Theoretical Study of the Initial Stages of Self-Assembly of a Carboxysome's Facet

J. P. Mahalik,[*,†] Kirsten A. Brown,[‡] Xiaolin Cheng,[¶,§] and Miguel Fuentes-Cabrera[*,‖,⊥]

Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN-37831, Chemistry Department, Mercer University, 1501 Mercer University Drive, Macon, GA-31207, UT/ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, P. O. Box 2008, Oak Ridge, TN-37831, Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, M407 Walters Life Sciences, 1414 Cumberland Avenue, Knoxville, TN-37996, Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, TN-37831, and Computing and Computational Sciences Directorate, Oak Ridge National Laboratory, TN-37831

E-mail: mahalikjp@ornl.gov; fuentescabma@ornl.gov

Abstract

Bacterial microcompartments, BMCs, are organelles that exist within certain type of bacteria and act as nano-factories. Among the different types of known BMCs, the

[*]To whom correspondence should be addressed

[†]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN-37831

[‡]Chemistry Department, Mercer University, 1501 Mercer University Drive, Macon, GA-31207

[¶]UT/ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, P. O. Box 2008, Oak Ridge, TN-37831

[§]Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, M407 Walters Life Sciences, 1414 Cumberland Avenue, Knoxville, TN-37996

[‖]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, TN-37831

[⊥]Computing and Computational Sciences Directorate, Oak Ridge National Laboratory, TN-37831

carboxysome has been studied the most. The carboxysome plays an important role in the light-independent part of the photosynthesis process, where its icosahedral-like proteinaceous shell acts as a membrane that controls the transport of metabolites. Although a structural model exists for the carboxysome shell, it remains largely unknown how the shell proteins self-assemble. Understanding the self-assembly process can provide insights into how the shell affects the carboxysome's function and how it can be modified to create new functionalities, such as artificial nanoreactors and artificial protein membranes. Here, we describe a theoretical framework that employs Monte Carlo simulations with a coarse grain potential that reproduces well the atomistic potential of mean force; employing this framework, we are able to capture the initial stages of the 2D self-assembly of CcmK2 hexamers, a major protein-shell component of the carboxysome's facet. The simulations reveal that CcmK2 hexamers self-assemble into clusters that resemble what was seen experimentally in 2D layers. Further analysis of the simulation results suggests that the 2D self-assembly of carboxysome's facets is driven by a nucleation-growth process, which in turn could play an important role in the hierarchical self-assembly of BMCs shells in general.

Keywords

Carboxysome, Protein Self-assembly, Potential of Mean Force, All-atomistic, Coarse grain model, Nucleation-growth

Bacterial micro-compartments (BMCs) are organelles that exist within cyanobacteria and enteric bacteria.[1] BMCs have a proteinaceous shell that encapsulates several enzymes that perform particular reactions. Three types of BMCs have been studied in detail: the carboxysome (CB), the propanediol- (Pdu) and the ethanolamine-utilization (Eut). The CB has been studied the most and is considered the prototype BMC.[1]

Two types of CBs, known as α and β, exist and they are predominantly present in oceanic and freshwater cyanobacteria, respectively.[2] Both have nanometer (nm) size icosahedral-like shells comprised of proteins that self-assemble into hexagons (referred to as hexamers from now on) and pentagons. The hexamers, which are found in the domain Pfam00936, form the facets of the icosahedron; the pentagons form the vertices, and they are found in the domain Pfam03319.[3] The shell encapsulates the enzymes carbonic anhydrase (CA) and ribulose biphosphate carboxylase oxygenase (RuBisCO).

The CB acts as a nano-factory inside of cyanobacteria, playing a major role in the light-independent part of the photosynthesis process by fixing CO_2 in its interior as follows: cytoplasmic bicarbonate (HCO_3^-) and ribulose-1-5-biphosphate (RuBP) enter the CB through pores in its shell; CA subsequently converts (HCO_3^-) into CO_2, which RuBisCO combines with RuBP to produce 3-phosphoglycerate (3-PGA); this product then exits the CB through its pores. By performing CO_2 fixation inside of the CB, RubisCO's efficiency is enhanced, photorespiration is prevented, and CO_2 leakage is ameliorated. The shell of the CB thus acts as a membrane that controls the transport of molecules in and out of the CB.

Consequently, much experimental work has been devoted to understanding the structure of the CB shell. Despite these efforts, many questions remain open. For instance, while it is known that several types of proteins form hexamers, the composition of the shell remains unknown;[2] while it is known that the hexamers have a concave and a convex side,[4] it is unclear how they are oriented in the CB shell, *i.e.* which side faces the cytosol and which the interior of the CB; while it has been observed that hexamers self-assemble into a 2D layer,[5,6] whether self-assembly occurs spontaneously or is driven by a nucleation-growth

process is unknown. Answering these questions is crucial for understanding how the shell affects the CB's function and how it can be modified to create new functionalities such as artificial nanoreactors and artificial protein membranes. Modeling and simulation can shed light on these questions but, to the best of our knowledge, they have not yet been used to investigate the CB shell.

The absence of simulations studies is probably a consequence of the CB's large size and the resulting complexity in its 3D self-assembly. Modeling 3D self-assembly of a system comprised of thousands of proteins and a diameter of approximately 100 nm is simply too demanding. To overcome this limitation, our strategy is to first focus on parts of the shell that are structurally relevant and computationally treatable. The facets fulfill these requirements because they constitute most of the shell's surface and their self-assembly, although still computationally expensive, can be studied in 2D. In this work, we use modeling techniques to investigate the 2D self-assembly of a proteins known as CcmK2, which belongs to the Pfam00936 domain and exists in the form of a hexamer in physiological conditions.

CcmK2 was crystallized by Kerfeld *et al.*[4] and was found to be composed of six protein subunits that form a hexamer with a central pore of about 7 Å in radius. The hexamers, in turn, self-assemble into a 2D layer where they are all uniformly oriented, *i.e.* in a layer all the hexamers have the same side (convex or concave) oriented in the same direction. These layers were suggested to represent the facets of icosahedral shell of β-carboxysomes. Dryden *et al.*[5] supported this hypothesis by showing that CcmK2 hexamers deposited on a lipid monolayer arrange themselves into a 2D layer with hexagonal symmetry. Thus, CcmK2 is a crucial component of the facets of the shell. Motivated by this, we investigate the initial stages of 2D self-assembly of CcmK2 hexamers employing Metropolis Monte Carlo simulations with a coarse-grained potential. The coarse-grain potential was selected on the basis of its potential of mean force (PMF) fit with that obtained from atomistic umbrella sampling molecular dynamics (MD) simulations for the binding of two hexamers. We find that the self-assembly is driven by a nucleation process, and further determine the minimum

size of the nucleus and the minimum concentration of hexamers needed for the self-assembly to occur.

