Self-Organized Morphology and Multiscale Structures of CoVE Proteins

PORNTHEP SOMPORNPIST¹ and R.B. PANDEY ^{2,3}

1.—Center of Excellence in Computational Chemistry, Department of Chemistry, Chulalongkorn University, Bangkok 10330, Thailand. 2.—School of Mathematics and Natural Sciences, University of Southern Mississippi, Hattiesburg, MS 39406-5043, USA. 3.—e-mail: ras.pandey@usm.edu

Self-organizing structures of CoVE proteins have been investigated using a coarse-grained model in Monte Carlo simulations as a function of temperature (T) in a range covering the native (low T) to denatured (high T) phases. The presence of even a few chains accelerates the very slow dynamics of an otherwise free protein chain in the native phase. The radius of gyration depends nonmonotonically on temperature and increases with the protein concentration in both the native and denatured phase. The density of organized morphology over residue-to-sample length scales (λ) is quantified by an effective dimension (D) that varies between ~ 2 at high to ~ 3 at low temperatures at $\lambda \sim R_g$ with an overall lower density ($D \sim 2$) on larger scales. The magnitude of D depends on temperature, length scale, and concentration of proteins, i.e., $D \sim 3.2$ at $\lambda \sim R_g$, $D \sim 2.6$ at $\lambda > R_g$, and $D \sim 2.0$ at $\lambda \gg R_g$, at $T = 0.024$.

INTRODUCTION

Exploiting the self- and directed assembly of proteins and peptides has attracted enormous interest for years, particularly for the design of smart biomaterials that exhibit desirable multiscale (e.g., nano, micro, meso, etc.) properties; For example, Loose et al.¹ provided an overview on lessons from proteins with “their ability to self-organize in space and time on different length scales.” The importance of the self-assembly of peptides and proteins in manipulating structures at a range of length scales for nanomaterials^{2,3} is not new.^{4–8} The unique evolution of multiscale structures for the self-assembly of specific proteins is also critical for understanding their unique biological functions. Examples include the collective morphology of nucleosomes with a huge variation in length scales,⁹ fibril formation of amyloid in neurodegenerative diseases,¹⁰ the creation of a network of specific pathways for selective transport in ion channels,¹¹ etc. The specificity of residues, which enables the unique and versatile response of a protein, offers a

great opportunity to explore a diverse range of structures emerging from the interplay between residue–residue and segmental interactions as well as thermal noise and entanglement. The self-organizing morphology of each protein with its unique structures leads to a collective structure with specific multiscale characteristics. These diverse structures may, however, converge to a unique morphology with universal characteristic at high temperatures at which residue–residue interactions become irrelevant. Thus, varying the temperature and concentration provides an opportunity to manipulate the multiscale structures resulting from the self-assembly of each protein.

CoVE is the smallest structural membrane protein of the SARS-CoV-2 virus.^{12–17} It has been proposed that its biological function is to act as a viroporin and self-assemble in the host membrane.¹² The interaction of the virus with external agents (i.e., humans or animals) as well as a wide range of organisms of diverse sizes and shapes, environmental particulates such as soot, dust, pollen, smoke, water droplets, etc.^{13–16} may lead to its disintegration. As a result, its constitutive elements that diffuse, assemble, and bind may remain inert and/or interact with its surroundings at various

scales. Before considering the complexity of such interactions with the real environment, it is prudent to understand how the constitutive elements (e.g., proteins) organize in a simple environment. Because of its smallest size, it is interesting to examine how the structural protein CoVE organizes and disperses when many proteins interact and perform stochastic movements in a simulation domain. It is worth pointing out that some generic environmental effects are implicitly included via the type of residue–residue interactions considered in the computer model simulated here (see below).

Self-assembly of proteins^{1–11} may result in a matrix that can provide mechanical strength. Depending on its specific characteristics, it may add flexibility with a stable dynamical response to the underlying environment such as a membrane. On the other hand, dispersion of proteins into a fluid may enhance its viscosity due to its hierarchical morphology. It is believed that CoVE may provide a pathway in membrane channels^{14–16} for selective transport of viral components.

