Monte Carlo study of the molecular mechanisms of surface-layer protein self-assembly

Christine Horejs, Mithun K. Mitra, Dietmar Pum, Uwe B. Sleytr, and Murugappan Muthukumar

Citation: The Journal of Chemical Physics **134**, 125103 (2011); doi: 10.1063/1.3565457
View online: http://dx.doi.org/10.1063/1.3565457
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/134/12?ver=pdfcov
Published by the AIP Publishing

Monte Carlo study of the molecular mechanisms of surface-layer protein self-assembly

Christine Horejs,[1] Mithun K. Mitra,[2] Dietmar Pum,[1] Uwe B. Sleytr,[1] and Murugappan Muthukumar[2,a)]

[1]*Department for Nanobiotechnology, University of Natural Resources and Applied Life Sciences, 1190 Vienna, Austria*

[2]*Department for Polymer Science and Engineering, University of Massachusetts, Amherst, Massachusetts 01003, USA*

(Received 17 November 2010; accepted 23 February 2011; published online 22 March 2011)

The molecular mechanisms guiding the self-assembly of proteins into functional or pathogenic large-scale structures can be only understood by studying the correlation between the structural details of the monomer and the eventual mesoscopic morphologies. Among the myriad structural details of protein monomers and their manifestations in the self-assembled morphologies, we seek to identify the most crucial set of structural features necessary for the spontaneous selection of desired morphologies. Using a combination of the structural information and a Monte Carlo method with a coarse-grained model, we have studied the functional protein self-assembly into S(surface)-layers, which constitute the crystallized outer most cell envelope of a great variety of bacterial cells. We discover that only few and mainly hydrophobic amino acids, located on the surface of the monomer, are responsible for the formation of a highly ordered anisotropic protein lattice. The coarse-grained model presented here reproduces accurately many experimentally observed features including the pore formation, chemical description of the pore structure, location of specific amino acid residues at the protein–protein interfaces, and surface accessibility of specific amino acid residues. In addition to elucidating the molecular mechanisms and explaining experimental findings in the S-layer assembly, the present work offers a tool, which is chemical enough to capture details of primary sequences and coarse-grained enough to explore morphological structures with thousands of protein monomers, to promulgate design rules for spontaneous formation of specific protein assemblies. © *2011 American Institute of Physics*. [doi:10.1063/1.3565457]

I. INTRODUCTION

Self-assembly is one of nature's strategies to organize matter on a large scale and thereby create order from disorder.[1–3] The process is ubiquitous for a great variety of biological molecules, such as lipids,[4,5] DNA,[6] polymers,[7,8] self-assembled monolayers,[9] and viruses,[10] and understanding the driving forces behind this process is one of the central challenges in biological physics.[11] Proteins, however, tend to stay soluble in solution or aggregate into various structures rather than self-assemble into defined patterns. This is because of their complex structure exhibiting different conformations and a close-knit relationship between structure and function. Aggregation into three-dimensional composites thus generally leads to a loss of functionality.[12,13] However, S(surface)-layer proteins represent a remarkable exception to this general trend.[14] Another exception is the formation of essentially one-dimensional fibers from amyloidogenic proteins.[15] The S-layer proteins crystallize into monomolecular arrays on the cell surface of a great variety of bacterial and all archaeal cells thereby providing the outermost cell envelope (S-layer). The crystallization of this kind of proteins then facilitates their function rather than forming a nonfunctional state as is the case for most of the ill-aggregated complex proteins. The self-assembly process is also remarkably robust with the protein monomers forming patterns with defined lattice symmetries not only in their natural environment but also in solution and on a variety of surfaces and interfaces, which makes them an interesting object for the investigation of the design of biomolecular self-assembly processes that do not lead to aggregation.[16]

There has been a wide range of studies over the past few decades in order to investigate the genetics, structure, function, and nanotechnological applications of S-layer proteins. While this has provided considerable insight into the structure–function relationship of several S-layer proteins, a fundamental understanding of the molecular mechanisms guiding the self-assembly process has been elusive. Recent studies on the crystallization of an S-layer protein, which self-assembles into structures exhibiting lattices with a p4 symmetry, have given an insight into the kinetics and pathways of the S-layer self-assembly on surfaces. This work was based on high-resolution atomic force microscopy in combination with computer simulations.[17,18] However, only recently, a combination of simulation and experimental techniques has enabled the determination of an atomistic structural model of one particular S-layer protein,[19,20] SbsB from *Geobacillus stearothermophilus* pV71/p2, which assembles into two-dimensional sheets in solution exhibiting a lattice with one

a)Author to whom correspondence should be addressed. Electronic mail: muthu@polysci.umass.edu.

monomer per unit cell (p1 symmetry).[21] This structural model provides now the opportunity to investigate the specific interactions between protein monomers that lead to self-assembly as opposed to aggregation in solution.

Coarse-grained modeling of proteins has become an important tool to address various questions regarding their structure and function. Depending on the lengthscale of interest, different strategies of coarse-graining have been implemented.[22–24] Further, depending on the questions that are of interest, these models can then be simulated using either Monte Carlo[25] or Langevin dynamics[26–28] to investigate the properties of the system. In a recent study,[29] we have investigated the kinetics of amyloid fibrillization by combining a coarse-grained model of folded polypeptides and the lattice Monte Carlo method. The results obtained by this modeling methodology are in remarkable agreement with all phenomenological results. Propelled by this advance, we present here an analogous coarse-grained model for the assembly of S-layer proteins.

In the present study, we use a coarse-graining strategy in which every amino acid is represented by a single coarse-grained bead placed at the center of mass of the α-carbon and the side chain. We use off-lattice Monte Carlo simulations of this coarse-grained model to determine the interaction energies for a pair of monomers in terms of their relative orientation and separation distance. Using these energies as an input, we adopt the method of Ref. 29 to elucidate the molecular mechanisms leading to the lattice formation and to follow the kinetics of morphological assembly from hundreds of protein molecules. Our simulations afford an amino acid level understanding of the interactions between monomers and enable us to identify the essential residues required for self-assembly. At a broader level, the simulations also offer valuable insight into the role of specific interactions, such as hydrophobic or electrostatic interactions in the overall self-assembly process which might be applicable in a wider context.

The simulation results are in excellent agreement with known experimental results, while additionally offering the opportunity for predicting the behavior of possible recombinant proteins and different features of the self-assembly process in different environmental conditions. Our results demonstrate the important role that simple techniques, such as Monte Carlo simulations can play in understanding the complex phenomenon of protein–protein interactions. To the best of our knowledge this is the first theoretical study of the self-assembly processes leading to S-layers in solution based on atomistic structural details of one S-layer protein. This example might help to better understand the aspects of protein–protein interactions that are critically required for the formation of well-defined functional morphologies in solution.

II. MODEL AND SIMULATION METHOD

The modeling of the formation of a crystalline two-dimensional lattice from hundreds of protein monomers is carried out in three stages. In the first stage, a coarse-grained model of a monomer is generated from the atomistic details of all amino acid residues of the monomer. In this united-atom coarse-grained model for the monomer, the polymer sequence and charge decoration on the amino acid residues, and their structural correlation with the tertiary structure are maintained. In the second stage, two such monomers are allowed to undergo rotation and translation, relative to each other, and the pairwise interaction energy was computed as a function of rotational and translational degrees freedom using a continuum Monte Carlo method. Only a couple of configurations of dimers dominate the energy landscape among numerous possibilities allowed by rotations and proximity. In the third stage, these two configurations and their energies are taken as an input in a lattice Monte Carlo method to follow the kinetics and morphology of spontaneously assembling structures from hundreds of the coarse-grained protein monomers. This three-stage multiscale simulation protocol allows the modeling of self-assembly from a collection of large numbers of protein monomers and at the same time maintaining the details related to the sequence and charge decoration of the protein monomer. The above mentioned three stages are described below.

