Structure and thermodynamics of colloidal protein cluster formation: Comparison of square-well and simple dipolar models

Teresa M. Young and Christopher J. Roberts^{a)}

Department of Chemical Engineering, University of Delaware, 150 Academy St., Newark, Delaware 19716, USA

(Received 3 July 2009; accepted 7 September 2009; published online 30 September 2009)

Reversible formation of weakly associated protein oligomers or clusters is a key early step in processes such as protein aggregation and colloidal phase separation. A previously developed cell-based, quasichemical model for lattice fluids [T. M. Young and C. J. Roberts, *J. Chem. Phys.* **127**, 165101 (2007)] is extended here to treat continuous-space systems. It is illustrated using two simplified limiting cases for globular proteins at the isoelectric point: spherical square-well (SW) particles with an isotropic short-ranged attraction and screened dipolar particles with SW attractions and square-shoulder repulsions. Cluster free energies (ΔA_i) and structures are analyzed as a function of the reduced second virial coefficient b_2^* . ΔA_i values and the average structures of clusters up to pentamers have distinct differences due to the anisotropic nature of the dipolar interactions. However, ΔA_i values can be mapped semiquantitatively between the two cases if compared at common values of b_2^* . Free energy landscapes of oligomerization are constructed, illustrating significant differences in landscape ruggedness for small clusters of dipolar versus SW fluids, and suggesting a possible molecular interpretation for empirical models of nucleation-dependent aggregation of proteins. © 2009 American Institute of Physics. [doi:10.1063/1.3238569]

I. INTRODUCTION

Reversible self-association of proteins in dilute solution plays an important role in native and non-native protein polymerization,^{1,2} fibril formation, and protein phase separation.^{3,4} The thermodynamics of this process determines the concentration of small oligomers or clusters that serve as prenuclei or nuclei in the overall kinetics of polymerization and phase separation.^{5–8} It is difficult to experimentally study such prenuclei because they are thermodynamically disfavored under the relevant experimental solution conditions, and thus exist at prohibitively low concentrations for direct experimental characterization by conventional means.^{7,9–12} It may be possible to determine nucleus stoichiometry based on experimental kinetics and regression versus mass-action models,^{5,7,8,10–12} but such approaches yield only limited insights into prenuclei structures or free energies.

Despite impressive advances in algorithms and computational architecture, atomistic molecular simulations remain limited in their ability to sample configuration space sufficiently well to provide accurate free energies for association of multiple large molecules such as globular proteins.^{13–15} Coarse-grained models, while losing atomistic detail of oligomer formation, allow for investigation of larger species¹⁶ and their self-association.¹⁷ The present report employs coarse-grained models of colloidal protein-protein interactions to solve for the free energy and structure of clusters using a cell-based, quasichemical (QC) theory.^{18–25} Previous work showed that the QC theory was quantitatively accurate

for predicting cluster free energies in simple lattice systems up to typical experimental volume fractions ($\sim 10^{-2}$) for protein solutions.²⁶

The osmotic second virial coefficient (B_{22}) has been used as a tool to quantify the average magnitude and sign of interactions between protein molecules in dilute solution,^{27–30} although it is not, per se, a direct predictor of standard free energies of association.³¹ Previous work has shown that to quantitatively predict B_{22} , a large amount of atomistic detail and careful treatment of the anisotropy of protein-protein interactions is necessary.²⁷ However, experimental B_{22} values intrinsically incorporate the effects of anisotropic or orientation-dependent protein-protein interactions. As such, it was proposed²⁶ that cluster free energies could be mapped between these limiting cases of isotropic and anisotropic interactions if systems are considered at similar values of reduced second virial coefficient $b_2^* = (B_{22}/B_{22,HS}) - 1$, with HS denoting the equivalent hard-sphere value for a given protein. If the earlier results with lattice models²⁶ can be generalized, it suggests that coarse-grained protein interaction models can be effectively combined with experimental virial coefficient measurements to provide accurate cluster free energies.

The present report focuses on two simple models of short-ranged colloidal interactions: the square-well (SW) model with short-ranged attractions and a dipole-like (DP) model in which each monomer is represented as two partially overlapping spheres or domains, with equal positive and negative charges polarized to opposite spheres. These represent limiting simple models of globular proteins at their isoelectric point, i.e., when they carry no net charge but the existing charges may or may not be uniformly distributed

^{a)}Author to whom correspondence should be addressed. Electronic mail: cjr@udel.edu. Telephone: 302-831-0838. Fax: 302-831-1048.

across the protein surface. The SW model approximates that the existing charges are perfectly uniformly distributed across the surface; the only interactions between proteins are therefore steric repulsions and weak, relatively nonspecific attractions due to van der Waals and solvophobic forces. The DP model approximates that the charges are heavily polarized to either end of the protein, making the net protein-protein interactions highly anisotropic. These relatively simple models are employed to allow for a broad range of conditions to be explored, and as a first step toward future work employing more detailed protein structural information in the model potentials.

Finally, the previous work²⁶ employed only lattice models. As a result, clusters necessarily had relatively simple morphology or structure that oversimplified the conformational degrees of freedom and packing constraints for particles in the clusters. It was therefore unclear the degree to which neglecting anisotropic protein-protein interactions would also result in realistic structures for equilibrium clusters, or how the free energy landscape for a given cluster depended on cluster stoichiometry and anisotropy of particle-particle interactions. This report addresses these questions by qualitatively and quantitatively comparing the equilibrium structure or morphology of clusters (up to pentamers) based on the SW and DP models. Together, the results provide insights into simplifications and limitations of coarse graining the surface heterogeneity of proteins when considering different aspects of weak protein self-association.

