	Alpha Proteobacteria	38	17	8	9	EutN
23	Beta Proteobacteria	41	10	6	4	EutN
24	Gamma Proteobacteria	100	100	0	100	EutN
25	Delta Proteobacteria					
26	Epsilon Proteobacteria	3	0	0	0	-
27	Spirochaetes	11	8	0	8	Hypothetical protein, EutN
28	Synergistetes	15	11	1	10	EutN
29	Tenericutes	15	3	0	3	Hypothetical protein, EutN
30	Thermodesulfobacteria	0	0	0	0	-
31	Thermotogae	8	0	0	0	-

Table SIII: pBLAST hits for amino acid sequence of CcmM from *Gloeobacter violaceus PCC 7421* from various phyla of Eubacteria and archaeabacteria

S. No	Phylum	Total Hits	E<10^{-5}	CcmM and E<10^{-5}	Other proteins and E<10^{-5}	Names of other proteins
1	Archaeabacteria	100	100	0	100	Carbonic anhydrase, acetyltransferase, ferripyochelin binding protein
2	Actinobacteria	100	100	0	100	Carbonate dehydratase, anhydrase, hypothetical protein, siderophore binding protein, phenylacetic acid degradation protein PaaY, 2,3,4,5-tetrahydropyridine-2,6-dicarboxylate-N-acetyltransferase
3	Aquificae	60	20	0	20	Acetyltransferase, transferase, hypothetical protein, putative carbonic anhydrase
4	Armatimonadetes	20	0	0	0	-
5	Bacteroidetes/Chlorobi	100	100	0	100	Acetyltransferase, transferase, hypothetical protein, carbonic anhydrase, hexapeptide repeat containing protein
6	Caldiserica	12	1	0	1	Hypothetical protein
7	Chlamydiae	30	0	0	0	-

8	Chloroflexi	73	13	0	13	Anhydrase, anhydratase, Transferase hexapeptide repeat containing protein, hypothetical protein
9	Chrysiogenetes	16	2	0	2	Transferase, anhydratase
10	Cyanobacteria	>100	>100	49	51	Carbonic anhydrase, cytochrome C biogenesis protein ccmM
11	Deferribacteres	33	8	0	8	Hypothetical protein, acetyltransferase
12	Deinococcus-thermus	76	22	0	22	Carbonate dehydratase, carbonic anhydrase, ferripyochelin binding protein, hypothetical protein, NUDIX protein
13	Dictyoglomi	15	2	0	2	Hypothetical protein
14	Elusimicrobia	4	1	0	1	Hypothetical protein
15	Fibrobacteres/Acidobacter	43	20	0	20	Transferase, Hypothetical protein, carbonic anhydrase
16	Firmicutes	100	100	4	96	Carbonic anhydrase, carbonate dehydratase, hypothetical protein, transferase
17	Fusobacteria	77	14	0	14	Bacterial transferase hexapeptide repeat containing protein, acetyltransferase, Hypothetical protein
18	Gemmatimonadetes	38	9	0	9	Transferase hexapeptide repeat containing protein, phenylacetic acid degradation protein PaaY, Hypothetical protein, transferase
19	Nitrospinae	22	3	0	3	Hypothetical protein, putative transferase, hexapeptide repeat protein
20	Nitrospirae	42	10	0	10	Carbonic anhydrase, putative transferase, acetyltransferase, Hypothetical protein
21	Planctomycetes	63	24	0	24	Ferripyochelin binding protein, Hypothetical protein, anhydrase, , phenylacetic acid degradation protein PaaY
22	Alpha Proteobacteria	100	100	0	100	Hexapeptide repeat containing transferase, carbonate dehydratase, acetyltransferase
23	Beta Proteobacteria	100	100	0	100	Carbonate dehydratase, acetyltransferase
24	Gamma Proteobacteria	100	100	0	100	Carbonate dehydratase, UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase
25	Delta Proteobacteria	101	101	0	101	Sulfate permease, carbonic anhydrase, transferase, protein YrdA, phenylacetic acid degradation protein PaaY, UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase
26	Epsilon Proteobacteria					-
27	Spirochaetes	100	100	0	100	Transefrase hexapeptide repeat protein, acetyltransferase, carbonic anhydrase
28	Synergistetes	33	7	0	7	Hypothetical protein, anhydrase, transferase
29	Tenericutes	11	1	0	1	Hypothetical protein
30	Thermodesulfobacteria	9	0	0	0	-
31	Thermotogae	74	2	0	2	Hypothetical protein, acetyltransferase

