Supporting Online Material for

Atomic-Level Models of the Bacterial Carboxysome Shell
Shiho Tanaka, Cheryl A. Kerfeld, Michael R. Sawaya, Fei Cai, Sabine Heinhorst, Gordon C. Cannon, Todd O. Yeates*

*To whom correspondence should be addressed. E-mail: yeates@mbi.ucla.edu

DOI: 10.1126/science.1151458

This PDF file includes

Materials and Methods
SOM Text
Figs. S1 to S4
Tables S1 and S2
References
1. MATERIALS AND METHODS

CcmL Cloning, Expression, and Purification
The *Synechocystis sp. PCC 6803* CcmL gene was amplified from *Synechocystis sp. PCC 6803* genomic DNA using the following primers: Forward, ccm1NdeIfwd: 5’-GGGGCATATGCAACTTGCCAAAGTTTTGGGGACCG-3’ and reverse, ccm1NotIrev: 5’-GGGGGCGGCCGTCTCCCATCATCCCTCTTTTTGAGAG-3’ containing an NdeI (C^ATATG) and NotI (GC^GGCCGC) restriction sites respectively. PCR products were digested with NdeI and NotI, and ligated into the multiple cloning site of pET22b (Novagen), which adds a C-terminal His6 tag to the protein coding sequence. The sequences of expression vectors were confirmed via plasmid sequencing (DNA Sequencing Facility, University of California, Los Angeles). For expression of recombinant protein, the constructs were transformed into BL21-Gold(DE3) cells (Novagen).

The cells were grown to an OD<sub>600</sub> of ~0.6 at 37°C, then pelleted and resuspended in a growth medium containing selenomethionine (SeMet). The SeMet protein expression was induced with 1.0 mM isopropyl-D-thiogalactoside (IPTG) for 4 hours. Cells were harvested by centrifugation, and cell pellets were resuspended in 20mM Tris, pH 8.0, 300mM NaCl, 10% glycerol, 10mM Imidazole, 2mM BME, 2μg/mL DNase, and 1:100 protease inhibitor mixture (Sigma). Resuspended cells were lysed by lysozyme treatment and French press. Cell lysates were centrifuged at 30,000 g for 30 minutes, and the supernatant was loaded onto a Ni-NTA column (Qiagen) pre-equilibrated with 20mM Tris (pH 8.0), 300mM NaCl, 10% glycerol, and 10mM imidazole. The bound proteins were eluted with 300mM imidazole. The purified SeMet CcmL protein was verified by SDS gel electrophoresis. Fractions containing SeMet CcmL were then pooled and dialyzed at 4 °C for overnight against a buffer containing 20mM Tris (pH 8.0), 100mM NaCl, and 10% glycerol. The dialyzed sample was concentrated to a final concentration of 11mg/mL for crystallization.

OrfA Cloning, Expression, and Purification
The expression clone pOrfA-pET22b was generated by PCR amplification of the *orfA* sequence using the following primers: Forward, oAfNdeI: 5’-GGTGTCATATGAAAATCATGCAAGTTGAG-3’ and reverse, oArXhoIns: 5’-GGTCCTCGAGCTCACATTCC-3’ containing an NdeI and XhoI (C^TCGAG) restrictions site, respectively.
PCR products were digested with NdeI and XhoI, and ligated into the multiple cloning site of pET22b (Novagen), which adds a C-terminal His₆ tag to the protein coding sequence. The sequences of expression vectors were confirmed via plasmid sequencing (DNA Sequencing Facility, University of Maine). For expression of recombinant protein, the constructs were transformed into BL21(DE3) cells (Novagen).

OrfA expression was induced by the addition of IPTG to a culture grown at 37 ºC with shaking (225 rpm) while they had reached an OD₆₀₀ reading of 0.6. After adding IPTG, cell growth continued at 25 ºC with shaking (225 rpm) for 3 hours before the culture was harvested by centrifugation (10k rpm, 10 min). Cells collected from 1 liter culture were resuspended in 30 ml B-PER II Bacterial Protein Extraction Reagent (Pierce) and shaken for 10 min at room temperature. 600 μl freshly prepared 10 mg/ml lysozyme was added to the mixture, followed by another 5 min shaking prior to disruption (three 30 s pulses) with an Ultrasonics Sonicator (model W-220F, Branson). Cell lysates were used for recombinant protein purification by affinity chromatography on a Ni-NTA column (Invitrogen) according to the manufacturer’s protocol. Purified OrfA was concentrated and dialyzed in 10 mM Tris pH 8.0 to a final concentration of 2 mg/ml.