A. Coarse-grained model

The coarse-grained model used in this work is based on the atomistic structure that we recently determined using *ab initio* molecular dynamic simulations[19] and small angle x-ray scattering.[20] The protein monomer is made up of 920 amino acid residues. Figure 1(a) shows the surface structure of the protein. The S-layer protein is L-shaped, and contains an N-terminal alpha-helical part that is responsible for anchoring the protein to the cell surface [Fig. 1(a) red part], a C-terminal part that contains mainly beta-sheet structures [Fig. 1(a) blue part], and an unfolded part that links the two domains of the protein. The L-shaped form as well as the architecture of the surface accessible amino acids are supposed to be mainly responsible for the way these proteins self-assemble.

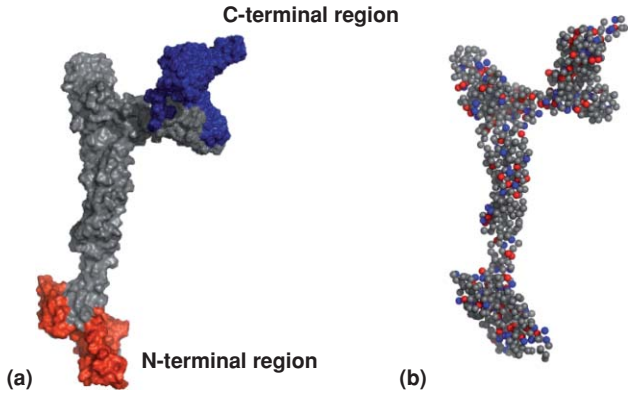

FIG. 1. S-layer protein SbsB from *Geobacillus stearothermophilus* pV72/p2. (a) Surface of an atomistic model calculated by molecular dynamics simulations (Ref. 19). The N-terminal region, which anchors the protein on the cell surface, is colored in red and the C-terminal region in blue. (b) Coarse-grained model used for the calculation of the self-assembly process. Each amino acid residue is represented by one bead of 0.65 nm diameter, which is located in the center of mass of the C_α and the side chain. Negatively charged beads are colored in red and positively charged beads are colored in blue. The protein is treated as a rigid body.

In order to model this protein, we represent every amino acid residue by a single coarse-grained bead placed at the center of mass of the α-carbon and the side chain, as shown in Fig. 1(b). This scheme thus preserves some information of the specific location of the side chain of the particular amino acid, and produces a faithful structural replica of the original protein. The folded protein monomer is treated as a rigid object, motivated by experimental observations that showed that the tertiary structure of the protein remains stable and is the same for monomers in solution as in self-assembled sheets.[20] We also account for the charges of the amino acid side chains, thus enabling us to study the effects of electrostatic interactions in the self-assembly process.

B. Dimer simulations

In order to determine the interprotein interaction energies, we performed off-lattice Monte Carlo simulations with two protein monomers. The two monomers were randomly placed inside a simulation box with hard walls, taking care only to ensure that the monomers did not overlap with each other. The interaction potential between the two monomers was initially taken as a sum of two contributions, a modified Miyazawa–Jernigan (MJ) contact potential, U_{contact} (Ref. 30) and an electrostatic energy term $U_{\text{electrostatic}}$ that accounts for the interactions between charged amino acids.

The MJ contact potential was developed to accurately model the interaction between amino acids based on a knowledge-based statistical analysis of known proteins in the protein database. These potentials have been shown to accurately capture the behavior of interacting amino acids in different situations.[30,31] In the original form, every pair $\{ij\}$ of amino acids is assigned a contact energy denoted by M_{ij} (in units of k_BT), which is defined as the energy difference accompanying the formation of contacts between i and j types of amino acids from those amino acids exposed to the solvent. Thus, the contact energy is given by M_{ij} if the two amino acids are closer than some threshold distance, and is zero if they are larger than this threshold. However, physically, since a complex amino acid residue is represented by a single coarse-grained bead, there is no sharp boundary region for the bead, and to model this effect, we softened this contact potential by a sigmoidal function (S_{ij}) over an interaction range defined by a lower bound (r_{LB}) and an upper bound (r_{UB}). The contact potential can then be written as the product of two terms,

$$U_{\text{contact}} = \sum_{i\in A}\sum_{j\in B} M_{ij}S_{ij}, \tag{1}$$

where the sum runs over all possible amino acid pairs between the two monomers A and B. The sigmoidal function S_{ij} is defined such that the potential varies smoothly between M_{ij} at the lower bound and zero at the upper bound,

$$S_{ij} = \left(\frac{1}{1+e^{r_{ij}}} - \frac{1}{1+e^{r_{\text{UB}}}}\right) \Bigg/ \left(\frac{1}{1+e^{r_{\text{LB}}}} - \frac{1}{1+e^{r_{\text{UB}}}}\right). \tag{2}$$

In our simulations, the lower cutoff was chosen to be $r_{\text{LB}} = 0.65$ nm. This was chosen as an optimum value since smaller values resulted in unphysical highly interpenetrating conformations, whereas larger values did not yield any stable ground state conformations. This lower cutoff can also be interpreted as the radius of the coarse-grained bead. For the upper cutoff, a value of $r_{\text{UB}} = 0.9$ nm was chosen to ensure conformity with previous work on coarse-grained models of proteins.[28] This softening of the potential is not expected to significantly alter the results,[30] but was nevertheless introduced to take into account the coarse-grained nature of the amino acid residues in our model. The MJ values for the contact potential were chosen after a comparison with alternative formulations of the contact energies by Thomas and Dill.[32] Their classification of amino acids in 3, 5, 10, and 20 groups was tested. It was found that for our protein, only the MJ values for the potential led to stable dimer conformations in the simulations.

For the electrostatic interaction, we used the Debye-Hückel potential (in units of k_BT):

$$U_{\text{electrostatic}} = \sum_{i\in A}\sum_{j\in B} \frac{q_iq_j\ell_B \exp(-r_{ij}/\kappa)}{r_{ij}}, \tag{3}$$

where $\kappa = 1$ nm is the Debye length and $\ell_B = 1$ nm is the Bjerrum length, corresponding to the dielectric constant of 56 for the ambient solution. The choice of this value for the dielectric constant is simply to take ℓ_B and κ as 1 nm. A slight variation in the value of ℓ_B is not expected to change the physical conclusions described below. The charges on the individual amino acid residues are denoted by q_i (± 1 or 0), and are taken to be their standard charges at physiological pH, relative to the individual pK_a values of the residues.[33] For the purpose of our simulations, the histidine residues in the monomer sequence were taken to be positively charged.