Table SIV: pBLAST hits for amino acid sequence of CcmN from Gloeobacter violaceus PCC 7421 from various phyla of Eubacteria and Archaeabacteria

S. No	Phylum	Total Hits	E<10^{-5}	CcmN and E<10^{-5}	Other proteins and E<10^{-5}	Names of other proteins
1	Archaeabacteria	107	15	0	15	Hexapeptide repeat containing transferase, carbonic anhydrase, acetyltransferase
2	Actinobacteria	103	51	0	51	Carbonic anhydrase, isoleucine patch superfamily enzyme, putative siderophore binding protein
3	Aquificae	11	0	0	0	-
4	Armatimonadetes	20	0	0	0	-
5	Bacteroidetes/Chlorobi	101	17	0	17	Carbonic anhydrase, bacterial transferase hexapeptide repeat protein, acetyltransferase
6	Caldiserica					
7	Chlamydiae	34	1	0	1	Carbonic anhydrase
8	Chloroflexi	26	7	0	7	Transferase hexapeptide repeat containing protein, anhydratase, acetyltransferase
9	Chrysiogenetes	15	0	0	0	-
10	Cyanobacteria	100	100	21	79	Hypothetical protein, hexapeptide repeat containing transferase, transferase, 2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-acetyltransferase
11	Deferribacteres	18	0	0	0	-
12	Deinococcus-thermus	66	11	0	11	Carbonic anhydrase, hypothetical protein, ferripyochelin binding protein, Isoleucine patch superfamily enzyme
13	Dictyoglomi	9	0	0	0	-
14	Elusimicrobia	13	0	0	0	-
15	Fibrobacteres/Acidobacter	36	0	0	0	-
16	Firmicutes	102	10	0	10	Hypothetical protein, transferase, carbonic anhydrase
17	Fusobacteria	43	1	0	1	Bacterial transferase hexapeptide repeat protein
18	Gemmatimonadetes	27	1	0	1	Transferase hexapeptide repeat containing protein
19	Nitrospinae	16	0	0	0	-
20	Nitrospirae	28	0	0	0	-
21	Planctomycetes	37	0	0	0	-
22	Alpha Proteobacteria	102	33	0	33	Anhydrase, bacterial transferase hexapeptide repeat protein, acetyltransferase
23	Beta Proteobacteria	105	11	0	11	Anhydrase, Transferase hexapeptide repeat containing protein
24	Gamma Proteobacteria	114	7	0	7	Carbonic anhydrase, hypothetical protein, acetyltransferase, bacterial transferase hexapeptide repeat protein
25	Delta Proteobacteria					
26	Epsilon Proteobacteria	29	0	0	0	-
27	Spirochaetes	100	0	0	0	-
28	Synergistetes	35	0	0	0	-
29	Tenericutes	9	0	0	0	-
30	Thermodesulfobacteria	8	0	0	0	-
31	Thermotogae	39	0	0	0	-

Table SV: pBLAST hits for amino acid sequence of CcmO from *Gloeobacter violaceus PCC 7421* from various phyla of Eubacteria and Archaeabacteria

S.No	Phylum	Total Hits	$E<10^{-5}$	CcmO and $E<10^{-5}$	Other proteins and $E<10^{-5}$	Names of other proteins
1	Archaeabacteria	10	0	0	0	-
2	Actinobacteria	>100	98	3	95	Carboxysome shell protein, EutM, PduA, PduK
3	Aquificae	4	0	0	0	-
4	Armatimonadetes	8	0	0	0	-
5	Bacteroidetes/Chlorobi	26	9	0	9	EutM, carboxysome shell protein, hypothetical protein PduT
6	Caldiserica					
7	Chlamydiae	50	20	0	20	EutM, carboxysome shell protein, hypothetical protein, microcompartment protein
8	Chloroflexi	12	3	0	3	EutM, carboxysome shell protein
9	Chrysiogenetes	9	0	0	0	-
10	Cyanobacteria	100	100	50	50	Hypothetical protein, EutM
11	Deferribacteres	7	0	0	0	-
12	Deinococcus-thermus	10	0	0	0	-
13	Dictyoglomi	9	0	0	0	-
14	Elusimicrobia	8	0	0	0	-
15	Fibrobacteres/Acidobacter	29	13	1	12	EutM carboxysome shell protein, microcompartment protein PduT
16	Firmicutes	>100	>100	7	93	EutM, PduA, BMC domain protein, carboxysome shell protein
17	Fusobacteria	>100	67	2	65	EutM, carboxysome shell protein
18	Gemmatimonadetes	14	5	0	5	EutM
19	Nitrospinae	11	0	0	0	-
20	Nitrospirae	5	0	0	0	-
21	Planctomycetes	85	35	1	34	EutM, carboxysome shell protein, microcompartment protein, hypothetical protein
22	Alpha Proteobacteria	97	31	21	10	EutM, PduA
23	Beta Proteobacteria	49	43	28	11	EutM, PduA, Hypothetical protein, PduT
24	Gamma Proteobacteria	100	100	21	79	EutM, PduJ, detox protein, PduA, EutN, hypothetical protein
25	Delta Proteobacteria					
26	Epsilon Proteobacteria	2	0	0	0	-
27	Spirochaetes	57	14	0	14	EutM, carboxysome shell protein, hypothetical protein, microcompartment protein
28	Synergistetes	96	27			EutM, PduT, carboxysome shell protein
29	Tenericutes	17	4	0	4	EutM, Propanediol utilization protein
30	Thermodesulfobacteria	11	0	0	0	-
31	Thermotogae	3	0	0	0	-

