Identification and Structural Analysis of a Novel Carboxysome Shell Protein with Implications for Metabolite Transport

Identification and Structural Analysis of a Novel Carboxysome Shell Protein with Implications for Metabolite Transport

Michael G. Klein, [1] Peter Zwart, [2] Sarah C. Bagby, [3] Fei Cai, [4] Sallie W. Chisholm, [3] Sabine Heinhorst, [4] Gordon C. Cannon[4] and Cheryl A. Kerfeld[1,5]

[1]US Department of Energy - Joint Genome Institute, Walnut Creek, CA 94598

[2]US Department of Energy – Lawrence Berkeley National Laboratory, Berkeley CA 94720 USA

[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139 USA

[4]Department of Chemistry and Biochemistry, The University of Southern Mississippi, Hattiesburg, MS 39406 USA

[5]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720 USA

Corresponding author: ckereld@lbl.gov Phone: (925) 296-5691 FAX: (925) 296-5752

Running Title: Structure of a Novel Carboxysome Shell Protein

Summary

Bacterial microcompartments are polyhedral bodies composed entirely of protein that function as organelles in bacteria; they promote subcellular processes by encapsulating and co-localizing targeted enzymes with their substrates. The best-characterized bacterial microcompartment is the carboxysome, a central part of the carbon concentrating mechanism that greatly enhances carbon fixation in cyanobacteria and some chemoautotrophs. Here we report the first structural insights into the carboxysome of *Prochlorococcus*, the numerically dominant cyanobacterium in the world's oligotrophic oceans. Bioinformatic methods, substantiated by analysis of gene expression data, were used to identify a new carboxysome shell component, CsoS1D, in the genome of *Prochlorococcus* strain MED4; orthologs were subsequently found in all cyanobacteria. Two independent crystal structures of *Prochlorococcus* MED4 CsoS1D reveal three features not seen in any BMC-domain protein structure solved to date. First, CsoS1D is composed of a fused pair of bacterial microcompartment domains. Second, this double-domain protein trimerizes to form a novel pseudohexameric building block for incorporation into the carboxysome shell, and the trimers further dimerize, forming a two-tiered shell building block. Third, and most strikingly, the large pore formed at the 3-fold axis of symmetry appears to be gated. Each dimer of trimers contains one trimer with an open pore and one whose pore is obstructed due to sidechain conformations of two residues that are invariant among all CsoS1D orthologs. This is the first evidence of the potential for gated transport across the carboxysome shell and reveals

a new type of building block for bacterial microcompartment shells.

Keywords: Carboxysome, bacterial microcompartment, gated transport, *Prochlorococcus*, carbon concentrating mechanism

Introduction

The cyanobacterium *Prochlorococcus* is the numerically dominant photosynthetic organism in much of the world's oceans. As such it is responsible for a significant fraction of biological carbon fixation in these systems. Like all cyanobacteria, *Prochlorococcus* is presumed to rely on a carbon concentrating mechanism to facilitate CO_2 capture and fixation [1; 2; 3]. Central to this mechanism is a microcompartment in the cell, the carboxysome (Figure 1A) which is the site of CO_2 concentration and fixation.

Carboxysomes are the foremost example of the polyhedral subcellular inclusions that have been termed bacterial microcompartments (Figure 1A), self-assembling protein shells that encapsulate enzymes and other functionally related proteins. In addition to carboxysomes, two other types of bacterial microcompartments are relatively well characterized (reviewed in [4; 5; 6]); they function in propane-diol utilization (encoded by the pdu operon) and ethanolamine utilization (encoded by the eut operon) in heterotrophic bacteria.

Carboxysomes have been observed in all cyanobacteria and many chemoautotrophs. They contain the key enzyme of carbon fixation, ribulose1,5-bisphosphate carboxylase/oxygenase (RubisCO) [7], which has a relatively high K_M and as such is not a kinetically efficient enzyme. Its efficiency is further compromised because it catalyzes the apparently unproductive fixation of O_2 in competition with CO_2. Cyanobacteria compensate for these deficiencies by encapsulating RubisCO within the carboxysome. In *Prochlorococcus*, apparently all

of the cellular RubisCO is located in the carboxysomes [8]; this is likely to be particularly important for this genus as the K_{CO2} value of its RubisCO (750 umol L^{-1}; [3]) is the highest known among cyanobacterial RubisCOs. Carbonic anhydrase (CA) activity, which converts the abundant intracellular bicarbonate to CO_2 is tightly associated with the carboxysome shell [9]. RubisCO derives a catalytic advantage not only from its proximity to CA activity in the carboxysome but also because the carboxysome shell impedes diffusive loss of HCO_3^- and CO_2 [10; 11].

There are two types of carboxysome, alpha and beta, distinguished by the form of RubisCO they encapsulate. Alpha (or cso-type) carboxysomes, containing Form 1A RubisCO, are found in the dominant oceanic picocyanobacteria, *Prochlorococcus* and *Synechococcus*, and in chemoautotrophs, such as *Halothiobacillus neapolitanus*; beta carboxysomes contain Form 1B RubisCO and are broadly distributed among cyanobacteria, occurring in *Trichodesmium, Thermosynechococcus, Synechocystis, Gloeobacter*, and *Crocosphaera*, and others [12; 13]. The two types of carboxysomes have very different genomic architecture; the alpha carboxysome is encoded by an operon (Figure 1B) that includes genes for the large and small subunits of RubisCO, the carbonic anhydrase CsoSCA (formerly CsoS3), CsoS2, a protein of unknown function and two small shell proteins that form the vertices of the carboxysome [14]. In contrast, the components of the beta carboxysome are encoded by several small, scattered gene clusters (the *ccm* genes;[4; 12; 15]).

This disparity at the operon level belies the similarity of the proteins that constitute alpha and beta carboxysome shells. The most common structural motif in

carboxysome shell proteins is the bacterial microcompartment (BMC) domain (Pfam 00936), which occurs in multiple paralogous proteins in both alpha and beta carboxysomes. Among these paralogs, CcmK1, 2 and 4 from *Synechocystis* sp PCC 6803 and CsoS1A from *H. neapolitanus* have been structurally characterized [14; 16; 17], revealing that they form cyclic hexamers. Pores formed at the six-fold symmetry axis of individual hexamers are thought to permit metabolite flow across the carboxysome shell. Hexamers of CcmK1, CcmK2 and CsoS1A tend to tile into uniformly oriented layers that presumably constitute the facets of the carboxysome shell.

Interestingly, the alpha carboxysome operons of cyanobacteria and chemoautotrophs differ in the number and position of BMC domains each encodes (Figure 1B). Whereas all chemoautotrophs contain 2-3 highly homologous single-BMC polypeptides (CsoS1A, B and C), at the end of the operon, in the *Prochlorococcus* species these are absent; instead there is a single, ~100 amino acid BMC protein gene (*csoS1*) preceding the gene for the large subunit of RubisCO. Not only the difference in position, but the difference in number of BMC domain-containing proteins in the genome is notable, for in all other organisms expected to form carboxysomes or related bacterial microcompartments, there are always at least two BMC domains present (Kerfeld et al., unpublished); the reason for this redundancy is at present, unknown. However, closer inspection of the high light-adapted *Prochlorococcus* genomes reveals that this shortfall is superficial. Upstream of *csoS1* and separated from it by a gene encoded on the opposite strand,

there is a gene for a 256 amino acid protein predicted to contain a BMC domain in its C-terminus. Both the sequence of this protein (PMM0547 in *Prochlorococcus* MED4; Figure 1B) and its genomic position relative to the other carboxysome genes are conserved in all sequenced *Prochlorococcus* strains, and indeed in all sequenced alpha carboxysome-containing cyanobacteria.

Here we report two independent crystal structures of PMM0547 at 2.2 and 2.3Å resolution, and a model for its interaction with CsoS1 in the *Prochlorococcus* carboxysome shell. The structural models and several complementary lines of evidence suggest that PMM0547 is a carboxysome shell protein, therefore we have provisionally dubbed it CsoS1D.

Results

Identification of a New Cyanobacterial Carboxysome Shell Protein

The gene for CsoS1D was detected by scanning the *Prochlorococcus* MED4 genome for bacterial microcompartment (BMC) domains. Subsequently, a search of the Pfam database with the *Prochlorococcus* MED4 CsoS1D amino acid sequence detected a single BMC domain in residues 190-256 and no other recognizeable domains. Putative orthologs (defined as bidirectional best BLASTP hits) to CsoS1D are found in all cyanobacteria, suggesting that this protein could be a part of both alpha and beta carboxysome shells. The primary structure is highly conserved (~80% identical) among the cyanobacteria that form alpha carboxysomes. Orthologs

found in the genomes of cyanobacteria that form beta carboxysomes, such as the product of *slr0169* in *Synechocystis* sp PCC6803, lack the first 50 residues and are ~ 45% identical to residues 52-256 of CsoS1D. In addition, orthologs to CsoS1D are also found in several other organisms that do not form carboxysomes, but based on the presence of BMC-domain containing proteins, do appear to have the potential to form bacterial microcompartments of unknown function.

The alpha carboxysome genes are arranged similarly in the 22 genomes of marine *Synechococcus* and *Prochlorococcus* species sequenced to date, with *csoS1D* upstream of and separated from the carboxysome operon by one gene that is transcribed from the opposite strand (Figure 1B). The gene that separates *csoS1D* from the cluster of carboxysome genes, annotated as encoding a Ham1 protein (PMM0548 in *Prochlorococcus* MED4), is widespread among bacterial genomes but is of unknown function. To determine the likelihood that *csoS1D* is a component of the *Prochlorococcus* carboxysome, we analyzed the RNA expression pattern of *csoS1D* and its neighbors in axenic *Prochlorococcus* MED4 cells over the diel cycle, using data collected for a study of the entire *Prochlorococcus* transcriptome in synchronously dividing cells entrained to a 24-h light-dark cycle (Zinser *et al.*, manuscript submitted). Like the genes in the carboxysome operon (*csoS1-csoS4B*, PMM0549-PMM0553; Figure 1B), *csoS1D* (PMM0547) is expressed with 24-h periodicity with peak expression occurring at sunrise. We subjected all expressed, cycling genes (a total of 1405; Zinser *et al.*, manuscript submitted) to "fuzzy" clustering to examine more closely the association of *csoS1D* expression with

expression of *csoS1-csoSCA* (PMM0549-PMM0553), known carboxysome-related genes. Unlike "hard" clustering, which assigns each gene to exactly one cluster, "fuzzy" clustering determines the fractional membership of each gene in each cluster, permitting relationships between clusters that share genes to be recognized [18]. Following the model of Zinser et al., (manuscript submitted), our initial analysis, used 16 clusters and "fuzzification" parameter $m = 1.25$. In 100 out of 100 runs, *csoS1, cbbL, cbbS,* and *csoS2* clustered together, each gene with very high fractional membership in this cluster (mean $\mu = 0.991 \pm 0.004$, 0.990 ± 0.006, 0.957 ± 0.031, 0.913 ± 0.058, respectively). To check whether the association of *csoS1D* with the carboxysome-containing cluster was robust, we repeated the clustering analysis with m ranging from 1.15 ("harder") to 1.35 ("fuzzier") and the number of clusters ranging from 10 to 18, performing 100 runs with each pair of input parameters. Cluster stability began to diminish at $m \geq 1.30$ (data not shown); nonetheless, across the entire range, *csoS1D* was more strongly clustered with *csoS1-csoS2* than with genes in any other cluster (Figure S1). Indeed, *csoS1D* was more stably clustered with *csoS1-csoS2* than were the known carboxysome components *csoSCA* and *csoS4A* and *B* (formerly known as *orfA* and *orfB*); at 16 clusters and $m = 1.25$, for instance, these shell proteins have mean carboxysome-cluster membership of just 0.262 ± 0.063 and 0.308 ± 0.095, respectively (Figures S1, S2)

We also examined the published expression patterns of the *csoS1D* ortholog *slr0169*, found in the genome of the beta carboxysome-forming cyanobacterium *Synechocystis* sp PCC6803. In these organisms, the carboxysome

genes (*sll1028–1032* (*ccmK2, ccmK1, ccmL, ccmM, ccmN*) *slr0009*, *slr0011*, and *slr0012* (RubisCO-related genes), *slr0436* (*ccmO*), *slr1838–1839* (*ccmK3, ccmK4*), and *slr1347*(*ccaA*) are scattered throughout the genome. Eisenhut *et al.* [19] examined the whole-genome transcriptional response of *Synechocystis* sp. PCC 6803 to long-term inorganic carbon limitation, a stress condition that is expected to affect expression of carboxysome-related genes. In this experiment, 24 h after WT 6803 cells were shifted from 5% CO_2 to 0.03% CO_2, expression of *slr0169* was found to be significantly (>2-fold) downregulated, in tandem with several known carboxysome genes: *ccmK2*, *ccmK3*, *ccmK4*, and *ccmLMN*. Expression levels of the remaining carboxysome genes (*ccmK1, rbcLXS*, and *ccaA*) dropped by a factor of 1.2–1.8, or remained unchanged (*ccmO*). Under these conditions, then, expression of *slr0169* has more in common with expression of most genes encoding carboxysome shell proteins than have genes as central to carboxysome function as *rbcLS*. The same phenomenon is observed in the light-stress study of Singh *et al*. [20].

