Published in final edited form as:
J Phys Chem B. 2013 December 19; 117(50): 16013–16028. doi:10.1021/jp409300j.

Coarse-Grained Model for Colloidal Protein Interactions, B_{22}, and Protein Cluster Formation

Marco A. Blanco[†], **Eric Sahin**[†], **Anne S. Robinson**[‡,†], and **Christopher J. Roberts**[*,†]

[†]Department of Chemical and Biomolecular Engineering and Center for Molecular and Engineering Thermodynamics, University of Delaware, Newark, Delaware 19176, United States

[‡]Department of Chemical and Biomolecular Engineering, Tulane University, New Orleans, Louisiana 70118, United States

Abstract

Reversible protein cluster formation is an important initial step in the processes of native and non-native protein aggregation, but involves relatively long time and length scales for detailed atomistic simulations and extensive mapping of free energy landscapes. A coarse-grained (CG) model is presented to semi-quantitatively characterize the thermodynamics and key configurations involved in the landscape for protein oligomerization, as well as experimental measures of interactions such as the osmotic second virial coefficient (B_{22}). Based on earlier work, this CG model treats proteins as rigid bodies composed of one bead per amino acid, with each amino acid having specific parameters for its size, hydrophobicity, and charge. The net interactions are a combination of steric repulsions, short-range attractions, and screened long-range charge-charge interactions. Model parametrization was done by fitting simulation results against experimental values of the B_{22} as a function of solution ionic strength for α-chymotrypsinogen A and γD-crystallin (gD-Crys). The CG model is applied to characterize the pairwise interactions and dimerization of gD-Crys and the dependance on temperature, protein concentration, and ionic strength. The results illustrate that at experimentally relevant conditions where stable dimers do not form, the entropic contributions are predominant in the free-energy of protein cluster formation and colloidal protein interactions, arguing against interpretations that treat B_{22} primarily from energetic considerations alone. Additionally, the results suggest that electrostatic interactions help to modulate the population of the different stable configurations for protein nearest-neighbor pairs, while short-range attractions determine the relative orientations of proteins within these configurations. Finally, simulation results are combined with Principal Component Analysis to identify those amino-acids / surface patches that form inter-protein contacts at conditions that favor dimerization of gD-Crys. The resulting regions agree with previously found aggregation-prone sites, as well as suggesting new ones that may be important.

[*]To whom correspondence should be addressed cjr@udel.edu.

Supporting Information Available

Details on the calculations of B_{22} and the PMFs and principal components are considered. Additionally, results regarding the response surface for aCgn and the density maps for P_1 vs. r and P_2 vs. P_3 are provided, as well as illustrative examples of the structures of the protein dimer in the dominant orientations at low temperatures. This material is available free of charge via the Internet at http://pubs.acs.org/.

Keywords

Coarse-grained models; protein interactions; virial coefficient; protein self-association; γD-crystallin

Introduction

Protein–protein interactions in solution are important mediators of protein phase behavior,[1,2] non-native aggregation,[3,4] increases in viscosity at high concentrations,[5,6] and *in vivo* self-assembly in biological systems.[7,8] In dilute solution *in vitro*, most soluble proteins fall into one of two categories: (i) they form stable dimers or oligomers, with equilibrium dissociation constants (K_d) on the order of μM or smaller values;[9,10] (ii) they are natively monomeric, and their protein–protein interactions are instead described in terms of the second osmotic virial coefficient B_{22}.[11,12] For proteins in category (i), K_d is easily related experimentally to the net free-energy of "binding" or self-association, and the enthalpic and entropic contributions that are involved.[13,14] In contrast, for proteins in category (ii) it is not possible to measure the equilibrium between monomer and "dimer" or oligomers at experimentally tractable concentrations. Therefore, it is less straightforward to experimentally determine the relative roles of "interactions" and enthalpic / entropic contributions to B_{22} in those cases.