Table SVI: List of *ccmK* gene complements in β cyanobacteria arranged according to the phylogenetic topology.

Clade	Name of organism	CcmK gene complements
A1	*S. elongatus PCC 6301*	*Syc0134 (ccml)-**syc0135 (ccmk1)**//syc1227 (ccmk4)-syc1228 (ccmk3)*
	S. elongatus PCC 7942	*Synpcc7942_1422 (ccml)-**Synpcc 7942_1421(ccmk**)//Synpcc7942_0284 (ccmk3)-Synpcc 7942_0285 (ccmk4)*
	Synechococcus sp JA-2-3B'a(2-13)	*CYB_1795 (ccml)-CYB_1796 (ccmk1)**-CYB_1797 (ccmk2)***
	Synechococcus sp JA-3-3Ab	*CYA_1613 (ccml)-CYA_1612 (ccmk2)-**CYA_1611 (ccmk1)***
	T. elongatus BP-1	*tll0945 (ccml)-tll0946 (ccmk1)-**tll0947 (ccmk2**)//tlr0954 (ccmk3)//tll 1596 (ccmk4)*
	G. violaceus PCC 7421	*gll2094 (ccml)-gll2095 (ccmk)-**gll2096 (ccmk)***
	T. erythreaum IMS 101	*Tery_3850 (ccml)-Tery_3851 (MCP)-**Tery_3852 (MCP)**//Tery_4328 (MCP)-Tery_4329 (MCP)*
	N. punctiforme STCC 29133	*Npun_F4293 (ccml)-Npun_F4292 (MCP)-**Npun_F4291 (MCP)**//Npun_F2745 (MCP)-Npun_2744 (MCP)*
	Anabaena PCC 7120	*all0866 (ccml)-all0867 (ccmk)-**all0868 (ccmk)**//alr0317 (ccmk)-alr0318 (ccmk)*
	A. variabilis ATCC 29413	*Ava_4470 (ccml)-Ava_4471 (MCP)-**Ava_4472 (MCP)**//Ava_4709 (MCP)-Ava_4710 (MCP)*
	Synechococcus sp PCC 7002	*SYNPCC7002_A1801 (ccml)-SYNPCC7002_A1802 (ccmk1)-**SYNPCC7002_A1803 (ccmk2)** //SYNPCC7002_A2612 (ccmk)-SYNPCC7002_A2613 (ccmk)*
	Synechocystis sp PCC 6803	*sll1030 (ccml)-sll1029 (ccmk1)-**sll1028 (ccmk2)** //slr1838 (ccmk3)-slr1839 (ccmk4)*
	Cyanothece sp PCC 7424	*PCC7424_1370 (ccml)-PCC7424_1371 (MCP)-PCC7424_1372 (MCP)-**PCC7424_1373 (MCP)** // PCC7424_2157 (MCP)-PCC7424_2158 (MCP)*
	Cyanothece sp ATCC 51142	*cce_4281 (ccml)-cce_4282 (ccmk1)-**cce_4283 (ccmk2)** //cce_2433 (ccmk4)-cce_2434 (ccmk3)*
	Cyanothece sp PCC 8801	*PCC8801_1598 (ccml)-PCC8801_1597 (MCP)-**PCC8801_1596 (MCP)** //PCC8801_1859 (MCP)-PCC8801_1860 (MCP)*
	A. marina MBIC 11017	*AM1_5382 (ccml)-AM1_5381 (ccmk)-**AM1_5380 (ccmk)** //AM1_0655 (ccmk)-AM1_0656 (ccmk)//AM1_3280 (ccmk)//AM1_5778 (ccmk)*
	Cyanothece sp PCC 7425	*Cyan7425_1616 (ccml)-Cyan7425_1617 (ccmk)-**Cyan7425_1618 (MCP)** //Cyan7425_2087 (MCP) // Cyan7425_2386 (MCP)*
	Synechocystis sp PCC 6803	*sll1030 (ccml)-**sll1029 (ccmk1)**-sll1028 (ccmk2) //slr1838 (ccmk3)-slr1839 (ccmk4)*