Using these two potential terms, we then simulated two protein monomers to obtain the minimum energy conformations. The two monomers were placed in a simulation box of size $8L$, where $L \simeq 227$ nm is the typical size of the protein, defined as the length of the space diagonal for the minimal rectangular cuboid that can contain the protein monomer. It was ensured that the monomers did not overlap with each other in the initial configuration. We then performed a standard Metropolis Monte Carlo simulation.[34] A trial move for the simulation consists of a combination of rotation and translation moves. A translation move involves shifting all the amino acid residues in the monomer by a randomly chosen displacement, with a maximum value of the displacement being given by $0.1L$, to ensure that the two monomers did not undergo large translations in a single move. A rotation move consists of randomly choosing a set of Euler angles (ψ, θ, ϕ), where $\psi, \phi \in [-\pi, \pi]$ and $\theta \in [-\pi/2, \pi/2]$. The whole monomer is then rotated in accordance with the rotation matrix defined by these Euler angles $R(\psi, \theta, \phi)$. The trial move is considered valid if no amino acid residues overlap with any other in the resultant conformation, and if both the monomers lie within the simulation box. A valid move is then accepted or rejected according to the standard Metropolis rules.

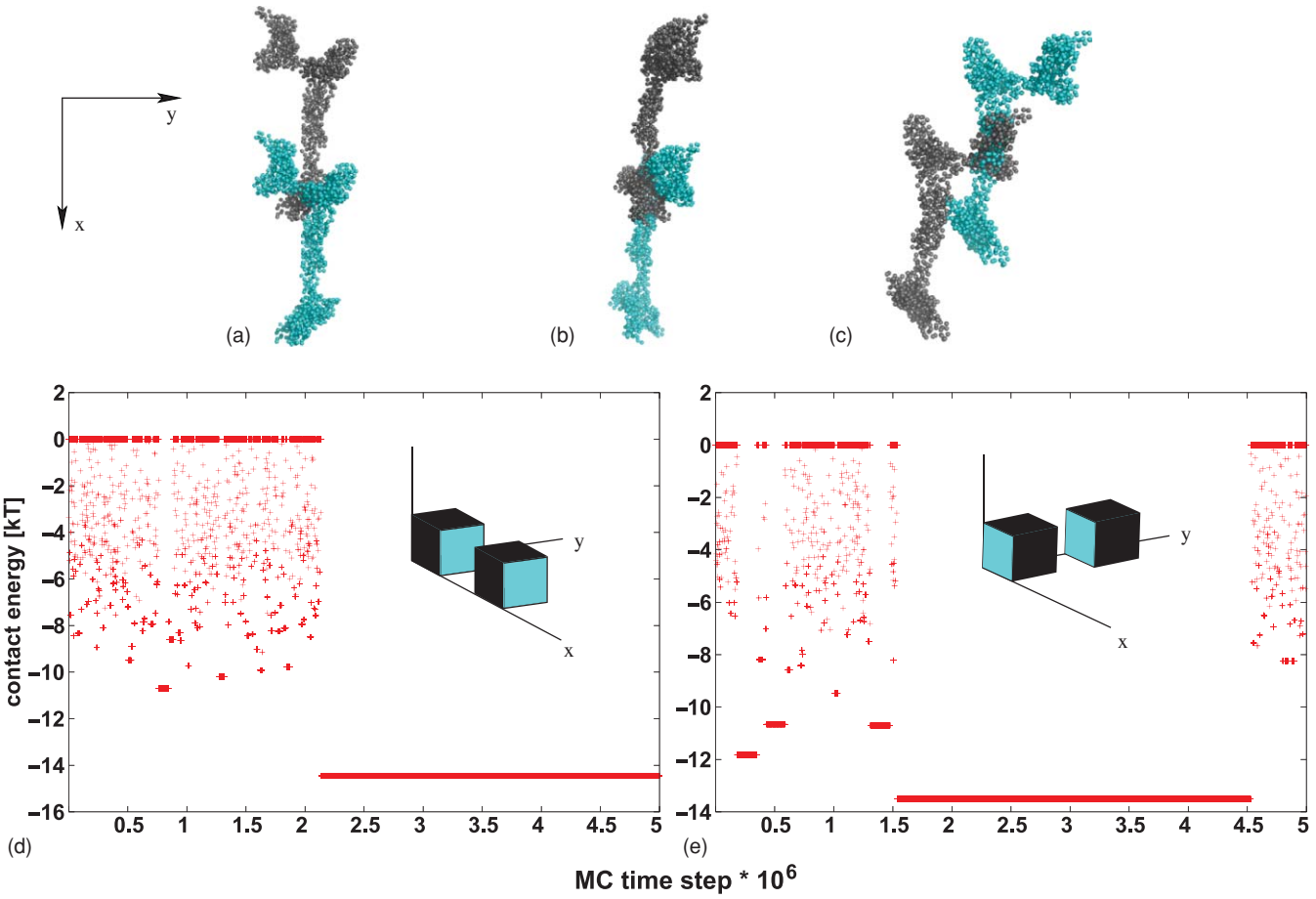

FIG. 2. Monte Carlo simulation of the interaction of two monomers resulting in two stable dimer conformations. (a, b) Hook conformation: one monomer is hooked to the other one by its N-terminus along the *x*-direction. (c) Parallel conformation: The two monomers associate along the *y*-direction and form a charged hydrophilic pore. (d, e) Sample contact energy profiles. (d) The stable energy minimum at $-14.5k_BT$ corresponds to the hook conformation. (e) The stable energy minimum at $-13.8k_BT$ corresponds to the parallel conformation. In these simulations, the amino acid residues were assigned their standard charges at physiological *p*H with histidine taken to be positively charged. Insets: Dimer conformations are represented by unit cubes (as a mnemonic representation) to be used in the lattice Monte Carlo simulation of the large-scale assembly process. These are provided to explain the transition from the off-lattice Monte Carlo scheme for two monomers to the lattice Monte Carlo scheme for simulating hundreds of monomers. Blue colored faces represent the interaction directions.

As described in the next section, we found that there were two minimal energy conformations for the dimers (see Sec. III A and Fig. 2). In addition, we also observed a third highly interpenetrating conformation, in which the two monomers lie almost on top of each other, but are slightly displaced, ensuring that the self-avoidance constraint of the individual amino acid beads is satisfied. This is clearly an unphysical conformation, which arises due to the coarse-grained nature of the model where we replace spatially extended amino acid side-chains by a finite bead, which leaves spurious empty spaces where another monomer can penetrate. Interestingly, it was observed that in all these minimum energy conformations, the two monomers lie in the same plane with the same orientation. It thus appeared that rotation moves were not important in determining the ground state conformations for our particular protein. As a result, in all subsequent runs, a trial move was taken to consist of a simple random translation, which considerably decreases the running time of our simulations, since scanning the general phase space of all rotational angles is unnecessarily expensive in computational time.

In order to exclude the unphysical interpenetrating conformation described above, we introduce an additional cutoff term, U_{cutoff}, which specifies the minimum approach distance between the centers of mass of the two monomers. The form of this term is considerably simplified by virtue of the fact that the minimum energy conformations were found to lie in a plane, and hence the effect of rotations is not important, allowing us to use a single value of the cutoff parameter to exclude highly overlapping conformations. This effect is then modeled by

$$U_{\text{cutoff}} = \begin{cases} 0 & r_{CM}^{A} - r_{CM}^{B} > \eta L \\ \infty & r_{CM}^{A} - r_{CM}^{B} < \eta L \end{cases}, \tag{4}$$

where, $r_{CM}^{(A,B)}$ denotes the center of masses of the two monomers and η is the cutoff parameter. Multiple trial runs were performed using different values of η, and we chose the minimum value which did not result in the unphysical third state (described above) as one of the ground state conformations. This threshold is given by $\eta = 0.15$ in our simulations.