**Protein Crystallization**

Diffraction quality protein crystals of the SeMet CcmL protein were grown using the hanging drop vapor diffusion method, in 0.1M Bis-Tris, pH5.5, 0.1M NaCl, and 30% PEG3350 at room temperature. Crystals were observed within a few days. 30% glycerol was used as an additional cryoprotectant for data collection. CcmL crystals belong to space group P2₁ with unit cell dimensions a=64.43, b=107.44, c=72.26 α=γ=90.0, β=96.509, and two CcmL pentamers in the asymmetric unit. Similarly, OrfA protein crystals were grown using the hanging drop vapor diffusion method in 20% PEG 3350 and 0.2M sodium formate at room temperature. OrfA crystals belong to P2₁ with unit cell dimensions a=64.5, b=107.6, c=72.4, α=γ=90.00, β=96.51, and one OrfA pentamer in the asymmetric unit.

**CcmL Structure determination and refinement**

A standard three wavelength anomalous dispersion data set was collected on a CcmL selenomethionyl derivative at the Advance Light Source, beamline 8.2.2. An ADSC quantum 315 CCD detector was used to record the data. Data were processed using DENZO/SCALEPACK (S1). All 20 possible Se sites were identified with the programs SHELXD (S2) using only the anomalous differences from the peak wavelength. Initial phases were calculated with MLPHARE and later improved by density modification with DM (S3). Initial estimates of the NCS operators were obtained from the relationships between heavy atom positions. NCS averaging was accomplished in two phases. First two-fold averaging was applied between the two pentamers in the asymmetric unit. The resulting electron density map was then used to refine NCS operators for ten-fold symmetry averaging. The experimental electron density was of sufficient quality to be interpreted by eye, but automated chain tracing programs succeeded in tracing only a small portion of the chain. An initial trace was manually built using the graphics
program O (S4) and combined with segments from a model built by RESOLVE. The model was refined using conjugate gradient and simulated annealing algorithms as implemented by the program CNS (S5). Strong NCS restraints were used throughout. This model was further refined with REFMAC (S6), in order to introduce TLS parameters in the refinement. Later rounds of model building were performed with the graphics program COOT (S7).

This data set was strongly anisotropic, with diffraction limits of 2.5 Å along the a* direction, 2.3 Å along the b* direction and 2.7 Å along the c* direction. For this reason, data were truncated that fell outside an ellipse centered at the reciprocal lattice origin and having vertices at 1/2.5 Å, 1/2.3 Å, and 1/2.7 Å along a*, b*, and c*, respectively. The data was anisotropically scaled using the procedure described in Strong et al., 2006 (S8) then used for refinement with REFMAC.

After three rounds of refinement, the five chains were essentially complete, but R_work and R_free stalled at 41.7% and 44.8%. Strong difference density could be found in the crystal interstices of the NCS averaged maps. In unaveraged maps the residual density appeared to suggest that additional pentamers were partially overlapping the current model. It was reasoned that this residual density could be explained by a lattice translocation disorder. This diagnosis is consistent with the appearance of streaks in every third row of the reciprocal lattice and the observation of two peaks in the native Patterson map: a 70σ peak at UVW (0.500, 0.342, 0.500), and a 32σ peak at (0.00, 0.311, 0.00). Up to this point, the current model consisted of two pentamers: P0 and P1. P1 is related to P0 by the translation t1 = (0.5, -0.342, -0.5), corresponding to the 70σ Patterson peak. We reasoned that two additional pentamers P2 and P3 would be related to P0 by t2 = (0.5, +0.342, -0.5), t3 = (0.00, 0.311, 0.00). This hypothesis was verified by applying t2 and t3 in turn to the reference pentamer and observing that the newly positioned pentamers indeed accounted for the residual electron density. It was realized that the structure factors could be corrected to remove the disorder by following the procedures outlined in Wang et al. (S9). The only variables required for the correction are the relative translation vectors and occupancies of P0, P1, P2, and P3. The vectors (t1, t2, and t3) were determined by the procedure above. The occupancies were derived empirically by performing a series of rigid body refinements with the program REFMAC. The individual refinements differed in the relative occupancies of P0 and P1 relative to P2 and P3. The refinement with occupancies 0.75 for P0 and P1 and 0.25 for P2 and P3 yielded the lowest R_free and so were used in the structure factor correction for the lattice translocation disorder. R_work and R_free for the two-pentamer-model immediately dropped to 32.0% and 34.9% after the structure factors were corrected. The final R and Rfree values following refinement were 24.2% and 29.8%. The quality of the structure was substantiated by model validation programs, and by the subsequent higher resolution studies on the homologous protein, OrfA. The geometric quality of the model was assessed with the following structure validation tools: ERRAT (S10), PROCHECK (S11), and WHATIF (S12). PROCHECK reported 93.4% of the residues fall in the most favored region of the Ramachandran plot and 6.2% of the residues were in additionally allowed regions. ERRAT reported an overall quality factor of 98%. Protein structures were illustrated using the program PyMOL (S13).
OrfA Structural Determination and Refinement