This dichotomy between the cooperative and conflicting nature of the collective response of organized assembly of proteins provides enormous opportunities for designing materials with wide-ranging and specialized characteristics, as mentioned above; For example, the aggregation of proteins triggered by their conformational changes may lead to undesirable structures such as fibrils of amyloid β -proteins, whereas the self-assembly of proteins into a viral capsid is critical for preserving its genome from external factors.¹⁰ We focus on the effect of temperature on the self-assembly of CoVE proteins,¹⁷ which has not been investigated in depth to the best of the authors' knowledge.

MODEL

A coarse-grained model of a protein is used on a cubic lattice as in our previous studies.^{9–11} The CoVE protein is a chain consisting of $L_c = 76$ residues^{12,17} in a specific sequence. A residue is represented by a cubic node of size $(2a)^3$, where a is the lattice constant. The atomic-scale details of the residues are ignored, but their specificity is captured via unique interactions (see below). The covalent (peptide) bonds between consecutive nodes (residues) are flexible, with bond lengths varying between 2 and $\sqrt{10}$ in units of the lattice constant (a).¹⁸ There are ample degrees of freedom for each residue to perform stochastic movements and their covalent bonds to fluctuate (unlike minimalist lattice models with constant bond lengths), which can be further enhanced by fine-graining.⁹ It should be pointed out that such bond-fluctuation models have been used extensively to investigate the structure and dynamics (including the viscoelastic properties) of complex polymer systems for which it is not feasible to reach all scales with fine-grained models

in a continuum host space.¹⁸ The advantages and pitfalls of such approaches have been well explored. Protein chains are more complex than a polymer chain, so we have extended and developed this efficient method to capture the specificity of each residue to model proteins.⁹

A number (N_c) of protein chains are considered in a cubic box of size L^3 . The protein chains are placed, one at a time, on the trail of a random walk of a cubic node with excluded volume constraints. It becomes increasingly difficult to insert more chains^{9–11} as the number N_c increases. The protein chains are moved around with an excluded volume constraint to randomize their distribution when preparing the sample. Each residue interacts with the surrounding residues (both inter- and intra-chain) as described by a generalized Lennard–Jones (LJ) potential,

$$U_{ij} = \left[\epsilon_{ij} \left(\frac{\sigma}{r_{ij}} \right)^{12} + \epsilon_{ij} \left(\frac{\sigma}{r_{ij}} \right)^6 \right], r_{ij} < r_c \quad (1)$$

where r_{ij} is the distance between the residues at site i and j , $r_c = \sqrt{8}$ and $\sigma = 1$ in units of the lattice constant; the cutoff range of the interaction (r_c) is selected to cover most pertinent interactions as in a generic LJ potential. To capture the specificity of each residue, the interaction strength ϵ_{ij} is selected uniquely for each residue–residue interaction pair, with appropriate attractive and repulsive values using the knowledge-based residue–residue (KBRR) interactions.^{19–24} Note that a huge ensemble of protein structures from the Protein Data Bank (PDB) has been used to derive residue–residue contact interactions and applied extensively for modeling protein structures for decades.^{19–24} The KBRR interactions¹⁹ implemented here have been used to investigate the structure and dynamics of several proteins as well as the self-assembly of peptides and proteins.^{9–11,17} It is worth pointing out that some effects of generic environments are implicitly taken into account via the KBRR interactions, since the snapshots of proteins in the PDB are in their appropriate solvents.

Each residue in each protein chain (where both the residue and chain are selected randomly) performs stochastic movement via the Metropolis algorithm, provided that the excluded volume constraints are satisfied. That is, a residue from site i is moved to one of its randomly selected site j with the Boltzmann probability $\exp(-\Delta E_{ij}/T)$, where $\Delta E_{ij} = E_j - E_i$ is the change in energy between the proposed new (E_j) and the old (E_i) configuration, and T is the temperature in reduced units of the Boltzmann constant k_B . Attempts to move each residue once define the unit Monte Carlo step (MCS),¹⁸ i.e., the number $N_c \times L_c$ of attempts to move randomly selected residues defines the unit MCS. Each simulation is performed for a

sufficiently long time ($t = 10^7$ MCS) to identify the trend and estimate the average values of a number of local and global physical quantities.

We consider a number of protein chains of $N_c = 1$ to 200, mostly on a 150^3 lattice for a range of temperatures $T = 0.010 - 0.040$ covering the native to denatured phases. The number of independent samples varies from 10 to 500 (where a lower number has a larger number of protein chains). We investigate physical quantities such as the energy of each residue and protein chain, their mobility and contact profiles, the mean-square displacement of the center of mass of the protein, radius of gyration, and its structure factor to provide insight into the morphological evolution as a function of temperature and the number of protein chains.