RESULTS

To reduce the computational cost associated with investigating the kinetics of the 2D self-assembly of CcmK2 hexamers, it is convenient to use other methods besides atomistic molecular dynamics (MD). Atomistic MD describes the atomic interactions in a precise manner and allows for tracking the movements of all individual atoms in the system in real time. However, atomistic MD will be impractical for the study of the CcmK2 self-assembly. Coarse-grained (CG) Monte Carlo (MC) simulations are better suited for our purpose. In a CG method, a group of atoms are described as a bead instead of individual atoms, which reduces the number of computed interactions; in a MC simulation, the system evolves to locate its most energetically stable configuration. In this work we use a MC technique that employs a CG potential.

In the literature, there are many CG potentials available for representing proteins.[7–26,26,27] The question is, which one represents a CcmK2 hexamer and describe the hexamer-hexamer interactions the best? Our first task was to answer this question. To do this, we compared the interaction energy profiles, *i.e.* the PMF, of a system composed of two CcmK2 hexamers obtained with CG potentials with the corresponding PMF obtained from all-atom umbrella sampling MD simulations. Among the large list of available CG potentials, we chose knowledge-based potentials that were generated self-consistently. Two such CG potentials were evaluated here, and the potential that reproduced the atomistic PMF better was chosen for investigating, in conjunction with a MC formalism, the kinetics of 2D self-assembly of a collection of CcmK2 hexamers.

Umbrella sampling MD simulations were used to compute the PMF that results from separating the two hexamers in a dimer. Atomistic PMFs have been calculated before,[28,29,29–32]

and among the different methods available we used that employed in Ref.[29] In Ref.,[29] to avoid the high conformational entropy associated with proteins, constraints were imposed to the rotation, the mean square deviation, and the CoMs of the proteins. To compute the PMF, the reaction coordinate was taken as the distance between the CoMs of both hexamers. The initial configurations for individual umbrella windows were generated by pulling one hexamer gradually away from the other, which was kept immobile by fixing its CoM. A total of 42 sampling windows were used with a window width of 0.5 Å and a force constant of 25 kcal/(mol Å^2) so as to ensure sufficient overlap between neighboring windows. Each initial configuration was equilibrated for 1 ns followed by a 2 ns production umbrella run. In all the simulations, harmonic restraints (with force constant of 25 kcal/(mol Å^2)) were applied to five selected Cα atoms of the fixed hexamer to prevent translational and rotational motion. The Weighted Histogram Analysis Method (WHAM) was used for the PMF calculation and the error bars were generated with bootstrap analysis[33]

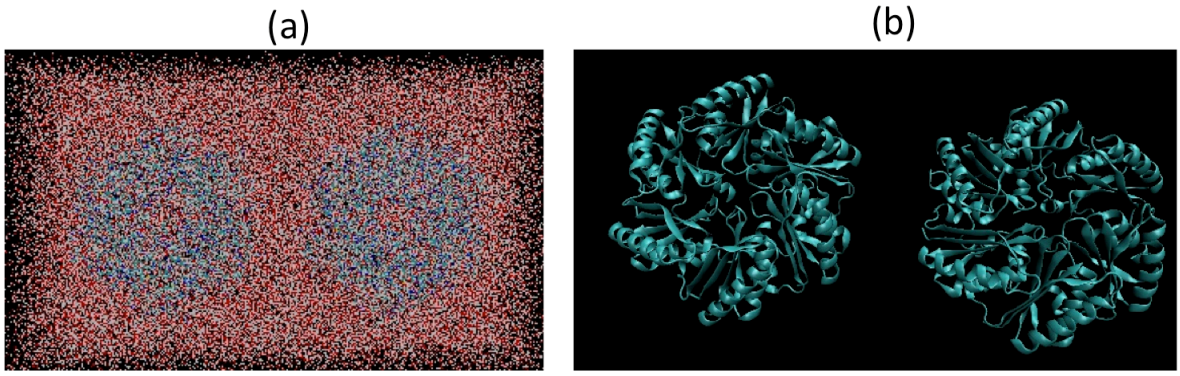

Figure 1: (a) Hydrated and (b) ribbon view of a dimer made of two CcmK2 hexamers.

For coarse-graining a CcmK2 hexamer, two knowledge-based statistical potentials were evaluated: the Miyazawa and Jernigan[34] (MJ) and the Thomas and Dill (TD).[35] Both restrict the movement of residues and solvent to a lattice and only take into account short-range nearest-neighbor interactions between the residues and solvent molecules.

To compute the PMF with both the MJ and the TD potentials, a dimer composed of two CcmK2 hexamers with their CoM at 86.5 Å was created. As in the case of the atomistic

PMF, the separation between their CoMs was taken as the reaction coordinate. Each CcmK2 hexamer was modeled as a rigid body, which means that the distance between every pair of residues within a hexamer remained fixed throughout the simulation. This assumption was justified because the atomistic calculations revealed that the hexamers are very stable irrespective of the distance between their CoMs (the RMSD of each individual CcmK2 hexamer was found to be about 1 Å only in 2 ns simulation time window). Moreover, as evidenced by the AFM experiments,[6] at large timescales the hexamers maintain their structural integrity. Conveniently, this assumption also sped up the calculations, since the interactions among the residues within a CcmK2 hexamer did not have to be computed. The PMF of two CcmK2 hexamers at a given separation distance and a fixed relative orientation was calculated by summing up all the pairwise interaction potentials between their corresponding residues. PMFs were computed by decreasing the distance between the hexamers from 86.5 to 66 Å in decrements of 0.2 Å.

The atomistic and CG PMFs are shown in Figure 2. In the atomistic PMF, the hexamers do not attract until they are 74 Å apart, repel each other below 68 Å, and attract most strongly at 70 Å with a strength of 2 k_BT. Of the two CG potentials considered here, the TD reproduces the atomistic PMF better. The shape, depth and range of interaction of the TD's PMF are all similar to the atomistic ones; the major difference is the strength of the repulsion, which is stronger in the TD than in the atomistic PMF. With the MJ potential, the PMF has a much shallower depth of only 0.8 k_BT. These results indicate that of these two potentials, TD is a better CG model for studying CcmK2 hexamer self-assembly.

To determine the PMF's dependence on the relative orientation, the PMFs were calculated with the TD potential for different relative orientations of the two hexamers. For this, one hexamer was kept fixed while the other was rotated around the y-axis by 10, 20, 30, 40 and 50 degrees (larger angles were not necessary since, for symmetrical reasons, a rotation of 0 and 60 degrees are equivalent). For each angle, the PMF was calculated by varying the distance between the CoMs of the hexamers. The PMF results revealed that the most stable

orientation occurs at a relative angle of 0 degrees. Increasing the angle lead to decrease in the PMF minimum due to fewer number of favorable contacts between the hexamers, as shown in Figure 3.