X-ray diffraction data were collected at the Advanced Light Source beamline 8.2.2 using an ADSC Quantum 315 CCD detector. Data were collected at 100 K using a single crystal that was cryo-protected by a quick dip in a solution containing 75% reservoir and 25% ethylene glycol. The structure was solved readily by molecular replacement using the CcmL pentamer as a search model. Based on a sequence alignment sidechains were trimmed to the last common atom between CcmL and OrfA. The five subunits comprising the asymmetric were placed by the program Phaser (S14). The model was refined to (R_work=18.5% and R_free=23.9%) using REFMAC and visualized with COOT.

Structural analysis

DALI was used for structural similarity search (S15). Buried surface area was calculated with AREAIMOL (S16), which is distributed in the CCP4 suite of crystallographic programs (S3).

Icosahedral Modeling of Pentamers into Folded Hexagonal Layers

The pentamer-hexamer fitting calculations were performed on coordinates from both model organisms, Syn. 6803 and H. neapolitanus. For the Syn. 6803 case, the pentamer coordinates were from the CcmL protein reported here [2QW7], while the hexamer coordinates were of the CcmK1 protein. The structure of the hexameric shell protein CcmK2 was reported earlier (S17), but a higher resolution structure of the very closely related protein, CcmK1, was determined during the course of this work. The CcmK1 coordinates were therefore used for the modeling calculation. That structure will be reported elsewhere (Tanaka and Yeates, personal communication), but the coordinates have been deposited [3BN4]. For the H. neapolitanus calculations, the pentamer coordinates were of the OrfA protein reported here [2RCF] while the hexamer coordinates were of the CsoS1A protein [2EWH] (S18).

In order to fold up a layer of hexagons so as to leave a pentameric vacancy at the center, the following procedure was applied. Two adjacent hexamers were generated based on the packing observed in crystal structures of the hexameric proteins. Then a 41.8° rotation was applied in order to create a bend between the two hexamers; this is the angle between adjoining faces in an icosahedron. The axis of rotation was taken to be perpendicular to the six-fold axes of symmetry through the hexamers, and passing as nearly as possible, in the least squares sense, through those atoms lying at the interface between the two hexamers. The same operation was then repeated in order to generate subsequent hexamers. The resulting group of five hexamers is arranged as required about a central pentameric hole. The procedure above was repeated with rotations of either +41.8° or -41.8° between adjacent hexamers, leading to two hexamer groupings, curved either inward or outward, for each species. The constructions described above make specific assumptions about how the hexameric layers should be folded up, i.e. by bending only between the hexamers. Whether this is an accurate approximation is not known at
In order to simplify the fitting calculations, the coordinate sets were first moved into canonical starting orientations and positions. The central pentamers (i.e. CcmL and OrfA) were put in starting orientations in which their axes of symmetry were along the z-axis, and their centers of mass were at the origin. For convenient viewing, they were further twisted so that the bottom edges of the pentamers were roughly horizontal when viewed in the x-y plane. The groupings of five hexamers were oriented and positioned in the same way. In addition, a second orientation of each pentamer, flipped upside-down, was generated by rotating 180 degrees about the y-axis.