RESULTS AND DISCUSSION

As described above, N_c chains of CoVE protein, each with a random configuration, are distributed randomly. The conformation of each protein undergoes configurational evolution through the time steps as its residues perform stochastic motion according to the Metropolis algorithm at a temperature T . The protein chains move as a result of the collective dynamics of their residues. The segments of one protein may come into contact (i.e., within the range of residue–residue interaction) with the segments of other proteins, and may even entangle depending on the number of protein chains (i.e., the volume fraction) in the simulation domain. Intra- and interprotein residue–residue interactions may result in a set of multiscale interacting segments that can become bound (stick together) due to their noncovalent attractive interactions or remain unbound due to thermal agitation. Some segments of proteins may remain isolated due to repulsive segmental interaction, while others may be trapped due to entanglement (i.e., geometrical constraints) in a frustrated state. Exploring such self-assembly and dispersion effects of a collection of specific proteins such as CoVE may lead to unique multiscale morphology due to the competition between the residue–residue interactions and thermal noise along with the entanglement.

Before considering the collective behavior of CoVE proteins, it is important to recall the effect of temperature on the conformational evolution of an isolated protein¹⁷ where intrachain residue–residue interactions compete with thermal noise. The global conformation of CoVE exhibits a non-monotonic transition¹⁷ from the native to denature phase with increasing temperature: the radius of gyration of the CoVE chain decays on heating in the native phase, in contrast to the continuous increase towards a random-coil conformation in the denatured phase. The contact profiles of CoVE from the simulations clearly reveal its distinct and unique segments¹⁷: N-terminal, transmembrane, and C-terminal. With these basic characteristics on the

thermal response of a CoVE chain, it will be easier to see the structural changes, if any, due to the presence of many proteins in a crowded environment in the following simulations.

Figure 1 shows snapshots of proteins in equilibrium in the native ($T = 0.020$) and denatured ($T = 0.028$) phase with a number of CoVE chains of $N_c = 200$. There is a clear distinction in the overall appearance of the morphology at these temperatures, with local (small-scale) aggregations in the native phase versus large-scale dispersion and little aggregation in the denatured phase. The distribution of local aggregates, albeit random, appears to be relatively uniform at the low temperature. The conformation of individual proteins appears to be globular in the native phase but opens up to a random-coil structure in the denatured phase. These snapshots are just two of an enormous ensemble of configurations at each temperature; they do however provide a glimpse of the effect of temperature on the self-organizing behavior of the CoVE proteins and their dispersion. Attempts are made to analyze a number of local and global physical quantities, i.e., the contact profile, variation of the root-mean-square (RMS) displacements, radius of gyration, and structure factor, which may provide some quantitative measures for identifying trends in the self-assembly of CoVE proteins.

RMS Displacement

How the proteins move is important to understand the evolution of their organizing structures. The multiscale dynamics of a homopolymer chain is relatively well understood. Because of the unique composition of a protein, i.e., the number of residues in a specific sequence and their specific interactions, its dynamics depends on the matrix in which it is embedded and the temperature. A free CoVE protein ($N_c = 1$) moves very slow (quasistatic) in its native phase (at low temperature) as it adopts globular configurations but becomes diffusive in the denatured phase (at higher temperatures). In a crowded environment with many proteins, the dynamics may be altered due to intra- and inter-chain residue–residue interactions, excluded volume constraints, and entanglement.

The variation of the average RMS displacement (R_c) of the center of mass of each protein with increasing time step (t) is presented in Fig. 2 for a range of low to high temperatures ($T = 0.020$ to 0.030) and a number of proteins 5 or 200. The range of the dynamics (slow to fast) can be characterized by estimating the power-law exponent (ν) with a scaling dependence of $R_c \propto t^\nu$, where $\nu = 1/2$ represents diffusion, $\nu = 1$ describes drift, and $\nu < 1/2$ indicates subdiffusive behavior.

The RMS displacement of a free chain ($N_c = 1$) at a low temperature ($T = 0.020$) is also included in the figure to act as a reference to show its slow ($\nu \rightarrow 0$) dynamics in comparison with the faster movements

Fig. 1. Snapshots of CoVE proteins at low ($T = 0.020$, left) and high ($T = 0.028$, right) temperatures with a number of protein chains of $N_c = 200$ on a 150^3 lattice at the end of 10 million time steps. Five representative protein chains are shown in different colors than the rest for easy visualization of their conformation.