	A platensis NIES 39	*NIES39_K04810 (ccml)-**NIES39_K04820 (ccmk1)**-NIES39_K04830 (ccmk2) //NIES39_A03150 (ccmk4)-NIES39_A03160 (ccmk3)*
	A platensis NIES 39	*NIES39_K04810 (ccml)-NIES39_K04820 (ccmk1)-**NIES39_K04830 (ccmk2)** //NIES39_A03150 (ccmk4)-NIES39_A03160 (ccmk3)*
	Cyanothece sp PCC 7425	*Cyan7425_1616 (ccml)-**Cyan7425_1617 (ccmk)-**Cyan7425_1618 (MCP) //Cyan7425_2087 (MCP) // Cyan7425_2386 (MCP)*
	T. elongatus BP-1	*tll0945 (ccml)-**tll0946 (ccmk1)**-tll0947 (ccmk2)//tlr0954 (ccmk3)//tll 1596 (ccmk4)*
	G. violaceus PCC 7421	*gll2094 (ccml)-**gll2095 (ccmk)-**gll2096 (ccmk)*
	Synechococcus sp JA-2-3B'a(2-13)	*CYB_1795 (ccml)-**CYB_1796 (ccmk1)-**CYB_1797 (ccmk2)*
	Synechococcus sp JA-3-3Ab	*CYA_1613 (ccml)-**CYA_1612 (ccmk2)-**CYA_1611 (ccmk1)*
	M. aeruginosa NIES 843	*MAE47920 (ccml)-MAE47930 (ccmk1)-**MAE47940 (ccmk2)** // MAE55390 (ccmk3)-mae55400 (ccmk4)*
	Cyanothece sp ATCC 51142	*cce_4281 (ccml)-**cce_4282 (ccmk1)-**cce_4283 (ccmk2) //cce_2433 (ccmk4)-cce_2434 (ccmk3)*
	Synechococcus sp PCC 7002	*SYNPCC7002_A1801 (ccml)-**SYNPCC7002_A1802 (ccmk1)-**SYNPCC7002_A1803 (ccmk2) //SYNPCC7002_A2612 (ccmk)-SYNPCC7002_A2613 (ccmk)*
	Cyanothece sp PCC 8801	*PCC8801_1598 (ccml)-**PCC8801_1597 (MCP)-**PCC8801_1596 (MCP) //PCC8801_1859 (MCP)-PCC8801_1860 (MCP)*
	M. aeruginosa NIES 843	*MAE47920 (ccml)-**MAE47930 (ccmk1)-**MAE47940 (ccmk2) // MAE55390 (ccmk3)-mae55400 (ccmk4)*
	Cyanothece sp PCC 7424	*PCC7424_1370 (ccml)-**PCC7424_1371 (MCP)-**PCC7424_1372 (MCP) // PCC7424_2157 (MCP)-PCC7424_2158 (MCP)*
	Cyanothece sp PCC 7424	*PCC7424_1370 (ccml)-PCC7424_1371 (MCP)-**PCC7424_1372 (MCP)-**PCC7424_1373 (MCP) // PCC7424_2157 (MCP)-PCC7424_2158 (MCP)*
	T. erythreaum IMS 101	*Tery_3850 (ccml)-**Tery_3851 (MCP)-**Tery_3852 (MCP)//Tery_4328 (MCP)-Tery_4329 (MCP)*
	A. marina MBIC 11017	*AM1_5382 (ccml)-**AM1_5381 (ccmk)-**AM1_5380 (ccmk) //AM1_0655 (ccmk)-AM1_0656 (ccmk)//AM1_3280 (ccmk)//AM1_5778 (ccmk)*
	Anabaena PCC 7120	*all0866 (ccml)-**all0867 (ccmk)-**all0868 (ccmk)//alr0317 (ccmk)-alr0318 (ccmk)*
	A. variabilis ATCC 29413	*Ava_4470 (ccml)-**Ava_4471 (MCP)**-Ava_4472 (MCP)//Ava_4709 (MCP)-Ava_4710 (MCP)*
	N. punctiforme STCC 29133	*Npun_F4293 (ccml)-**Npun_F4292 (MCP)-**Npun_F4291 (MCP)//Npun_F2745 (MCP)-Npun_2744 (MCP)*