A search was then conducted for the optimal fit of the pentamers into the pentameric holes in the centers of the hexamer groupings. The calculations were performed for both species, and for all four possible combinations created by the uncertainty regarding which side of the pentamers should be up, and which way the hexamer layers should be curved. In each search, the pentamer was rotated and translated about the z-axis. Owing to the tightness of the fit, it was only judged necessary to perform rotations with a range of -12° to 12° (sampled at 2° intervals), and translations with a range of 20 Å (sampled at 1 Å intervals). Candidate solutions were obtained by evaluating packing energies according to Rosetta-Dock ([S19]), using ellipsoidal representations of the amino acid residues. This level of detail was judge to be most appropriate owing to various uncertainties, for example regarding the exact identities of the homologous shell proteins that come into contact at the vertices, and the way in which the hexagonal layers should be folded at the vertices, as noted above.

In order to obtain additional measures of fit, the Rosetta-Dock energies were supplemented with calculations of surface complementary ([S20]) and buried surface area ([S16]). These values were calculated in the vicinity (e.g. +/- 4° and +/- 2 Å) of the high scoring docking solutions. It should be noted that the docking and surface calculations were only performed on a subset of the coordinates. Owing to the inherent symmetry of the system, it was only necessary to include one of the five hexamers, the others being essentially redundant. In addition, small shifts (+/- 2 Å) of the pentamer were allowed perpendicular to the z-axis, either towards or away from the hexamer. This variation was to allow for uncertainty in the actual spacing of the hexamers in the pentameric grouping. This was taken to be a point of minor uncertainty, since slightly different spacings of hexamers arise from slightly different choices of the position of the 41.8° rotation axis between hexamers, which was chosen heuristically as described above.

The results of the fitting calculations are given below. The model names refer to the species (either Syn. 6803 or H. neapolitanus) and the up vs down orientation of the hexamer layer and the pentamer. The model numbers correspond with those in the text.

<table>
<thead>
<tr>
<th>Model</th>
<th>Rosetta-Dock</th>
<th>Surf_complem.</th>
<th>Area Buried</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_syn</td>
<td>-35.62</td>
<td>0.496</td>
<td>996</td>
</tr>
<tr>
<td>2_syn</td>
<td>-35.90</td>
<td>0.387</td>
<td>937</td>
</tr>
</tbody>
</table>
*Areas are in square Angstroms and represent the area buried between the pentamer and one hexamer.

The results from the two species show a similar trend, but differ in detail. Greater confidence was placed on the Syn. 6803 modeling because the detailed packing of the hexameric layer in that case has been validated by multiple dependent crystal structures (CcmK1, CcmK2, and CcmK4). Based on the scores and on visual inspection, solution 4 was ruled out. Solution 3 was judged to be unlikely, whereas solutions 1 and 2 were both plausible and difficult to discriminate. Solution 1 places the N and C termini of the hexameric proteins on the interior of the shell and the termini of the pentameric proteins on the exterior. Solution 2 places the pentamers in the same orientation and the hexamer layer in the opposite orientations.

Candidate solutions 1 and 2 were used to generate complete icosahedral shells. The sizes of carboxysomes vary, and based on size ranges reported by Jensen, et al. (S21), triangulation numbers for carboxysome shells could vary from about 75 to 108. In this range, icosahedra with simple geometries – i.e. with icosahedral edges running either along or perpendicular to component hexagons – can be obtained with T=75, 81, and 108. We constructed models of the smallest plausible carboxysome cage using T=75 (h=5, k=5), based on the CcmL and CcmK1 structures from Syn. 6803.

The complete icosahedral shells were constructed from the optimal pentamer-hexamer packing solutions, and from the hexamer-hexamer packings that have been established by multiple crystal structures of hexagonal layers. Perfect icosahedral symmetry was imposed. Each triangular face of the T=75 shell contains 37 hexamers, one-third of which (i.e. 12 and one-third hexamers or 74 subunits) are unique. Minor defects are evident at the edges of the icosahedra (Figure 4c). The simple construction model adopted here introduces an abrupt bend at the edges of the icosahedron. This is likely to deviate appreciably from the real situation; detailed information is lacking on the structure of the edges of the carboxysome.