Fig. 2. Average RMS displacement (R_c) of center of mass of a protein chain versus time step (t) with a number of proteins of $N_c = 5$ or 200 on a 150^3 lattice with 10 to 500 independent samples for the temperature range of $T = 0.020$ to 0.030 . Data for a free chain ($N_c = 1$) at $T = 0.020$ are also included to illustrate the difference in the variation of the RMS displacement of the protein with the time step in the presence of more than one protein chain. The slopes of the asymptotic data are indicated at each temperature to characterize the dynamics.

in the presence of more chains, i.e., $\nu \approx 0.446$ for $N_c = 5$. The slow dynamics of a free protein chain at the low temperature is characteristics of an overall attractive residue-residue intrachain interaction. One would expect that adding chains might inhibit the dynamics of the protein due to steric and hardcore excluded volume interactions, but the opposite is observed in the native state (at the low temperature of $T = 0.02$). Indeed, the presence of even a few protein chains accelerates their motion due to the interference of interchain interactions with the intrachain interactions; a similar

phenomenon has also been observed in the self-organizing process of other proteins.^{9,10} The asymptotic dynamics of each protein chain in a dilute environment ($N_c = 5$) remains close to diffusion, with a slightly faster motion at high temperatures (e.g., $T = 0.028$ and 0.030). Addition of many protein chains to form a somewhat crowded environment ($N_c = 200$) does not seem to alter the asymptotic dynamics of each protein over the entire temperature regime ($T = 0.020$ to 0.030). This is perhaps because this relatively low concentration of proteins is not sufficient to encounter severe constraints,

e.g., excluded volume and entanglement, but enough to interact with surrounding chains for self-assembly in the relatively long time required to reach the asymptotic diffusive dynamics. Unfortunately, it is not feasible to consider higher protein concentrations at present. Despite the relatively stable (almost diffusive) dynamics over the entire temperature range, the structures of each protein on average do vary with temperature, as described in the next section.

Radius of Gyration

The average radius of gyration (R_g) of the protein chain at each temperature for a fixed number of protein chains can be estimated from the ensemble of their conformations in about last one-third of the time steps in each independent run. Figure 3 shows the variation of the average radius of gyration with temperature with a number of protein chains from $N_c = 1$ to 200. The radius of gyration of a single protein chain ($N_c = 1$) decays with increasing temperature in the native phase ($T = 0.010$ and 0.020) but increases with temperature on further heating in the range of $T = 0.020$ to 0.030 as the protein denatures to a random-coil conformation.¹⁷ Such a nonmonotonic thermal response appears to be characteristic of membrane proteins.¹¹ The protein chains conform to the smallest globular conformation at $T = 0.020$ and reach the most expanded random coil at $T = 0.030$.

This nonmonotonic dependence of the radius of gyration on temperature appears to persist with the addition of a few protein chains ($N_c = 5, 20$), where the global dynamics of the protein chain is much

faster than that of a free chain ($N_c = 1$) (Fig. 2). Further increasing the number of protein chains ($N_c = 50$ to 200) enhances their contacts and therefore protein–protein interactions, and the rate of decay of R_g with T in the native phase reduces with crowding. The minimum magnitude of the radius of gyration in the native state at $T = 0.020$ increases with the crowding (Fig. 3). The maximum magnitude of the radius of gyration in the denatured state at $T = 0.030$ also increases with the crowding. The self-organizing structure of each protein and the resulting morphology at larger length scales at extreme native and denatured states exhibit a systematic density distribution that can be analyzed by examining the structure factor (see below). Let us examine the average contact profile of each residue next before analyzing the multiscale structures as the size of proteins and their segments reorganize with temperature and crowding.

Contact Profile

The contact map of a protein is essentially a snapshot (Fig. 1) of the contacts of each residue with others along its contour. The profile of the contacts is the average number N_n of residues around each within the range of interaction along the backbone of each protein. The average is performed over the number of independent runs, number of protein chains, etc. For example, for a simulation with a free protein chain, the average involves the number of independent samples of the last snapshot at the end of the simulation in equilibrium. With many proteins, the average involves the number of independent samples as well as the number of proteins.