The Structure of CsoS1D: A Tandem BMC Domain Protein

CsoS1D was expressed and crystallized in two distinct crystal forms. The orthorhombic ($P2_12_12_1$) crystal form was solved by multiwavelength anomalous dispersion (MAD) methods using 2.3 Å diffraction data from a crystal prepared with selenomethionine substituted protein. This crystal form contains six molecules in the crystallographic asymmetric unit. The structure refined to final R and Rfree values of 18.5 and 21.6 %, respectively (Table S1, Figure S3). The second crystal

form (R3, rhombohedral) was subsequently solved by molecular replacement using the monomer from the orthorhombic crystal form structure as a search model. Although twinned, the rhombohedral crystal form refined to 2.2 Å with 20.4% and 25.9% for R and Rfree values, respectively. The twin fraction for the rhombohedral form refined to 0.45. For both crystal forms, the amino acid sequence was readily modeled with the exception of the N-terminus (~50 residues) of the protein which was absent from the electron density maps in both crystal forms. The stereochemistry of the final refined model in both crystal forms is excellent with 98.7 % and 94.3% of the residues in the most favored regions of the Ramachandran plot for the orthorhombic and rhombohedral crystal forms, respectively and no residues in the disallowed regions.

Carboxysome shell proteins are characterized by a small (~80 amino acid) domain known as the BMC domain (Pfam00936). To date, structures have been determined for BMC-domain containing proteins ranging from 94 to 116 amino acids in length; each containing a single-BMC domain. Sequence analysis of CsoS1D predicted that its C-terminus (residues 190-256) contained a BMC domain, but the N-terminal 150 amino acids lack sequence homology to any known domains. The crystal structures revealed that residues 50-150 in CsoS1D also contain a BMC domain; thus CsoS1D is composed of a tandem pair of BMC domains (referred to as N-and C-BMC; Figure 2). The N- and C-BMC share less than 18% sequence identity (Figure 2B) yet their alpha-carbon backbones superimpose with an RMS deviation of 1.27 Å over 95 residues (Figure 2B).

Overall, the fold of the N-BMC and C-BMC of CsoS1D are very similar to those of the single-BMC domain proteins that have been structurally characterized [14; 16; 17; 21]. The fold consists of three alpha helices (designated A, B and C) and four beta strands (designated β1, β2, β3, β4; Figures 2B, 2C). The C-BMC overlaps with CsoS1A, a shell protein from the alpha carboxysome of the chemoautotroph *H. neapolitanus*, over 81 alpha carbon atoms with a RMS deviation of 1.43Å. The N-BMC superimposes on CsoS1A with a RMS deviation of 1.67Å over 79 alpha carbon atoms. This consistency in domain fold is despite a difference in connectivity; both BMC domains of CsoS1D contain the permutation in secondary structure recently noted in PduU, a single-BMC domain shell protein of the bacterial microcompartment involved in 1,2-propanediol utilization [21]. Like PduU, the N-terminal ~34 residues of each BMC domain in CsoS1D contribute a beta-strand and a short alpha helix (β1 and α–helix A), to the overall fold of the domain; in the typical single-domain BMC proteins these secondary structure elements are instead contributed by the C-terminus.

The CsoS1D Trimer Forms a Pseudohexamer that Contains a Gated Pore at the Three-fold Symmetry Axis

The asymmetric unit of the orthorhombic crystal form of CsoS1D contains six molecules tightly associated with one another, to form a dimer of trimers that constitutes a hexamer (Figures 2D, 2E). In the rhombohedral form, the two CsoS1D molecules in the asymmetric unit form a dimer, with the crystallographic three-fold axis generating a similar dimer of trimers. The interactions between the trimers to form a dimer of trimers/hexamer are discussed below.

Because each CsoS1D monomer is composed of a tandem pair of BMC domains, CsoS1D trimers form pseudo-hexamers (Figure 2A) that recapitulate the single-BMC domain carboxysome shell protein hexamers that are thought to constitute the facets of the carboxysome shell [14; 16; 17]. One face of the trimer appears slightly broader than the other and more closely resembles a regular hexagon (Figures 2, 3). The two sides of the trimer also have distinctive electrostatic surfaces with the broader, more hexagonal face being less polar (Figure 3).

In addition to these differences between the two faces of each trimer, we observed a striking difference between the two trimers in each CsoS1D hexamer. In the single-BMC domain carboxysome shell proteins that have been structurally characterized, the hexamer's six-fold axis of symmetry is lined by conserved, positively charged sidechains. These pores are proposed to be conduits for diffusion of metabolites such as HCO_3^-, ribulose-1,5-bisphosphate (RuBP) and 3-phosphoglycerate (PGA) across the carboxysome shell. In each trimer of CsoS1D, the sidechains of the invariant residue Arg121 converge at the three-fold axis of symmetry; however the sidechains of Arg121 and of Glu120, which is also absolutely conserved among CsoS1D orthologs, adopt dramatically different conformations in the two trimers (Figures 4A, 4B; Supplementary animation). In one trimer, Arg121 and Glu120 sidechains are oriented to leave a pore of ~14Å at the three-fold axis of symmetry (Figures 2A, 3C, 3D Figure 4); we refer to this trimer as the open trimer. In contrast, in the other "closed" trimer, each CsoS1D monomer's Glu120 sidechain forms a salt bridge with Arg121 of the adjacent monomer, effectively closing the pore

(Figure 4A). Strikingly, in both crystal forms, the CsoS1D hexamer is a dimer of one open and one closed trimer. The open and closed trimers in the orthorhombic crystal superimpose with an RMS deviation of 1.43 Å over 599 alpha carbon atoms.

To quantify the differences between the monomers in the open and closed trimers, the six monomers in the asymmetric unit of the orthorhombic crystal forms were compared to one another (Table S2). Two classes of conformation of the monomer emerged, each distinctive of either the open or closed trimer. The three monomers in the open trimer (class 1) share the same conformation, as they superimpose with a RMS deviation ranging from 0.14 to 0.21 Å. The three monomers in the closed trimer (class 2) also share essentially the same conformation, as they superimpose with RMS deviations ranging from 0.19 to 0.27 Å. When compared to each other, however, the two classes of monomers appear to be significantly different, as they superimpose with RMS deviations ranging from 0.95 to 1.1 Å over 204 alpha carbon atoms.

This distinct conformational difference between the monomers in the two types of trimers is also observed in the rhombohedral crystal form. The asymmetric unit of the rhombohedral crystal form contains only two molecules of CsoS1D; one from the open and one from the closed trimer. They superimpose with an RMS deviation of 1.03 Å over 201 alpha carbon atoms. Further analysis (data not shown), indicates that the open trimer in the R3 form contains monomer similar to the class 1 and the closed trimer contains monomers similar to class 2. Although the crystal forms are clearly distinct and grown from very different crystallization conditions, CsoS1D

adopts the same two specific conformations that make up the open and closed trimers.

The structural differences between the monomers that compose the open and closed trimers are distributed across the polypeptide chain. The largest differences are localized to the loop that gates the pore (residues 120-123). The counterpart to the gating loop in the C-BMC (residues 224 to 227) also differs substantially between the monomers of the open and closed trimers. Likewise, in both the N- and C-BMC, the beta strand (β2) structurally adjacent to these loops is shifted concomitantly. Finally, part of the region of permutation in the primary structure (the A helix and the loop connecting the A helix to the β2 strand) in both the N- and C-BMC also differ substantially in conformation between the class 1 and class 2 monomers (rmsd between the class 1 and class 2 monomers alpha carbon backbones differ by ~0.68 Å and 0.36 Å in this region of the N- and C-BMC, respectively). The collective effect of these differences between the class 1 and class 2 monomers on the trimers can be viewed by toggling between the open and closed trimers in a superposition (Supplementary animation).

CsoS1D is a Dimer of an Open and a Closed Trimer

The CsoS1D structure from the orthorhombic crystal form contains six molecules arranged as a dimer of trimers in the crystallographic asymmetric unit (Figures 2D, 2E). Likewise, in the rhombohedral crystal form, a dimer of open and closed trimers is revealed by applying the crystallographic symmetry to the two molecules contained in the asymmetric unit. In both cases, the dimer interface is formed by interaction

between the relatively non-polar and hexagonal (Figures 3B, 3D) surface of the two trimers.

The interface between the two trimers buries 6,573 $Å^2$ of surface area (all surface areas are reported as the total surface area buried). The two surfaces fit snugly together (Figures 2E, 4C) as reflected by their shape correlation statistic, Sc. Values for Sc range between 0 and 1; 1 being a perfect fit and 0 no interaction, [22]; the Sc for the trimer-trimer interface of CsoS1D is 0.70 and 0.68 for the orthorhombic and rhombohedral crystal form structures, respectively. The trimer-trimer interface residues are primarily found on the A helix and the loop that connects the A helix to beta strand 2 in both the N- and C-BMC domains (residues 67-84 and 170-188; Figure 2C); one of the most structurally variable regions between the class 1 and class 2 monomers. These are also the regions of secondary structure in the N-BMC and C-BMC domains that are permuted relative to known single-BMC domain proteins. A closer inspection of the architecture of the CsoS1D hexamer reveals that the open and closed trimers are offset by a rotation of ~60° degrees across the dimerization interface. Thus, each monomer from one trimer thus interacts with two monomers from the opposite trimer (Figure 2E).

CsoS1D Models into a Carboxysome Shell Layer

In previous studies of the single-BMC domain carboxysome shell proteins, the proteins formed hexamers that tend to assemble into uniformly oriented layers in the crystals, suggesting a model for the facets of the carboxysome shell [14; 16; 17]. The size

and hexagonal shape of the CsoS1D pseudohexamer suggested that it could fit into such a model. To generate a model for a facet of the *Prochlorococcus* MED4 carboxysome shell, we constructed a layer of CsoS1 (a single-BMC domain protein) hexamers based on the layer structure of CsoS1A from *H. neapolitanus* [23]. Both *Prochlorococcus* MED4 and *H. neapolitanus* contain alpha carboxysomes and the sequence identity between CsoS1 of *Prochlorococcus* MED4 and *H. neapolitanus* CsoS1A is 80.4% overall. More than half of the sequence variation is confined to the C-terminal 10 residues of CsoS1, which are not part of the inter-hexamer interface in the CsoS1A layer; when these residues are omitted from the alignment the sequence identity is 93%.

Adjacent hexamers in the resulting CsoS1 model layer (Figure 6A) pack together with a Sc of 0.67 (compared to 0.68 for the adjacent CsoS1A hexamers observed in the crystal structure), burying 1,697Å^2 of surface area between adjacent hexamers (Table 1). We then substituted either an open or a closed CsoS1D trimer for a CsoS1 hexamer in the layer and examined the initial fit of each. The open trimer clashed somewhat with adjacent CsoS1 hexamers, and would require a slight adjustment to the mainchain in addition to some sidechain remodeling to obtain a realistic fit. In contrast, the closed trimer modeled readily into the CsoS1 layer (Figure 6) and required only rotamer adjustments in two amino acid sidechains to prevent any clashing. This CsoS1D-CsoS1 model was refined by energy minimization. The Sc and buried surface area in the resulting heterologous interface between the closed trimer of CsoS1D and the adjacent CsoS1 hexamers was

calculated; they indicate that modeled interaction is comparable to the interfaces among shell proteins that have been observed or modeled to-date (Table 1).