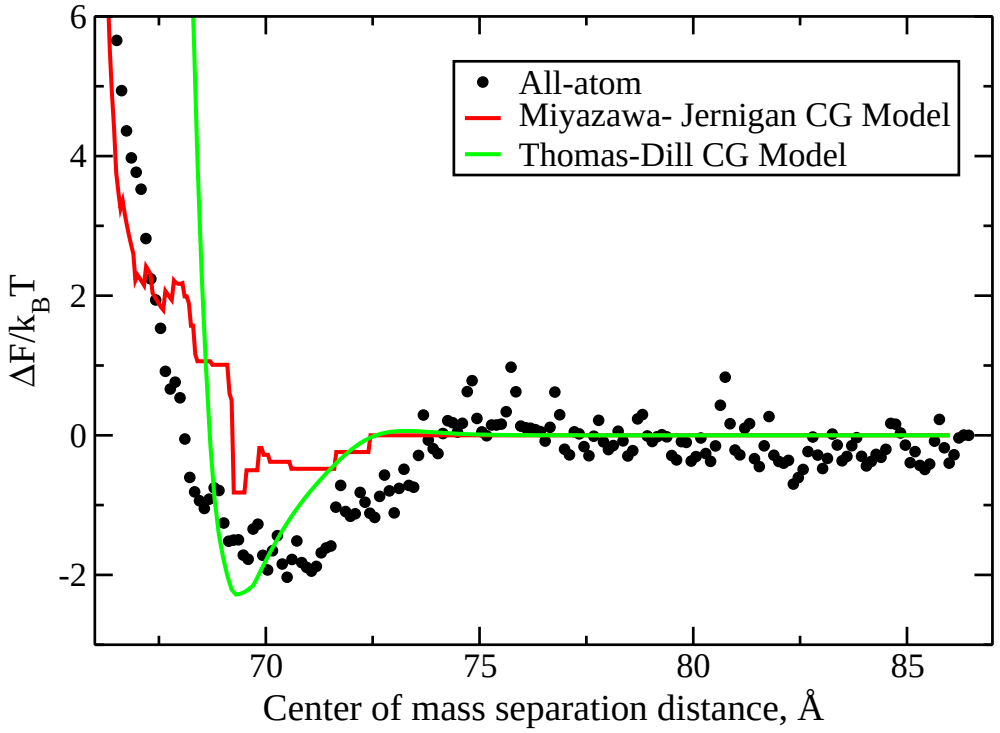

Figure 2: Atomistic and CG PMFs as computed with the MJ and TD models (see text).

Figure 5 shows 6 snapshots of the self-assembly process obtained with MC technique and the TD potential: initially, all the hexamers are isolated, but as the simulation proceeds they aggregate into clusters of different sizes. The clusters grow radially outward without any preferential direction, while in a cluster the relative orientation between hexamers and the distance between their CoMs are always about 0 degrees and 70Å, respectively. The largest cluster in Figure 5 has a dimension of 42 nm x 42 nm (21 hexamers).

In Figure 4 we show how 2D self-assembly varies with the concentration of CcmK2 hexamers. At each concentration, the growth kinetics were determined by plotting the

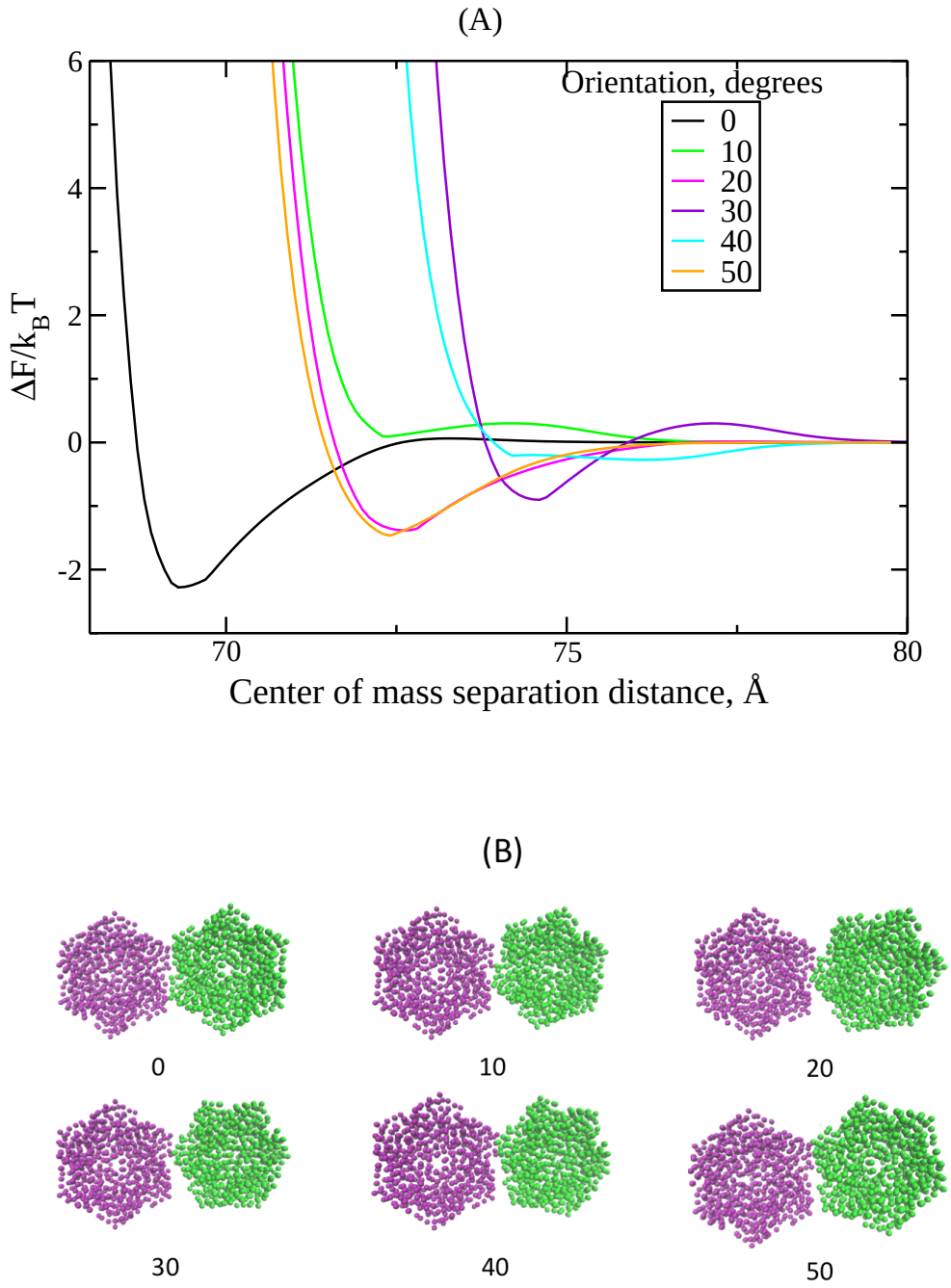

Figure 3: (A) The PMF of CcmK2 hexamers was computed for different relative orientation using the TD model (see text). The most favorable orientation occurs at zero degrees, as evidenced by the minimum of the PMF. (B) Snapshots of relative orientation of the two hexamers with their corresponding relative orientation angles.

size of the largest cluster *versus* the number of MC steps, averaged over 30 independent simulations. As seen in Figure 4, for the highest CcmK2 concentration (black curve, 0.015625 hexamers/nm^2), the size of the largest cluster increases from 1 to 14 in about 80000 MC steps. At 0.014172 and 0.013521 hexamers/nm^2 growth is still observable, but at the concentration of 0.01 hexamers/nm^2 there is hardly any growth over the simulation time window in this study (10^5 MC steps). Further analysis reveals that if a cluster with four hexamers forms and remains stable, continued growth ensues. This behavior is a signature of a nucleation-growth kinetics process, in which a free energy barrier for the formation of nucleus needs to be overcome. Nucleation-growth process typically occurs among units with low binding free energy (of the order of k_BT, as it's the case here) and requires a minimum concentration for self-assembly to occur. By contrast, spontaneous growth process is characterized by high binding free energy (at least an order of magnitude higher than k_BT) and by a self-assembly process that is independent of the concentration.