Fig. 3. Variation of average radius of gyration (R_g) of CoVE protein with temperature (T) for a number of proteins from $N_c = 1$ to 200. Inset shows variation of R_g with number of chains N_c in native ($T = 0.020$) and denatured ($T = 0.030$) phase to illustrate how crowding enhances the spread of the protein. Statistics the same as in Fig. 2.

For a free protein ($N_c=1$), it is easy to identify three distinct domains:¹⁷ the N-terminal (¹M–¹⁷V), transmembrane segment (¹⁸L–⁴⁴C), and C-terminal (⁴⁵N–⁷⁶V) of the CoVE protein chain from the variation of the contact profiles with temperature. In the native phase, the protein chain is globular with a high degree of residue–residue contacts in the transmembrane region and sporadic variations in the number of residue contacts N_n in the N- and C-terminals. At higher temperature, the globularization is reduced to vanishingly low (N_n) contacts in the N- and C-terminals while retaining an appreciable globularization in the transmembrane segment. Thus, the contact profile depends on temperature.

The contact profiles for a number of protein chains of $N_c = 1$ to 200 for the range of temperatures from $T = 0.020$ to 0.030 are presented in Fig. 4. Adding even a few protein chains (i.e., $N_c = 5$ or 20) alters the contact profile considerably from that of a single protein chain ($N_c = 1$); note the similar effect on the global dynamics (Fig. 2). Furthermore, there is very little change in the contact profile in the temperature range from $T = 0.020$ to 0.026 , i.e., in the native to denatured phase, with a relatively small number of protein chains ($N_c = 5$ or 20). The addition of protein chains appears to reduce the effect of temperature on the contact profile in a

certain range, perhaps due to interchain segmental organization unlike the intrachain segmental assembly with $N_c = 1$. Raising the temperature to $T = 0.028$ shows a systematic growth in the contact profile for the higher numbers of chains $N_c = 50$, 100 , and 200 , particularly in the C-terminal segment of the protein. The higher segmental mobility at higher temperature ($T = 0.028$) enhances the self-assembly as the proteins denature (Fig. 4). A similar trend regarding the effect of crowding on the contact profile at $T = 0.030$ is expected at higher protein concentrations, although this regime remains beyond our reach at present. The self-organizing structures depend on the temperature with enhanced crowding.

Structure Factor

The global structures of a collection of interacting proteins in a crowded environment evolve as their residues and segments perform stochastic motion due to thermal noise. Complex interactions at multiple length scales (residue, segments of proteins, and beyond), excluded volume constraints, and entanglement in the crowded domain lead to diverse structures that vary with the length scale. Multiscale structures due to competing and cooperative mechanisms such as aggregation, dispersion, trapping, etc. are difficult to characterize precisely.

Fig. 4. Average number of surrounding residues within the range of interaction around each residue along the backbone of the protein for a range of temperature from $T = 0.020$ to 0.030 with a number of protein chains of $N_c = 1$ to 200. Statistics the same as in Fig. 2 except the number of independent samples (10 to 500) is higher for fewer chains.

One can, however, gain some insight from analysis of the structure factor $S(q)$, defined as

$$S(q) = \left\langle \frac{1}{N} \left| \sum_{j=1}^N e^{-i\vec{q} \cdot \vec{r}_j} \right|^2 \right\rangle_{|\vec{q}|} \quad (2)$$

where r_j is the position of each residue in all the protein chains and $|q| = 2\pi/\lambda$ is the wavevector corresponding to a wavelength λ . By considering a power-law scaling of the structure factor with the wavevector, i.e.,

$$S(q) \propto q^{-1/\gamma} \quad (3)$$

one can study the spread of residues over the length scale λ by evaluating the exponent γ . The wavelength λ can vary from the size of a residue to the size of the simulation domain, spanning over the size of the protein. The radius of gyration (R_g) is a measure of the average size of each protein (Fig. 3) evaluated from the distribution of its residues in its conformational ensemble. For a length scale comparable to the average radius of gyration ($\lambda \sim R_g$), one can evaluate how the average radius of gyration of the protein scales with its number of residues N (a measure of the mass of the protein) through a power law, $R_g \propto N^\nu$, where $\nu = \gamma$ and the effective dimension (D) of the protein is $D = 1/\gamma$. The effective dimension (D) of the self-assembly can be estimated for the entire range of length scales (λ), i.e., $\lambda \geq R_g$.