A2	*A platensis NIES 39*	*NIES39_K04810 (ccml)-NIES39_K04820 (ccmk1)-NIES39_K04830 (ccmk2) //NIES39_A03150 (ccmk4)-**NIES39_A03160 (ccmk3)***
	M. aeruginosa NIES 843	*MAE47920 (ccml)-MAE47930 (ccmk1)-MAE47940 (ccmk2) // MAE55390 (ccmk3)-**MAE55400 (ccmk4)***
	Synechocystis sp PCC 6803	*sll1030 (ccml)-sll1029 (ccmk1)-sll1028 (ccmk2) //**slr1838 (ccmk3)**-slr1839 (ccmk4)*
	Cyanothece sp PCC 7424	*PCC7424_1370 (ccml)-PCC7424_1371 (MCP)-PCC7424_1372 (MCP)-PCC7424_1373 (MCP) // PCC7424_2157 (MCP)-**PCC7424_2158 (MCP)***
	Cyanothece sp ATCC 51142	*cce_4281 (ccml)-cce_4282 (ccmk1)-cce_4283 (ccmk2) //cce_2433 (ccmk4)-**cce_2434 (ccmk3)***
	Cyanothece sp PCC 8801	*PCC8801_1598 (ccml)-PCC8801_1597 (MCP)-PCC8801_1596 (MCP) //**PCC8801_1859 (MCP)**-PCC8801_1860 (MCP)*
	Synechococcus sp PCC 7002	*SYNPCC7002_A1801 (ccml)-SYNPCC7002_A1802 (ccmk1)-SYNPCC7002_A1803 (ccmk2) //**SYN-PCC7002_A2612 (ccmk)**-SYNPCC7002_A2613 (ccmk)*
	T. erythreaum IMS 101	*Tery_3850 (ccml)-Tery_3851 (MCP)-Tery_3852 (MCP)//**Tery_4328 (MCP)**-Tery_4329 (MCP)*
	Anabaena PCC 7120	*all0866 (ccml)-all0867 (ccmk)-all0868 (ccmk)//**alr0317 (ccmk)-**alr0318 (ccmk)*
	A. variabilis ATCC 29413	*Ava_4470 (ccml)-Ava_4471 (MCP)-Ava_4472 (MCP)//**Ava_4709 (MCP)-**Ava_4710 (MCP)*
	S. elongatus PCC 6301	*Syc0134 (ccml)-syc0135 (ccmk1)//syc1227 (ccmk4)-**syc1228 (ccmk3)***
	S. elongatus PCC 7942	*Synpcc7942_1422 (ccml)-Synpcc 7942_1421(ccmk)//**Synpcc7942_0284 (ccmk3)-**Synpcc 7942_0285 (ccmk4)*
	A. marina MBIC 11017	*AM1_5382 (ccml)-AM1_5381 (ccmk)-AM1_5380 (ccmk) //AM1_0655 (ccmk)-**AM1_0656 (ccmk)//** AM1_3280 (ccmk)//AM1_5778 (ccmk)*
	Anabaena PCC 7120	*all0866 (ccml)-all0867 (ccmk)-all0868 (ccmk)//alr0317 (ccmk)-**alr0318 (ccmk)***
	A. variabilis ATCC 29413	*Ava_4470 (ccml)-Ava_4471 (MCP)-Ava_4472 (MCP)//Ava_4709 (MCP)-**Ava_4710 (MCP)***
	Cyanothece sp ATCC 51142	*cce_4281 (ccml)-cce_4282 (ccmk1)-cce_4283 (ccmk2) //**cce_2433 (ccmk4)**-cce_2434 (ccmk3)*
	Cyanothece sp PCC 7424	*PCC7424_1370 (ccml)-PCC7424_1371 (MCP)-PCC7424_1372 (MCP)-PCC7424_1373 (MCP) // **PCC7424_2157 (MCP)**-PCC7424_2158 (MCP)*
	M. aeruginosa NIES 843	*MAE47920 (ccml)-MAE47930 (ccmk1)-MAE47940 (ccmk2) // **MAE55390 (ccmk3)-**MAE55400 (ccmk4)*
	Cyanothece sp PCC 8801	*PCC8801_1598 (ccml)-PCC8801_1597 (MCP)-PCC8801_1596 (MCP) //PCC8801_1859 (MCP)-**PCC8801_1860 (MCP)***
	Synechocystis sp PCC 6803	*sll1030 (ccml)-sll1029 (ccmk1)-sll1028 (ccmk2) //slr1838 (ccmk3)-**slr1839 (ccmk4)***
	Synechococcus sp PCC 7002	*SYNPCC7002_A1801 (ccml)-SYNPCC7002_A1802 (ccmk1)-SYNPCC7002_A1803 (ccmk2) //SYN-PCC7002_A2612 (ccmk)-**SYNPCC7002_A2613 (ccmk)***
	A. marina MBIC 11017	*AM1_5382 (ccml)-AM1_5381 (ccmk)-AM1_5380 (ccmk) //AM1_0655 (ccmk)-AM1_0656 (ccmk)// AM1_3280 (ccmk)//**AM1_5778 (ccmk)***
	A platensis NIES 39	*NIES39_K04810 (ccml)-NIES39_K04820 (ccmk1)-NIES39_K04830 (ccmk2) //**NIES39_A03150 (ccmk4)-**NIES39_A03160 (ccmk3)*
	T. erythreaum IMS 101	*Tery_3850 (ccml)-Tery_3851 (MCP)-Tery_3852 (MCP)//Tery_4328 (MCP)-**Tery_4329 (MCP)***
	A. marina MBIC 11017	*AM1_5382 (ccml)-AM1_5381 (ccmk)-AM1_5380 (ccmk) //**AM1_0655 (ccmk)**-AM1_0656 (ccmk)// AM1_3280 (ccmk)//AM1_5778 (ccmk)*
	A. marina MBIC 11017	*AM1_5382 (ccml)-AM1_5381 (ccmk)-AM1_5380 (ccmk) //AM1_0655 (ccmk)-AM1_0656 (ccmk)// **AM1_3280 (ccmk)**//AM1_5778 (ccmk)*
	S. elongatus PCC 6301	*Syc0134 (ccml)-syc0135 (ccmk1)//**syc1227 (ccmk4)-**syc1228 (ccmk3)*
	S. elongatus PCC 7942	*Synpcc7942_1422 (ccml)-Synpcc 7942_1421(ccmk)//Synpcc7942_0284 (ccmk3)-**Synpcc 7942_0285 (ccmk4)***