For verifying whether the self-assembly process is or not nucleation driven, we used classical nucleation theory.[36,37] In this theory, the nucleation time, τ_{nucl}, is related to concentration, c (in our case the concentration of hexamers), as follows

$$\tau_{nucl} = Aexp\left[\frac{B}{ln(c/c_m)}\right] \tag{1}$$

where A and B are constants, and c_m is the minimum concentration for nucleation; when c is above c_m the system is said to be supersaturated. According to Eq. 1, $\ln\tau_{nucl}$ *versus* $1/\ln(c/c_m)$ is linear, from which one can extract the value of c_m. As mentioned above, we observed that self-assembly ensues when the cluster reaches a size of four and remains stable. Therefore, we take τ_{nucl} as the first passage time to a cluster with five hexamers and c as the corresponding concentrations; τ_{nucl}= 11100, 22010, 32920 and c= 0.015625, 0.014172 and 0.013521 units nm^{-2}. Subsequently, we plotted $\ln\tau_{nucl}$ *vs.* $1/\ln(c/c_m)$ by fitting c_m (because

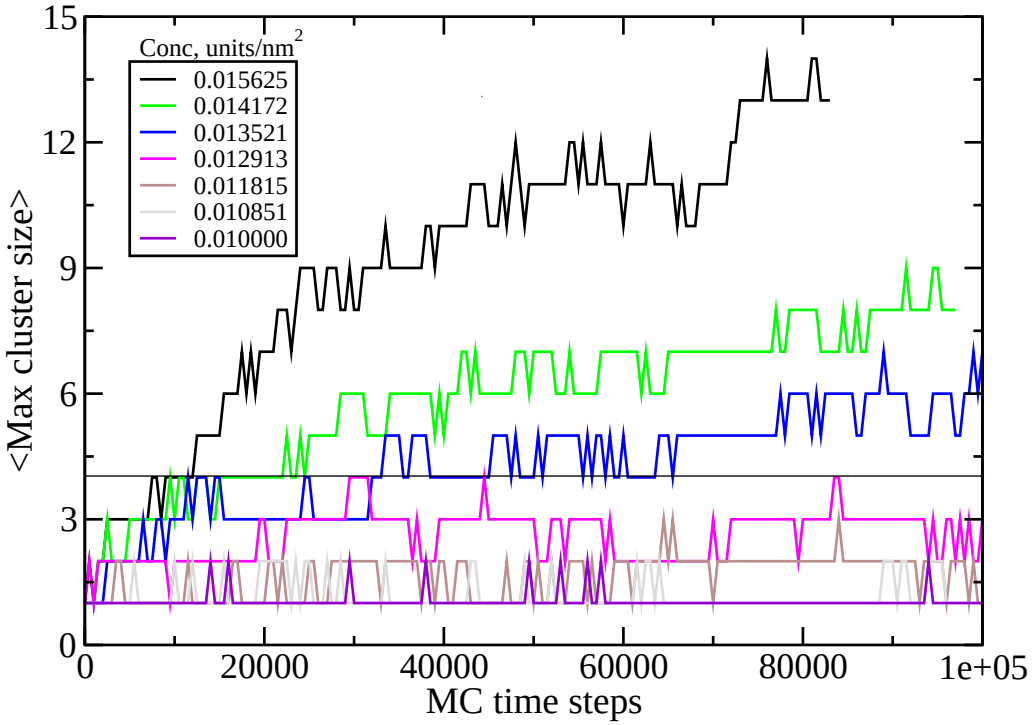

Figure 4: Size of the maximum cluster *versus* the number of MC steps, average over 30 different simulations. The curves correspond to different concentrations, which are given in the inset of the figure .

our purpose is to qualitatively validate whether the process is or not nucleation driven, the exact values of A and B are not relevant). In this way it was found that a value of c_m of 0.007 hexamers/nm^2 produced a straight line with a regression coefficient of 1. Thus, we estimate that 0.007 hexamers/nm^2 is the the minimum concentration needed for the 2D self-assembly of a CcmK2 layer. It might be seen as a contradiction that c_m is lower than the lowest concentration studied here, 0.01 hexamers/nm^2, for which no growth was actually observed, see Figure 4. It must be noted, however, that the simulations were carried out for 10^5 MC steps only. Presumably, if the simulations were to continue, growth–although slow–would also occur at the concentration of 0.01.