Figure 5 shows the variation of the structure factor (S) with the wavelength (λ) at representative temperatures from $T = 0.020$ to 0.030 for the most

crowded protein environment ($N_c = 200$). In the native phase ($T = 0.020$), the effective dimension of the protein chain is $D \sim 3.4$ at a length scale comparable to the radius of gyration ($\lambda \sim R_g$), which implies that each protein chain is globular. On large scale ($\lambda \geq R_g$), the effective dimension is $D \sim 1.9$, which suggests that the globular chains are randomly distributed over the entire sample in the native phase (see also Fig. 1). The variation of the structure factor of the self-assembly of the crowded proteins with the wavelength (λ) in the denatured phase ($T = 0.030$) on the other hand exhibits oscillation. The power-law exponent of $D \sim 1.9$ at $T = 0.030$ implies that extended protein chains are distributed randomly at all length scales. At an intermediate (denaturing) temperature, i.e., $T = 0.024$, the effective dimension D varies from ~ 3.2 at $\lambda \sim R_g$ to $D \sim 2.6$ at $\lambda > R_g$ and $D \sim 2.0$ at $\lambda \gg R_g$. Thus, from the variation of the structure factor with the wavelength, one can estimate the mass distribution of the assembly at all length scales at each temperature. The self-organizing structure of proteins is unique with a specific distribution of its residues over the length scales. Since each protein is unique with diverse and versatile conformations, their self-assembly should provide a great opportunity for designing biomaterials with desirable characteristics.

CONCLUSION

Large-scale Monte Carlo simulations are performed to investigate the self-organizing morphology of CoVE proteins using a coarse-grained model

Fig. 5. Variation of the structure factor S with the wavelength λ for 200 protein chains in the simulation domain at a range of temperatures from $T = 0.020$ to 0.030 . Top-left inset shows the structure factor at extreme temperatures in the native ($T = 0.020$) and denatured ($T = 0.030$) phase to highlight the extreme differences. The magnitude of the slopes with the segments of fit data are included for representative temperatures for clarity. Bottom-right inset shows the variation of R_g with temperature for a length scale guide.

for a range of temperatures (spanning from the native to denatured phase) with different numbers of protein chains ($N_c = 1$ to 200) on a 150^3 lattice. Their visualization shows clear differences in morphology with temperature from small to large scales, which are probed in depth by analyzing local and global physical quantities.

The dynamics and structure of a free CoVE protein chain ($N_c = 1$) depend on temperature: in the native phase, it moves very slowly and conforms to globular configurations, while in the denatured phase, it diffuses and conforms to random coils. The radius of gyration of the CoVE chain decays on raising the temperature in the native phase until a certain temperature ($T = 0.020$), reaching a minimum ($R_g = R_{g1}$) beyond which it increases to a random-coil conformation in the denatured phase with a maximum value ($R_g = R_{g2}$) at a temperature of $T = 0.030$, which is characteristic of a membrane protein. Proteins accelerate in the native phase with the addition of even a few protein chains to the domain due to cooperative and competing inter- and intrachain interactions. The rate of decay of the radius of gyration of the protein with temperature in the native phase reduces with the number of chains and almost vanishes with appreciable crowding ($N_c = 200$). The magnitude of both R_{g1} and R_{g2} in the native and denatured phase, respectively, increases with increasing crowding.

The contact profiles show three distinct segments of a free CoVE chain in the native phase, where the N- and C-terminals with least segmental organization are separated by a transmembrane segment with the most globularization, which decreases with temperature. Adding a few chains ($N_c = 5, 20$) retains the distinct features of the contact profile in the native phase to the low-temperature denaturing regime ($T = 0.020$ to 0.026), which suggests the onset of some segmental co-organization. The self-organized assembly with the addition of more chains ($N_c = 50$ to 200) not only retains these features but also induces segmental globularization even at higher temperature, i.e., $T = 0.028$. Segmental self-assembly appears to be enhanced with crowding due to increased protein-protein interactions.