DISCUSSION

The overall pseudohexameric structure of the CsoS1D trimer and its fit within a model layer of CsoS1 (Figure 6), suggest that it is part of the carboxysome shell of *Prochlorococcus*. The amount of surface area buried (Table 1) between CsoS1D and CsoS1 in the modeled interaction is comparable to that in experimentally observed the homo-oligomeric interactions between hexamers of single-BMC domain proteins. The shape complementarity of and the amount of surface area buried in the modeled CsoS1D-CsoS1 heterologous interfaces are comparable to the experimentally observed homo-hexameric interfaces and to the heterologous interfaces between the pentameric and hexameric shell proteins in the most plausible models for the carboxysome shell (Table 1).

The CsoS1D pseudohexamer exhibits features necessary for incorporation into proposed models for the facets of the carboxysome shell. The pseudohexamer edges are of appropriate length and each edge contains the most strongly conserved residue among all BMC domains, the lysine residue at position 23 in the canonical BMC domain (see Pfam00936 HMM Logo http://pfam.sanger.ac.uk/family?acc=PF00936). This amino acid is found at the two-fold symmetry axis between adjacent hexamers in the CcmK1, CcmK2 and CsoS1A layers (and the model layer of CsoS1, Figure 6); in each case its sidechain is positioned roughly parallel to the edge of the hexamer, held

in that conformation by hydrogen bonds to the main chain carbonyl oxygen of the lysine in the opposite BMC domain and to the sidechain of either an Asp or Asn residue (conserved at position 19 in the Pfam00936 HMM Logo) within the same monomer. The conformation and the hydrogen bonding interactions between these residues in CsoS1D (Asp104 and Lys108, and Asn 208 and Lys212 in the N-BMC and C-BMC, respectively), are present in both crystals forms. Thus the conformation of the sidechains of the conserved lysine residues in the N-BMC and C-BMC are disposed to interact with adjacent hexamers at the two fold-axis of symmetry (Figure 6). Presumably, if it was not a carboxysome shell component, CsoS1D would be released from selective pressure to retain these residues (and their conformation) and the overall size and shape of the pseudohexamer.

In addition to the structural evidence for a role in the carboxysome shell, several other lines of evidence substantiate the proposed role for CsoS1D in the carboxysome. From genomic sequence analysis, all bacteria that appear to have the potential to form bacterial microcompartments (Kerfeld et al., unpublished) contain two or more BMC domains within their genomes; high light-adapted *Prochlorococcus* strains appeared to be the exception. The detection of the tandem BMC domain protein CsoS1D increases the known number of BMC domains in the genomes of *Prochlorococcus* MED4 and other high light adapted *Prochlorococcus* strains from one to three.

At the level of gene expression, the co-expression of *csoS1D* and the known carboxysome proteins in synchronized *Prochlorococcus* MED4 cells is consistent with a role in the carboxysome for CsoS1D. In unperturbed, synchronously cycling

Prochlorococcus MED4 cells, the *csoS1D* expression pattern clusters more tightly with that of RubisCO and the major carboxysome shell genes than does expression of the known shell genes *csoS4A* and *csoS4B*. A similar pattern for the *csoS1D* ortholog in *Synechocystis* PCC 6803, *slr0169*, emerges under low-carbon and high-light stresses [19; 20]. CsoS1D even offers a rationalization of one seemingly discordant expression pattern uncovered in an examination of the transcriptional effects of deleting *glcD* [19]. The enzyme GlcD is involved in detoxifying the byproduct of photorespiration, long proposed to be a signal of C_i starvation [24]. Growing the Δ*glcD* mutant grown at high CO_2 does partially phenocopy the metabolic state of wild-type cells grown at low CO_2 [25], but analysis of wild-type and Δ*glcD* mutant gene expression under high and low CO_2 appears to preclude the possibility that a single regulatory mechanism coordinates expression of all *Synechocystis* PCC 6803 carboxysome genes, *slr0169* included [19]. Eisenhut *et al.* propose that governing carboxysome gene expression by multiple mechanisms would allow the cell to tune the composition of the carboxysome shell according to the circumstances of the cell's growth. The identification of CsoS1D as a carboxysome shell protein gives this argument credence in two regards. First, all alpha carboxysome-containing cyanobacteria maintain a slight but consistent genomic separation of *csoS1D* from the alpha carboxysome operon; this conserved genomic context is consistent with there being a selective advantage to the disjoint regulation of different shell proteins. (Such an argument is difficult to advance in the beta carboxysome-containing *Synechocystis* sp. 6803, as its operons are in general poorly

organized.) Second, the ability to tune carboxysome shell composition might be particularly useful if different shell proteins have different metabolite transport properties, as the novel pore structure in CsoS1D strongly suggests.

The presence of orthologs to *csoS1D* (defined as a reciprocal bi-directional BLAST hit) in all cyanobacterial genomes also is consistent with a structural role for this protein in the carboxysome. A *csoS1D* ortholog (PCC-0905) is also found in the genome of the chromatophore of the amoeba *Paulinella* [26]. The *Paulinella* chromatophore appears to have originated from a relatively recent endosymbiosis of a cyanobacterium closely related to *Synechococcus* WH5701, however the chromatophore genome has undergone a massive genome reduction (to 26% of the gene content of *Synechococcus* WH5701). Among the genes retained in the chromatophore genome is a complete set of photosynthesis genes [26], in which the cyanobacterial alpha carboxysome gene cluster (Figure 1B), including *csoS1D* and *ham1*, is conserved intact.

The pore formed at the three-fold axis of the CsoS1D pseudohexamer is unique among bacterial microcompartment shell proteins that have been structurally characterized to date. CsoS1D is the first BMC protein structure to reveal a potential for gating at the pore (Figures 3, 4). Because the sidechains of the conserved residues that line the pore in the CcmK1, 2 and 4 and CsoS1A hexamers are fully extended, these 4-7Å pores appear to be constitutively open [14; 16; 17]. In contrast, the CsoS1D pore is open or closed depending on the sidechain conformations of two residues, Glu120 and Arg121 (Figure 4), both absolutely conserved among all

CsoS1D orthologs. The open pore in CsoS1D is ~14Å in diameter, significantly larger than those of single-BMC domain carboxysome shell protein hexamers. A larger pore may be advantageous for the diffusion of the bulkier carboxysome metabolites (e.g. RuBP) which would be expected to pass through the CcmK1, CcmK2, CcmK4 and CsoS1A pores only with difficulty.

Because both of the CsoS1D structures determined here lack the N-terminal 50 amino acids, it is possible that there is an additional domain situated over the CsoS1D pore that may be involved in regulating transport across the trimer. This is an enigmatic region of primary structure as it is only found in the *Prochlorococcus* and *Synechococcus* CsoS1D orthologs and its primary sequence is relatively poorly conserved.

The structure of CsoS1D is the first model of a tandem BMC domain containing protein (Figure 2); the second BMC domain in the N-terminus of the protein was unexpected, since it was not predicted from the primary structure. The genomes of organisms that produce beta carboxysomes all contain a gene for a tandem-BMC domain protein, *ccmO*, but one had not been identified in cyanobacteria that form alpha carboxysomes. In CsoS1D, combining the two BMC domains and a trimeric oligomeric state results in a pseudohexamer, which presents two types of edges (Figures 2, 6) for interaction with neighboring (pseudo)hexamers within a model carboxysome shell layer. To date, only homohexamers of single-BMC domain carboxysome shell proteins have been structurally characterized; it is not yet known whether mixtures of single-BMC domain containing proteins such as

CcmK1-4 or CsoS1A,B,C form mixed hexamers under physiological conditions. Fusing two BMC domains has implications for their assembly into shell layers; fusion increases the effective concentration of protein domains with respect to each other [27] which could possibly enhance the rate of shell assembly. At the same time, the presence of two distinct edges to match reduces the number of degrees of freedom of fit. This would be expected to slow the assembly of heterohexamers into a layer.

Interestingly, there are no single-BMC domain homologs of the N-terminal BMC domain of CsoS1D in the sequence database. Because of this and the general observation that gene fusion is a later event in protein evolution, the most plausible scenario for the evolutionary history of CsoS1D includes a duplication of a permuted BMC domain, followed by fusion with subsequent divergence of the primary structure of the N-BMC domain. The underlying functional reason for the divergence in sequence of the N-BMC domain is presently unknown.

In both crystal forms of CsoS1D, the protein was arranged as a dimer of trimers (Figures 2D, 2E) formed by the dimerization of the broader, more hexagonal, faces of two trimers. Several lines of evidence suggest that a dimer of trimers is a physiologically relevant form of CsoS1D. In general, interfaces in obligate oligomers, such as most homodimers, are large and relatively hydrophobic [28]; in CsoS1D the dimerization buries a total of 6,573 $Å^2$. Moreover, the face of the CsoS1D trimer buried by the dimerization is less polar than the opposite face that is exposed to solvent (Figure 3); more than half of residues buried in dimerization are nonpolar (14 of 26 in each monomer). Furthermore, the fit between the two

interfaces is snug; the shape complementarity is 0.70 for the orthorhombic crystal form and 0.68 for the rhombohedral crystal form (a typical value for an antibody: antigen complex is 0.64-0.68 [22]). Finally, the residues interacting across the trimer-trimer interface also are strongly conserved; 21 of the 26 residues are identical among all *Prochlorococcus* and *Synechococcus* CsoS1D orthologs.

The stability of interaction between the trimers is likely enhanced by staggering the interactions among domains; each monomer in a trimer interacts with two different monomers in the opposite trimer (Figure 2E). Notably, the 26 residues each monomer contributes to the trimer-trimer interface are all found in the permuted segments of the primary structure (between residues 50-84 of the N-BMC and 156-190 of the C-BMC) in CsoS1D. Likewise, in the structure of PduU, the only single-BMC domain protein containing this permutation that has been structurally characterized, there is a similar dimerization of hexamers, which is likewise mediated by the permuted regions of primary structure [21]. The elution behavior of CsoS1D in size exclusion chromatography is also consistent with hexameric and larger assemblies (data not shown). Collectively these observations suggest that the dimer of trimers configuration, found in the two crystal forms of CsoS1D, is physiologically relevant.

A structural consequence of the tight appression of two CsoS1D trimers is a central channel formed by the continuity between the pores of the two trimers (Figure 4D) that is open on one end and closed on the other. The Glu120 and Arg121 residues involved in gating the opening are closer to the outer, solvent exposed

surface of each trimer. The channel is approximately 73 Å lengthwise and ranges in diameter between 0 and 18 Å (Figure 4D). The interior of the channel is mainly positively charged (Figure 4C). The volume of the channel is substantial, 13,613 $Å^3$, of sufficient size to be a microenvironment for ligand binding and catalysis. Accordingly, we screened CsoS1D for a range of enzymatic activities and attempted co-crystallization with metabolites that must cross the carboxysome shell (e.g. RuBP and bicarbonate) but these experiments did not yield conclusive results.

While the trimeric/pseudohexameric structure of CsoS1D is compatible with current models for the facets of the carboxysome shell, the dimerization of the CsoS1D trimers raises new questions about the details of shell architecture: for example, could the entire carboxysome shell be composed of a bilayer of BMC-domain containing proteins? In the structures of BMC proteins that form layers of uniformly oriented hexamers (CsoS1A, CcmK1 and CcmK2), adjacent layers in the crystals are aligned with respect to their pores, but the orientation of molecules between layers varies (i.e. in CsoS1A all of the layers are oriented similarly, in CcmK1 the convex surfaces of hexamers in adjacent layers appear to interact, in CcmK2 isologous interfaces are juxtaposed but any two adjacent layers were not as tightly appressed as in the CsoS1D dimer of trimers). Alternatively, it may be that only a subset of proteins making up the carboxysome shell are assembled in two tiers; this subset is perhaps restricted to those containing the secondary structure permutation observed in CsoS1D and PduU, which also forms dimers of hexamers [21]. Dyad shell building blocks may occur for a specific functional purpose and may be

differentially regulated; CsoS1D's position outside of the known carboxysome operon and the expression profile of its *Synechocystis* ortholog *slr0169* are consistent with this hypothesis. Based on the CsoS1-CsoS1D model, the second trimer could occur as a protrusion from the surface of the carboxysome (where, for example, it could be a part of an adjacent carboxysome). Alternatively, the second trimer could face the interior of the carboxysome, where it could play a role in organizing the interior enzymes RubisCO and/or CsoSCA.