2. SUPPORTING TEXT

Structure of CcmK1
The structure of CcmK2 was determined previously at a resolution of 2.9 Å (S17). The homologous protein CcmK1 is 93% identical in sequence to CcmK2. The crystal structure of CcmK1 was determined here at a higher resolution of 2.0 Å, and was used for
modeling the CcmL (pentamer) – CcmK (hexamer) assembly because of its higher resolution. The structure of CcmK1 is highly similar to CcmK2. The details of that structure will be reported elsewhere (Tanaka and Yeates, personal communication), but the coordinates of CcmK1 have been deposited (PDB code 3BN4).

**Pore radii**
The pore radii for the CcmL and OrfA pentamers were determined as follows. The distance between atom centers across the pore was measured, and the vanderWaals radii of those atoms (~3.4 Å) were subtracted. In the intact shell, these pores are not expected to be important for molecular diffusion, owing to their small size and small number.

**Protein spacings in the hexagonal layer**
The spacings in the CcmK hexagonal layers have been established here and in earlier studies (S17) to be very nearly equal to 70 Å between the centers of adjacent hexagons sharing an edge. The edge length of a hexagon is given by this distance times the square root of 3, divided by 3, which is 40.4 Å.

**The number of shell subunits in the carboxysome**
The number of shell subunits in a carboxysome depends on its size, which varies between organisms and even within individual cells. In the present work, the number of subunits is based on a carboxysome of about 1100 Å - 1200 Å in diameter, which would have a triangulation number between about 75 and 80 (S18, S21). The number of subunits (including hexameric and pentameric subunits) in an icosahedral shell is given by 60*T. This would be 4500 subunits for a T=75 shell. Carboxysomes of the beta type may be even larger than this, while smaller carboxysomes are observed in H. neapolitanus (S22, S23), which contains alpha type carboxysomes. Experimental estimates of the number of shell subunits in that organism are around 3500 (S22, S23). Based on the reported spacing between hexagonal subunits (S17, S18), this would imply a shell having a diameter of around 850 Å, which is within the range observed by electron microscopy. Finally, the models presented here presume that no other proteins constitute a major component of the outer shell itself. This requires that the hexagonal layers extend all the way to the edges of the icosahedral shell. The possibility cannot be ruled out at the present time that other measurable but minor constituents of the carboxysome, such as the carbonic anhydrase, might form part of the shell, including the edges.

3. SUPPORTING TABLES

**Table S1. CcmL diffraction data collection and refinement statistics**

<table>
<thead>
<tr>
<th>Data Collection</th>
<th>CcmL (peak)</th>
<th>CcmL (inflection)</th>
<th>CcmL (remote)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>P2₁</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unit cell dimensions (Å)</td>
<td>a=64.5, b=107.6, c=72.4, α=γ=90.0, β=96.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.97944</td>
<td>0.979617</td>
<td>0.975593</td>
</tr>
<tr>
<td>Resolution limit (Å)</td>
<td>19.8 - 2.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R_{sym} (%)</td>
<td>9.4 (19.1)</td>
<td>9.6 (20.7)</td>
<td>9.3 (23.9)</td>
</tr>
<tr>
<td>I/σ(last shell)</td>
<td>2.5</td>
<td>2.2</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>174,955</td>
<td>84,734</td>
<td>172,570</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>Total observations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique reflections (Bijvoet separate)</td>
<td>52,537</td>
<td>26,627</td>
<td>52,314</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>98.0 (84.2)</td>
<td>97.3 (80.7)</td>
<td>97.1 (79.1)</td>
</tr>
</tbody>
</table>

**Phase Determination**

- **Rcullis** (%) (20-2.4 Å, acentric/centric, isomorphous):
  - 0.98/0.96
- **Rcullis** (%) (20-2.4 Å, anomalous):
  - 0.93
- Phasing power (20-2.6 Å, acentric/centric):
  - 0.30/0.21

**Model Refinement**

- Rwork (54 Å-2.4 Å):
  - 24.2 %
- Rfree (54 Å-2.4 Å):
  - 29.8 %
- Number of residues (protein/water):
  - 961/90
- Average B (Å²) (main chain/side chain):
  - 32.0/32.2
- Rmsd bonds (Å):
  - 0.018
- Rmsd angles (°):
  - 2.38
- Rmsd B-values (Å² bonded):
  - 2.68

---

*Rsym = Σ|I-I>|²/ΣI²*

*The inflection data set was treated as a reference for phasing. Statistics are reported to 3.3Å resolution.