II. MODEL DESCRIPTION AND DERIVATIONS

Two different potentials of mean force ($W_{i,j}$) between particles i and j were tested. A SW model was used to approximate particles interacting with uniform, isotropic short-ranged attractions,

$$W_{i,j}(r_{i,j}) = \begin{cases} \infty, & \text{if } r_{i,j} \leq \sigma \\ -\varepsilon, & \text{if } \sigma < r_{i,j} < \lambda\sigma \\ 0, & \text{if } r_{i,j} \geq \lambda\sigma \end{cases}, \quad (1)$$

$r_{i,j}$ is the center-to-center distance between molecules i and j , ε is the well depth, σ is the effective hard-sphere diameter of the molecule, and $(\lambda-1)\sigma$ is the width of the attractive well. For the calculations using this model and the one below, $\lambda=1.2$. All distances were scaled by σ , and $T^*=k_B T/\varepsilon$ (k_B is Boltzmann's constant).

A DP model was used to approximate particles with highly anisotropic distributions of surface charge, in the limit that one pole contains all positive charge, and the other contains all negative charge. For simplicity, each molecule was treated as two partially overlapping hard spheres (A and B) of equal diameter σ . The A - B bond distance was 0.5σ , giving an end-to-end distance or effective HS diameter (σ_{eff}) of 1.5σ . To approximate short-ranged, screened charge-charge intermolecular attractions and repulsions, each A and B sphere was given a SW attractive well or a square-shoulder (SS) repulsion. The potential of mean force for this model is then

$$W_{i,j} = \sum_{\eta, \eta' = A, B} W_{i,j}^{\eta, \eta'}(r_{i,j}^{\eta, \eta'}), \quad (2)$$

with $W_{i,j}^{\eta, \eta'}$ denoting the interaction between sphere η of molecule i and sphere η' of molecule j , and $r_{i,j}^{\eta, \eta'}$ denoting the center-to-center distance between spheres i and j . The functional form of each $W_{i,j}^{\eta, \eta'}$ is the same as Eq. (1) when $\eta=\eta'$, and with ε replaced by $-\varepsilon$ when $\eta \neq \eta'$. The possible values of interaction energies between molecules i and j range from -2ε to $+2\varepsilon$ and depend on both spatial position and molecular orientation. The value of λ was set as 1.2, in order to approximate a typical charge screening length for a protein with $\sigma \approx 4$ nm in dilute 1:1 electrolyte ($\sim 10^{-2}M$) at ambient temperature.

The derivation of the QC cell theory for a continuous-space system is similar to that previously described for lattice systems.²⁶ The canonical partition function (Q) for N molecules in a three-dimensional system of volume V , and at temperature T , can be written formally as

$$Q(N, V, T) = \frac{1}{N! \Lambda^{3N}} \sum_{\mathbf{n}}^* \frac{N!}{n(1)! n(2)! \cdots n(N_c)!} \times \int_v d\tau_{n(1)} \int_v d\tau_{n(2)} \cdots \int_v d\tau_{n(N_c)} \times \prod_{i=1}^{N_c} \exp\left(-\beta E_{n(k)} - \frac{\beta}{2} \sum_{j \neq i} E_{\sigma(i,j)}\right). \quad (3)$$

In Eq. (3), Λ is the de Broglie wavelength, $\beta=1/k_B T$, and V has been partitioned into N_c equally sized abutting cubic subcells of volume v . $n(k)$ is the occupancy number or number of molecules in the k th subcell and $\tau_{n(k)}$ denotes a particular configuration of the $n(k)$ molecules in the k th subcell. $E_{n(k)}$ denotes the sum over all interactions between only the $n(k)$ molecules, while $E_{\sigma(j,k)}$ denotes the sum over all interactions between pairs of molecules such that one of the molecules lies in subcell j , and the other lies in subcell k . Interactions are defined in the present report in terms of the potentials of mean force in Eq. (1) or Eq. (2). Any particular distribution of subcell occupancy numbers is denoted by the vector $\mathbf{n}=\{n(1), n(2), \dots, n(N_c)\}$. The asterisk on the summation over all possible $\mathbf{n}$ in Eq. (4) indicates each $\mathbf{n}$ is constrained by

$$\sum_{k=1}^{N_c} n(k) = N. \quad (4)$$

For a dilute system, so long as the range of the interactions is not large compared to $v^{1/3}$, the boundary energy between subcells ($E_{\sigma(j,k)}$) is neglected, making the subcell configurations independent of each other. Equation (3) then simplifies to

$$Q(N, V, T) = \frac{1}{(\Lambda^3/v)^N} \sum_n^* \left(\int_{v(1)} d\tau_{n(1)} \frac{e^{-\beta E_{n(1)}}}{n(1)! v^{n(1)}} \cdots \right. \\ \left. \times \int_{v(N_c)} d\tau_{n(N_c)} \frac{e^{-\beta E_{n(N_c)}}}{n(N_c)! v^{n(N_c)}} \right). \quad (5)$$

For a classical system, the number of distinguishable ways of placing i molecules in a subcell with volume v and a value of energy within $[E_n, E_n + dE_n]$ is denoted as $\omega_v(n, E_n) dE_n$ and is defined as

$$\omega_v(n, E_n) dE_n = \frac{1}{n! v^n} \int_v d\tau_n [H(E_n) - H(E_n + dE_n)], \quad (6)$$

where $H(E_n)$ is the heavy-side function, $\int_v d\tau_n [H(E_n) - H(E_n + dE_n)]$ denotes the set of the configurations of n molecules with an energy in the range $E_n + dE_n$, and ω_v is normalized relative to the ideal gas result (v^n).