Despite this limitation, B_{22} has traditionally been used to assess the magnitude and sign of colloidal interactions between protein molecules at dilute conditions from either computational or experimental methods.[15–17] B_{22} and other osmotic virial coefficients play central roles in different models and theories relating colloidal interactions to protein self-assembly processes, including phase separation,[18–20] aggregation,[21–23] and cluster formation.[24–26] Molecular simulations have the potential to provide insights into these processes, but require extensive simulation of two or more proteins to assess both energetic and en-tropic contributions to the monomer–dimer/oligomer free energy "landscape". This is extremely computationally expensive if one uses an all-atom description, and could benefit from a coarse-grained (CG) modeling approach, in which key physics of the interactions between proteins are retained while making larger length- and time-scale simulations more tractable.[27,28]

CG models are somewhat context specific, depending on the length and/or time scales of interest. In the context of protein-protein interactions in dilute solution, previous work showed that a minimal set of key physics to include in a CG model to accurately capture B_{22} at high ionic strength where electrostatics are heavily screened are:[29,30] steric interactions via the positions and sizes of the amino acids that make up the folded protein structure; and short-ranged attractions that combine the effects of van der Waals contacts (relative to water-protein interactions) and hydrophobic attractions. Here we also include the effect of long-ranged repulsions and attractions due to screened charge–charge interactions. For the purposes of this work, only native or folded proteins are considered, where keeping the molecule rigid (i.e., not permitting unfolding) is reasonable for comparison with experiments.[31,32] If one considers natively unfolded proteins or assembly steps that involve

large conformational / folding changes of the proteins involved, then the flexibility of the protein backbone and side chains should instead be included.[33]

A number of different CG models have been developed to characterize several protein processes, including protein folding,[34,35] protein fibril formation,[36,37] and protein–protein/ligand binding.[38,39] A recent review from Tozzini[40] is recommended for those interested in a detailed discussion on CG models. In the case of protein oligomerization, CG models have focused primarily on the overall thermodynamics and kinetics for the formation of protein clusters, rather than characterization of possible specific amino-acids and/or "hot-spots" within the protein sequence that are important to the oligomerization of folded proteins.[25,32] Only those CG models that were developed to study peptide aggregation and amyloid formation have been implemented both to identify aggregation-prone regions and to characterize the thermodynamics of the process.[37,41]

The present report provides a CG model that conserves physically relevant details at the amino-acid level while remaining simple enough to be tractable for extensive simulations of protein oligomerization. As a first application of the model, it is used here primarily to study reversible dimerization and protein–protein interactions of human γD-crystallin (gD-Crys), a two-domain globular protein of ca. 20 kDa that is present *in vivo* in the eye lens. gD-Crys is a Greek-key beta-sheet protein, with structural motifs similar to immunoglobulins, and whose folded and non-native aggregates have been associated with cataracts.[42,43] gD-Crys is a natively monomeric and highly-aggregation-resistant protein at physiological conditions (pH near neutral, ionic strength ~ 100 mM). However, sequence mutations[44–49] or different solution conditions (e.g., acidic pH or higher salt concentration)[50–52] have shown dramatic changes in terms of aggregation propensity, as well as the conformational and colloidal stability of gD-Crys. As such, gD-Crys serves as a model system to test the effects of perturbations of protein–protein interactions on the thermodynamics of protein oligomerization. The present report focuses on changes in solution conditions (e.g., protein and salt concentration), while longer term goals include identifying and altering aggregation-prone "hot-spots" regions in the protein sequence, and extending to more aggregation prone molecules such as antibodies.

Model and Methods

Model Description

Interactions between proteins are treated as occurring in an implicit solvent, and being a pair-wise sum of the interactions between amino acid residues on different proteins. Previous work[53] showed that a 1 bead-per-amino-acid model was equivalent to a 4-bead-per-amino acid model if one is interested in B_{22} and interactions between relatively rigid, folded proteins with short-ranged attractions and repulsions. The model here builds from that work, and also includes long-range electrostatic interactions between amino acids. Interactions between amino acids on the same protein are neglected, as each protein is treated as a rigid body that can translate its center of mass, and rotate as a rigid body about that center. The interaction potential, $u_{ij}(r)$, between residue i of one protein and residue j of another consists of: (i) a short-ranged attraction that accounts for a combination of hydrophobic attraction (for hydrophobic residues) and van der Waals attractions relative to

the corresponding solvent-protein attractions; (ii) steric repulsions; and (iii) screened electrostatic attractions and repulsions. The first two are combined in terms of a contribution from short-ranged attractions and repulsions (u_{ij}^{sh}), while the third is represented by a screened electrostatic interaction (u_{ij}^{el}) with a range that depends on the effective ionic strength.