*RCullis = Σε/Σ|FPH-FP|, where ε=lack of closure.

*RCullis = Σε/Σ|F⁺-F⁻|, where ε=lack of closure.

*Phasing power = <FH/ε>.

*Rwork = Σ|Fobs-Fcalc|/ΣFobs

*Rfree = Σ|Fobs-Fcalc|/ΣFobs, where all reflections belong to a test set of 5% randomly selected data.

Numbers in parentheses refer to values calculated over just the highest resolution shell.

---

Table S2: OrfA diffraction data collection and refinement statistics

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Collection</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Space group</td>
<td>P2₁</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unit cell dimensions (Å)</td>
<td>a=61.9, b=86.8, c=35.2, α=γ=90, β=90.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution limit (Å)</td>
<td>43.4-2.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rsym(%)</td>
<td>15.6 (41.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I/σ(last shell)</td>
<td>3.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total observations</td>
<td>143760</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique observations</td>
<td>20056</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completeness (% last shell)</td>
<td>99.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Model Refinement</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rwork(54 Å-2.4 Å)</td>
<td>18.4 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter</td>
<td>Value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------------------------------------------------</td>
<td>---------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R_{\text{free}}$ ($54 \text{ Å} - 2.4 \text{ Å}$)</td>
<td>24.3 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of residues (protein/solvent)</td>
<td>408/273</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average $B$ (Å$^2$) (main chain/side chain)</td>
<td>15.1/19.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rmsd bonds (Å)</td>
<td>0.006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rmsd angles (°)</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rmsd $B$-values (Å$^2$ bonded)</td>
<td>0.78</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^aR_{\text{sym}} = \Sigma |I - <I>|^2 / \Sigma I^2$

$^bR_{\text{work}} = \Sigma |F_{\text{obs}} - F_{\text{calc}}| / \Sigma F_{\text{obs}}$

$^cR_{\text{work}} = \Sigma |F_{\text{obs}} - F_{\text{calc}}| / \Sigma F_{\text{obs}}$, where all reflections belong to a test set of 5% randomly selected data.

Numbers in parentheses refer to values calculated over just the highest resolution shell.
Figure S1. Diagram showing the two ways a sheet of hexamers can be folded up, after removing one sector, to leave a pentagonal opening. The two alternate scenarios lead to different inside vs. outside orientations for the hexamer sheet. The hexagonal proteins shown are CcmK1 from Syn. 6803. Note the size of the pentagonal opening, with edges of approximately 40 Å, is compatible with insertion of a shell pentamer (i.e. CcmL), whose pentagonal base has an edge of about 42 Å at its widest point.
Figure S2. Contour plots of the calculated energies for fitting the CcmL pentamer into the grouping of five CcmK1 hexamers. The highest contours correspond to the lowest (most favorable energies). The three plots represent different up vs. down orientations of the pentamer and hexamers. Plot (A) corresponds to model 1 in the text, (B) is model 2, and (C) is model 3. The three plots are contoured the same.
Figure S3. Calculated electrostatic potentials (blue indicating positive and red indicating negative) for the four known carboxysome hexamer structures (S17, S18) and the two carboxysome pentamers presented here (CcmL and OrfA). In each panel, one of the amino acid residues lining the pore is highlighted. Figures colored using the Ezprot program suite (S24).
Figure S4. Illustration of the hydrophobic character of the two sides of the carboxysome hexamers (blue indicating polar and red indicating highest hydrophobicity). The N and C-termini of the protein chains both sit on the side labeled 2 in this figure. Some of the differences on side 2 between homologues are due to different lengths and positions of the C-terminal alpha helix. Figures colored using the Ezprot program suite (S24).
5. REFERENCES FOR SUPPORTING MATERIAL

S20. M. C. Lawrence, P. M. Colman, J Mol Biol 234, 946 (Dec 20, 1993).