Because each subcell is independent, the integrals over subcells containing the same number of molecules are identical, and subcells can be grouped according to occupancy number n , and the spatial distribution of n will be random. In addition, the integral over all configurations $\tau_{n(i)}$ in a given subcell is formally replaced with an integral over the energy E_n at fixed n . Taking the maximum term for Q and a random distribution for $\mathbf{n}$ gives

$$Q(N, V, T) = \frac{1}{(\Lambda^3/v)^N} \left[\frac{N_c!}{\prod_{i=0}^{n_{\max}} (x_i N_c)!} \times \prod_{i=0}^{n_{\max}} \left(\int \omega_v(i, E_n) dE_n \exp(-\beta E_n) \right)^{x_i N_c} \right]. \quad (7)$$

In Eq. (7), $n_{\max}$ denotes the maximum number of molecules which can be sterically placed into v , and x_i is the fraction of the N_c subcells that contain i molecules. The remainder of the derivation is analogous to that in Ref. 26, with only key results reproduced here for continuity and completeness of presentation and discussions below.

The Helmholtz free energy density of the system is defined as

$$a = \frac{A}{V} = - \frac{k_B T \ln Q(N, V, T)}{v N_c}. \quad (8)$$

Applying Stirling's approximation to Eq. (7) and substituting into Eq. (8) gives

$$a = k_B T \rho \ln(\Lambda^3/v) + \frac{k_B T}{v} \sum_{i=0}^{n_{\max}} x_i (\ln x_i - \alpha_i), \quad (9)$$

with

$$\alpha_i \equiv \ln \left[\int \omega_v(i, E_n) \exp(-\beta E_n) dE_n \right]. \quad (10)$$

The equilibrium value of the Helmholtz free energy density (a) is obtained by solving for the distribution of x_i values that

minimizes the free energy at fixed T^* and bulk density (ρ). At fixed ρ , there are two constraints on the x_i distribution: $\sum_{i=0}^{n_{\max}} (i/v) x_i = \rho$ and $\sum_{i=0}^{n_{\max}} x_i = 1$. This results in two independent x_i values ($i = \delta, \gamma$). The equilibrium distribution of cell density probabilities $\{x_i\}$ is found by solving $(\partial a / \partial x_i)_{T, \rho, x_j \neq \delta, \gamma} = 0$ for each $i \neq \delta, \gamma$, subject to the constraints. The final expression for $x_{i \neq \delta, \gamma}$ is

$$\ln \left(\frac{x_i}{x_\gamma} \right) = \frac{\rho_\gamma - \rho_i}{\rho_\gamma - \rho_\delta} \ln r + \alpha_i - \frac{\rho_i - \rho_\delta}{\rho_\gamma - \rho_\delta} \alpha_\gamma - \frac{\rho_\gamma - \rho_i}{\rho_\gamma - \rho_\delta} \alpha_\delta, \quad (11)$$

with $r = x_\delta / x_\gamma$ and $\rho_i = i/v$. At a desired ρ and T , r is obtained by solving

$$\rho(r, T) = \frac{[\rho_\gamma + r \rho_\delta + \sum_{i \neq \delta, \gamma} (x_i / x_\gamma) \rho_i]}{1 + r + \sum_{i \neq \delta, \gamma} x_i / x_\gamma}, \quad (12)$$

with x_i / x_γ from Eq. (11). Normalization of x_i gives

$$x_\gamma = \frac{1}{1 + r + \sum_{i \neq \delta, \gamma} x_i / x_\gamma}, \quad (13)$$

and the value of x_δ is obtained from the definition of r .

K_i is defined as the equilibrium constant for the conversion of i monomers (M_1) to a cluster or oligomer containing i molecules (M_i), with a monomer and an oligomer defined below. On a number-concentration basis (K_i) is given as

$$K_i = \exp \left(\frac{-\Delta A_i^0}{k_B T} \right) = \frac{[M_i] / c_0}{[M_1]^i / c_0^i} = \frac{N_v^{(i)} / v}{(N_v^{(1)} / v)^i} c_0^{i-1}, \quad (14)$$

where ΔA_i^0 denotes the free energy of formation of M_i from i monomers at a standard state number concentration c_0 . $N_v^{(i)}$ denotes the equilibrium average number of oligomers of size i per subcell of volume v for a given T , ρ , and set of subcell degeneracies. $N_v^{(i)}$ is given mathematically in the present model as

$$N_v^{(i)} = \sum_{n=0}^{n_{\max}} \sum_j \int \lambda_j^{(i)} p_v(n, E_n, \lambda_j) dE_n. \quad (15)$$

In Eq. (15), $\lambda_j^{(i)}$ denotes the i th entry in the vector λ_j . The cluster vector λ_j is defined as follows. Every subcell configuration for a given n and E_n can be characterized by a vector $\lambda_j = [\lambda_j^{(1)}, \lambda_j^{(2)}, \dots, \lambda_j^{(n_{\max})}]$ with $\lambda_j^{(k)}$ denoting the number of M_k oligomers in that configuration (including monomers, M_1); j is an index that distinguishes the finite number of possible realizations of λ . The probability of observing λ_j in a subvolume v for a given n and E_n is given by

$$p_v(n, E_n, \lambda_j) dE_n = \frac{\omega'_v(n, E_n, \lambda_j)}{\omega_v(n, E_n)} p_v(n, E_n) dE_n, \quad (16)$$

where $\omega'_v(n, E_n, \lambda_j) dE_n$ is the number of distinguishable ways of observing λ_j in a subvolume v for a given n and E_n , and $p_v(n, E_n) dE_n$ is the overall probability of observing a subvolume v with a given n and E_n ,

$$p_v(n, E_n) dE_n = \frac{x_n \omega_v(n, E_n) \exp(-\beta E_n) dE_n}{\exp(\alpha_n)}. \quad (17)$$

The procedure for identifying oligomers and obtaining $\omega'_v(n, E_n, \lambda_j) dE_n$ is described below.