B	*T. elongatus BP-1*	*tll0945 (ccml)-tll0946 (ccmk1)-tll0947 (ccmk2)//tlr0954 (ccmk3)//**tll 1596 (ccmk4)***
	Cyanothece sp PCC 7425	*Cyan7425_1616 (ccml)-Cyan7425_1617 (ccmk)-Cyan7425_1618 (MCP) //Cyan7425_2087 (MCP) // **Cyan7425_2386 (MCP)***
	Cyanothece sp PCC 7425	*Cyan7425_1616 (ccml)-Cyan7425_1617 (ccmk)-Cyan7425_1618 (MCP) //**Cyan7425_2087 (MCP)** // Cyan7425_2386 (MCP)*
	T. elongatus BP-1	*tll0945 (ccml)-tll0946 (ccmk1)-tll0947 (ccmk2)//**tlr0954 (ccmk3**)//tll 1596 (ccmk4)*

Bibliography

1. Jensen T and Bowen C. "Organization of the centroplasm in *Nostoc punctiforme"*. *Proceedings of the Iowa Academy of Science* 68 (1961): 89-96.

2. Kerfeld CA., *et al*. "Bacterial microcompartments". *Annual Review of Microbiology* 64 (2010): 391-408.

3. Penrod J T and Roth JR. "Conserving a volatile metabolite: a role for carboxysome - like organelles in *Salmonella enteric"*. *Journal of Bacteriology* 188 (2006): 2865-2874.

4. Sampson EM and Bobik T A. "Microcompartments for B-12 dependent 1,2-propanediol degradation provide protection from DNA and cellular damage by a reactive metabolic intermediate*"*. *Journal of Bacteriology* 190 (2008): 2966-2971.

5. Huseby D L and Roth JR. "Evidence that a metabolic microcompartment contains and recycles private cofactor pools". *Journal of Bacteriology* 195 (2013): 2864-2879.