Our final potential energy is then given by a sum of these three terms,

$$U_{\text{total}} = U_{\text{contact}} + U_{\text{electrostatic}} + U_{\text{cutoff}}, \tag{5}$$

where the monomers are assumed to undergo only translation moves (as rotations are found to be irrelevant for the energy minima for a pair of monomers studied here). The ground state conformations of the dimer system were then obtained using this potential via the standard Metropolis Monte Carlo algorithm. Indeed, the energies of the stable states of the dimer are the same as the results without the cutoff and fully allowed rotations of the monomers.

C. Lattice Monte Carlo

In order to gain an insight into the kinetics of the growth process and the morphology of the final assembled structures, we also performed a lattice Monte Carlo using the results of the dimer simulation. The detailed off-lattice dimer simulations revealed that there are two ground state conformations that are dominant, and in these conformations, the two monomers lie in the same plane and with the same orientation. Then, given that rotations were not significant for the interaction among monomers for this protein, we use a mnemonic representation, whereby the protein monomers are represented as cubes, randomly distributed in a cubic lattice which constitutes the simulation box. This is only a mapping for computational ease, and maintains the off-lattice pairwise interaction energies of the states with energy minima. The interaction energies are extracted from the dimer simulations as the ground state energies, and in the lattice simulations these are assigned according to the direction of interaction of the two monomers. This is explicitly shown in the insets of Figs. 2(d) and 2(e), where the blue faces represent the interaction directions corresponding to the two ground states (as discussed below in Sec. III). A trial move then consists of selecting a monomer at random and displacing it to a random position on the cubic lattice. The trial move is considered valid if the attempted new position is unoccupied. If in the new position, the monomer is a neighbor to another monomer, whether in isolation or as a part of a cluster, it is assigned an interaction energy depending on which two faces are in contact. The values of these different interaction energies are obtained as an input from the dimer simulations, as mentioned above. The trial move is then accepted or rejected according to the standard Metropolis rules. The moves are only the nearest neighbor motions, where the self-assembled structures are not allowed to make concerted moves. For the lattice Monte Carlo, we have taken the simulation box to be a cube of linear size 50 in units of the size of each monomer.

We have performed two sets of lattice Monte Carlo simulations for a given protein concentration. In the first, a seed is placed at the center of the simulation box, and the monomers are allowed to adhere only to the seed and its continuously growing growth surfaces with the pairwise interaction energies as obtained above. Here, there is only one growing cluster and there is a continuous depletion of the monomers from the solution without any competition from any other growing cluster (as this is deliberately suppressed in the first set of simulations). We have monitored the growth rate as a function of protein concentration. In the second set, there are no seeds, and the monomers adhere to each other to form multiple clusters which then coarsen among themselves competitively. We have monitored the time evolution of cluster size distributions in the second set of simulations.

III. RESULTS AND DISCUSSION

A. Interaction of two monomers

The S-layer protein under investigation typically assembles into two-dimensional sheets in aqueous solutions, with the monomers arranged in a p1 lattice symmetry.[21] To analyze possible conformations between protein monomers that finally lead to the two-dimensional sheets, we first simulated the interaction between two monomers. In addition to establishing the primary stable conformations of a pair of interacting monomers, we have monitored the relative weights of hydrophobic and electrostatic interactions by deliberate manipulation of charges on the amino acid residues. Furthermore, we have also investigated the role of different segments of the primary sequence of the monomer in determining the most stable conformations of the dimer by, for example, cutting out a segment in the N-terminus of the monomer.

The off-lattice Monte Carlo simulations show the presence of two dominant ground state conformations (Fig. 2). In the first conformation [Figs. 2(a) and 2(b)] the proteins are hooked into each other and in the second conformation [Fig. 2(c)] the monomers are arranged parallel to each other while establishing a pore. As marked in Fig. 2, we assign the x and y directions to correspond to the direction of hooking and the direction of parallel arrangement, respectively. Two typical contact energy plots are given in Figs. 2(d) and 2(e). The energy minimum in Fig. 2(d) corresponds to the hooked conformation, and the one in Fig. 2(e) corresponds to the parallel state. As explained in Sec. II, the insets in Figs. 2(d) and 2(e) denote, respectively, the most favorable approaches along the x-axis and y-axis corresponding to the hooklike and parallel conformations of the dimer.

In addition to the two ground states, the energy plots also show the existence of many metastable states with higher contact energy [Figs. 2(d) and 2(e)]. Some of these metastable conformations are shown in detail in Fig. 3. Most conformations resemble the ground state conformations but are slightly shifted toward x- or y-direction, suggesting that they may play a role in further stabilizing the spontaneously formed lattice toward its final minimum energy conformation. We know from transmission electron microscopy studies in solution as well as from atomic force microscopy investigations on the S-layer formation on surfaces that these proteins do not tend to form three-dimensional aggregates under appropriate conditions as is the case in our simulations. The self-assembled sheets rather form patches with defined edges on solid or lipid surfaces[35,36] or various different structures like sheets, cylinders, or tubes in solution.[37] To investigate the role of those metastable states in the process of the lattice formation, future studies using replica exchange and other simulation methods might be interesting to get information about the underlying free energy landscape. However, for the present work, these metastable states are not included in the large-scale simulations as the main focus of our work lies

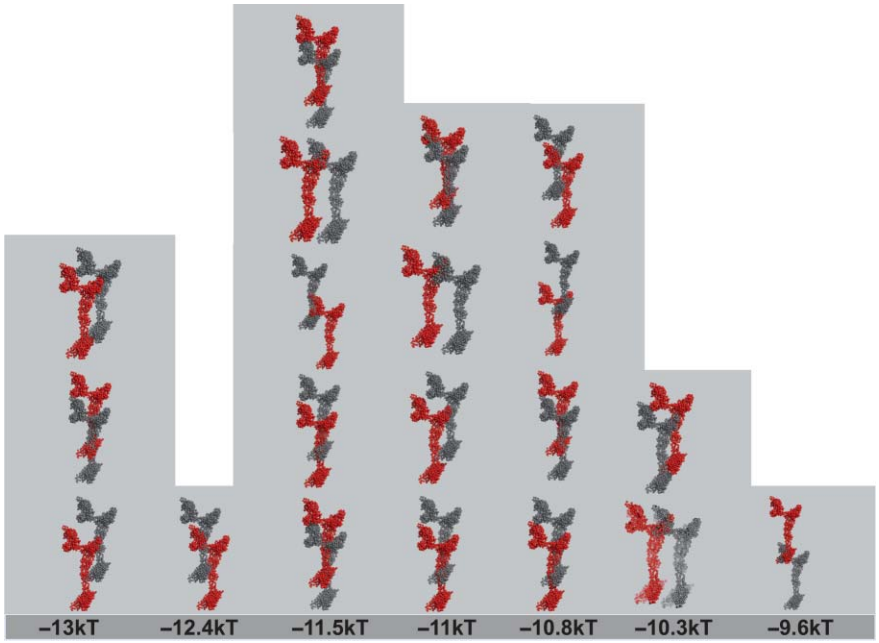

FIG. 3. Metastable states as obtained from the dimer simulations with different contact energy values. Most conformations are similar to the hook or to the parallel ground state, but slightly shifted in *x*- or *y*-direction.

on the structural details of the lattice in its minimum energy state.

It has been shown experimentally that the N-terminal region of the protein, comprising up to 208 amino acid residues, can be truncated without hindering the formation of the p1 lattice.[38] In order to find whether our coarse-grained model has the capacity to reproduce this experimental fact, we carried out the dimer simulations of the coarse-grained model without this N-terminal region, and indeed the ground state conformations were the same as that found for simulations of the entire protein. As a result, removal of the N-terminus with 208 amino acid residues is not detrimental to the assembly of a layerlike structure.