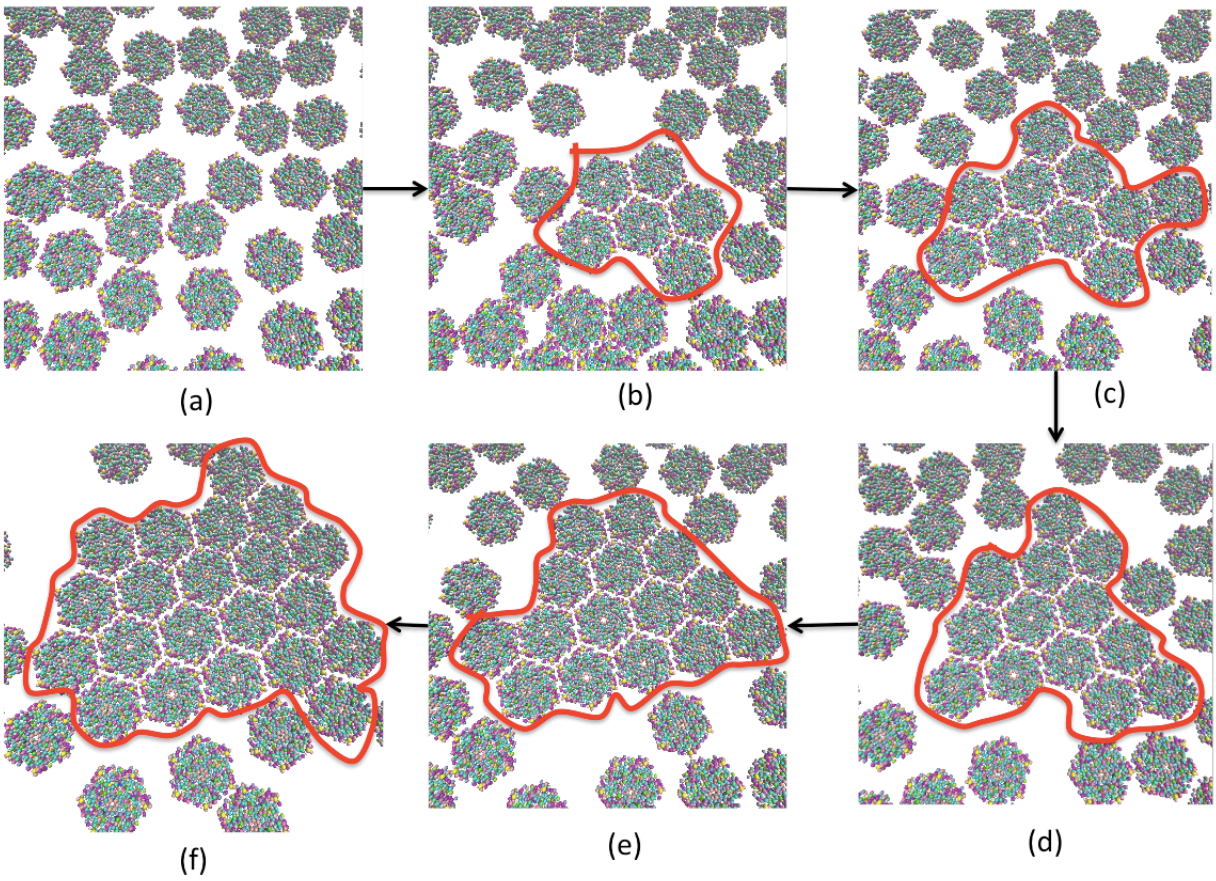

Figure 5: Snapshots of self-assembly at room temperature and at a concentration of 0.015625 hexamers/nm^2. (a)-(f) correspond to MC timestep= 0, 10000, 19000, 35000, 64000, 80000, respectively. For clarify, the largest cluster is outline in red.

DISCUSSION

It is instructive to put our results in the context of previous studies, and in particular those of Dryden *et al.*,[5] Cameron *et al.*[38] and a very recent study by Sutter *et al.*[6] Dryden *et al.*[5] studied the self-assembly of His-tagged CcmK2 hexamers on a monolayer of nickelated lipid molecules at the air-water interface. Because the proteins were tagged at the C-terminus, which resides on one side of the hexamer, the hexamers were expected to be uniformly oriented, *i.e.* with their C-terminus side facing the lipid monolayer. The layers formed in this way were transferred to a carbon grid and imaged with transmission electron microscopy. Dryden *et al.*[5] observed 2D layers with dimensions of approximately 500 nm x 300 nm, where the hexamers were arranged in an hexagonal manner with their CoMs about 68 Å apart.

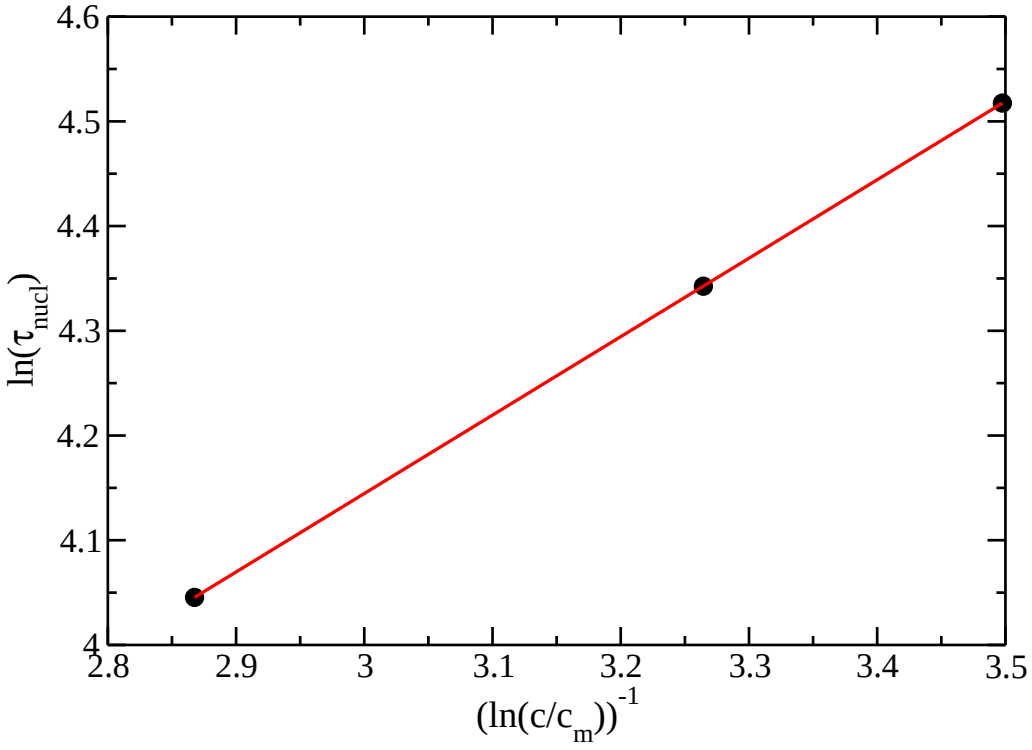

Figure 6: Nucleation time follows the classical nucleation theory: $\ln(\tau_{nucl})$ *versus* $1/\ln(c/c_m)$ is linear with the best fit value c_m=0.007 hexamers/nm^2 . The regression coefficient is R^2=1.0.

In our calculations, we imposed two conditions only, namely that the hexamers remain in one layer and are oriented uniformly with their C-terminus facing in the -y-direction. With only these two conditions, the hexamers were found to self-assemble in a hexagonal manner similar to that seen experimentally by Dryden *et al.*[5] The samples considered in our calculations, however, are not as large as those made by Dryden *et al.*[5] For example, the largest cluster we observed has dimensions 42 nm x 42 nm. Yet, our goal was not to approach the experimental size, which would be prohibitively expensive even for the coarse grain technique used here, but to explore the mechanism of the early-stage self-assembly of a CcmK2 layer. By studying smaller samples, we observed the formation of clusters, in which the hexamers are hexagonally arranged with their CoMs separated at 70 Å. The arrangement

of hexamers in these clusters resembles very much that observed by Dryden *et al.*[5] This lends support to the 2D self-assembly mechanism captured in the calculation, which is found to be governed by a nucleation-growth process that requires the formation of a stable cluster with at least four hexamers and a minimum concentration of 0.007 hexamers/nm^2. It is possible that this nucleation-growth process helps CcmK2 to avoid kinetically trapped pathways. For other biological protein self-assembly processes, such as viral-capsid self-assembly, weak inter-subunit interactions results in nucleation-growth mechanism, which leads to correctly assembled structures. On the other hand, kinetically trapped pathways have been reported for stronger inter-subunit attractions.[37,39–45]

The use of tags in Dryden *et al.*[5] brings up an interesting question: how do these tags affect self-assembly? For self-assembly to take place, the hexamers need to be able to diffuse, to move about the 2D plane so they can interact with other hexamers. From Dryden *et al.*[5] it is unclear how strongly the hexamers are bound to the lipid monolayer, if all the hexamers or only a few are bound to the lipid molecules, or how those hexamers–bound or not–diffuse. It stands to reason that the barrier to diffusion of a hexamer should be different depending on whether it is bound or unbound to a lipid monolayer. In our simulations, no lipid bilayer was involved. Thus there was no barrier for the hexamer diffusion, yet we obtained self-assembled patterns that resembled those seen by Dryden *et al.*[5] From this we can infer that in Ref.,[5] for whatever reason (*i.e.* weak hexamer-lipid binding, or most likely mobility of the lipid layer) the hexamers are mobile and interact with each other in a similar fashion as that simulated here. Therefore, the tags used in Dryden *et al.*[5] do not have a major impact on the CcmK2 self-assembly thermodynamics.