$$u_{ij}(r_{ij}) = u_{ij}^{sh}(r_{ij}) + u_{ij}^{el}(r_{ij}) \quad (1)$$

where r_{ij} is the center-to-center distance between the *i* residue of one protein and the *j* residue of another protein. In the present work, only two proteins will be considered at a time, therefore indices denoting the different proteins are understood in what follows. The total effective potential energy, or solvent averaged potential of mean force (W), is then given by:

$$W = \sum_{i<j} u_{ij}(r_{ij}) \quad (2)$$

Attractive interactions between amino acids depend, at least in part, on their hydrophobicity and hydrogen-bonding capability (i.e., their water "affinity").[54–56] Building from the coarse-grained model for W that was developed and tested by Bereau and Deserno,[57] the magnitude of the attractions between residues on adjacent proteins is described primarily in terms of the relative hydrophobicity of the two residues that are interacting. This level of specificity is achieved by considering two parameters in the model. The first parameter provides a relative hydrophobicity score, ε_i, which is dimensionless and ranges from 0 for the most hydrophilic residue to 1 for the most hydrophobic. The values of ε_i are those used by Bereau and Deserno[57] based on Miyazawa and Jernigan's statistical analysis[58] of residue-residue contacts within the crystal structures of multiple proteins. The second free parameter, ε_{hp}, accounts for translating the strength of the attractive interaction into an absolute scale -i.e., ε_{hp} has units of energy. Thus, short-ranged interactions are treated as

$$u_{ij}^{sh}(r_{ij}) = \begin{cases} 4\epsilon_{hp}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \epsilon_{hp}(1-\epsilon_{ij}) & \text{if} \quad r_{ij} \le r_c, \\ 4\epsilon_{hp}\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] & \text{otherwise.} \end{cases} \quad (3)$$

where $\sigma_{ij} = (\sigma_i + \sigma_j)/2$, with σ_i being the van der Waals diameter of the *i*-th residue, which is calculated here based on the van der Waals area[59] by assuming a spherical shape for the single bead that represents a given amino acid. $r_c = 2^{1/6}\sigma_{ij}$ is the distance at which the interaction potential switches from being repulsive to attractive, in a Weeks-Chandler-Anderson type of treatment. This value is such that both the potential and its first derivative are continuous. The use of this form for the potential allows all types of residues to have the same strength for the short-ranged steric repulsion, whereas the attractive force depends on the relative affinity ε_{ij} between the *i*-th and *j*-th amino acids, and this form is also amenable to molecular dynamics simulations. The value of ε_{ij} is calculated from the geometric average

of the relative hydrophobic score of the residues *i* and *j* (i.e. $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$). Table 1 provides the values for σ_i and ε_i used here.

The long-range electrostatic interactions are modeled with a Yukawa-type potential:

$$u_{ij}^{el}(r_{ij}) = \epsilon_{cc} q_i q_j \frac{exp[-\kappa(r_{ij} - \sigma_{ij})]}{r_{ij}/\sigma_{ij}} \quad (4)$$

where q_i is the charge of residue *i* at a given pH, and is located in the center of the corresponding bead. The charge is assigned based on the pK_a of the side chain of the specific residue, and it adopts values of +1 or −1; that is, only conditions where the pH does not lie close to any side chain pKa values are considered. κ is an adjustable parameter representing the inverse of the effective interaction distance or screening length. When one considers the case of salts being treated as simple primitive ions, it is then expected to only be a function of the ionic strength. ε_{cc} is a parameter that accounts for the magnitude of the charge-charge interactions. It includes the effects of solvent dielectric constant, and is adjusted relative to the magnitude of the short-range attraction energy, so as to reasonably mimic experimental behavior. The use of a Yukawa-type potential provides flexibility to the model in order to capture the effect of media conditions (e.g. pH, ionic strength, temperature, etc.) on the residue–residue interactions. Thus, the model used here contains a total of three adjustable parameters (ε_{hp}, κ and ε_{cc}) that are determined by fitting against experimental values of the osmotic second virial coefficient (See below for details).