III. SECOND VIRIAL COEFFICIENT

The second virial coefficient (B_2) can be expressed using the above nomenclature as^{26,32,33}

$$B_2(T) = -\frac{vN_c}{Q_1^2} \left(Q_2 - \frac{1}{2}Q_1^2 \right), \quad (18)$$

with Q_1 and Q_2 representing $Q(N=1, V, T)$ and $Q(N=2, V, T)$, respectively. For the SW potential, Eq. (18) can be solved analytically to give³³

$$B_2(T) = \frac{2\pi\sigma^3}{3} [1 - (\lambda^3 - 1)(\exp(-\beta\epsilon) - 1)]. \quad (19)$$

For the DP potential, Eq. (18) is solved using the method introduced in Ref. 26. The canonical partition function is expressed as

$$Q(N, V, T) = \left(\frac{v}{\Lambda^3} \right)^N \int \Omega(N, V, E) \exp(-\beta E) dE, \quad (20)$$

with $\Omega(N, V, E)$ denoting the degeneracy for the entire system. The $N=1$ system has 1 subcell with $n=1$ and $E_n=0$, and N_c-1 subcells with $n=0$ and $E_n=0$. Thus,

$$\Omega(1, V, 0) = N_c [\omega_v(1, 0)]. \quad (21)$$

For the $N=2$ and $E=0$ case, there is either one subcell with $n=2$ and $E_n=0$, two abutting subcells with $n=1$ and $E_\sigma=0$, or two nonabutting subcells with $n=1$. This gives

$$\begin{aligned} \Omega(2, V, 0) = & N_c \omega_v(2, 0) + \frac{N_c!}{2! (N_c - 2)!} [\omega_v(1, 0)]^2 \\ & - \frac{N_c}{2} [\omega_v(1, 0)]^2 \int_{E_\sigma \neq 0} P(E_\sigma) dE_\sigma, \end{aligned} \quad (22)$$

with $P(E_\sigma) dE_\sigma$ introduced as the probability of placing two abutting subcells, each with $n=1$, such that the value of the boundary energy is E_σ . It is defined as

$$P(E_\sigma) = 6p_s(E_\sigma) + 12p_e(E_\sigma) + 8p_c(E_\sigma). \quad (23)$$

The subscripts s , e , and c distinguish the probabilities of two abutting subcells which share a side, edge, or corner, respectively, and have boundary energy between E_σ and $E_\sigma + dE_\sigma$. The values of $p_s(E_\sigma)$, $p_e(E_\sigma)$, and $p_c(E_\sigma)$ were determined via normalized histograms for the frequency of observing a given value of E_σ during a random walk simulation of two abutting subcells, with each subcell containing one molecule. Separate simulations were used for boxes sharing a side, edge, or corner.

For the $N=2$ and $E_n=E'$ case, there is either one subcell with $n=2$ and $E_n=E'$, or two abutting subcells with $n=1$ and $E_\sigma=E'$. This gives

$$\Omega(2, V, E') = N_c \omega_v(2, E') + \frac{N_c}{2} [\omega_v(1, 0)]^2 P(E') dE'. \quad (24)$$

Applying Eq. (20) to Eqs. (21), (22), and (24), and substituting into Eq. (18), gives

$$\begin{aligned} B_2(T) = & v \left[\frac{1}{2} - \frac{\omega_v(2, 0)}{[\omega_v(1, 0)]^2} - \int_{E' \neq 0} \frac{\omega_v(2, E')}{[\omega_v(1, 0)]^2} e^{-\beta E'} dE' \right. \\ & \left. - \frac{1}{2} \int_{E' \neq 0} P(E') (e^{-\beta E'} - 1) dE' \right]. \end{aligned} \quad (25)$$

This approach was also tested for the SW model (results not shown) and found to give equivalent numerical results to those from Eq. (19). B_2 is reported with respect to the hard sphere value ($B_2^{\text{HS}} = 2\pi\sigma^3/3$), as $b_2^* \equiv (B_2 - B_2^{\text{HS}})/B_2^{\text{HS}}$.

IV. DEGENERACY CALCULATIONS

Degeneracies $\omega_v(n, E_n) dE_n$ were determined numerically as a function of n and E_n using $v = (5\sigma)^3$ for the SW model, and $v = (5\sigma_{\text{eff}})^3$ for the DP model. The fractional or relative degeneracy $f_v(n, E_n) dE_n$ was determined with the Wang–Landau (WL) algorithm³⁴ as a function of E_n for a given value of n . Moves which caused an overlap of molecules were immediately rejected and $E_n = \infty$ was not a factor in the WL convergence. To account for this when converting from f_v to ω_v , a random walk simulation for each value of n was performed to estimate the probability of randomly selecting a configuration with an infinite energy ($p_n(\infty)$) due to one or more particle overlaps. The value of $\omega_v(n, E_n) dE_n$ is then given by

$$\omega_v(n, E_n) dE_n = \frac{[1 - p_n(\infty)] f_v(n, E_n) dE_n}{n!}. \quad (26)$$