Additionally, the form of our interaction potential allows us to investigate separately the role of hydrophobic and electrostatic interactions in the self-assembly process. To this end, we performed simulations in which all amino acid side chains were kept uncharged. Interestingly, this did not affect the two ground state conformations obtained in the presence of electrostatic interactions. This result suggests that hydrophobic interactions are the dominant mechanism responsible for the self-assembly of the proteins.[39–41] Analysis of the ground state conformations also suggests that it is the amino acid residues on the surface of the monomer that are responsible for the dimer formation. Further support for this conjecture is provided by simulations performed with models with a second level of coarse-graining, where only hydrophobic amino acid side chains located on the surface of the monomer were taken into account, which resulted in the same two conformations for the ground state as the full protein. This stripped coarse-grained model contains only 137 amino acids. This small subset of necessary amino acid residues then explains the robustness of the S-layer self-assembly process with respect to genetic engineering and fusion proteins.

Experimentally, it is also known that S-layer proteins self-assemble in aqueous solutions independent of environmental conditions, such as *p*H, temperature, and salt concentration, which is necessary in order for them to carry out their function *in vivo*. Although charged groups are present on the surface of the monomers, our results show that the hydrophobic interactions are much stronger than the electrostatic interactions and only a few hydrophobic surface sites are involved in the required protein–protein interactions. This dominance of the hydrophobic interactions then presents an explanation for the robustness of the assembly process across the *p*H range, since this affects only the electrostatic interactions, which do not play a crucial role in the layer formation.

B. Large scale assembly

In order to investigate the growth kinetics of the large-scale assembly of S-layer sheets, we performed lattice Monte Carlo simulations using the energies of the two dominant ground state interactions obtained from the dimer simulation. The lattice simulations were performed in two ways. In order to investigate the growth rate, we performed seeded growth studies at different monomer concentrations. In these simulations, one protein monomer, the seed, was held fixed in the center of the simulation box, and the monomers are allowed to interact with the ground state energies only if they are part of the cluster which also comprises the seed monomer. The simulations were started from a random initial configuration [Fig. 4(a)], and the final configuration is a single

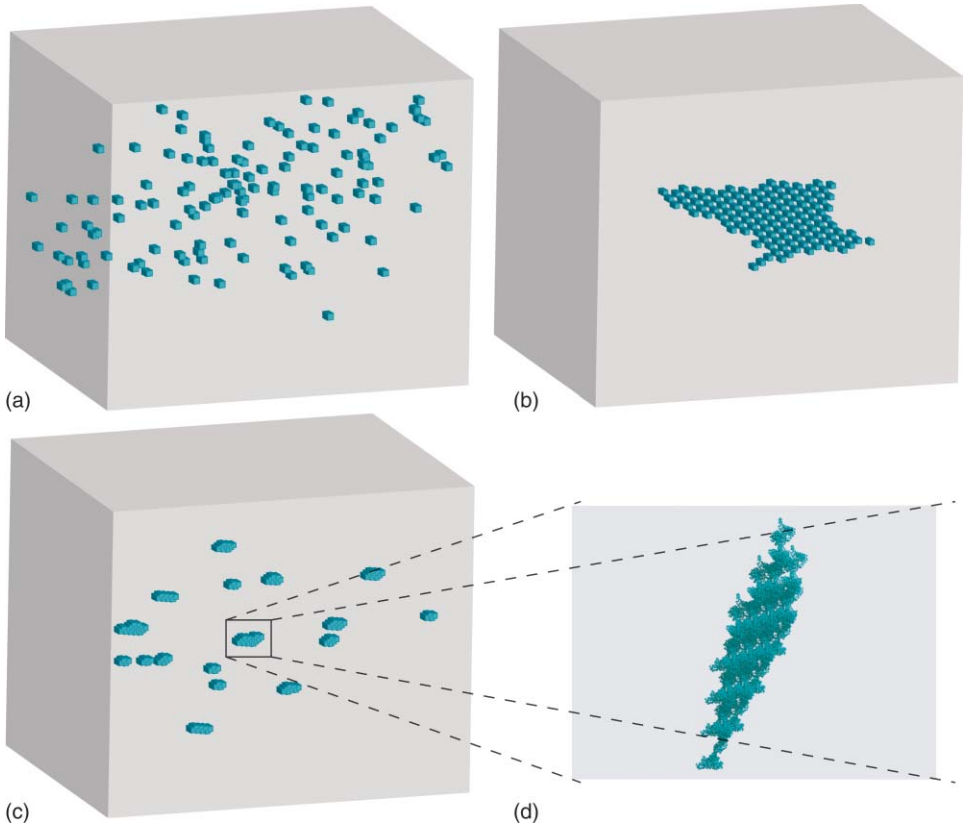

FIG. 4. Representation of the lattice Monte Carlo simulations of the large-scale self-assembly process. Energy values were taken from the interaction of two monomers. Each protein is represented by a unit cube, which is only a mnemonic representation. (a) Initial configuration: cubes are randomly distributed in the simulation box. (b) Single-seeded growth study. One monomer was kept fixed as the seed. Proteins interact with the seed resulting in a self-assembled two-dimensional sheet. (c) Competitive growth study. No seed was introduced to the system. Multiple sheets start to grow during the initial period. (d) Magnified view of the corresponding S-layer sheet using the coarse-grained model.

assembled sheet [Fig. 4(b)]. The number of monomers in the growing cluster increases with time before it reaches the saturation limit defined by the number of monomers in the system. This growth curve is shown for three different concentrations of monomers in Fig. 5(a). The growth curves are characterized by the absence of any lag time, which implies spontaneous assembly into layers. The growth curves show the typical features of a brief sigmoidal part for very short times, a linear dependence for a substantial time duration, and a saturation regime at very long times. The slope in the linear regime is taken as the growth rate. For the monomer concentrations $c = 0.001$, 0.0015, and 0.002 (corresponding to 125, 187, and 250 protein monomers, respectively), the growth rates are found to be 0.473×10^{-2}, 0.765×10^{-2}, and 1.18×10^{-2}, respectively (in units of number of monomers in the cluster per unit time). Thus, we have found that the growth rate increases linearly with the concentration of the monomer.

We have also studied the competitive growth of different S-layer sheets in the general case when no seed is introduced into the system. Sheets can now nucleate at random throughout the system, and multiple sheets start to grow during the initial period. It must be noted that we do not observe any nucleation barrier for the present situation, unlike the situation with the amyloid fibrillization.[29] As free monomers become depleted from the system, the different sheets compete with each other, and the growth of larger sheets takes place at the expense of dissolution of the smaller sheets, through the familiar Ostwald ripening or coarsening process. A sample configuration of such a system is shown in Fig. 4(c). We have also studied the distribution of cluster sizes in this case, and the distribution goes from a sharply peaked distribution about unity (corresponding to the monomers) for short times to a broader distribution peaked about a higher cluster size at later times [Fig. 5(b)]. As an example, at the Monte Carlo time of 10^6, the cluster size distribution is peaked at 6 with a very long tail extending to the number of monomers inside the lattice to be even more than 20. A decade of time later, the cluster size distribution is still peaked at 6, but now there are more clusters with higher number of monomers in them. This is analogous to our previous study on amyloid growth.[29] In the present study, we do not attempt to establish the time-exponent for the coarsening kinetics,[42] by relegating this to a future report.