Computational methods

In order to parameterize the potential energy, several computational techniques were applied to identify and characterize important contributions to the protein self-association process, as well as to analyze thermodynamic properties. These methods include Metropolis Monte Carlo (MC) and Molecular Dynamics (MD), which are described in detail in the Supporting Information, with only key details and definitions given below. All lengths are measured in units of $\mathscr{L} = 1$ Å, and energies are related to the thermal energy $\mathscr{E} = k_B T_r \approx 0.6 \quad kcal \quad mol^{-1}$, where k_B is the Boltzmann constant and T_r = 300 K is a reference temperature. Masses are denoted in units of $\mathscr{M}$, which corresponds to the average weight of an amino acid (~ 110 Da ≈ 1.84 × 10^{-25} Kg). This choice of units leads to a characteristic time τ for molecular dynamics of $\tau = \mathscr{L}\sqrt{\mathscr{M}/\mathscr{E}} \approx 0.7 \quad ps$. As noted above, γD-crystallin (gD-Crys) is used as the model system, and its structure is taken from the Protein DataBank (PDB code: 1HK0). In addition, protein–protein interactions are assessed for bovine α-chymotrypsinogen (aCgn; PDB code: 1EX3) as a means to test the transferability of the present coarse-grained model.

Theoretical osmotic second Virial coefficient

Experimental data for the osmotic second virial coefficient, B_{22}, was used to determine the adjustable parameters in the potential energy function. B_{22} is formally related to protein–protein interactions in the limit of low protein concentration, averaged over the spatial

degrees of freedom of the solvent and any co-solute or co-solvent species -i.e., the protein–protein potential of mean force W in a grand-canonical ensemble[60] via

$$B_{22} = -\frac{1}{2}\int_r \int_\Omega \left(e^{-W/k_B T} - 1\right) dr d\Omega \quad (5)$$

where T is the absolute temperature, r is the centers-of-mass distance between two protein molecules, and Ω denotes the orientational degrees-of-freedom of one protein with respect to the other. W is represented here by Eq. 2, and is an explicit function of the center-to-center distance (r_{ij}) between all pairs of amino acids between two proteins; however, r_{ij} changes with r and Ω. Thus, W is also a function of the position and orientation of each protein.

To obtain theoretical values of B_{22}, the direct Mayer Sampling method[61,62] was applied. This method calculates the integral in Eq. 5 using a biased MC approach, performing importance-sampling based on those configurations relevant to the integral for the system of interest and a reference system for which B_{22} is known. Additional details are provided in Supporting Information.

Thermodynamics and Free-Energy Landscapes of Protein Self-Association

To determine the dominant configurations and thermodynamic properties of a pair of proteins, Replica Exchange Molecular Dynamics[63] (REMD) simulations were performed on a constant volume ensemble with two proteins ($N = 2$). This method was combined, for efficiency, with a symplectic quaternion scheme,[64] and coupled to a Nosé-Hoover Chain thermostat[65] to generate the correct canonical distribution for each replica in the simulation. REMD is a suitable method to determine accurate ensemble properties, since it helps to prevent simulations from becoming trapped in local basins on the free-energy landscape at low temperatures. Furthermore, combining REMD with weighted histogram analysis methods (WHAM)[66,67] allows one to reconstruct the density of states of the system, and calculate thermodynamic observables over a continuous range of temperatures.

All REMD simulations presented here were performed for $N = 2$, with proteins confined into a cubic box of length L with periodic boundary conditions. The box-length is set to simulate a desired protein concentration. A set of replicas in REMD were distributed between 80 K and 400 K, with the total number of replicas adjusted for a given set of model parameters (ε_{hp}, ε_{cc}, and κ), to assure acceptance ratios for the replica swaps between 40% and 60%. An integration time step $\delta t = 0.005\,\tau$ (≈ 3.5 fs) was used for all the REMD simulations, and MC swaps between replicas were attempted every $1\,\tau$ for REMD steps. The initial configuration of each replica (given by center-of-mass position and the four elements on the quaternion vector of each molecule) was chosen randomly. Each simulation employed two short warming-up periods for thermal equilibration of $5 \times 10^4\,\tau$ each, using standard MD simulations and REMD, respectively.[63] Thereafter, a sampling period of $1 \times 10^6\,\tau$ was performed using REMD, where the configurations and energy values of each replica were stored every $1\,\tau$ for further structural and thermodynamic analysis.

In order to characterize the thermodynamics of protein self-association, a series of order parameters Γ_i^α was collected during REMD simulations. These order parameters account for

the relative orientation of the *i*-th residue belonging to the α protein (α = 1, 2) with respect to the center-to-center distance between proteins, and they are calculated as

$$\Gamma_i^\alpha = \frac{r_i^\alpha \cdot r}{|r_i^\alpha||r|}$$

where r_i^α is the vector from the center-of-mass of the α protein to the *i* residue, *r* is the vector from the COM of the α protein to that of the other protein, and | . . . | denotes the magnitude of a given vector.