The factor of $n!$ is needed in Eq. (26) to correct for the fact that the WL algorithm treats particles as being distinguishable when histograms are recorded at a given n value. Details of the numerical implementation of the WL algorithm were the same as those described previously.²⁶ As done previously, the degeneracies were obtained via WL for $n \leq 12$. This range of n provides a large enough subcell density for the dilute solution conditions here, and is much larger than the minimum n needed to accurately predict cluster probabilities for $i \leq 5$.²⁶

V. OLIGOMER DISTRIBUTIONS

An oligomer was defined as a contiguous set of molecules such that $r_{i,j} \leq \lambda\sigma$ for molecule i and at least one other molecule j in the oligomer. The size of the oligomer is given by the number of monomers that it contains. An algorithm to determine λ_j for any given configuration of n particles is described in detail in Appendix C of Ref. 26. Briefly, the algorithm makes use of a nearest neighbor or cluster array $\mathbf{C}$, which is constructed from every pair of molecules i and j , such that

$$C_{ij} = \begin{cases} 1, & \text{if } i = j \\ 1, & \text{if } r_{i,j} \leq \lambda\sigma \\ 0, & \text{if } r_{i,j} > \lambda\sigma \end{cases}. \quad (27)$$

The algorithm recursively steps through $\mathbf{C}$ to determine the oligomer to which each particle belongs, including isolated particles or “monomers,” and the total number of molecules in each oligomer. The result is used to update histograms for

the λ_j vectors, as a function of n and E_n . The histograms are later normalized to provide $\omega'_v(n, E_n, \lambda_j) / \omega_v(n, E_n)$ after sampling from a random walk in E_n for each n , using the previously determined $\omega_v(n, E_n)$ to bias the configurations and provide a random walk.

VI. ORDER PARAMETERS

For the SW model, the order parameter characterizing the structure of a given cluster was taken as the radius of gyration (R_g). For the DP model, R_g and an orientational order parameter (χ , defined below) were calculated for each cluster; alternative order parameters were considered but did not provide additional insights beyond R_g and χ .

The orientational order parameter χ was defined based on the number and types of contacts within a cluster M_i . The two types of contacts for DP particles are repulsive (A-A or B-B) and attractive (A-B). For a given arrangement of molecules in a cluster, the number of repulsive contacts is n_{rep} and the number of attractive contacts is n_{att} . The value of χ for that cluster arrangement is defined as

$$\chi = \frac{n_{\text{rep}} - n_{\text{att}}}{n_{\text{rep}} + n_{\text{att}}}. \quad (28)$$

Therefore, the value of χ lies between 1 (all repulsive contacts) and -1 (all attractive contacts), with $\chi=0$ in the limit of large randomly arranged clusters.

During the random walk simulations described above to determine the distribution of λ_j values, histograms of each order parameter ($y=R_g$ or χ) as a function of E_n were obtained for each cluster stoichiometry (i). A given histogram was normalized at each value of i , n , and E_n to determine $P_{i,n,E_n}(y)$. The distribution of y at each i for a given choice of T^* and ρ was found by

$$P_{i,T^*,\rho}(y) = \frac{\sum_n \sum_{E_n} P_{i,n,E_n}(y) x_n \omega_v(n, E_n) \exp(-\beta E_n)}{\sum_n \sum_{E_n} x_n \omega_v(n, E_n) \exp(-\beta E_n)}, \quad (29)$$

with T^* and ρ setting the equilibrium x_n values.

The average value of y was found at a given T^* and ρ for each cluster size i , $\langle y \rangle = \sum_y y P_{i,\rho,T^*}(y)$. A free energy landscape was determined for y at constant values of i , T^* , and ρ , such that $\Delta A_i(y) / k_B T = -\ln P_{i,\rho,T^*}(y) - \Delta A_{i,\text{ref}} / k_B T$. The value for $\Delta A_{i,\text{ref}}$ was found by equating $\Delta A_i(\langle y \rangle) / k_B T$ with $\Delta A_i / k_B T$ from Eq. (14). In the case of χ , which is a discrete quantity, $\langle \chi \rangle$ did not always correspond to a possible χ value. To keep the distributions consistent across cluster sizes, the free energy landscapes in those cases are therefore reported as $\Delta \Delta A_i^*$, $\Delta \Delta A_i^* \equiv -T^* \ln P_{i,T^*,\rho}(\chi)$.

VII. CLUSTER PROBABILITIES FROM CANONICAL DIRECT MONTE CARLO SIMULATION

Canonical ensemble Monte Carlo (MC) simulations with a Metropolis algorithm^{35,36} were performed for the isotropic (SW) model for selected state points to compare the cluster probabilities with those determined from the QC theory and subcell degeneracies. A volume of $(20\sigma)^3$ was used for a range of $\rho(0.001-0.03)$ and $T^*(0.3-1.0)$, respectively. Standard periodic boundary conditions and random single-

FIG. 1. Comparison between QC and MC for the probability of observing an oligomer composed of particles. Each symbol represents a different T^* and volumetric $\rho(=N\pi\sigma^3/V)$: fixed $\rho=10^{-3}$ with T^* of 0.40 (●), 0.50 (■), 0.60 (◆), and 0.65 (▲); fixed $T^*=1.0$ with $\rho=0.01$ (○), 0.02 (□), and 0.05 (◇). The line is a guide to the eyes for an ideal 1:1 correlation.