CONCLUSIONS

There is compelling evidence that PMM0547 in *Prochlorococcus* (provisionally dubbed CsoS1D) is a carboxysome shell protein. Here we report two independent crystal structures of this protein, and a model for its interaction with CsoS1 in the *Prochlorococcus* carboxysome shell. The structural models are complemented by genomic and transcriptomic evidence which together suggest that PMM0547 is a carboxysome shell protein. Strikingly, although bioinformatic methods predicted that CsoS1D would contain a single BMC domain, the crystal structure reveals it to have two tandem BMC domains, making it the first double BMC-domain protein to be structurally characterized. In addition to uncovering the second, cryptic, BMC domain, structural analysis of CsoS1D revealed several other unexpected features. All characterized single-BMC-domain proteins form a hexameric ring; CsoS1D trimerizes, but the tandem BMC domains make it a pseudohexamer that is compatible with the hexameric layers that are proposed to form the facets of the carboxysome

shell. Moreover, CsoS1D trimers form tightly appressed dimers, sandwiching two rings together to form a novel two-tiered carboxysome shell building block. The structures of CsoS1D also reveal an unusually large pore that can adopt both open and closed conformations. No other carboxysome shell protein has been observed to possess a two-state pore; CsoS1D thus is the first shell protein to present a means for gating metabolite flow across the carboxysome shell. Collectively, these features are likely to have a significant impact on carboxysome ultrastructure and function in *Prochlorococcus*. Moreover, the universality of CsoS1D orthologs alongside canonical shell proteins in other cyanobacteria suggests that, should the environment demand it, the functions made possible by these novel carboxysome features may be readily available to all cyanobacteria

Materials and Methods

PMM0547 Cloning, Expression, and Purification

Prochlorococcus marinus subsp. MED4 genomic DNA was used as template in polymerase chain reaction using the oligonucleotide primers with the following sequences: CGACCATGGAACCAACTTCTAGC and CGAAGCTTAATAATTTGATATTTGATCAATTGC, which contain *NcoI* (C^CATGG) and *HindIII* (A^AGCTT) restricstion sites. The PCR product was digested with *NcoI* and *HindIII*, and ligated into the multiple cloning site of pProEX-HTb (Life Technologies, Inc.), which adds a His_6 to the N-terminus of the PMM0547 protein coding sequence. The sequence of the protein coding region of the

recombinant plasmid was confirmed by DNA sequencing (The Univ. of Maine DNA Sequencing Facility).

For expression of recombinant protein, the constructs were transformed into *Escherichia coli* BL21 cells (Novagen). For Seleno-methionine labeling of the PMM0547 gene product, the plasmid was transformed into the *E. coli* methionine auxotroph B834 strain (Novagen). PMM0547 was purified from fresh BL21 cultures grown in LB (ALDRICH-SIGMA) to $O.D._{600}$=0.8 at 37 ºC, and induced at 25 ºC for 18 h with 0.4 mM IPTG (ALDRICH-SIGMA). Following centrifugation, the cell pellets were lysed by sonication (model W-220F, Branson) in buffer containing 500 mM NaCl, 50 mM Tris-HCl (pH 8.0), 10% glycerol and 2 mM Beta-ME. Poly-histidine tagged proteins were purified using the standard purification protocol recommended for use with Ni-TA resin (QIAGEN). To obtain higher purity preparations for use in SAXS, the recombinant PMM0547 protein was purified using TALON resin (CLONTECH). Proteins were eluted from these resins with 200 mM imidazole, 500 mM NaCl, 100 mM Tris-HCl (pH 8.0). The purified recombinant 0547 protein (5-10 mg/ml) was dialyzed overnight in 10 mM Tris-HCl (pH 8.0) at 4 ºC. Purified protein was stored at 4 ºC and diluted to 1 mg/ml with dH_2O prior to setting up crystallization trials.

Crystallization and Structure Determination

Orthorhombic crystals were grown at 18 ºC by hanging drop vapor diffusion by mixing with a solution containing 12 % Tacsimate (HAMPTON RESEARCH)

adjusted to pH 6.2 with HCl. The rhombohedral crystal form was grown at 18 ºC by hanging drop vapor diffusion by mixing with 0.7M ammonium phosphate (HAMPTON RESEARCH), adjusted to pH 4.2 with 100 mM CHES (ALDRICH-SIGMA) (pH 9). Crystallization drops were prepared with either 4+1 or 2+1 μl protein-to-crystallization buffer drop ratios for the Tacsimate and ammonium phosphate conditions, respectively.

Native and selenomethionine (ALDRICH-SIGMA) derived crystals were soaked rapidly in crystallization solutions supplemented with 30% ethylene glycol (HAMPTON RESEARCH) prior to freezing. Tacsimate-grown crystals were plunged into liquid nitrogen and crystals from the ammonium phosphate condition were frozen by placing directly into the cryo-stream of nitrogen gas. Diffraction from the crystals prepared with Tacsimate was consistent with orthorombic space group $P2_12_12_1$ (unit cell dimensions: A=122.41, B=131.286, C=131.762, α,β,γ=90) and the diffraction from crystals grown from the ammonium phosphate condition was Rhombohedral Space Group R3 (unit cell dimensions: A=117.163, B=117.163, C=101.463 α=β=90, γ=120). Diffraction data were collected at Lawrence Berkeley's Advanced Light Source beamline 8.2.2 and processed with DENZO and SCALEPACK [29].

We prepared crystals of seleniomethionine substituted protein to assist in solving the phases for the orthorhombic crystal form. Forty heavy atom selenium positions were identified from a dataset collected at the peak wavelength for selenium using the program Phenix.hyss using diffraction data in resolution ranges 30 to 3.5 Å,

where the anomalous signal appeared to be strong based on results from the program xtriage in the Phenix software suite [30]. Heavy atom refinement and phasing was performed using autoSHARP [31]. Density modification using SOLOMON [32] produced exceptionally high quality electron density map calculated without NCS averaging to 2.3 Å. Using the density modified phases from SOLOMON, the program ARP/wARP [33], implemented in CCP4 [34], constructed an initial model for the six monomers in the crystallographic asymmetric unit, including fitting the amino acid sequence into the correct registry. Model building was completed using the graphics program COOT [35] and a well defined NCS averaged electron density map (Figure S3) produced using the phases calculated in SHARP after applying solvent flattening and density averaging procedures in the program RESOLVE [36]. Positional and B factor refinements and water-picking were performed using Phenix.refine. The stereochemistry of the final refined model was analyzed with the program MolProbity [37]. The 6 chains that comprise the hexamer found in the asymmetric unit have connectivity throughout the polypeptide chain yet the residues absent at the N-and C-terminal are variable. Residues 50 to 256, 48 to 256, 50 to 256, 53 to 256, 52 to 252, and 52 to 256 are present in chains A-F of the final model, respectively.

Diffraction data from the rhombohedral crystal form was consistent with twinning as detected in xtriage from the phenix suite. Nevertheless, the structure was solved by molecular replacement methods using the monomeric form from the orthorhombic crystal form using the program Phaser [38] and refined to 2.2 resolution using phenix. The structure was refined using a twin fraction of 0.45. In R3 chain

B the loop residues 120 to 122 are refined with occupancies of zero, as the electron density map that corresponds to this region of the polypeptide chain was not well defined.

Analysis of structure surface complementarity and surface area was performed with CCP4 programs SC [39] and AREAIMOL [40], respectively. Least squares RMS deviation was calculated in CCP4 using program LSQKAB [41]. The model of CsoS1 was generated with SWISSMODEL [42]. The docking model of CsoS1D-CsoS1 was refined using CNS energy minimization program [43].

Clustering analysis of diel expression data

For details of sample collection and data processing, see Zinser *et al.* (submitted). In their RMA-averaged data, they identified 1405 genes in *Prochlorococcus* strain MED4 whose expression oscillated when cells were grown in light:dark (14:10) synchronized cells doubling once per day. The expression profiles of these genes were subjected to soft c-means clustering using the R package Mfuzz [18; 44]. Clustering was performed with 25 input parameter pairs, each combination of 5 cluster numbers *c* (10, 12, 14, 16, and 18) and 5 "fuzzification" paramters *m* (1.15, 1.20, 1.25, 1.30, and 1.35). This analysis was repeated 100 times with each set of input parameters as a check on the variability of the output fractional cluster membership (μ) values. Figure S1 was generated using the R package Heatplus, taking as input the output from one randomly selected clustering run for each pair of input parameters.

REFERENCES

1. Dobrinski, K. P., Longo, D. L. & Scott, K. M. (2005). The carbon-concentrating mechanism of the hydrothermal vent chemolithoautotroph *Thiomicrospira crunogena*. *J Bacteriol* **187**, 5761-5766.
2. Price, G. D., Sultemeyer, D., Klughammer, B., Ludwig, M. & Badger, M. R. (1998). The functioning of the CO_2 concentrating mechanism in several cyanobacterial strains: a review of general physiological characteristics, genes, proteins and recent advances. *Can J Bot* **76**, 973-1002.
3. Scott, K. M., Henn-Sax, M., Harmer, T. L., Longo, D. L., Frame, C. H. & Cavanaugh, C. M. (2007). Kinetic isotope effect and biochemical characterization of form IA RubisCO from the marine cyanobacterium *Prochlorococcus marinus* MIT9313. *Limnol Oceanogr* **52**, 2199-2204.
4. Yeates, T. O., Kerfeld, C. A., Heinhorst, S., Cannon, G. C. & Shively, J. M. (2008). Protein-based organelles in bacteria: carboxysomes and related microcompartments. *Nat Rev Microbiol*.
5. Bobik, T. A. (2006). Polyhedral organelles compartmenting bacterial metabolic processes. *Appl Environ Microbiol* **70**, 517-525.
6. Heinhorst, S., Cannon, G. C. & Shively, J. M. (2006). Carboxysomes and carboxysome-like inclusions. In *Complex Intracellular Structures in Prokaryotes* (Shively, J. M., ed.), Vol. 2, pp. 141-164. Springer, Berlin/Heidelberg.
7. Shively, J. M., Ball, F., Brown, D. H. & Saunders, R. E. (1973). Functional organelles in prokaryotes: Polyhedral inclusions (carboxysomes) in *Thiobacillus neapolitanus*. *Science* **182**, 584-586.
8. Lichtle, C., Thomas, J. C., Spilar, A. & Partensky, F. (1995). Immunological and ultrastructural characterization of the photosynthetic complexes of the prochlorophyte *Prochlorococcus* (Oxychlorobacteria). *J. Phycology* **31**, 934-941.
9. So, A. K., Espie, G. S., Williams, E. B., Shively, J. M., Heinhorst, S. & Cannon, G. C. (2004). A novel evolutionary lineage of carbonic anhydrase (epsilon class) is a component of the carboxysome shell. *J Bacteriol* **186**, 623-630.
10. Dou, Z., Heinhorst, S., Williams, E. B., Murin, C. D., Shively, J. M. & Cannon, G. C. (2008). CO_2 fixation kinetics of *Halothiobacillus neapolitanus* mutant carboxysomes lacking carbonic anhydrase suggest the shell acts as a diffusional barrier for CO_2. *J Biol Chem* **283**, 10377-10384.
11. Heinhorst, S., Williams, E. B., Cai, F., Murin, C. D., Shively, J. M. & Cannon, G. C. (2006). Characterization of the carboxysomal carbonic anhydrase CsoSCA from *Halothiobacillus neapolitanus*. *J Bacteriol* **188**, 8087-8094.
12. Badger, M. R. & Price, G. D. (2003). CO_2 concentrating mechanisms in cyanobacteria: molecular components, their diversity and evolution. *J Exp Bot* **54**, 609-622.
13. Badger, M. R., Price, G. D., Long, B. M. & Woodger, F. J. (2006). The environmental plasticity and ecological genomics of the cyanobacterial CO_2 concentrating mechanism. *J Exp Bot* **57**, 249-265.
14. Tanaka, S., Kerfeld, C. A., Sawaya, M. R., Cai, F., Heinhorst, S., Cannon, G. C. & Yeates, T. O. (2008). Atomic-level models of the bacterial carboxysome shell. *Science* **319**, 1083-1086.