6. Chen P., *et al*. "The control region of the pdu/cob regulon in *Salmonella typhimurium"*. *Journal of Bacteriology* 176 (1994): 5474-5482.

7. Kofoid E., *et al*. "The 17-gene ethanolamine (eut) operon of *Salmonella typhimurium* encodes five homologues of carboxysome shell proteins*"*. *Journal of Bacteriology* 181 (1999): 5317-5329.

8. Badger MR., *et al*. "The diversity and coevolution of rubisco, plastids, pyrenoids and chloroplast based CO_2 concentrating mechanisms in algae". *Canadian Journal of Botany* 76 (1998): 1052-1071.

9. Kaplan A and Reinhold L. "CO_2-concentrating mechanisms in photosynthetic microorganisms". *Annual Review of Plant Physiology and Plant Molecular Biology* 50 (1999): 539-570.

10. Badger M R., *et al*. "Evolution and diversity of CO_2 concentrating mechanisms in cyanobacteria". *Functional Plant Biology* 29 (2002): 161-173.

11. Cannon GC., *et al*. "Carboxysome genomics: a status report". *Functional Plant Biology* 29 (2002): 175-182.

12. Tanaka S., *et al*. "Atomic level models of the bacterial carboxysome shell". *Science* 319 (2008): 1083-1086.

13. Price G D., *et al*. "Advances in understanding the cyanobacterial CO_2-concentrating mechanism (CCM): Functional components, Ci transporters, diversity, genetic regulation and prospects for engineering into plants". *Journal of Experimental Botany* (2007).

14. Alber BE and Ferry JG. "A carbonic anhydrase from the archaeon *Methanosarcina thermophila*". *Proceedings of the National Academy of Sciences of the United States of America* 91 (2013): 6909-6913.

15. Pena KL., *et al*. "Structural basis of the oxidative activation of the carboxysomal gamma carbonic anhydrases, CcmM". *Proceedings of the National Academy of Sciences of the United States of America* 107 (2010): 2455-2460.

16. Espie GS and Kimber M S. "Carboxysomes: cyanobacterial Rubisco comes in small packages". *Photosynth. Research* 109 (2011): 7-20.

17. Price GD., *et al*. "Analysis of a genomic DNA region from the cyanobacterium *Synechococcus sp* str PCC 7942*"*. *Journal of Bacteriology* 175 (1993): 2871-2879.

18. Long B M., *et al*. "Analysis of carboxysomes from *Synechococcus* PCC 7942 reveals multiple RuBisCO complexes with carboxysomal proteins CcmM and CcaA". *Journal of Biological Chemistry* 282 (2007): 29323-29335.

19. Kinney J N., *et al*. "Elucidating essential role of conserved carboxysomal protein CcmN reveals common feature of bacterial microcompartment assembly". *Journal of Biological Chemistry* 287 (2012): 17729-17736.

20. Havemann GD and Bobik T A. "Protein content of polyhedral organelles involved in coenzyme B12-dependent degradation of 1,2-propanediol in *Salmonella enetrics serovar typhimurium* LT2*"*. *Journal of Bacteriology* 185 (2003): 5086-5095.

21. Rae BD., *et al*. "Functions, Compositions and Evolution of the two types of carboxysomes: Polyhedral Microcompartments that facilitate CO_2 fixation in Cyanobacteria and some Proteobacteria". *Microbiology and Molecular Biology Reviews* 77 (2013): 357-379.

22. Menon B B., *et al*. "*Halothiobacillus neopolitanus* carboxysomes sequester heterologous and chimeric rubisco species". *pLOS One* 3 (2008): e3570.

23. Fan C G., *et al*. "Interactions between the termini of lumen enzymes and shell proteins mediate enzyme encapsulation into bacterial microcompartments". *Proceedings of the National Academy of Sciences of the United States of America* 109 (2012): 14995-15000.

24. McClelland M., *et al*. "Complete genome sequence of *Salmonella enteric serovar typhimurium* LT2". *Nature* 413 (2001): 852-856.

25. Price G D., *et al*. "The cyanobacterial CCM as a source of genes for improving photosynthetic CO_2 fixation in crop species". *Journal of Experimental Botany* 64 (2013): 753-768.

26. Corchero J L and Cedano J. "Self-assembling, protein-based intracellular bacterial organelles: emerging vehicles for encapsulating, targeting and delivering therapeutical cargoes". *Microbial Cell Factories* 10 (2011): 92.

27. Gaurav V., *et al*. "Sequence Matrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information". *Cladistics* 27 (2010): 171-180.