molecule translations were employed.^{35,36} The system was equilibrated and then simulated for between 5×10^5 and 10^6 MC cycles for ensemble averaging (1 cycle= N trial moves). During ensemble averaging, ghost subcells of size $(5\sigma)^3$ were randomly placed and a histogram was recorded of the number of each size of cluster within those subcells. The normalized histogram provided the probability of observing an oligomer of i molecules [$p(i)$] for each set of T^* and ρ . Subcells which contained oligomers that crossed over the subcell boundary were discarded, so as to not undercount the size of the oligomers within the ghost subcells.²⁶

VIII. RESULTS AND DISCUSSION

Figure 1 shows a comparison between the probabilities of observing a given oligomer size from the QC theory and from canonical MC simulations at a range of T^* and ρ for the SW model. For reasons discussed in Ref. 26, only oligomers up to the size of the shortest linear dimension of the subcell were included. The subcell volume is $(5\sigma)^3$ for all studies reported here, and therefore the results in Fig. 1 include clusters from $i=2$ to $i=5$. The QC predictions quantitatively reproduce the results from direct MC simulation, similar to what was found using lattice systems.²⁶

Figure 2(a) shows K_i values ($i=2-5$; ref. state= $1 \mu\text{M}$; $\sigma=4 \text{ nm}$) from QC theory as a function of the corresponding value of b_2^* for a given T^* value. $\log K_i$ values are large and negative and become more so with increasing cluster size. The K_i values also asymptotically approach plateau values as b_2^* becomes more negative. This highlights that cluster formation is thermodynamically disfavored even under conditions with strong favorable attractions, as expected for typical volume fractions ($\phi \sim 0.01$ or less) of protein solutions or dilute colloidal suspensions (see also discussion below regarding Fig. 3). The results for both the SW and DP interaction models are qualitatively similar, with a roughly constant difference in $\Delta A_i / k_B T$ ($\sim \log K_i$) between the two cases. Figure 2(b) further illustrates this result by comparing the values

FIG. 2. (a) Oligomer equilibrium constants based on a standard state concentration of $1 \mu\text{M}$, with a HS or effective HS diameter of 4 nm , as a function of the reduced second virial coefficient. Black and gray lines represent the SW and the DP models, respectively. The label on each curve denotes the oligomer stoichiometry i . (b) Free energy of formation comparison between the DP and SW models. Each symbol represents a different cluster size: dimer ($\bullet$), trimer ($\blacksquare$), tetramer ($\blacklozenge$), and pentamer ($\blacktriangle$). There are 12 points for each cluster size, representing a range of b_2^* values from -20 to -0.01 . The solid line is fit to the data with a slope of 0.895 ± 0.014 and a y -intercept of -0.11 ± 0.33 (95% confidence intervals). The dashed line is provided to illustrate an ideal 1:1 correlation.

of ΔA_i for the SW and DP models at a given b_2^* value. The solid line is fit to cluster sizes 2–5 across a range of b_2^* values.

Figure 2(b) suggests that differences in cluster free energies for the isotropic (SW) and anisotropic (DP) cases can be reasonably mapped between each other. This is perhaps not surprising for the case of dimers ($i=2$), because b_2^* is a measure of the orientationally averaged interactions between two monomers. It is more interesting that this holds reasonably well for the larger clusters, as this suggests that results from simple models that neglect the details of orientation-dependent interactions may be sufficient if the corresponding virial coefficient value is known from experiment or from more structurally detailed calculations.

The underlying energetic (ΔU_i) and entropic ($T\Delta S_i$) contributions to the average cluster free energies are shown in Fig. 3, on a per monomer basis for easier comparison between curves. ΔU_i and ΔS_i were calculated from the temperature derivatives of ΔA_i at fixed ρ for a given T^* and b_2^* . The SW model shows little to no change in internal energy as a function of b_2^* [Fig. 3(a)], with the net unfavorable ΔA_i

FIG. 3. (a) Internal energy change per monomer and (b) entropy change per monomer for cluster formation as a function of the reduced second virial coefficient. Black and gray lines represent the SW and DP models, respectively. The arrow on each plot marks the direction of increasing i .

values arising from large unfavorable entropy of cluster formation. The DP model shows similar behavior to the SW model in terms of average cluster free energy (Fig. 2), however there is a pronounced difference when comparing the energetic [Fig. 3(a)] and entropic [Fig. 3(b)] contributions. Dimers through pentamers show a weak dependence of ΔU_i on b_2^* under strongly attractive conditions (low b_2^*) but under weakly attractive conditions ($b_2^* \approx -2.5$ and higher) there is a sharp increase in ΔU_i . The entropy of cluster formation is large and unfavorable for the DP model and is even more so than for the SW model. This reflects the orientational degrees of freedom lost by molecules involved in clusters (see also discussion below).

All of the above results were based on averages across all clusters of a given stoichiometry, but did not distinguish between clusters with different structures or order parameters at fixed stoichiometry. One natural order parameter is the radius of gyration (R_g), as this informs upon differences between extended and globular cluster morphologies. Figures 4(a) and 4(b) show the free-energy landscapes (ΔA_i versus R_g and i) for cluster formation of the SW and DP models at selected values of b_2^* . The curves for different i values are offset by the corresponding differences in the average cluster ΔA_i values in Figs. 4(a) and 4(b) for a given b_2^* value. The largest R_g values for each i correspond to essentially linear

FIG. 4. Free energy landscape as a function of R_g/σ for each cluster stoichiometry for (a) $b_2^* = -10$ and (b) $b_2^* = -2.5$. Panel (c) shows the average $\langle R_g \rangle/\sigma$ vs b_2^* for each cluster stoichiometry. In all panels, black and gray lines indicate SW and DP results, respectively.