As defined above, Γ_i^α takes values between 0 and 1 for amino acids positioned on the "face" of the protein at near contact with the other protein, and it is smaller than 0 otherwise. Thus, these order parameters are directly related to the association-prone orientations/ configurations between two proteins. However, from a practical perspective, the analysis of these order parameters presents several complications as even for small proteins the total number of these order parameters can become very large (on the order of hundreds). Furthermore, they carry redundant information regarding the orientation of the self-assembled proteins, since at a given protein orientation the position of any residue is not independent of the others because the model treats both proteins as rigid bodies. In order to reduce this high dimensionality without losing the relevant information to the protein self-association process, Principal Component Analysis (PCA) was applied.[68]

PCA allows decomposing a given data set to look for most of the variability in the data by finding the principal components (i.e. the eigenvectors) of its covariance matrix. For the present case, this data set consists of the set of $\{\Gamma_i^\alpha\}$ from both proteins obtained during the sampling period of a REMD simulation at a given temperature. That is, there is a unique set of principal components for each simulated temperature, ionic strength, and protein concentration, over the full time course of the simulation. Since Γ_i^α is related to the ensemble of protein orientations relevant to protein–protein interaction, these principal components correspond to a set of orthogonal orientational axes along the path for protein self–association. Similarly, previous studies have successfully used PCA as a tool to identify and reduce different order parameters for processes such as protein folding,[69] protein aggregation,[70] and phase separation.[71]

Due to the statistical nature of PCA, the resulting new set of orthogonal orientational axes strongly depends on aspects such as the number of sampling points and the underlying distribution for the different configurations/orientations, and thus they may vary between data sets at different simulated conditions (e.g., temperature, ionic strength, and protein concentration). In order to overcome this limitation, the principal components of an arbitrary reference system were used to evaluate all sets of $\{\Gamma_i^\alpha\}$. This reference system is such that it contains most of the variability with respect to the information of interest (i.e. the relative orientation between both proteins).

Experimental Methods

Protein preparation and static light scattering

gD-Crys was expressed and purified as described previously.[49] Prior to measurements of light scattering, gD-Crys was dialyzed into 5 mM acetate buffer, pH 5.5, and used within 2 days of dialysis. Protein concentration was determined by absorbance at 280 nm using an extinction coefficient of 41040 M^{-1} cm^{-1}. Protein samples for light scattering were prepared at protein concentrations ranging from 1 to 10 mg/mL, and different NaCl concentrations to adjust ionic strength, and filtered prior to use. Static light scattering (SLS) measurements were performed using a Brookhaven Instruments Corporation (Holtsville, New York) instrument equipped with a Lexel (Fremont, California) model 95 argon-ion laser (λ = 532 nm), a BI9000AT correlator, and a BI200SM goniometer. For each salt concentration, scattered intensities of each sample, solvent, and toluene were recorded at a temperature of 25°C and scattering angle of 90°. The apparent weight-averaged molecular weight (M_w) and protein-protein Kirkwood-Buff integral (G_{22}) were obtained by regressing the data against[72]

$$\frac{R_{ex}}{K}=M_w c+G_{22}c^2 \quad (6)$$

where c is the concentration (mass/vol) of protein, R_{ex} is the excess Rayleigh ratio, and K is an optical constant that is a function of the refractive index of the solution.[72] The refractive index (n) and the differential index of refraction (dn/dc) were measured for all solution conditions (data not shown) and were found to be independent of ionic strength (n = 1.333 and average dn/dc = 0.187 ± 0.002 cm^3/g). Eq. 6 provides a more general functional form to relate protein–protein interactions with Rayleigh scattering compared to those that only apply at dilute conditions.[16,29,30]

In the "dilute" limit, G_{22} can be formally replaced by $-2B_{22}$,[60] and the resulting expression is equivalent to traditional models to analyze SLS data.[72] The "dilute" limit is such that the product of $cG_{22} \rightarrow 0$,[72] that is, low protein concentration and/or weak protein–protein interactions. For the present case, the range of protein concentration and the strength of the intermolecular interactions for gD-Crys ensures that the resulting fitted values of G_{22} are equivalent to the osmotic second virial coefficient B_{22}. For later analysis, B_{22} was scaled by the theoretical value of the hard sphere (HS) second virial coefficient using $B_2^{HS}=2\pi d^3/3$, where d is the effective HS diameter for a monomer. d was taken as 4 nm in this work for both gD-Crys and aCgn, yielding $B_2^{HS}=80.69 \quad \mathrm{L}/mol$.[3,72]