28. Thompson JD., *et al*. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". *Nucleic Acids Research* 22 (1994): 4673-4680.

29. Sela I., *et al*. "GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters". *Nucleic Acids Research* (2015): W7-W14.

30. Landan G and Graur D. "Local reliability measures from sets of co-optimal multiple sequence alignments". *Pacific Symposium on Biocomputing* 13 (2008): 15-24.

31. Tamura K., *et al*. "MEGA6: Molecular Evolutionary Genetics Analysis version 6.0". *Molecular Biology and Evolution* 30 (2013): 2725-2729.

32. Nei M and Kumar S. "Molecular Evolution and Phylogenetics". Oxford University Press, New York (2000).

33. Altschul S F., *et al*. "Basic local alignment search tool". *Journal of Molecular Biology* 215 (2000): 403-410.

34. Altschul S F., *et al*. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Nucleic Acids Research* 25 (1997): 3389-3402.

35. Geer LY., *et al*. "CDART: Protein Homology by Domain Architecture". *Genome Research* 12 (2002): 1619-1623.

36. Tajima F and Nei M. "Estimation of evolutionary distance between nucleotide sequences". *Molecular Biology and Evolution* 1 (1984): 269-285.

37. Jukes T H and Cantor C R. "Evolution of protein molecules". In Munro HN, editor, *Mammalian Protein Metabolism*, Academic Press, New York (1969).

38. Nelissen B., *et al*. "An early origin of plastids within the cyanobacterial divergence is suggested by evolutionary trees based on complete 16S rRNA sequences". *Molecular Biology and Evolution* **12** (1995): 1166-1173.

39. Memon D., *et al*. "A global analysis of adaptive evolution of operons in cyanobacteria". *Antonie. Van. Leeuwenhoek*. 103 (2013): 331-346.

40. Gupta R S and Mathews D W. "Signature proteins for the major clades of Cyanobacteria". *BMC Evolution and Biology* 10 (2010): 24.

41. Dvořák P., *et al*. "Morphological and molecular studies of Neosynechococcus sphagnicola, gen. et sp. nov. (Cyanobacteria, Synechococcales)". *Phytotaxa* 170.1 (2014): 024-034.

42. Soo RM., *et al*. "An expanded genomic representation of the phylum cyanobacteria". *Genome Biology and Evolution* 6 (2014): 1031-1045.

43. Di Rienzi SC., *et al*. "The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria". *Elife* 2 (2013): e01102.

44. Rippka R., *et al*. "A cyanobacterium which lacks thylakoids". *Archives of Microbiology* 100 (1974): 419-436.

45. Guglielmi G., *et al*. "The structure of *Gloeobacter violaceus* and its phycobilisomes". *Archives of Microbiology* 129 (1981): 181-189.

46. Salstam E and Campbell D. "Membrane lipid composition of the unusual cyanobacterium *Gloeobacter violaceus PCC 7421*, which lacks sulfoquinovosyl diacylglycerol". *Archives of Microbiology* 166 (1996): 132-135.

47. Frank S., *et al*. "Bacterial microcompartments moving into a synthetic biological world". *Journal of Biotechnology* 163 (2013): 273-279.

48. Vogel C., *et al*. "Structure, function and evolution of multidomain proteins". *Current Opinion in Structural Biology* 14 (2004): 208-216.

49. Enright A J., *et al*. "Protein interaction maps for complete genomes based on gene fusion events". *Nature* 402 (1999): 86-90.

50. Marcotte E M., *et al*. "A combined algorithm for genome-wide prediction of protein function". *Nature* 402 (1999): 83-86.

51. Snel B., *et al*. "Genome evolution: Gene fusion versus gene fission". *Trends in Genetics* 16 (2000): 9-11.

52. Yanai I., *et al*. "Evolution of gene fusions: horizontal transfer versus independent events". *Genome Biology* 3 (2002): research0024.

53. Abdul-Rahman F., *et al*. "The distribution of polyhedral bacterial microcompartments suggests frequent horizontal transfer and operon reassembly". *Journal of Phylogenetics and Evolutionary Biology* 1 (2013): 1-7.

54. Price MN., *et al*. "The life-cycle of operons". *PLoS Genetics* 2 (2006): e96.

55. Kumar S. "Molecular clocks: four decades of evolution". *Nature Reviews Genetics* 6 (2005): 654-662.

56. Csaba Pál., *et al*. "An integrated view of protein evolution". *Nature Reviews Genetics* 7 (2006): 337-348.