15. Cannon, G. C., Bradburne, C. E., Aldrich, H. C., Baker, S. H., Heinhorst, S. & Shively, J. M. (2001). Microcompartments in prokaryotes: carboxysomes and related polyhedra. *Appl Environ Microbiol* **67**, 5351-5361.
16. Tsai, Y., Sawaya, M. R., Cannon, G. C., Cai, F., Williams, E. B., Heinhorst, S., Kerfeld, C. A. & Yeates, T. O. (2007). Structural analysis of CsoS1A and the protein shell of the Halothiobacillus neapolitanus carboxysome. *PLoS Biol* **5**, e144.
17. Kerfeld, C. A., Sawaya, M. R., Tanaka, S., Nguyen, C. V., Phillips, M., Beeby, M. & Yeates, T. O. (2005). Protein structures forming the shell of primitive bacterial organelles. *Science* **309**, 936-8.
18. Futschik, M. E. & Carlisle, B. (2005). Noise-robust soft clustering of gene expression time-course data. *J Bioinform Comput Biol* **3**, 965-88.
19. Eisenhut, M., von Wobeser, E. A., Jonas, L., Schubert, H., Ibelings, B. W., Bauwe, H., Matthijs, H. C. P. & Hagemann, M. (2007). Long-term Response Towards Inorganic Carbon Limitation in Wild Type and Glycolate Turnover Mutants of the Cyanobacterium Synechocystis sp. Strain PCC 6803 10.1104/pp.107.103341. *Plant Physiol.*, pp.107.103341.
20. Singh, A. K., Elvitigala, T., Bhattacharyya-Pakrasi, M., Aurora, R., Ghosh, B. & Pakrasi, H. B. (2008). Integration of carbon and nitrogen metabolism with energy production is crucial to light acclimation in the cyanobacterium synechocystis. *Plant Physiol* **148**, 467-78.
21. Crowley, C. S., Sawaya, M. R., Bobik, T. A. & Yeates, T. O. (2008). Structure of the PduU Shell Protein from the Pdu Microcrocompartment of *Salmonella*. *Structure* **16**, 1324-1332.
22. Lawrence, M. C. & Colman, P. M. (1993). Shape Complementarity at Protein/Protein Interfaces. *J. Mol Biol.* **234**, 946-950.
23. Tsai, Y., Sawaya, M. R., Cannon, G. C., Cai, F., Williams, E. B., Heinhorst, S., Kerfeld, C. A. & Yeates, T. O. (2007). The structure of the shell protein CsoS1A from *Halothiobacillus neapolitanus* and its implications for carboxysome function. *PLoS Biology* **5**, 1345-1354.
24. Marcus. Y., Harel, E. & Kaplan, A. (1983). Adaptation of the cyanobacterium *Anabaena varaibilis* to low CO2 concentration in the environment. *Plant Physiology* **71**, 208-210.
25. Eisenhut, M., Huege, J., Schwarz, D., Bauwe, H., Kopka, J. & Hagemann, M. (2008). Metabolome phenotyping of inorganic carbon limitation in cells of the wild type and photorespiratory mutants of the cyanobacterium *Synechocystis* sp Strain PCC6803. *Plant Physiology* **in press**.
26. Nowack, E. C. M., M., M. & Glockner, G. (2008). Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. *Current Biology* **18**, 410-418.
27. Kuriyan, J. & Eisenberg, D. S. (2007). The origin of protein interactions and allostery in colocalization. *Nature* **450**, 983-990.
28. Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **93**, 13-20.
29. Otwinowski, Z., and Minor, W. (1997). Processing of X-ray Diffraction Data Collected in Oscillation Mode. In *Methods in Enzymology: Macromolecular Crystallography, part A* (Carter, C. W., and Sweet, R.M., ed.), Vol. 276, pp. 307-326. Academic Press, New York.
30. Adams, P. D., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). PHENIX: building new software for automated crystallographic structure determination. *Acta*

Crystallogr D Biol Crystallogr **58**, 1948-54.

31. Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. (2007). Automated structure solution with autoSHARP. *Methods Mol Biol* **364**, 215-30.
32. Abrahams, J. P. & Leslie, A. G. (1996). Methods used in the structure determination of bovine mitochondrial F1 ATPase. *Acta Crystallogr D Biol Crystallogr* **52**, 30-42.
33. Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). ARP/wARP's model-building algorithms. I. The main chain. *Acta Crystallogr D Biol Crystallogr* **58**, 968-75.
34. (1994). The CCP4 suite: programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr* **50**, 760-3.
35. Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126-32.
36. Terwilliger, T. C. (2000). Maximum-likelihood density modification. *Acta Crystallogr D Biol Crystallogr* **56**, 965-72.
37. Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., 3rd, Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* **35**, W375-83.
38. Read, R. J. (2001). Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr D Biol Crystallogr* **57**, 1373-82.
39. Lawrence, M. C. & Colman, P. M. (1993). Shape complementarity at protein/protein interfaces. *J Mol Biol* **234**, 946-50.
40. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55**, 379-400.
41. Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr A* **A32**, 922-923.
42. Peitsch, M. C. (1996). ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem Soc Trans* **24**, 274-9.
43. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* **54**, 905-21.
44. Team, R. D. C. (2008). A language and environment for statistical computing. *http://www.R-project.org*.
45. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876-4882.
46. DeLano, W. L. (2002). DeLano Scientific, Palo Alto, CA.
47. Petrey, D. & Honig, B. (2003). GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol* **374**, 492-509.
48. Smart, O. S., Goodfellow, J. M. & Wallace, B. A. (1993). The Pore Dimensions of Gramicidin A. *Biophysical Journal* **65**, 2455-2460.

ACKNOWLEDGEMENTS

We would like to thank Sandra Schwarte, Martin Hagemann, Hermann Bauwe Michael Sawaya and Todd Yeates for helpful discussions and Edwin Kim and Jay Kinney for figure preparation and technical assistance. SCB is a Howard Hughes Medical Institute Predoctoral Fellow. GRC and SH acknowledge support from the National Science Foundation (MCB-0818680 and DMR-0213883). The work of CAK is performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract number DE-AC02-05CH11231, Lawrence Livermore National Laboratory under contract number DE-AC52-07NA27344.

FIGURE LEGENDS

Figure 1. (A) Negatively stained transmission electron micrograph of *Prochlorococcus* MED4 showing a single carboxysome (arrow). (Electron Micrograph courtesy of Luke Thompson (MIT) and Nikki Watson, Keck Imaging Center, Whitehead Institute for Biomedical Research). (B) Genomic organization of the carboxysome gene clusters of the cyanobacterium *Prochlorococcus* MED4 and the chemoautotroph *Halothiobacillus neapolitanus* indicating the positions of the bacterial microcompartment (BMC) domains. Arrows representing genes are

proportional to gene size. Panel 2C prepared with CLUSTAL_X [45]

Figure 2. (A) The CsoS1D trimer. (B) Superposition of the N-BMC (green) and C-BMC (blue) domain structures of CsoS1D. (C) Sequence alignment and secondary structure of the N-BMC and C-BMC domains of CsoS1D. Arrows correspond to beta strands and heavy blue lines represent alpha helices, colored as in (B). (D) The CsoS1D dimer of trimers in ribbons and (E) space-filling representation. The N-BMC and C-BMC are colored as in (B). Panels A, C-E and Figures 4-6 prepared with PyMOL [46].

Figure 3. Electrostatic surface representation of the (A and B) closed CsoS1D trimer and the open (C and D) CsoS1D trimer. The surfaces shown in A and C are exposed to solvent in the crystal; those in (B) and (D) are involved in the dimerization of the open and closed trimer. Figures are colored according to electrostatic potential with blue positive and red negative. This figure and figures 4A-C prepared with GRASP2 [47].

Figure 4. A close-up view (from the solvent exposed side of each trimer) of the gated pore formed at the three-fold axis of symmetry in its closed (A) and open (B) conformations. The electrostatic potential is colored as in Figure 3. The conserved amino acids (Glu120 and Arg121) that gate the pore are highlighted. (C) The central channel created by the dimerization of the open and closed CsoS1D trimers. (D)

Radius of the central channel in the dimer of trimers plotted as a function of vertical position along the channel (calculated with HOLE2, [48]).

Figure 5. Superposition of the open (gray) and closed (blue) trimers (A) and of (B) individual monomers distinctive of the open (gray) and closed (blue) trimers.

Figure 6. (A) Model of the interaction of the closed CsoS1D trimer (colored as in Figure 2A) and a layer of CsoS1 hexamers (blue). A single CsoS1D monomer interacts with three CsoS1 hexamers, labeled in reference to the interface analysis reported in Table 1. (B and C) Close-up views of the conserved Lys sidechains (circled) at the intermolecular interface of adjacent BMC domains between CsoS1D and CsoS1. (B) and between adjacent CsoS1 hexamers.

Table 1. Properties of protein-protein interfaces among bacterial microcompartment shell proteins

Data bank accession codes

The coordinates and structure factors for both crystal forms of CsoS1D have been deposited in the RCSB PDB (http://www.rcsb.org/pdb/home/home.do) under the accession numbers 3F56 (orthorhombic crystal form) and 3FCH (rhombohedral crystal form).

Supplementary Material

Figure S1. Heat map depicting the clustering patterns (see Methods) of gene expression from 25 time points over two diel cycles in synchronized *Prochlorococcus* MED4 cells for a range of clustering parameters (c, number of clusters; m, "fuzziness" of clustering). Clustering was run on the entire set of 1405 expressed, cycling genes; output is shown for only those genes in the immediate vicinity of the carboxysome operon. The heat map gives the output of one arbitrarily selected clustering run for each set of parameters. Genes *PMM0549-0555* constitute the known carboxysome operon in *Prochlorococcus* MED4. *PMM0547* (*csoS1D*) is separated from the main carboxysome operon by *PMM0548*, encoding a non-carboxysome-related Ham1 protein. Genes not labeled in the operon diagram encode hypothetical proteins. Coloring indicates the strength, μ, of each gene's association with each cluster under a given pair of clustering parameters.

Figure S2. Fractional membership, μ, in the cluster containing known carboxysome genes *CsoS1*, *CbbL*, *CbbS*, and *CsoS2* for the 20 genes surrounding *CsoS1D*. Genes not labeled in the operon diagram encode hypothetical proteins. Each panel plots the output μ values for these 20 genes from 100 soft c-means clustering runs using the given pair of input parameters and diel expression data for all 1405 *Prochlorococcus* MED4 genes that oscillate in synchronized cells grown on a 24-h light-dark cycle (Zinser et al, submitted). Run-to-run variability, seen as the vertical smearing of

each gene's cluster membership values, increases with cluster number and "fuzziness", but for each set of input parameters, the distribution of membership values for *CsoS1D* (highlighted in light blue) in the cluster containing the carboxysome operon (orange) is higher than that of shell proteins *CsoS4A* and *CsoS4B* (red).

Figure S3. Stereo image showing residues 55 to 65 from Chain A of the refined model inside the experimental electron density map prepared in RESOLVE with density modification using solvent flattening and NCS averaging.

Table S1. Data collection and refinement statistics.