clusters ($R_{g,\max}/\sigma = 0.60, 0.98, 1.34$, and 1.70 for $i=2, 3, 4$, and 5 , respectively), while the lowest R_g values correspond to tightly packed, globular structures for the SW model. The obvious exception is $i=2$, for which clusters are necessarily linear for the SW model. Some notable qualitative features of the SW free energy landscapes are that more attractive conditions [larger negative b_2^* , Fig. 4(a)] lead to relatively shallow local minima or a broad global minimum, while less attractive conditions [higher b_2^* , Fig. 4(b)] flattens the energy landscapes to some extent. Overall, the SW landscapes are typically relatively flat, with no pronounced local barriers or ruggedness, indicating that there are not large entropic bottlenecks for rearranging the internal “packing” of clusters. The higher stoichiometry clusters become less compact [larger average R_g , Fig. 4(c)] with increasing b_2^* , as expected

FIG. 5. Average orientational order parameter (χ), as a function of b_2^* . Cluster stoichiometry is denoted by the number on each curve.

based on the higher entropy and energy of more extended clusters for the SW model, and in keeping with the results in Fig. 3.

A similar analysis of the anisotropic oligomers reveals somewhat contrasting results. As b_2^* is increased, the average R_g [Fig. 4(c)] shows little or no change. The free energy landscapes of R_g produce similar, if not exact results at low [Fig. 4(a)] and high [Fig. 4(b)] b_2^* values. This implies that the free energy differences as a function of b_2^* (Fig. 2) for the DP model arise from changes in the internal arrangement of particles in a cluster, rather than simply how extended or globular the cluster morphology is under strongly versus weakly attractive conditions.

An additional conclusion from Fig. 4 is that R_g is not a sufficient order parameter for the DP clusters. An additional order parameter [χ , defined in Eq. (29)] is based on the balance of nearest-neighbor contacts that are repulsive or attractive in a cluster, i.e., the relative orientations of particles within a cluster. $\chi < 0$ suggests attractive contacts are the prominent type in the cluster, while $\chi > 0$ indicates more repulsive contacts. In terms of spatial arrangements, clusters with large negative (positive) values of χ , yet compact R_g values, require antiparallel (parallel) side-to-side stacking of the dipoles. Figure 5 shows $\langle \chi \rangle$ as a function of increasing b_2^* . As b_2^* is increased, the types of contacts within the oligomers change from predominantly attractive to a random mix of attractive and repulsive. Figure 6 shows free energy landscapes ($\Delta \Delta A_i^* = \Delta A_i/\epsilon$) as a function χ at b_2^* values of -10 (left column) and -2.5 (right column) for $i=2$ (first row), 3 (second row), 4 (third row), and 5 (fourth row). Figure 6 is drawn such that the bottom of the gray bars indicates the free energy value for a given χ value. χ values with only white bars are ones with essentially infinitely high free energy.

Figure 6 shows an evolution in terms of differences in free energy landscapes between dimers and larger clusters, and in terms of the effects of weakly versus strongly attractive conditions. For dimers, the most probable state is for $\chi=0$ at both low and high b_2^* values, with large free energy barriers separating the higher local minima at lower and higher values of χ . The differences between low and high b_2^* are relatively minor for the dimer. The population of $\chi=-1$ becomes significantly more favorable for strongly attractive

FIG. 6. Free energy landscapes vs χ at $b_2^* = -10$ (left column) and $b_2^* = -2.5$ (right column). Each row represents a different cluster size: dimer (first row), trimer (second row), tetramer (third row), and pentamer (fourth row).

b_2^* values, but the dominant population still remains near $\chi=0$. The landscapes for trimer, tetramer, and pentamer are qualitatively different from those for the dimer. For these larger oligomers, there is a broad rugged basin centered on slightly negative χ values, with a number of relatively shallow local minima within the basin. There is a high barrier separating this broad basin from a narrow well of similar depth at $\chi=-1$. Another high barrier separates a narrow, much higher local minimum near $\chi=1$.

Comparing these landscapes as cluster size increases and/or b_2^* decreases, the minimum at slightly negative χ values becomes broader and less rugged. This suggests that transitions between different arrangements of DP molecules in a given cluster become more facile as one increases the cluster size, and to a lesser extent as the interparticle attractions become larger compared to $k_B T$. Although the DP model uses only simple SW and SS interactions, this broadening of and smoothing of the landscape suggests a possible molecular interpretation for the empirical observation that aggregation of some proteins appears to require a conformational rearrangement step within a preformed, reversible “prenucleus,” and that these prenuclei often have stoichiometries larger than a dimer, but still within the range $i \sim 10$ or less.³⁷ That is, if internal rearrangements of particles within a cluster are more facile for larger clusters, this may then increase the rates at which additional conformational “misfolding” and formation of strong interprotein contacts occur.

Comparing the results for the energy landscapes of the SW and DP models, it also is apparent that while the average free energy values for different cluster stoichiometries map reasonably well between the two models if one considers them at comparable b_2^* values, the detailed structural attributes and free-energy landscapes for the clusters are markedly different between the models, and this is a direct result of the lack of orientation dependence for the SW model. Translating this to colloidal and protein cluster formation, these results indicate that it may be possible to obtain reasonable estimates of overall free energy changes for cluster formation without structurally detailed models for protein-protein interactions. However, it is perhaps not surprising that b_2^* alone does not account well for differences in structural aspects and details of the free energy landscapes.