The self-organizing (collective) morphology of proteins in bioinspired materials becomes more important as the protein concentration is increased, i.e., with higher levels of crowding, where multi-scale interactions (residue-residue to protein-protein) play a significant role. The largest number of CoVE protein chains considered in this study is $N_c = 200$, for which multiscale structures are analyzed in depth based on the scaling of the structure factor $S(q)$. In the native phase at $T = 0.020$, the protein chains are globular with effective dimension $D \sim 3$ on a scale comparable to their radius of gyration. At larger scales ($\lambda \geq R_g$), the global morphology becomes relatively sparse with an effective dimension of $D \sim 2$ and a uniformly random distribution

of globular proteins. The global morphology in the denatured phase at $T = 0.030$ is also a uniform random distribution of interacting proteins, each of which conform to a random-coil configuration. At an intermediate temperature of $T = 0.024$ at which the protein chain is partially denatured, the effective dimension varies with the length scale with $D \sim 3.2$ at $\lambda \sim R_g$, $D \sim 2.6$ at $\lambda > R_g$, and $D \sim 2.0$ at $\lambda \gg R_g$.

The self-assembly depends on the temperature, where the density of the protein material varies with the length scale. However, the nature of the dependence of the physical properties such as the density distribution on the length scale and temperature presented here is specific to self-assembly of CoVE and will differ considerably from that of other proteins such as lysozyme.¹⁰ Thus, understanding the self-organizing structures of proteins including CoVE provides an enormous opportunity to design customizable materials with a desirable density distribution over specific length scales to achieve fine-tuned properties.

DECLARATIONS

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. M. Loose, K. Kruse, and P. Schwill, *Annu. Rev. Biophys.* 40, 315. (2011).
2. K. Matsuura, *RSC Adv.* 4, 2942. (2014).
3. E. De Santis, and M.G. Ryadnov, *Chem. Soc. Rev.* 44, 8288. (2015).
4. S. Ahn, L.F. Deravi, S.-J. Park, B.E. Dabiri, J.-S. Kim, K. Kit Parker, and K. Shin, *Adv. Mater.* 27, 2838. (2015).
5. Q. Luo, C. Hou, Y. Bai, R. Wang, and J. Liu, *Chem. Rev.* 116, 13571. (2016).
6. H. Sun, Q. Luo, C. Hou, and J. Liu, *Nano Today* 14, 16. (2017).
7. P. Katyal, M. Meleties, J.K. Montclare, and A.C.S. Biomater, *Sci. Eng.* 5, 4132. (2019).
8. G. Thakur, and T. Thundat, *J. Nanomed. Res.* 6, 00154. (2017).
9. R.B. Pandey, and B.L. Farmer, *J. Chem. Phys.* 141, 175103. (2014).
10. R. B. Pandey, B. L. Farmer, and Bernard S. Gerstman, *AIP Adv.* 5, 092502 (2015).
11. W. Rangubpit, S. Kitjaruwankul, P. Boonamnaj, P. Sompornpisut, and R.B. Pandey, *AIMS Biophys.* 6, 68. (2019).
12. For sequence, see <https://www.uniprot.org/uniprot/P59637> as of February 24, 2021.
13. D. Schoeman, and B. C. Fielding, *Virol. J.* 16, 69. (2019).
14. Y.L. Siu, K.T. Teoh, J. Lo, C.M. Chan, F. Kien, N. Escriou, S.W. Tsao, J.M. Nicholls, R. Altmeyer, J.S.M. Peiris, R. Bruzzone, and B. Nal, *J. Virol.* 82, 11318. (2008).
15. L. Wilson, C. Mckinlay, P. Gage, and G. Ewart, *Virol.* 330, 322. (2004).
16. J. Torres, J. Wang, K. Parthasarathy, and D. Xiang Liu, *Biophys. J.* 88, 1283. (2005).
17. R.B. Pandey. arXiv:2009.00049 (2020).
18. K. Binder, *Monte Carlo and Molecular Dynamics Simulations in Polymer Science* (Oxford University Press, Oxford, 1995).

19. M.R. Betancourt, and D. Thirumalai, *Prot. Sci.* 2, 361. (1999).
20. S. Miyazawa, and R.L. Jernigan, *Macromolecules* 18, 534. (1985).
21. S. Miyazawa, and R.L. Jernigan, *J. Mol. Biol.* 256, 623. (1996).
22. S. Tanaka, and H.A. Scheraga, *Macromolecules* 9, 945. (1976).
23. A. Godzik, A. Kolinski, and J. Skolnick, *Proteins* 4, 363. (1996).
24. S.-Y. Huang, and Z. Xiaoqin, *Proteins* 79, 2648. (2011).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.