Results and Discussion

Parametrization of the residue-residue interactions

The potential energy model developed here contains three free-parameters, ε_{hp}, ε_{cc}, and κ, which provide flexibility to the model and simulate the effect of media conditions on protein–protein interactions. In the example below, these parameters are adjusted to semi-quantitatively capture experimental behavior of the osmotic second virial coefficient. B_{22} is a function of the spatial average protein-protein interactions at the conditions of the protein solution, and thus its value is a function of the protein structure/sequence, temperature, pH,

and ionic strength. At low ionic strengths, for instance, electrostatic interactions are relatively unscreened, and the value of B_{22} includes relatively large contributions from attractive and repulsive electrostatic interactions. As salt concentration increases, charge-charge interactions become screened, and non-electrostatic attractive interactions are the dominating term.[73] In choosing values for ε_{cc} and ε_{hp}, one may then expect that experimental data is needed that spans from low to high ionic strength at fixed pH, with κ treated as a function of ionic strength.

Experimental values of B_{22} were measured for γD-crystallin at pH = 5.5 and ionic strength (I) of 4.2, 14.2, 54.2, 254.2 and 504.2 mM , and used to refine the adjustable model parameters. Figure 1a shows the experimentally measured B_{22} values for gD-Crys as a function of the square root of I (or alternatively, κ) together with an illustrative curve obtained from the coarse-grained model at fixed values of ε_{hp} and ε_{cc} (details on the calculation of the theoretical curve are discussed below). The value of B_{22} at the plateau at high ionic strength is determined by the value of ε_{hp} (i.e., the strength of the non-electrostatic attractive interactions), with larger values of ε_{hp} corresponding to more negative values of B_{22}. Interestingly, the qualitative shape of the B_{22} vs. ionic strength curve is the same for any choice of ε_{hp} and ε_{cc} with the model used here (results not shown), and therefore one can use experimental data from conditions that correspond to fully screened charges (e.g, I ~300–500 mM for a 1:1 electrolyte) in order to set the value for the strength of the short-range interactions.

Thus, the value of B_{22} at the highest ionic strength (I = 504.2 mM) is used to find an appropriate value for ε_{hp} for gD-Crys. A similar approach is applied below for aCgn. At this high salt concentration, it is approximated that the contribution of electrostatic interactions to W is negligible in Eq. 2. Fig. 1b shows the simulated values of B_{22} relative to the hard-sphere second virial coefficient B_2^{HS} as a function of ε_{hp} at 300 K and pH=5.5, assuming ε_{cc} = 0. In order to assure that the CG model produces the correct experimental behavior when charge-charge interactions are highly screened, one must select the value of ε_{hp} in Figure 1b that gives a good semi-quantitative match to the corresponding experimental values. In the case of gD-Crys this leads to a value of ε_{hp} = 0.375 $\mathscr{E}$ as the short-range attraction parameter per amino acid. This may not be the only reasonable choice for the value of this parameter, given the statistical uncertainty in the experimental B_{22} values. Nevertheless, for the purposes of this work, this choice of ε_{hp} provides a good semi-quantitative representation of protein-protein interactions, and therefore it is used in the remainder of this report.

For the remaining parameters associated with electrostatic interactions, a more detailed analysis of their effect in B_{22} is required. Given the nature of B_{22} as an average measurement of the protein–protein interactions, a strong coupling between these parameters is expected since multiple choices of ε_{cc} and κ may lead to the same value of B_{22}. Traditionally, when this issue arises using a Yukawa-type potential, it has been resolved by adopting the Debye-Huckle model, which provides a mathematical relationship between κ and ionic strength.[27,32] However, it is known that the Debye-Hückel theory may underestimate the role of preferential solvation or preferential accumulation of different ions, for example, within the Hofmeister series.[74,75] As such, the present work also

considered the effects of κ and ε_{cc} on the calculated B_{22} values without first assuming a relationship such as a Debye-Huckel model.