In Cameron *et al.*,[38] fluorescence microscopy was used to monitor the stages of carboxysome formation. It was found that when a protein called CcmN was present, CcmK2 hexamers were able to associate to form a functional carboxysome. Cameron *et al.*[38] proposed that CcmN acts as a bridge that connects the interior of the carboxysome (termed the proto-carboxysome, PC) to the shell, and that the shell formation depends on its presence.

The CcmN proteins can be seen as the equivalent to the tags used in Dryden *et al.*:[5] they bind CcmK2 hexamers and enable their self-assembly. But as in Ref.,[5] in Cameron *et al.*[38] it is unclear how strong the CcmN-CcmK2 binding is, if all the hexamers or only a few are bound to the CcmN proteins, or how those hexamers–unbound or not–move about the PC's surface.

In a very recent study, Sutter *et al.*[6] used high-speed atomic force microscopy (HS-AFM) to investigate the 2D self-assembly of the *Haliangium ochraceum*'s BMC-H protein. This protein, which resembles a C-terminal deletion mutant of CcmK2, also forms a hexamer with a concave and convex side. Sutter *et al.*[6] deposited samples of BMC-H proteins on top of mica but did not employ tags. They found that BMC-H hexamers self-assemble into 2D sheets with a clear sidedness, concave or convex; individual sheets in which the concave and convex side were facing the same direction were not observed. HS-AFM further allowed Sutter *et al.*[6] to observe that the self-assembly of BMC-H sheets is a very dynamic process. In particular, they characterized several type of movements: translational motion of individual hexamers from one sheet to another, sometimes attaching or detaching themselves; translational motion of individual hexamers across the edges of one sheet; motion of individual hexamers inside the sheets; translational motion of individual patches, each path made of several hexamers, in which the patches detach from or attach to other sheets. As shown in the movie included in the supporting information, in our simulations individual hexamers and small short-lived patches move in a manner that mimics very much that observed by Sutter *et al.*:[6] the movie clearly shows CcmK2 hexamers translating along the edge of a large cluster, moving from one cluster to another, and short-lived clusters that break apart into individual hexamers. This process is very dynamic and resembles very much that seen by Sutter *et al.*[6] for BMC-H proteins, which suggest that BMC-H proteins might also self-assemble driven by a nucleation-growth process. It remains to be seen whether nucleation-growth drives the self-assembly of other carboxysome's shell proteins.

CONCLUSIONS

With the goal of understanding how the facets of the carboxysome's shell might form, a Monte Carlo technique was used to investigate the self-assembly of CcmK2 hexamers in 2D. The interactions between the hexamers were described with a coarse-grained potential fitted to a potential of mean force (PMF) derived from atomistic molecular dynamics simulations. The hexamers were allowed to translate and rotate about a plane with no barrier to diffusion. The simulations revealed that CcmK2 hexamers self-assemble into clusters following a nucleation process that requires the formation of a nucleus with at least four hexamers and a minimum concentration of 0.007 hexamers/nm^2. When these results were put in the context of previous experimental studies, it was found that CcmK2 hexamers are ordered in a similar fashion as in the 2D layers made experimentally by Dryden *et al.*[5] Thus, it is concluded that the cluster formation seen here represents the initial stage of 2D self-assembly of CcmK2 layers. One wonders how the initial stages of self-assembly might be affected by including more than one type of hexamer, *i.e.* CcmK2 along with CCmK4–another hexamer. The theoretical framework presented here can be extended to address this type of question, potentially providing important insights into the structure-function relationship in this important class of bacterial microcompartments. This theoretical framework can be applied to other biological protein-assembly processes, such as viral-capsid self-assembly,[41] provided that the constituent proteins are conformationally stable.

METHODS

Atomistic Potential of Mean Force

The crystal structure of the CcmK2 hexamer was taken from the protein databank (PDB ID: 2A1B)[4] and used to create a dimer made of two CcmK2 hexamers. For each hexamer, the protonation states of the titratable residues were determined by calculating the

pKa, using the Karlsberg webserver,[46] and by manually checking for local hydrogen bonding residues. The non-polar hydrogens were added using the Visual Molecular Dynamics[47,48] (VMD) utility program psfgen. Different dimers were created by varying the distance between the center-of-mass (CoM) of the hexamers. Each dimer was solvated with the TIP3P water molecules with a minimum of 15 Å water on each side of a cubic box. Charge neutralization was accomplished by adding Na^+ and Cl^- ions, resulting in a 0.1 M solution. The resultant system contained 142908 atoms. This system subsequently underwent two equilibration steps: 10000 steps of minimization and 2 ns equilibration with decreasing positional restraints on the Cα atoms. All these MD simulations were performed with the NAMD program and the CHARMM27 force field (with the backbone CMAP correction) for the protein.[49] A short-range cutoff of 12 Å was used for non-bonded interactions, and long-range electrostatic interactions were treated with the particle mesh Ewald method with a grid spacing of 1.0 Å. Langevin dynamics and a Langevin piston algorithm were used to maintain the temperature at 310 K and the pressure of 1 atm. The r-RESPA multiple-time-step method was employed, with time steps of 2 fs for bonded, 2 fs for short-range non-bonded, and 4 fs for long-range electrostatic forces. The bonds between hydrogen and heavy atoms were constrained with the SHAKE algorithm.

Coarse-grained Potential of Mean Force

In the MJ potential,[34] each residue of the six proteins that make a CcmK2 hexamer was represented by a bead centered at the CoM of the residue's side chain; the only exception is the Glycine residue, in which case the bead is centered at the Cα position. The short-range pairwise interaction potential between the 20 types of naturally occurring residues was taken from Table **V** of Ref.[34] The cut-off distance for interaction between a pair of residues was set at 6.5 Å, as recommended.[34] Because we are using a non-lattice model, a second cut-off was introduced to avoid any substantial overlap between residues at short distances; we chose this cut-off as 3.0 Å. The pairwise interaction potential $U_{ij}(d)$ (in units of k_BT) between a

pair of residues of types i and j separated by a distance d (in Å units) is then given by the expression

$$\frac{U_{ij}(d)}{k_BT} = \begin{cases} 0, & \text{if } d > 6.5\text{Å} \\ V_{ij}, & \text{if } 6.5\text{Å} \geq d \geq 3.0\text{Å} \\ V_{ij} + 0.5\left(\left[\frac{3.0}{d}\right]^9 - 1\right), & \text{if } d < 3.0\text{Å} \end{cases}$$

where V_{ij} is the interaction potential between residues i and j and 3.0 Å is the cut-off. Below this cut-off a soft-repulsive potential between residues sets in.