C. Morphology of the assembled sheet

Figure 6 shows the resulting lattice structure. The architecture of the lattice exhibits identical pores with a diameter of 3.25 nm, which is in good agreement with

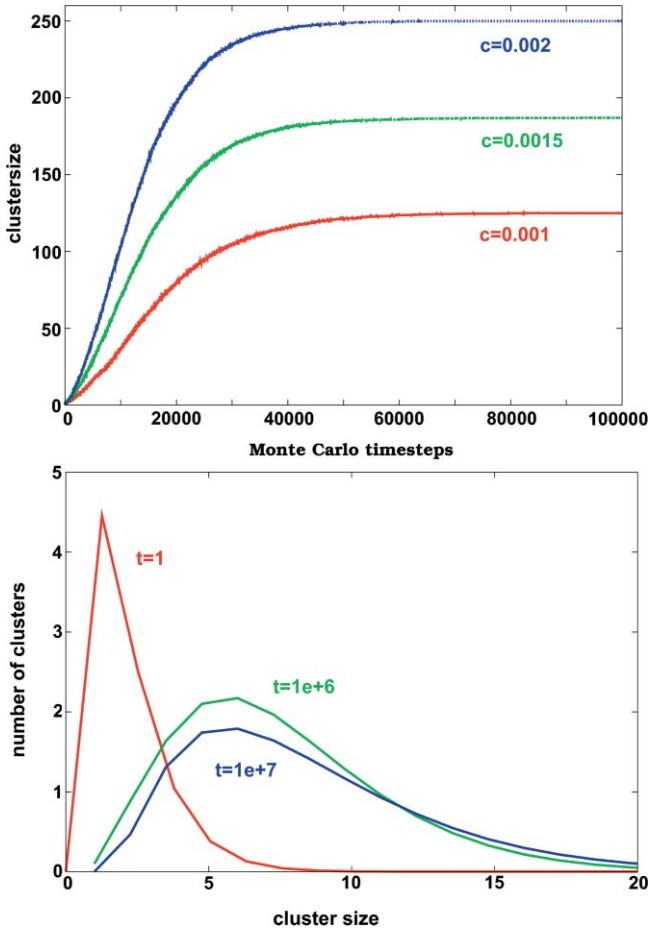

FIG. 5. Large-scale self-assembly. (a) Growth rate of the seeded cluster for different monomer concentrations. (b) Number distribution of the cluster size in the competitive growth case. The plots were smoothed using a Bezier curve. The initial distribution at $t = 1$ is scaled down by a factor of 10.

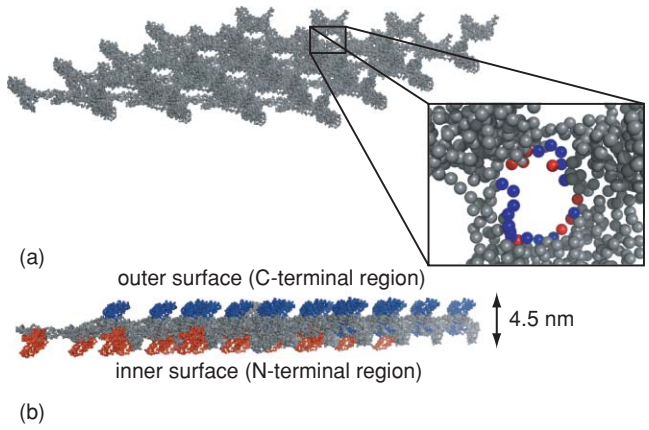

FIG. 6. Morphology of the self-assembled S-layer lattice as obtained by lattice Monte Carlo simulations. (a) Lattice made up of 16 protein monomers, represented by the coarse-grained model. The highly ordered structure is arranged as a p1 lattice and exhibits pores of identical size and morphology. The magnified view shows details of the pore. Hydrophilic residues are colored in blue while hydrophobic residues are colored in red. The pore is mainly hydrophilic with a net positive charge. Altogether 24 residues contribute to the pore. (b) Cross section of the lattice. The lattice has a width of 4.5 nm. The outer surface is composed of the C-terminal region (blue), the inner surface (*in vivo* anchored to the cell membrane) is composed of the N-terminal region (red). This architecture leads to the anisotropy of the lattice due to the different charge composition of the two terminal regions.

experimentally determined values.[21] The pores consist of mainly hydrophilic residues (magnified view in Fig. 6: blue beads), among which six are positively charged and three are negatively charged. Altogether 24 residues contribute to the pore. These residues are R100, V104, R187, G188, D189, Q192, T332, S333, S334, and D335 of the first monomer (mainly the N-terminal part), and S476, K478, A479, S480, F481, V483, F485, D487, K490, R491, T492, F493, K746, and L747 of the second monomer (mainly the C-terminal part). These pores are responsible for the transport of ions and other molecules into and out of the underlying cell, thereby functioning as selective ion gates, both with respect to charge as well as size, as has been experimentally determined.[43] The resulting lattice is 4.5 nm thick, which is consistent with experimentally measured values.[21] The lattice parameters as determined by small angle x-ray scattering and TEM experiments are $a = 9.9$ nm, $b = 7.6$ nm, and $\gamma = 81^\circ$.[21] Our simulations yield parameter values of $a = 11.1$ nm, $b = 6.4$ nm, and $\gamma = 88^\circ$. This difference might arise due to the local conformational adaptivity of the monomers, which was not included in this study. The parameters are in reasonable agreement given the simple nature of our model.

The final morphology of the p1 lattice clearly shows that the N-terminal region [Fig. 6(b) in red] and a large part of the C-terminal region [Fig. 6(b) in blue] extend out of the lattice plane and are not directly involved in the interactions responsible for the formation of the lattices. S-layer fusion proteins have been produced by fusing the partner molecule to the N-terminus of the protein.[44,45] These recombinant proteins did not lose their ability to self-assemble. The location of the N-termini in the lattice explains this behavior, because possible fusion partners do not disturb the necessary interactions directly.

The fact that both the N-terminal and the C-terminal regions are separated from the lattice surface leads to an insight about intersheet interactions. Based on the pK values[33] of the amino acid residues and the pH of the solution in which the monomers assemble, different levels of stacking of layers can be expected. Although the individual protein monomer has essentially a zero net charge under physiological pH conditions, the C-terminal region (300 aa) has a net negative charge (58 negatively charged and 26 positively charged residues) with mainly hydrophilic residues, whereas the N-terminal region (210 aa) has a very small net positive charge (28 positively charged and 23 negatively charged) with mainly hydrophobic amino acids [Fig. 1(a)].[35,46–49] This anisotropy of charge distributions gives an insight about the formation of bilayers and even multilayers. The interactions between two layers in solution can change depending on the pK values of the specific residues, and hence can affect the assembly along the direction perpendicular to the sheets. Below a pH of 3.5, both surfaces of the S-layer lattice are positively charged. This implies that the individual S-layer lattices repel each other to yield individual monolayers as the final assembled structure. Conversely, at a $p\mathrm{H} \gtrsim 13$, both surfaces of the lattice are negatively charged, and again we expect monolayers. Near physiological pH, all polar residues are fully charged, and hence

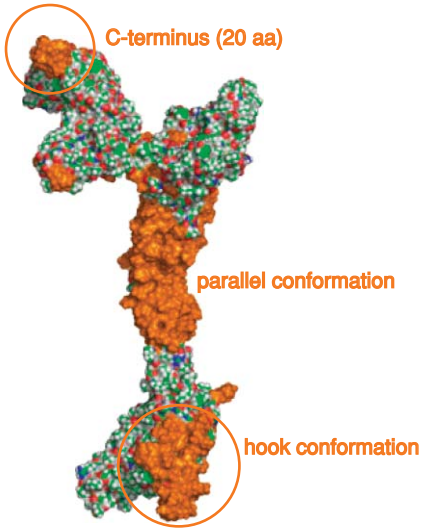

FIG. 7. Structural model of the S-layer protein SbsB. Amino acid residues involved in interactions in the p1 lattice are colored in orange. Twenty residues at the C-terminus are part of interactions as well as 35 residues near the N-terminus, which are involved in the formation of the hook conformation. The mainly unfolded domain in the middle part of the protein is part of the interactions required for the formation of the parallel dimer conformation. All orange colored residues are essential for establishing a stable lattice structure.