In order to obtain an unbiased estimation of these parameters, the values for B_{22} as a function of $\varepsilon_{cc}/\varepsilon_{hp}$ and κ were calculated using Mayer sampling with the CG protein model at a fixed value of ε_{hp} (determined above), and then interpolated using Response Surface methods,[76] with $\varepsilon_{cc}/\varepsilon_{hp}$ ranging between 0 and 1, and $1/\kappa$ between 5 $\mathscr{L}$ and 40 $\mathscr{L}$. The bounds for the ranges of κ and ε_{cc} were selected to cover a physically realistic range of B_{22} values for proteins in aqueous solutions. Thus, the range for κ, for instance, corresponds to conditions where the effective electrostatic interaction distance (i.e., the screening length) spans from the size of an average amino-acid, to conditions where each charge in one protein can interact effectively with all the charges on the other protein -i.e., $1/\kappa$ of the order of the protein diameter. The response surface was obtained from a multivariate regression of simulated B_{22} values to a full-quadratic polynomial (i.e., including second-order and cross terms). Details about the fitted model and the best-fitted values for the corresponding coefficients are provided in the Supporting Information. Fig. 2 shows the resulting response surface for the calculated values of B_{22}/B_2^{HS} vs. $\varepsilon_{cc}/\varepsilon_{hp}$ and $1/\kappa$ using the value of ε_{hp} determined above. Fig. 2 includes the values of B_{22} calculated from Mayer Sampling method (circles on the figure), which are used to construct the B_{22}-surface.

Fig. 2 clearly illustrates the coupling between ε_{cc} and κ that was qualitatively anticipated in the discussion above. The profiles for B_{22} versus screening length differ slightly as one changes ε_{cc}. As expected, the largest quantitative differences occur for small values of κ (low ionic strength). To find an appropriate set of parameters, one must select a value of ε_{cc} which provides a physically reasonable description of protein–protein interactions for large screening lengths (i.e. small κ). Experimentally, B_{22}/B_2^{HS} values for proteins typically lie between approximately ±10 unless one considers: (i) extremely highly charged proteins; or (ii) conditions corresponding to such strong protein–protein attractions/repulsions that they are thermodynamically unstable or highly metastable with respect to phase separation.[72,77] By inspection of Fig. 2, one can see that quantitatively reasonable B_{22} profiles are obtained for $0.2 < \varepsilon_{cc}/\varepsilon_{hp} < 0.4$, since values of ε_{cc} outside that region provides either too large or too small B_{22} values at very long screening lengths, conditions that correspond to essentially just protein, water, and counterions needed for electronegativity (i.e. for κ^{-1} larger than the protein diameter). Any value of ε_{cc} within the range described above will give a reasonable semi-quantitative description of protein–protein interactions with respect to salt concentration provided that a good correlation between the screening length and the ionic strength is given. For simplicity and concreteness, a value of $\varepsilon_{cc} = 0.125\,\mathscr{E}$ (i.e. $\varepsilon_{hp}/\varepsilon_{cc} = 3$) is chosen in the remainder of the results reported here.

Finally, the set of experimental B_{22} values in Fig. 1a is used to estimate a scaling relationship for κ with ionic strength. Based on standard theoretical arguments for dilute and semi-dilute salt conditions,[78] κ was scaled with the square root of ionic strength (I), -i.e., $\kappa = \alpha\sqrt{(I)}$. Combining this scaling relationship with the response surface for B_{22}, and fitting against the experimental B_{22} vs I provides a value for α. Fig. 1a shows a comparison

between the resulting theoretical B_{22} values and those obtained from light scattering experiments for γD-crystallin. The fitted value for α was found to be 3.5 ± 0.5 $M^{-1/2}nm^{-1}$.

Figure 1a illustrates a good semi-quantitative agreement between the coarse-grained model results, and those from light scattering for gD-Crys. As noted above, it is anticipated that a different set of values might be required for different proteins for a quantitative description. Although the choice of values for the set of parameters allows one to capture the behavior of B_{22} with respect to changes on media conditions, the statistical uncertainty on the experimental data and some of the intrinsic characteristics of each protein such as the concentration of counterions and the distribution of surface residues make it unlikely that exactly the same model parameters will globally apply to different proteins. If one were interested in applying the working coarse-grain model for quantitative analysis, both ε_{hp} and α would need to be readjusted to a similar experimental profile for the protein of interest. In the case of the ratio $\varepsilon_{cc}/\varepsilon_{hp}$, it is expected to be more general since it is calculated based on the physical behavior of colloidal interactions, and any error introduced from the use of this value should be damped by the scaling of κ with the ionic strength.