Table S2. Summary of RMS deviations between CsoS1D monomers

Supplemental Movie Comparison of the open and closed CsoS1D trimers

Figure 1

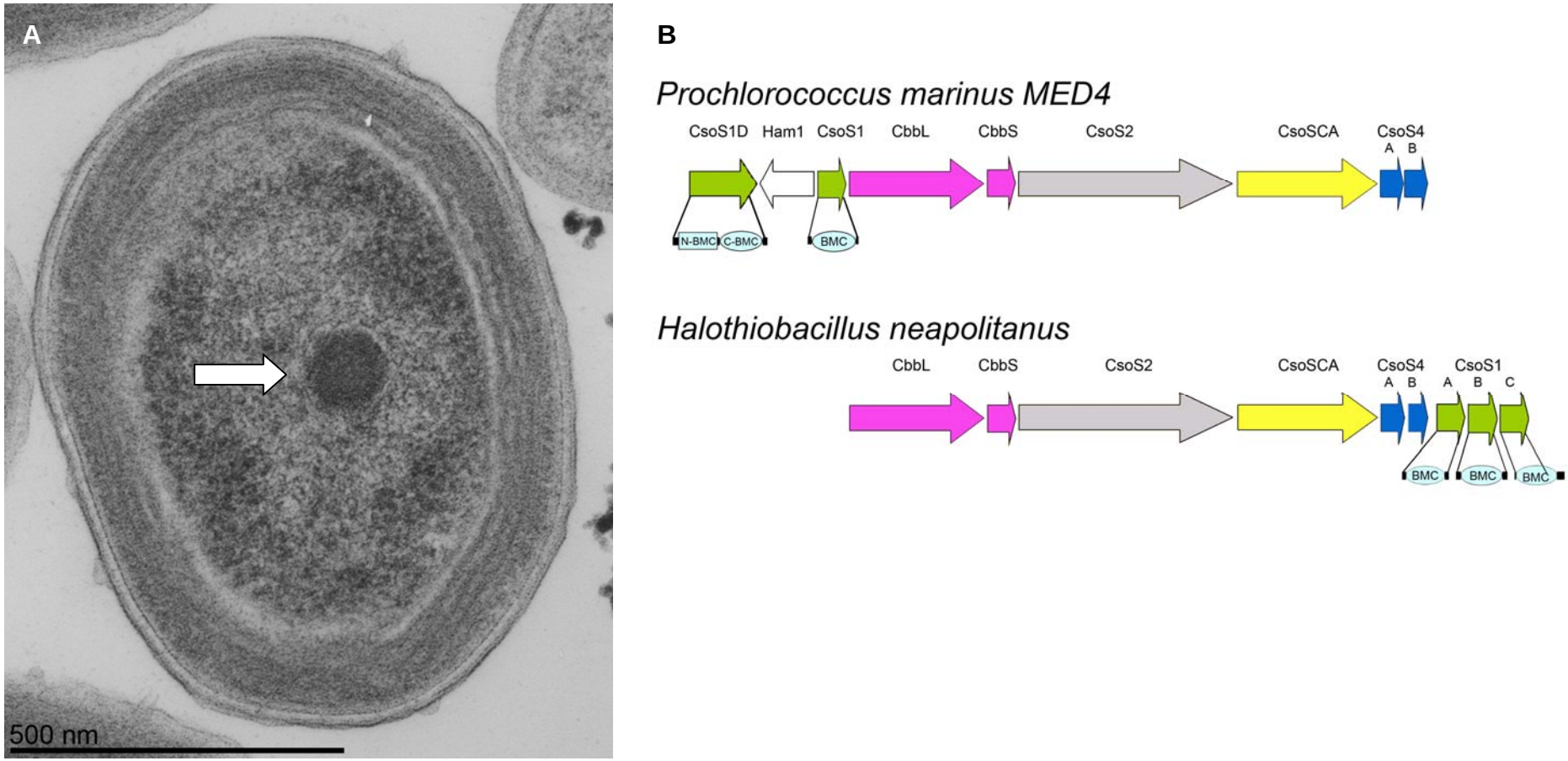

A
500 nm
B
Prochlorococcus marinus MED4
CsoS1D
Ham1
CsoS1
CbbL
CbbS
CsoS2
CsoSCA
CsoS4
A
B
N-BMC
C-BMC
BMC
Halothiobacillus neapolitanus
CbbL
CbbS
CsoS2
CsoSCA
CsoS4
A
B
CsoS1
A
B
C
BMC
BMC
BMC

Figure 2

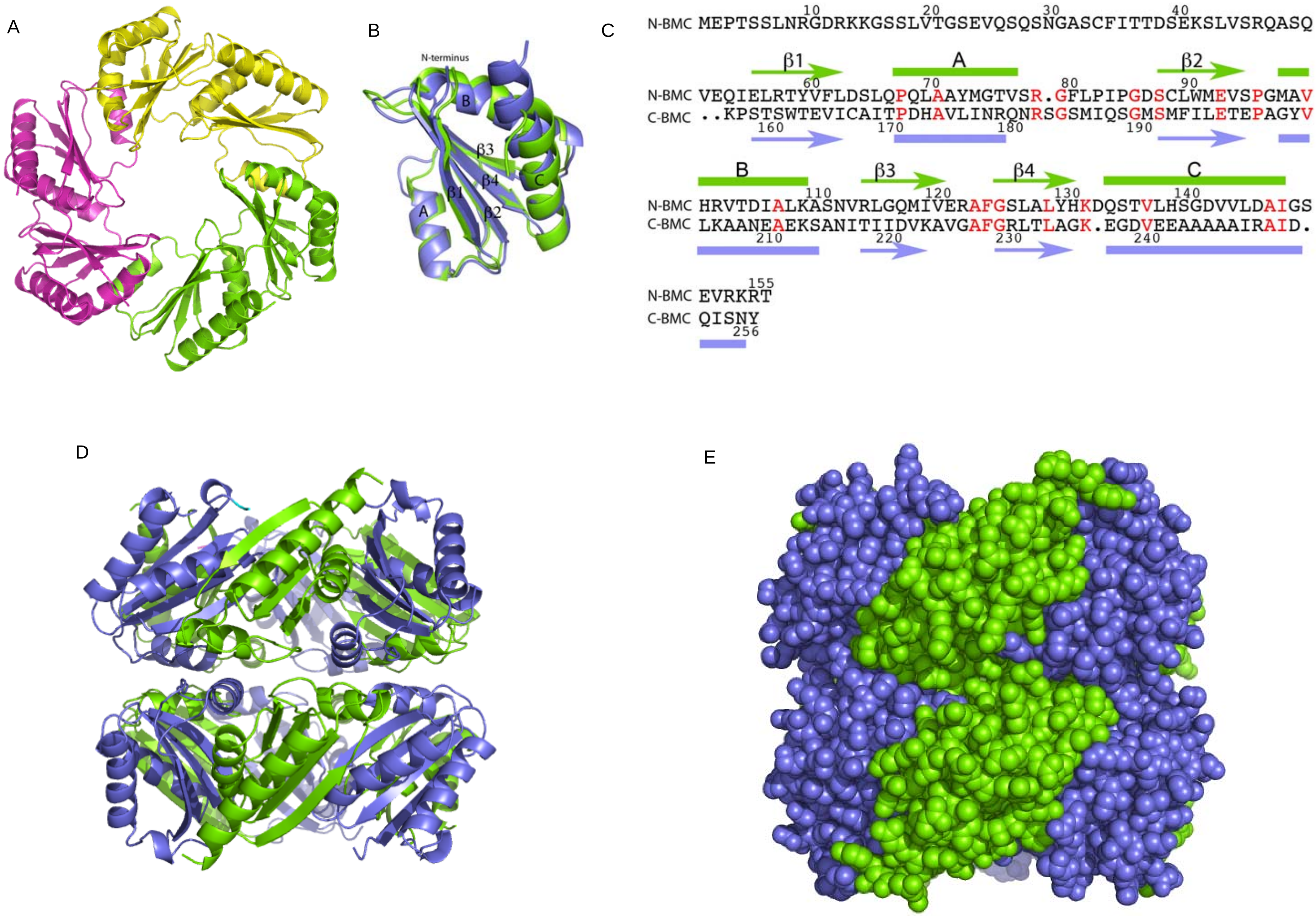

A
B
N-terminus
A
B
C
β1
β2
β3
β4
C
N-BMC MEPTSSLNRGDRKKGSSLVTGSEVQSQSNGASCFITTDSEKSLVSRQASQ
β1 A β2
N-BMC VEQIELRTYVFLDSLQPQLAAYMGTVSR.GFLPIPGDSCLWMEVSPGMAV
C-BMC ..KPSTSWTEVICAITPDHAVLINRQNRSGSMIQSGMSMFILETEPAGYV
B β3 β4 C
N-BMC HRVTDIALKASNVRLGQMIVERAFGSLALYHKDQSTVLHSGDVVLDAIGS
C-BMC LKAANEAEKSANITIIDVKAVGAFGRLTLAGK.EGDVEEAAAAAIRAID.
N-BMC EVRKRT
C-BMC QISNY
D
E