In the TD potential,[35] each residue is represented by a bead centered at the Cβ position of the corresponding residue; once again, the only exception is Glycine, for which case the bead is centered at the Cα atom. The short range pairwise interaction potential between the 20 different naturally occurring residues were taken from Table **1** of Ref.[35] The pairwise interaction potential $U_{ij}blue(d)$ (in units of k_BT) between a pair of residues of types i and j separated by d (in Å units) can be expressed as

$$\frac{U_{ij}(d)}{k_BT} = \begin{cases} V_{ij}S_{ij}, & \text{if } d \geq 6.0\text{Å} \\ V_{ij}S_{ij} + 0.5\left(\left[\frac{6.0}{d}\right]^9 - 1\right), & \text{if } d < 6.0\text{Å} \end{cases}$$

where V_{ij} corresponds to the interaction between residues i and j and $S_{ij} = 0.5(tanh(9.0-d) + tanh(9.0+d))$ was added here to ensure that U_{ij} smoothly decays to zero at 10.5 Å. As in the MJ potential, a soft-repulsive potential was added to avoid substantial overlap between residues at short distances below a cutoff of 6.0 Å.

The MJ and TD potential also differ in another aspect, namely the inclusion or not of bonds between residues. In the MJ potential these bonds are neglected, whereas they are included in the TD potential.

Monte Carlo simulation of 2D self-assembly

Monte Carlo, MC, simulations were performed in a box with dimensions D Å x 100 Å x D Å and with periodic boundary conditions in the xz direction. D was varied between 400 Å and 600 Å to simulate different concentrations of CcmK2 hexamers. Initially, 25 hexamers were placed randomly inside the box with random orientations; the CoM of each hexamer in the y-direction was kept fixed to a constant value of 0.0, which meant the hexamers were allowed to translate and rotate in the xz plane only. After this initial setup, the simulation proceeded as follows. A hexamer was randomly selected and subjected to a random translation and rotation. Translation was performed along a vector in the xz plane with each component varying anywhere between -3.0 and 3.0 Å per step. Rotation was defined around the y-axis with an angle varying between -5.0 and 5.0 degrees per step. The translational and rotational movement of a hexamer was accepted if the change in the total potential energy of the system was smaller than 0, *i.e.* $\Delta U \leq 0$. Otherwise, a random number between 0 and 1 was generated and compared with $exp(-\Delta U)$. If this number was smaller than $exp(-\Delta U)$, the movement was accepted, otherwise rejected. For each concentration, 30 independent MC simulations were performed to obtain enough statistics. The concentrations considered, in units of the number of hexamers/nm^2, were 0.015625, 0.014172, 0.013521, 0.012913, 0.011815, 0.010851, and 0.01. All the MC simulations were performed at around room temperature (310 K) to mimic the experimental conditions in Dryden *et al.*[5]

During the 2D self-assembly process, some hexamers remained isolated while others clustered. For deciding whether a hexamer was isolated or clustered with others, we use the following criterion: two hexamers are clustered if more than 25 pairs of residues are closer than 9 Å. (This criterion was established visually by noting the minimum pair of interacting residues needed to keep a dimer stable.) The maximum cluster size, averaged over 30 simulations, was then plotted against the MC number of steps to determine the kinetics of the 2D self-assembly.

Acknowledgments

This research was conducted at the Center for Nanophase Materials Sciences (CNMS), which is a U.S. Department of Energy Office of Science User Facility. K. Brown was supported by an appointment under the Science Undergraduate Laboratory Internships (SULI), administered by the Oak Ridge Institute for Science and Education between the U.S. Department of Energy and Oak Ridge Associated Universities. The computations were performed using resources of the CNMS and the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory. The authors would like to thank Mitch Doktycz, David Garcia, Pat Collier, Scott Retterer, Rajeev Kumar and Bobby Sumpter for insightful discussions.

Supporting Information Available

A movie demonstrating the self-assembly of CcmK2 hexamers. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

1. Yeates, T. O.; Kerfeld, C. A.; Heinhorst, S.; Cannon, G. C.; Shively, J. M. Protein-Based Organelles in Bacteria: Carboxysomes and Related Microcompartments. *Nat. Rev. Microbiol.* **2008**, *6*, 681–691.

2. Badger, M. R.; Price, G. D. CO2 Concentrating Mechanisms in Cyanobacteria: Molecular Components, Their Diversity and Evolution. *J. Exp. Bot.* **2003**, *54*, 609–622.

3. Kinney, J. N.; Axen, S. D.; Kerfeld, C. A. Comparative Analysis of Carboxysome Shell Proteins. *Photosynth. Res.* **2011**, *109*, 21–32.

4. Kerfeld, C. A.; Sawaya, M. R.; Tanaka, S.; Nguyen, C. V.; Phillips, M.; Beeby, M.; Yeates, T. O. Protein Structures Forming the Shell of Primitive Bacterial Organelles. *Science* **2005**, *309*, 936–938.

5. Dryden, K. A.; Crowley, C. S.; Tanaka, S.; Yeates, T. O.; Yeager, M. Two-Dimensional Crystals of Carboxysome Shell Proteins Recapitulate the Hexagonal Packing of Three-Dimensional Crystals. *Protein Sci.* **2009**, *18*, 2629–2635.

6. Sutter, M.; Faulkner, M.; Aussignargues, C.; Paasch, B. C.; Barrett, S.; Kerfeld, C. A.; Liu, L. Visualization of Bacterial Microcompartment Facet Assembly Using High-Speed Atomic Force Microscopy. *Nano Lett.* **2015**,

7. Arkhipov, A.; Freddolino, P. L.; Schulten, K. Stability and Dynamics of Virus Capsids Described by Coarse-Grained Modeling. *Structure* **2006**, *14*, 1767–1777.

8. Chu, J. W.; Voth, G. A. Coarse-Grained Modeling of the Actin Filament Derived from Atomistic-Scale Simulations. *Biophys. J.* **2006**, *90*, 1572–1582.

9. Clementi, C. Coarse-Grained Models of Protein Folding: Toy Models or Predictive Tools? *Curr. Opin. Struct. Biol.* **2008**, *18*, 10–15.

10. Doruker, P.; Jernigan, R. L.; Bahar, I. Dynamics of Large Proteins Through Hierarchical Levels of Coarse-Grained Structures. *J. Comput. Chem.* **2002**, *23*, 119–127.