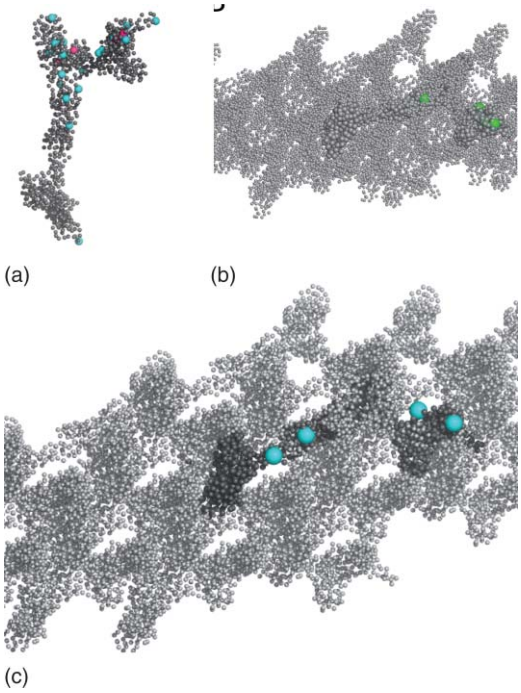

FIG. 8. Comparison of the calculated monomer and lattice structure with experimental results addressing the accessibility of individual residues (Ref. 51). (a) Twenty three residues have been experimentally determined as being accessible in the monomer structure. In the calculated structure 19 out of these 23 amino acids are located on the surface, which are colored in blue. Four residues are not accessible in the simulated monomeric model, which are represented by pink beads. (b) Residues that were experimentally determined to be part of intermolecular cross-links in the S-layer lattice are also seen in our model as represented by green beads. (c) Out of the five residues that were experimentally determined to be very accessible at the outer surface of the lattice, four are given by our simulations as shown in blue color.

the C-terminal face of the assembled lattice has a net negative charge, while the N-terminal face has a net positive charge. In this scenario, oppositely charged faces of individual lattices attract each other, and the final structure is a multilayer of individual S-layer lattices. For *p*H ranges close to the *p*I of one protein (*p*H $\sim$ 5), the inner surface of the lattice is nearly neutral, while the outer surface is net negatively charged. In this situation, the two neutral faces can come together to form bilayers. We therefore conjecture that the S-layer sheets will assemble as monolayers in the very basic and very acidic regimes, double layers close to the *p*I of the protein (5.2) and multilayers close to the physiological *p*H (7.0).

In order to quantitatively estimate how many amino acid residues are essential for the dimer formation, we define a sphere of radius four times the bead size around every amino acid residue. If any amino acid residue from the second monomer lies within this "overlap sphere," we defined the two residues to be part of overlaps. The overlap criterion as characterized by the radius of this sphere was determined as the one which gave physically realistic answers for the overlapping amino acid residues. It was found that altogether 241 amino acid residues of one monomer are involved in overlaps in the lattice, which corresponds to only 26% of the total number of residues. Figure 7 shows all required amino acid residues in detail (colored in orange). Analysis of the overlapping amino acid residues required for self-assembly shows that the very end of the C-terminus is required for the stability of the lattice structure (Fig. 7), which is in agreement with experimental observations.[44] This C-terminal region required for self-assembly is composed of 20 amino acid residues, with 15 of these being hydrophobic, three positively charged, and two negatively charged. This region plays a crucial role in the formation of the parallel dimer conformation, and cannot be truncated without loss of the ability to self-assemble correctly. The predominance of hydrophobic amino acid residues (75%) in this interacting part supports the hypothesis of hydrophobic interactions being regarded as the driving force for the self-assembly of S-layers. Thirty five amino acid residues near the N-terminal region (outside the 208 residues which can be deleted without affecting the self-assembly) and a part in the middle of the protein (Fig. 7 orange amino acids) are also essential parts of the interactions within the lattice by providing hydrophobic side chains for the formation of the two dimer ground states.

We also compared our lattice architecture with the results obtained by recent experimental studies addressing the surface accessibility of amino acid side chains of this S-layer lattice,[50–53] which is shown in Fig. 8. In this study, single-cysteine mutants of the protein were generated and then analyzed using the substituted cysteine accessibility method. Using this method, the experiments identified 23 residues that were accessible in the monomer, which implies that they are located on the surface of the monomer. In our coarse-grained model, 19 out of these 23 amino acid residues were found to be located on the surface [Fig. 8(a)]. The small discrepancy can arise either due to minor errors in the original structural model, or because of the nature of the coarse-graining strategy. Since the amino acid side chains are extended objects, some of these side chains might be accessible in the

experiments, even though the center of masses of these amino acids, corresponding to the position of our coarse grained beads, can lie just underneath the surface. Nevertheless, the success rate is appreciable, given the coarse-grained nature of our model. Further, these 23 amino acid residues were subject to a chemical cross-linking screen to identify their position in the assembled lattice. It was found that five out of these 23 amino acids were very accessible on the outer surface of the assembled S-layer lattice. In our simulations, we found that four out of these five residues are located on the outer surface [Fig. 8(c) blue beads], with the single residue that did not match being a part of the original four (out of 23) amino acid residues that did not match the surface accessibility data of the monomer. Out of the remaining 18 residues that are not on the outer surface of the assembled lattice, it was postulated experimentally that four are located at the interface between the interacting monomers in the lattice (intermolecular cross-links). All four of these were found to lie at the interface in our model also [Fig. 8(b) green beads]. These results demonstrate that the correct location of amino acid side chains was captured in the simple coarse-grained model of the protein, and thus offers a simple strategy to identify surface-accessible residues in the S-layer assemblies in place of extensive mutagenesis and chemical cross-linking studies.

IV. CONCLUSIONS

Our simulations of the self-assembly process of S-layer proteins using a coarse-grained model and Monte Carlo algorithms reveal the molecular structure of an S-layer lattice and the nonbonded interactions driving the self-assembly of these proteins. Our results reproduce all of the essential features observed experimentally: the anisotropy of the highly ordered lattice structure, the presence of hydrophilic pores bearing a net positive charge, the significant difference of the C- and N-terminal regions regarding the stability and interactions in the lattice, and the location of individual residues on the surface or at the interface of the two-dimensional sheets.

The underlying molecular mechanisms leading to the self-assembly of S-layer proteins are guided by the interaction of only few hydrophobic amino acid residues located on the surface of the proteins. The importance of hydrophobic interactions for the self-assembly of molecules is a well known phenomenon for lipids, small peptides, and other simple inorganic molecules. Our simulations show that the same basic principles also hold true for S-layer proteins, despite being larger in size and much more complicated in their tertiary structure, and can explain their self-assembly into functional structures with a defined morphology. S-layer proteins of different organisms differ enormously in their amino acid sequence and do not show any significant sequence homology to other protein classes. Nevertheless they all self-assemble into a variety of different lattice structures, and it would be interesting to study if the assembly in these proteins is also guided by the same hydrophobic interactions, and hence adhere to these common design rules.