Figure 3

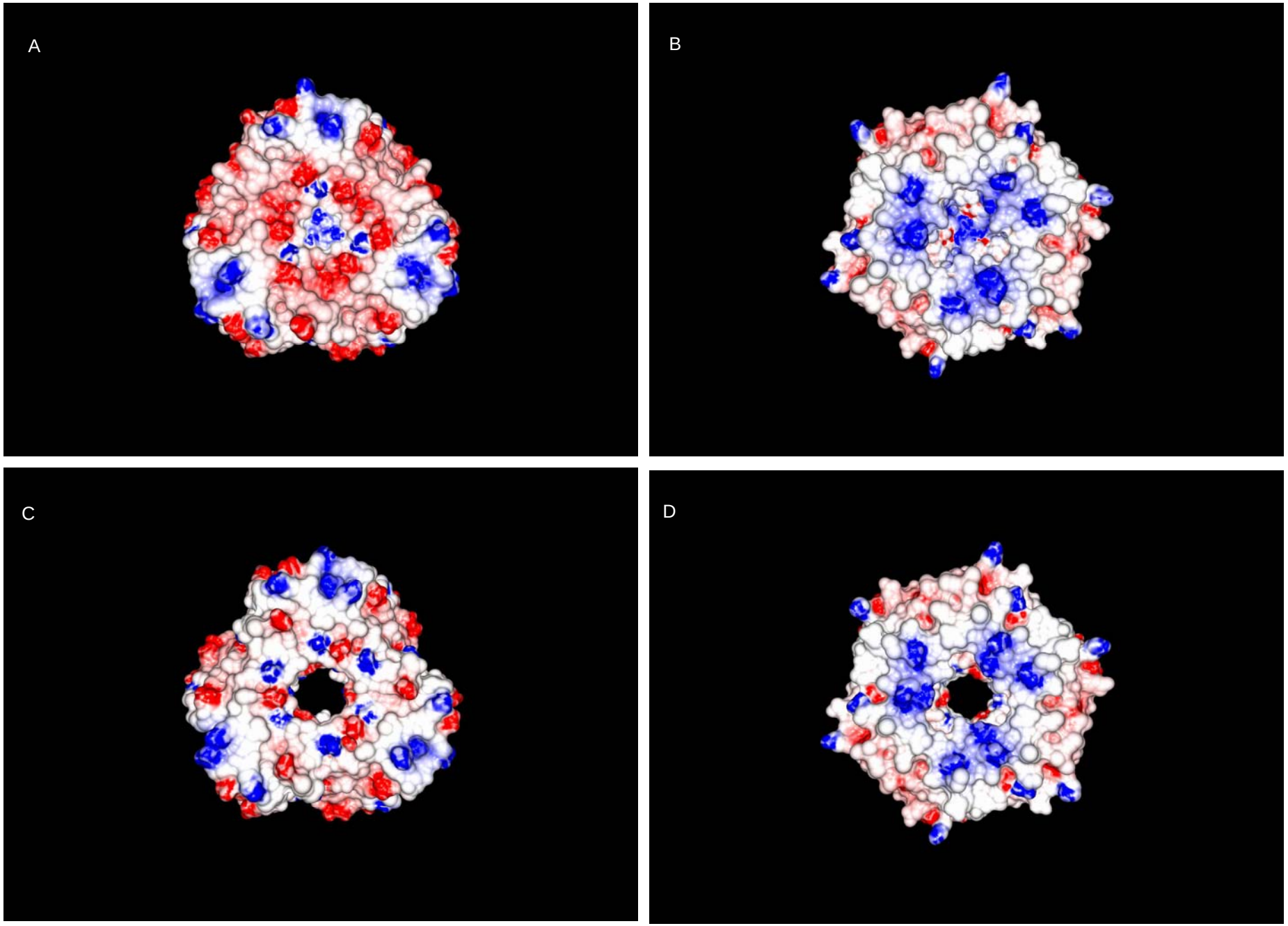
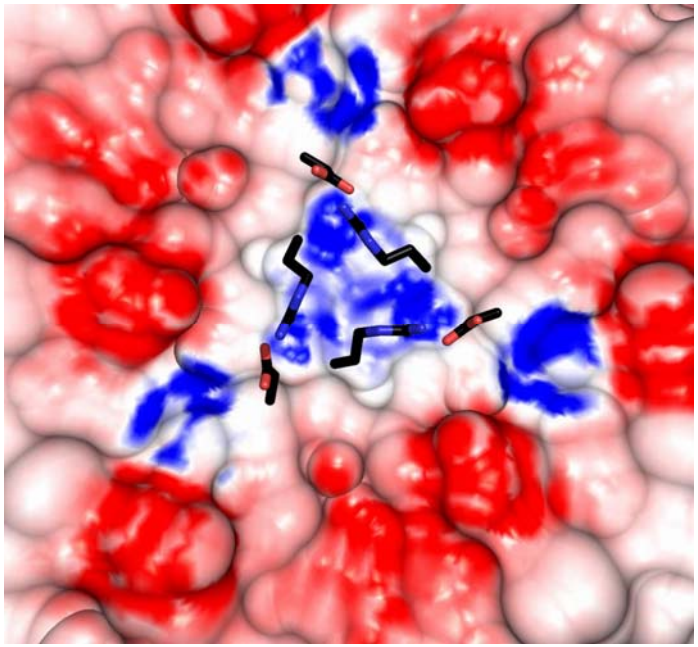
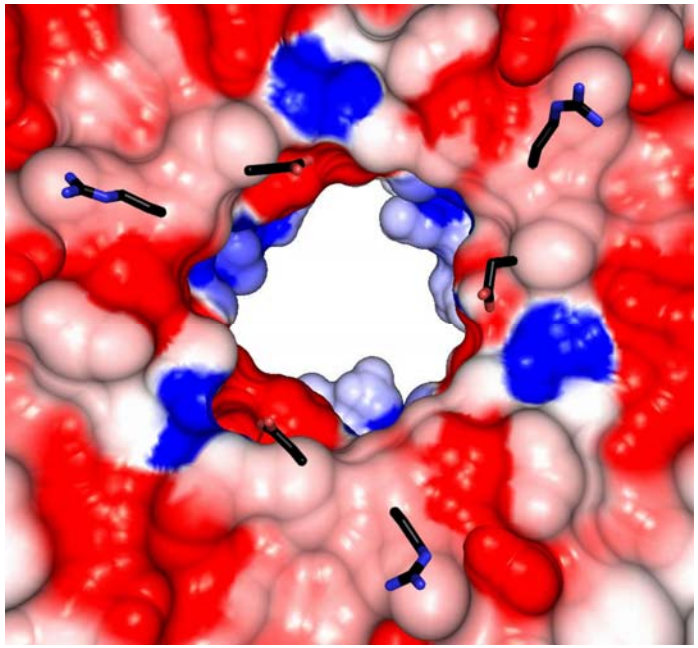


Figure 4

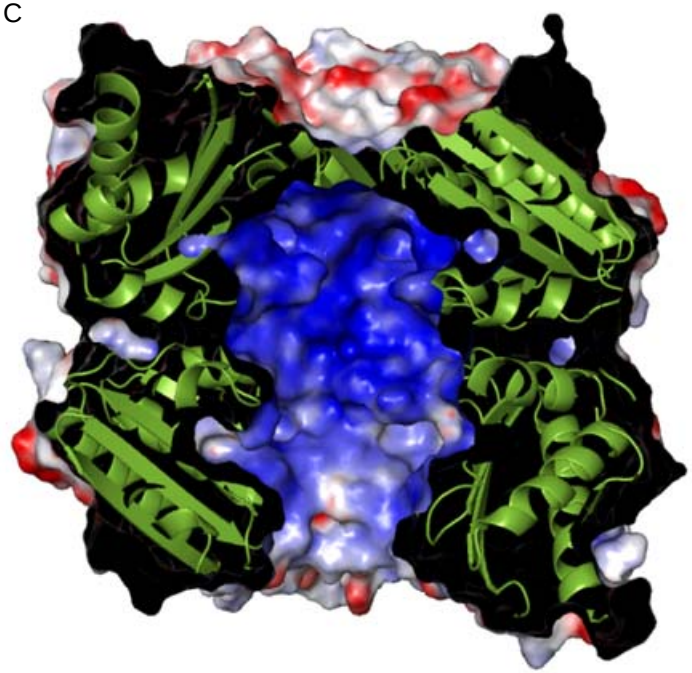
A



B



C



D

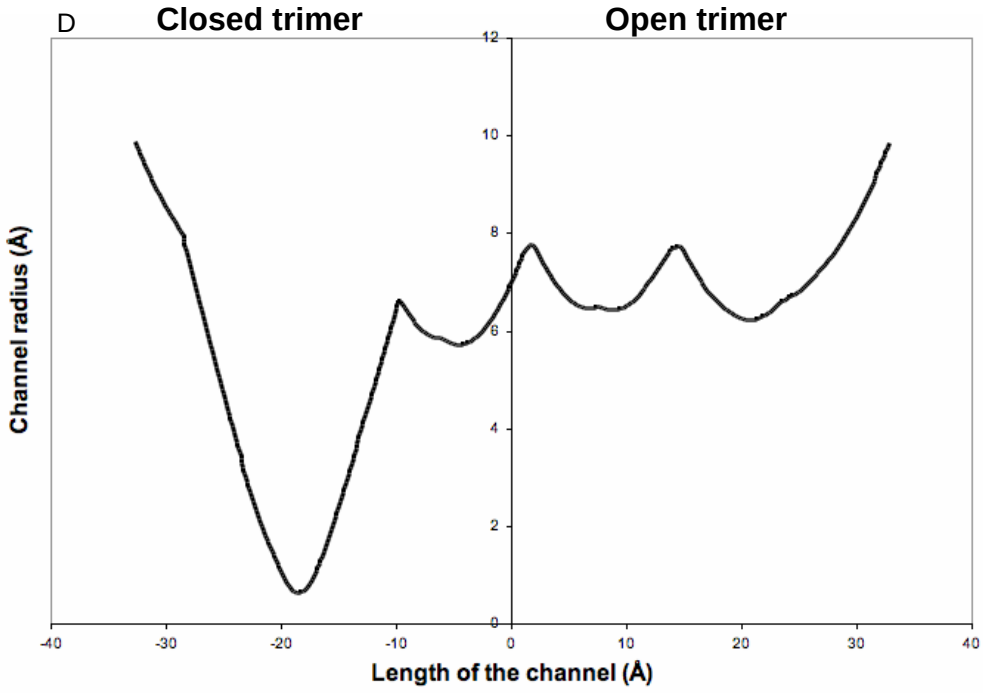
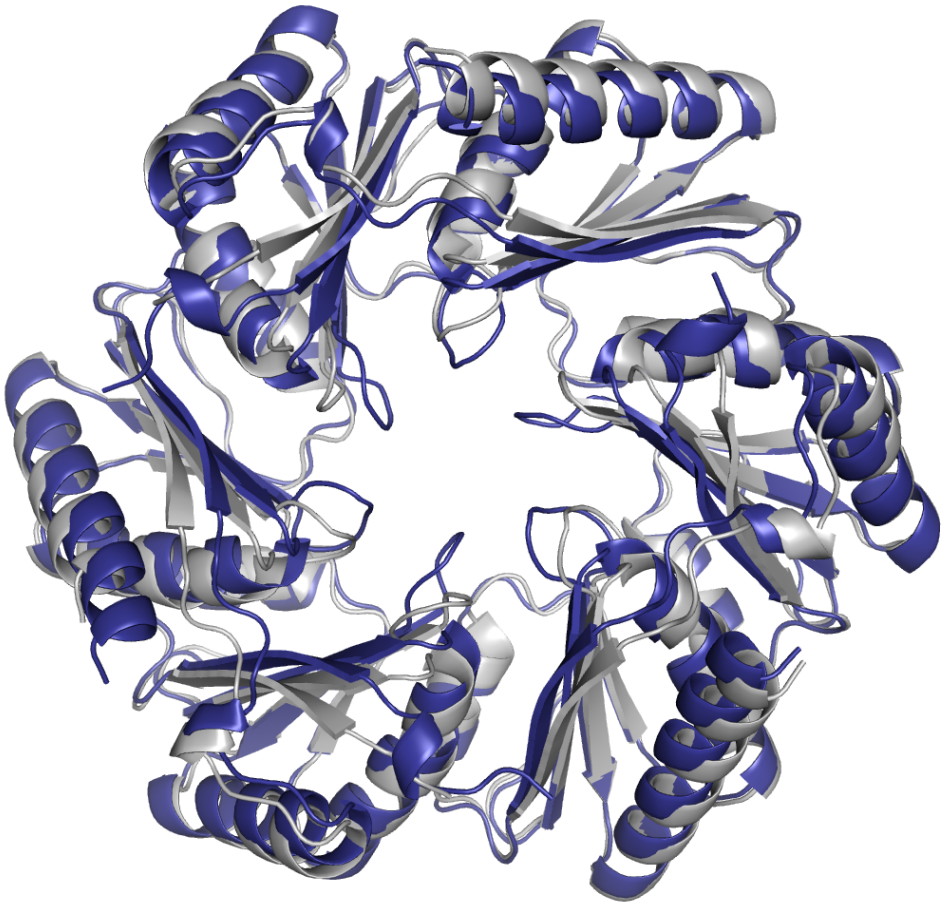