11. Head-Gordon, T.; Brown, S. Minimalist Models for Protein Folding and Design. *Curr. Opin. Struct. Biol.* **2003**, *13*, 160–167.

12. Hills, R. D.; Brooks, C. L. Insights from Coarse-Grained Go Models for Protein Folding and Dynamics. *Int. J. Mol. Sci.* **2009**, *10*, 889–905.

13. Maupetit, J.; Tuffery, P.; Derreumaux, P. A Coarse-Grained Protein Force Field for Folding and Structure Prediction. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 394–408.

14. Neri, M.; Anselmi, C.; Cascella, M.; Maritan, A.; Carloni, P. Coarse-Grained Model of Proteins Incorporating Atomistic Detail of the Active Site. *Phys. Rev. Lett.* **2005**, *95*.

15. Nguyen, H. D.; Reddy, V. S.; Brooks, C. L. Deciphering the Kinetic Mechanism of Spontaneous Self-Assembly of Icosahedral Capsids. *Nano Lett.* **2007**, *7*, 338–344.

16. Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. Coarse Grain Models and the Computer Simulation of Soft Materials. *J. Phys.: Condens. Matter* **2004**, *16*, R481–R512.

17. Noid, W. G.; Chu, J. W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. The Multiscale Coarse-Graining Method. I. A Rigorous Bridge Between Atomistic and Coarse-Grained Models. *J. Chem. Phys.* **2008**, *128*.

18. Pellarin, R.; Caflisch, A. Interpreting the Aggregation Kinetics of Amyloid Peptides. *J. Mol. Biol.* **2006**, *360*, 882–892.

19. Tozzini, V. Coarse-Grained Models for Proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.

20. Wei, G. H.; Mousseau, N.; Derreumaux, P. Computational Simulations of the Early Steps of Protein Aggregation. *Prion* **2007**, *1*, 3–8.

21. Shih, A. Y.; Freddolino, P. L.; Arkhipov, A.; Schulten, K. Assembly of Lipoprotein Particles Revealed by Coarse-Grained Molecular Dynamics Simulations. *J. Struct. Biol.* **2007**, *157*, 579–592.

22. Shen, V. K.; Cheung, J. K.; Errington, J. R.; Truskett, T. M. Coarse-Grained Strategy for Modeling Protein Stability in Concentrated Solutions. II: Phase Behavior. *Biophys. J.* **2006**, *90*, 1949–1960.

23. Matysiak, S.; Clementi, C. Minimalist Protein Model as a Diagnostic Tool for Misfolding and Aggregation. *J. Mol. Biol.* **2006**, *363*, 297–308.

24. Hills, R. D.; Lu, L. Y.; Voth, G. A. Multiscale Coarse-Graining of the Protein Energy Landscape. *PLoS Comput. Biol.* **2010**, *6*.

25. Cheon, M.; Chang, I.; Hall, C. K. Extending the PRIME Model for Protein Aggregation to All 20 Amino Acids. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 2950–2960.

26. Bratko, D.; Blanch, H. W. Competition Between Protein Folding and Aggregation: A Three-Dimensional Lattice-Model Simulation. *J. Chem. Phys.* **2001**, *114*, 561–569.

27. Bereau, T.; Deserno, M. Generic Coarse-Grained Model for Protein Folding and Aggregation. *J. Chem. Phys.* **2009**, *130*.

28. Gumbart, J. C.; Roux, B.; Chipot, C. Efficient Determination of Protein-Protein Standard Binding Free Energies from First Principles. *J. Chem. Theory Comput.* **2013**, *9*, 3789–3798.

29. Gumbart, J. C.; Roux, B.; Chipot, C. Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy? *J. Chem. Theory Comput.* **2013**, *9*, 794–802.

30. Wang, L.; Siu, S. W. I.; Gu, W.; Helms, V. Downhill Binding Energy Surface of the Barnase-Barstar Complex. *Biopolymers* **2010**, *93*, 977–985.

31. Woo, H. J.; Roux, B. Calculation of Absolute Protein-Ligand Binding Free Energy from Computer Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6825–6830.

32. Zeller, F.; Zacharias, M. Efficient Calculation of Relative Binding Free Energies by Umbrella Sampling Perturbation. *J. Comput. Chem.* **2014**, *35*, 2256–2262.

33. Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. Multidimensional Free-Energy Calculations Using the Weighted Histogram Analysis Method. *J. Comput. Chem.* **1995**, *16*, 1339–1350.

34. Miyazawa, S.; Jernigan, R. L. Self-Consistent Estimation of Inter-Residue Protein Contact Energies Based on an Equilibrium Mixture Approximation of Residues. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 49–68.

35. Thomas, P. D.; Dill, K. A. An Iterative Method for Extracting Energy-like Quantities from Protein Structures. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 11628–11633.

36. Kundagrami, A.; Muthukumar, M. Continuum Theory of Polymer Crystallization. *J. Chem. Phys.* **2007**, *126*.

37. Mahalik, J. P.; Muthukumar, M. Langevin Dynamics Simulation of Polymer-Assisted Virus-like Assembly. *J. Chem. Phys.* **2012**, *136*.

38. Cameron, J. C.; Wilson, S. C.; Bernstein, S. L.; Kerfeld, C. A. Biogenesis of a Bacterial Organelle: The Carboxysome Assembly Pathway. *Cell* **2013**, *155*, 1131–1140.

39. Ceres, P.; Zlotnick, A. Weak Protein-Protein Interactions Are Sufficient to Drive Assembly of Hepatitis B Virus Capsids. *Biochemistry* **2002**, *41*, 11525–11531.

40. Endres, D.; Zlotnick, A. Model-Based Analysis of Assembly Kinetics for Virus Capsids or Other Spherical Polymers. *Biophys. J.* **2002**, *83*, 1217–1230.

41. Hagan, M. F. Modeling Viral Capsid Assembly. *Adv. Chem. Phys.* **2014**, *155*, 1–67.

42. Hagan, M. F.; Chandler, D. Dynamic Pathways for Viral Capsid Assembly. *Biophys. J.* **2006**, *91*, 42–54.

43. Hagan, M. F.; Elrad, O. M.; Jack, R. L. Mechanisms of Kinetic Trapping in Self-Assembly and Phase Transformation. *J. Chem. Phys.* **2011**, *135*.

44. Zlotnick, A.; Aldrich, R.; Johnson, J. M.; Ceres, P.; Young, M. J. Mechanism of Capsid Assembly for an Icosahedral Plant Virus. *Virology* **2000**, *277*, 450–456.

45. Zlotnick, A.; Johnson, J. M.; Wingfield, P. W.; Stahl, S. J.; Endres, D. A Theoretical Model Successfully Identifies Features of Hepatitis B Virus Capsid Assembly. *Biochemistry* **1999**, *38*, 14644–14652.

46. `http://agknapp.chemie.fu-berlin.de/karlsberg/`.

47. Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.

48. http://www.ks.uiuc.edu/Research/vmd/.

49. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E. *et al.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.