In addition, our work also has potential applications in nanobiotechnology. Our analysis enables us to identify the essential amino acid residues required for the lattice formation. The rest of the sequence may be altered by introducing functionalized groups for different applications. Thus, this provides a tool in designing recombinant proteins for structures with desired features on the large scale by a specific variation of the primary sequence.

ACKNOWLEDGMENTS

Acknowledgment is made to AFOSR (FA9550-10-1-0159), AFOSR (FA9550-09-1-0342), NSF (DMR-0706454), and the Materials Research Science and Engineering Center (MRSEC) at the University of Massachusetts for support. C.H. holds a DOC-fFORTE fellowship of the Austrian Academy of Sciences.

[1] G. M. Whitesides and B. Grzybowski, Science **295**, 2418 (2002).
[2] G. M. Whitesides and M. Boncheva, Proc. Natl. Acad. Sci. U.S.A. **99**, 4769 (2002).
[3] J. Israelachvili, *Intermolecular and Surface Forces*, 2nd ed. (Academic, London, 1991).
[4] R. Lipowsky, Nature (London) **359**, 475 (1991).
[5] E. Sackmann, *Structure and Dynamics of Membranes: Generic and Specific Interactions* (Elsevier, Amsterdam, 1995).
[6] S. M. Douglas, H. Dietz, T. Liedl, B. Högberg, F. Graf, and W. M. Shih, Nature (London) **459**, 414 (2009).
[7] M. Muthukumar, C. K. Ober, and E. L. Thomas, Science **277**, 1225 (1997).
[8] J. T. Chen, E. L. Thomas, C. K. Ober, and G.-P. Mao, Science **273**, 343 (1996).
[9] D. K. Schwartz, Annu. Rev. Phys. Chem. **52**, 107 (2001).
[10] D. G. Angelescu and P. Linse, Soft Matter **4**, 1981 (2008).
[11] M. McCullagh, T. Prytkova, S. Tonzani, N. D. Winter, and G. C. Schatz, J. Phys. Chem. B **112**, 10388 (2008).
[12] C. M. Dobson, Nature (London) **426**, 884 (2003).
[13] C. M. Dobson, Nature (London) **418**, 729 (2002).
[14] U. B. Sleytr, Nature (London) **257**, 400 (1975).
[15] J. C. Rochet and P. T. Lansbury, Curr. Opin. Struct. Biol. **10**, 60 (2000).
[16] U. B. Sleytr, P. Messner, D. Pum, and M. Sára, Angew. Chem., Int. Ed. **38**, 1034 (1999).
[17] S. Chung, S.-H. Shin, C. R. Bertozzi, and J. J. De Yoreo, Proc. Natl. Acad. Sci. U.S.A. **107**, 16536 (2010).
[18] S. Whitelam, Phys. Rev. Lett. **105**, 088102 (2010).
[19] C. Horejs, D. Pum, U. B. Sleytr, and R. Tscheliessnig, J. Chem. Phys. **128**, 065106 (2008).
[20] C. Horejs, D. Pum, U. B. Sleytr, H. Peterlik, A. Jungbauer, and R. Tscheliessnig, J. Chem. Phys **133**, 175102 (2010).
[21] B. Kuen, A. Koch, E. Asenbauer, M. Sára, and W. Lubitz, J. Bacteriol. **179**, 1664 (1997).
[22] V. Tozzini, Curr. Opin. Struct. Biol. **15**, 144 (2005).
[23] J. K. Cheung and T. M. Truskett, Biophysical J. **89**, 2372 (2005).
[24] A. P. Heath, L. E. Kavraki, and C. Clementi, Proteins **68**, 646 (2007).
[25] *Monte Carlo Methods in Statistical Physics*, edited by K. Binder (Springer, Berlin, 1986).
[26] M. P. Allen and D. J. Tildesley, *Computer Simulation Of Liquids* (Clarendon, Oxford, UK, 1987).
[27] *Computer Simulations of Surfaces and Interfaces*, edited by B. Dünweg, D. P. Landau, and A. I. Milchev (Springer, Berlin, 2002).
[28] A. Y. Shih, A. Arkhipov, P. L. Freddolino, and K. Schulten, J. Phys. Chem. B **110**, 3674 (2006).
[29] J. Zhang and M. Muthukumar, J. Chem. Phys. **130**, 035102 (2009).
[30] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[31] J. Skolnick, L. Jaroszewski, A. Kolinski, and A. Godzik, Protein Sci. **6**, 676 (1997).
[32] P. D. Thomas and K. Dill, Proc. Natl. Acad. Sci. U.S.A. **93**, 11628 (1996).
[33] T. E. Creighton, *Proteins: Structures and Molecular Properties* (W. H. Freeman, San Francisco, 1992).
[34] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[35] D. Pum, M. Weinhandl, C. Hödl, and U. B. Sleytr, J. Bacteriol. **175**, 2762 (1993).

[36]J. L. Toca-Herrera, R. Krastev, V. Bosip, S. Küpcü, D. Pum, A. Fery, M. Sára, and U. B. Sleytr, Small **1**, 339 (2005).
[37]P. Messner, D. Pum, and U. B. Sleytr, J. Ultrastruct. Mol. Res. **97**, 73 (1986).
[38]D. Rünzler, C. Huber, D. Moll, G. Köhler, and M. Sára, J. Biol. Chem. **279**, 5207 (2004).
[39]C. Tanford, Science **200**, 1012 (1978).
[40]D. Chandler, Nature (London) **437**, 640 (2005).
[41]E. E. Meyer, K. J. Rosenberg, and J. Israelachvili, Proc. Natl. Acad. Sci. U.S.A. **103**, 15739 (2006).
[42]M. K. Mitra and M. Muthukumar, J. Chem. Phys. **134**, 044901 (2011).
[43]S. Sotiropoulou, S. S. Mark, E. R. Angert, and C. A. Batt, J. Phys. Chem. C **111**, 13232 (2007).
[44]D. Moll, C. Huber, B. Schlegel, D. Pum, U. B. Sleytr, and M. Sára, Proc. Natl. Acad. Sci. U.S.A. **99**, 14646 (2002).
[45]H. Tschiggerl, J. L. Casey, K. Parisi, M. Foley, and U. B. Sleytr, Bioconjugate Chem. **19**, 860 (2008).
[46]M. Sára and U. B. Sleytr, J. Bacteriol. **169**, 2804 (1987).
[47]D. Pum, M. Sára, and U. B. Sleytr, J. Bacteriol. **171**, 5296 (1989).
[48]M. Sára, D. Pum, and U. B. Sleytr, J. Bacteriol. **174**, 3487 (1992).
[49]C. Mader, S. Küpcü, M. Sára, and U. B. Sleytr, Biochim. Biophys. Acta **1418**, 106 (1999).
[50]S. Howorka, M. Sára, Y. Wang, B. Kuen, U. B. Sleytr, W. Lubitz, and H. Bayley, J. Biol. Chem. **275**, 37876 (2000).
[51]H. Kinns and S. Howorka, J. Mol. Biol. **377**, 589 (2008).
[52]T. O. Yeates and J. E. Padilla, Curr. Opin. Struct. Biol. **12**, 464 (2002).
[53]R. F. Service, Science **298**, 2322 (2002).