Figure 5

A



B

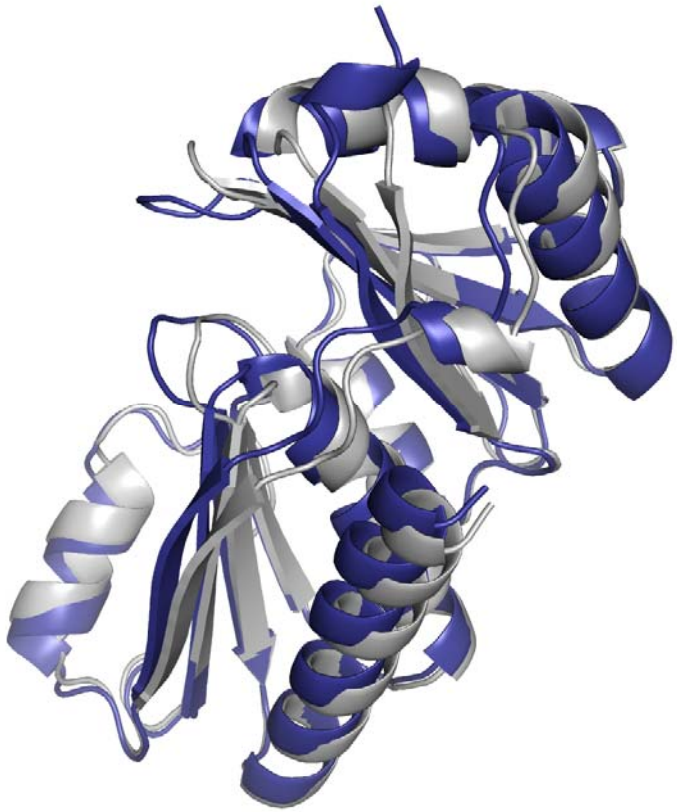


Figure 6

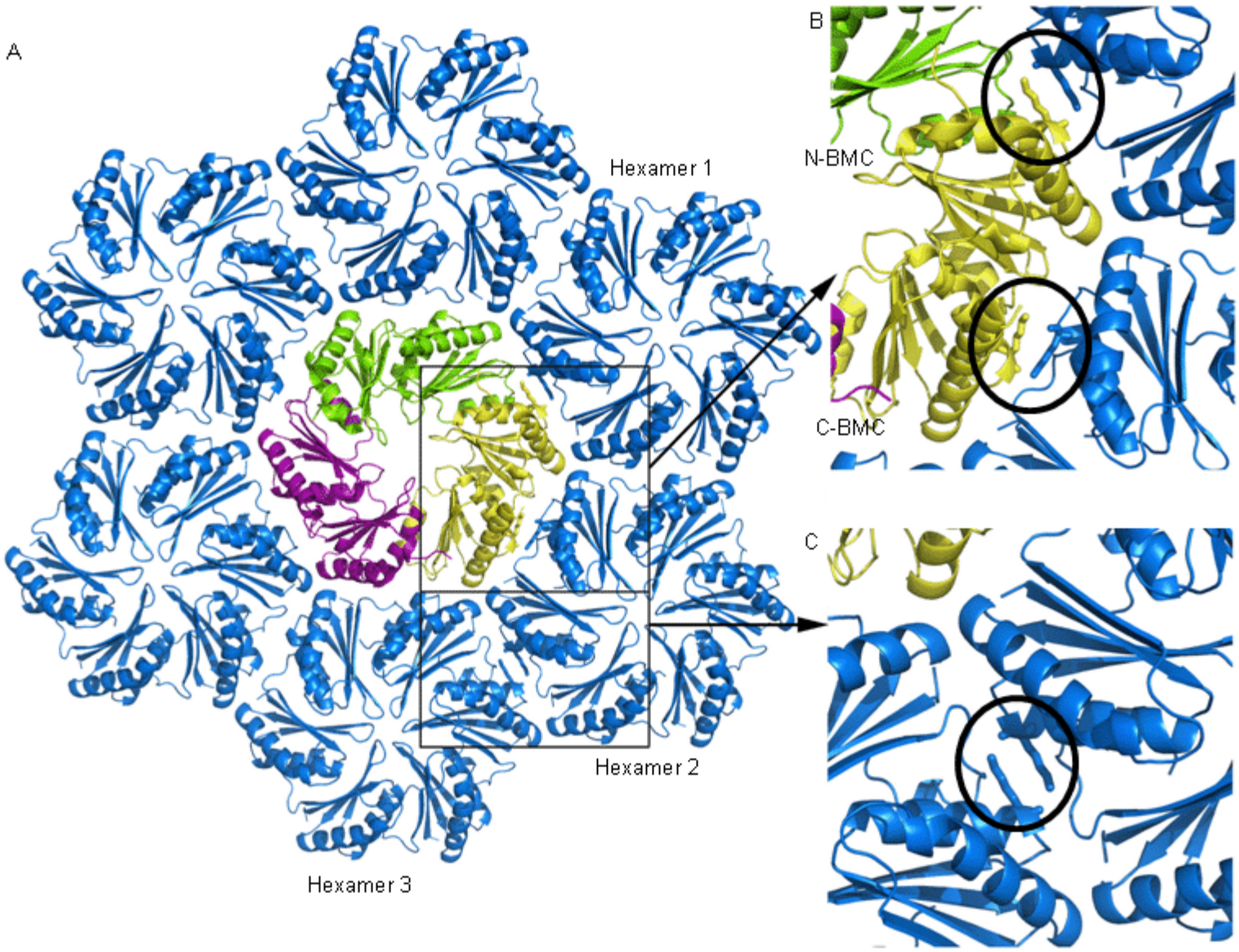


Figure S1

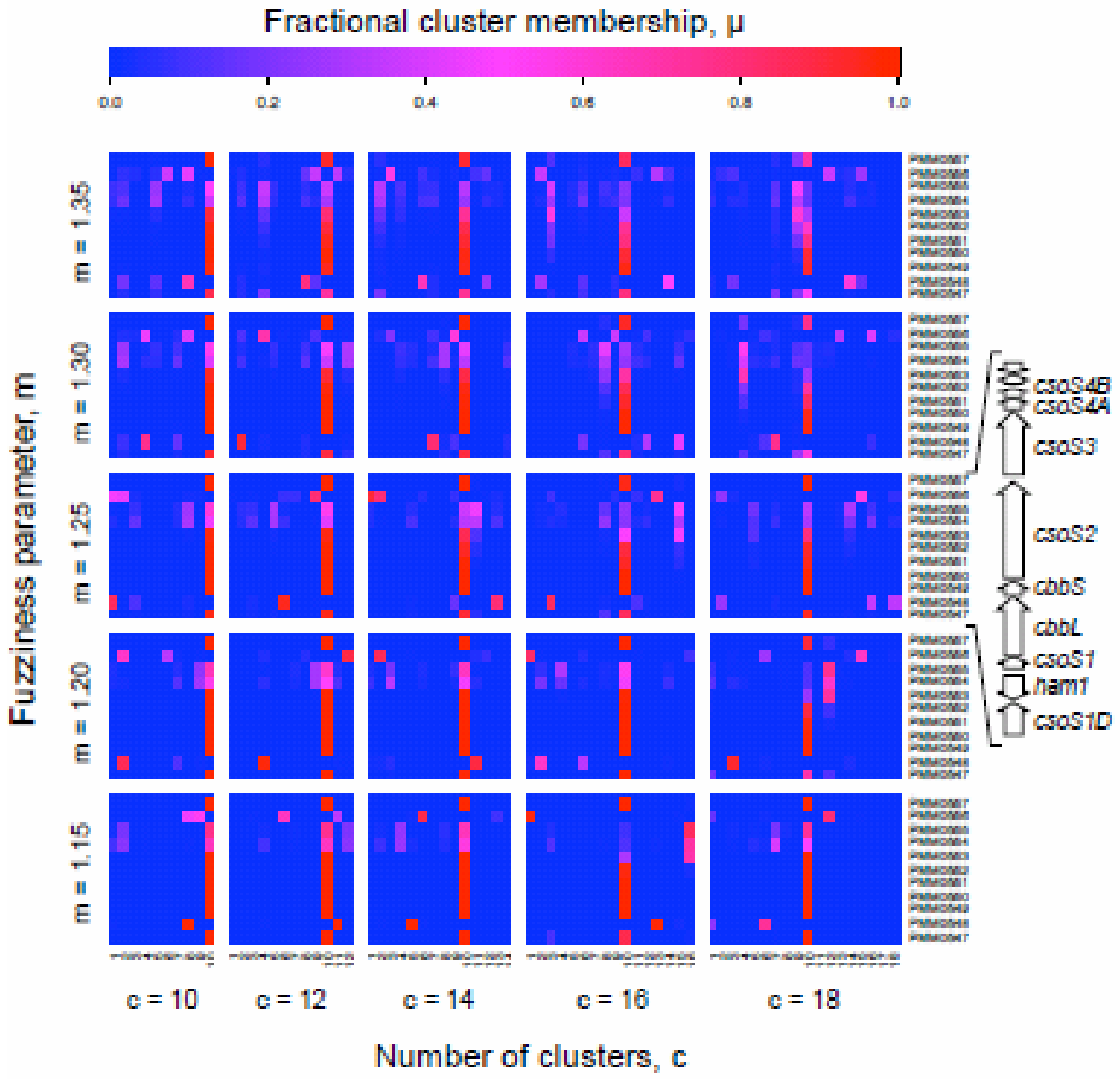

Fractional cluster membership, μ
0.0
0.2
0.4
0.6
0.8
1.0
Fuzziness parameter, m
m = 1.35
m = 1.30
m = 1.25
m = 1.20
m = 1.15
csoS4B
csoS4A
csoS3
csoS2
cbbS
cbbL
csoS1
ham1
csoS1D
c = 10
c = 12
c = 14
c = 16
c = 18
Number of clusters, c

Figure S2

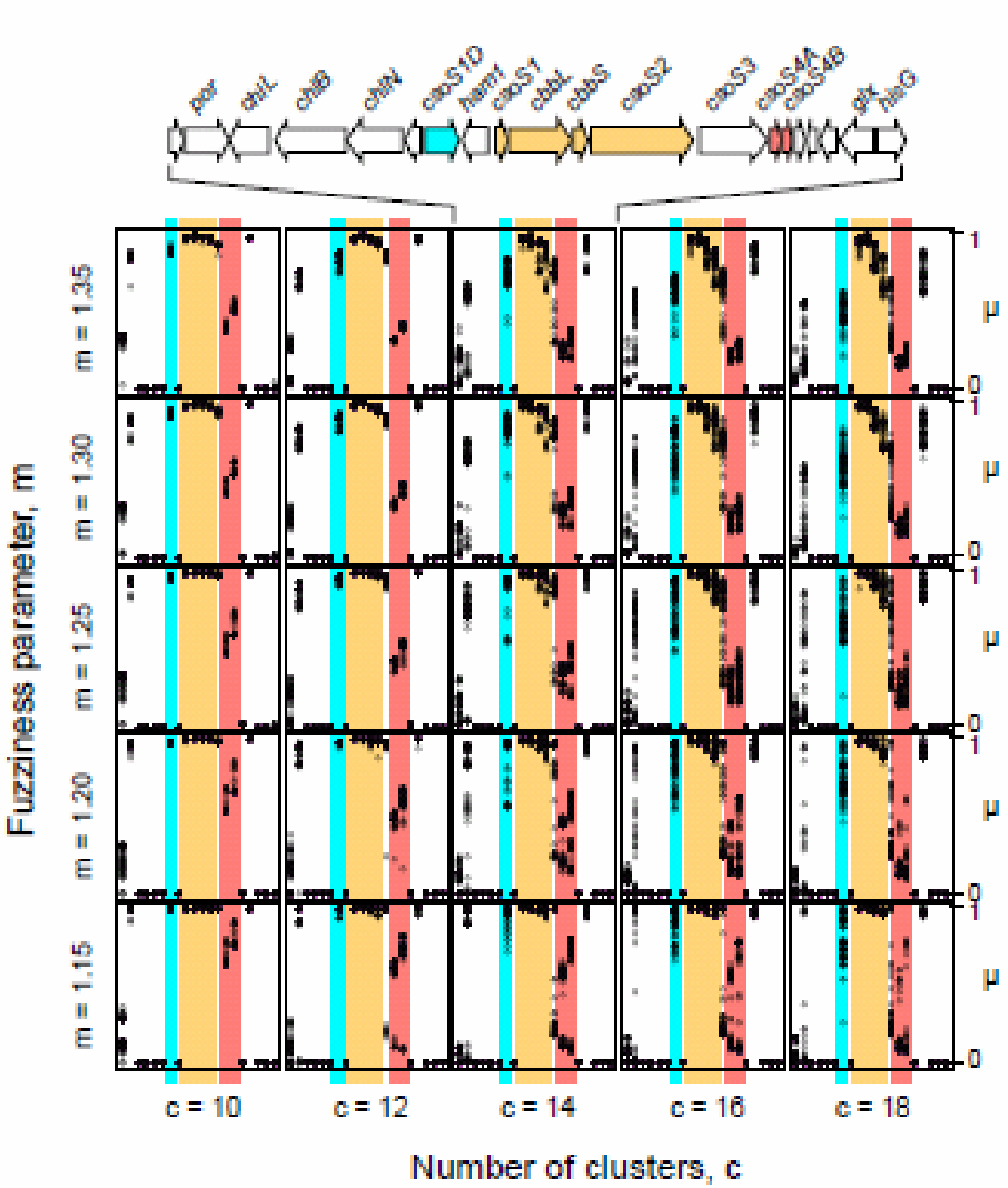

por
orf1
chlB
chlN
cbaS1D
hem1
cbaS1
cbbL
cbbS
cbaS2
cbaS3
cbaS4
cbaS4B
gtk
hisG
m = 1.35
m = 1.30
m = 1.25
m = 1.20
m = 1.15
Fuzziness parameter, m
1
0
μ
c = 10
c = 12
c = 14
c = 16
c = 18
Number of clusters, c

Figure S3