IX. SUMMARY AND CONCLUSIONS

A QC method for predicting cluster free energies of simple lattice systems²⁶ was adapted for systems with continuous-space coordinates, and tested using a SW potential and a DP potential to mimic extremes of behavior for proteins at their isoelectric point. The QC method provided quantitative agreement with direct Monte Carlo simulations, while also providing global behavior of the standard free energy of forming clusters up to pentamers, along with free energy landscapes as a function of different order parameters. The results indicate that the overall or average free energy to form a given cluster stoichiometry maps well between different potentials if they are compared at similar values of the reduced second virial coefficient. Clusters composed of particles with strongly anisotropic interactions clearly show rugged free-energy landscapes that become less rugged with increasing cluster size, in keeping with empirical models and interpretations for the experimental nucleation processes relevant to protein aggregation and to colloidal and protein phase behavior.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the University of Delaware Research Foundation and the National Institutes of Health (Grant No. 1 R01 EB006006) for support.

- ¹D. Thirumalai, D. Klimov, and R. Dima, *Curr. Opin. Struct. Biol.* **13**, 146 (2003).
- ²R. Wetzel, *Acc. Chem. Res.* **39**, 671 (2006).
- ³S. Tanaka, K. Ito, R. Hayakawa, and M. Ataka, *J. Chem. Phys.* **111**, 10330 (1999).
- ⁴P. G. Vekilov, *Cryst. Growth Des.* **4**, 671 (2004).
- ⁵F. Oosawa and S. Asakura, *Thermodynamics of the Polymerization of Protein* (Academic, New York, 1975).
- ⁶P. G. Debenedetti, *Metastable Liquids: Concepts and Principles* (Princeton University Press, Princeton, 1996).
- ⁷F. Ferrone, *Methods Enzymol.* **309**, 256 (1999).
- ⁸J. M. Andrews and C. J. Roberts, *J. Phys. Chem. B* **111**, 7897 (2007).
- ⁹A. L. Fink, *Folding Des.* **3**, R9 (1998).
- ¹⁰M. M. Pallitto and R. M. Murphy, *Biophys. J.* **81**, 1805 (2001).
- ¹¹C. J. Roberts, R. T. Darrington, and M. B. Whitley, *J. Pharm. Sci.* **92**, 1095 (2003).
- ¹²J. M. Andrews and C. J. Roberts, *Biochemistry* **46**, 7558 (2007).

- ¹³H.-H. G. Tsai, M. Reches, C.-J. Tsai, K. Gunasekaran, E. Gazit, and R. Nussinov, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 8174 (2005).
- ¹⁴S. Jang and S. Shin, *J. Phys. Chem. B* **110**, 1955 (2006).
- ¹⁵A. Melquiond, N. Mousseau, and P. Derreumaux, *Proteins: Struct., Funct., Bioinf.* **65**, 180 (2006).
- ¹⁶B. Urbanc, J. Borreguero, L. Cruz, and H. E. Stanley, *Methods Enzymol.* **412**, 314 (2006).
- ¹⁷J. K. Cheung and T. M. Truskett, *Biophys. J.* **89**, 2372 (2005).
- ¹⁸A. Lomakin, D. B. Teplow, D. A. Kirschner, and G. B. Benedek, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 7942 (1997).
- ¹⁹E. A. Guggenheim and M. McGlashan, *Proc. R. Soc. London, Ser. A* **206**, 448 (1951).
- ²⁰R. Goldstein and L. Stryer, *Biophys. J.* **50**, 583 (1986).
- ²¹P. G. Debenedetti, V. Raghavan, and S. S. Borick, *J. Phys. Chem.* **95**, 4540 (1991).
- ²²C. J. Roberts and P. G. Debenedetti, *J. Chem. Phys.* **105**, 658 (1996).
- ²³L. R. Pratt and R. A. LaViolette, *Mol. Phys.* **94**, 909 (1998).
- ²⁴L. R. Pratt, R. A. LaViolette, M. A. Gomez, and M. E. Gentile, *J. Phys. Chem. B* **105**, 11662 (2001).
- ²⁵C. J. Roberts, *J. Phys. Chem. B* **108**, 19843 (2004).
- ²⁶T. M. Young and C. J. Roberts, *J. Chem Phys.* **127**, 165101 (2007).
- ²⁷B. L. Neal, D. Asthagiri, and A. M. Lenhoff, *Biophys. J.* **75**, 2469 (1998).
- ²⁸E. Y. Chi, S. Krishnan, B. S. Kendrick, B. S. Chang, J. F. Carpenter, and T. W. Randolph, *Protein Sci.* **12**, 903 (2003).
- ²⁹J. J. Valente, K. S. Verma, M. C. Manning, W. W. Wilson, and C. S. Henry, *Biophys. J.* **89**, 4211 (2005).
- ³⁰W. Liu, T. Cellmer, D. Keerl, J. M. Prausnitz, and H. W. Blanch, *Biotechnol. Bioeng.* **90**, 482 (2005).
- ³¹W. F. Weiss IV, T. M. Young, and C. J. Roberts, *J. Pharm. Sci.* **98**, 1246 (2009).
- ³²A. Ben-Naim, *Statistical Thermodynamics for Chemists and Biochemists* (Plenum, New York, 1992).
- ³³D. A. McQuarrie, *Statistical Mechanics* (University Science Books, Sausalito, CA, 2000).
- ³⁴F. Wang and D. P. Landau, *Phys. Rev. Lett.* **86**, 2050 (2001).
- ³⁵M. P. Allen and D. Tildesley, *Computer Simulation of Liquids* (Oxford Science Publications, Oxford, United Kingdom, 1987).
- ³⁶D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithm to Applications* (Academic, New York, 1996).
- ³⁷C. J. Roberts, *Biotechnol. Bioeng.* **98**, 927 (2007).