The transferability of the present coarse-grained model was considered by performing a re-adjustment of the model parameters against a different protein, α-chymotrypsinogen (aCgn), following a similar procedure to the one described above. Experimental values from a separate study for B_{22} at a pH=3.5 and different ionic strengths[73,79] were used for this analysis. The results of this re-parametrization of the CG model are summarized in the Supporting Information (Fig. S2 and S3) in terms of ε_{hp} and the response surface for aCgn. The B_{22} response surface for aCgn shows the same qualitative and semi-quantitative behavior as that of gD-Crys, suggesting that the parameters obtained above can be used for qualitative analysis of protein–protein interactions and experimental trends for different proteins. Additionally, for physically reasonable B_{22} values, the response surface in Fig. S3 shows that the range of $\varepsilon_{cc}/\varepsilon_{hp}$ is the same as the one observed for gD-Crys (i.e., $0.2 < \varepsilon_{cc}/\varepsilon_{hp} < 0.4$). ε_{hp} was found to be slightly larger (= $0.408\mathscr{E}$) than that obtained for gD-Crys. By analogy with the procedure used for gD-Crys, a value of $\varepsilon_{hp}/\varepsilon_{cc} = 3$ was used in order to obtain the proportionality between κ and the square-root of I -i.e., the value of α, which was obtained as $\alpha = 3.08 \pm 0.1$ $M^{-1/2}nm^{-1}$ for aCgn. Fig. 3 compares the values of B_{22} for aCgn obtained from light scattering experiments and the resulting theoretical curve. For comparison, simulated values of B_{22} using the same set of parameters obtained for gD-Crys are shown in the figure.

The results in Fig. 3 indicate that some factors for different proteins and/or counterions in solution may cause the CG model parameters to not be quantitatively universal, although the values are quite similar, and so, many of the qualitative and semi-quantitative conclusions below are expected to hold more widely than just for the proteins considered here. Similarly, the procedures illustrated above can be used to determine an appropriate new set of parameters, if needed, for the most quantitative assessment of the behavior of different proteins. If one is only interested in a qualitative or semi-quantitative analysis of protein–protein interactions and B_{22}, this CG model and the set of values for ε_{hp}, ε_{cc}, and κ (or α) found for gD-Crys will provide a reasonable approximation, where large deviations are expected only at very low ionic strengths. This is not completely surprising as Grünberger et

al.[53] showed that one can obtain the same qualitative behavior of B_{22} versus the strength of residue–residue interactions for different structural levels of coarse-graining, while quantitative equivalence between different CG model requires an appropriate re-scaling of the strength of the interactions between amino-acids.

Of course there is no unique methodology for optimizing the model parameters. For instance, one can alternatively consider a Debye-Hückel approach to set the proportionality between κ and ionic strength, and use the experimental data to fit ε_{hp} and ε_{cc}. Such approach yields a proportionality constant of 3.24 $M^{-1/2}nm^{-1}$, which is very close to the a values found for gD-Crys and aCgn. Similarly, one could fit the model to the experimental data for both proteins simultaneously. Inspection of the values above and in the figure caption show that the resulting parameters are reasonably close to one another. For concreteness in characterizing the self-association behavior of gD-Crys, the values above for gD-Crys were used in the remainder of the work below.

Thermodynamics of protein self-association

In what follows, the coarse-grained model was applied to characterize and analyze the effect of intermolecular forces, in particular electrostatic interactions, on the thermodynamics of protein self-association for γD-crystallin. A series of REMD simulations were performed at a effective pH = 5.5 and a screening length of 40 $\mathscr{L}$ (i.e. $I \approx 5.5$ mM). At this condition, all acid or basic residues are charged, and each of these amino acids will feel the electrostatic forces coming from all the other charged residues once the two molecules are at contact. The simulations are carried out for three different theoretical protein concentrations: 5, 10, and 20 mg/mL (i.e. box sizes of 233.5, 180, and 147.1 $\mathscr{L}$, respectively). These concentrations correspond to 0.25, 0.5, and 1 mM protein, respectively. For completeness and comparison, an additional simulation is performed for the case when the concentration is 10 mg/mL and all charge-charge interactions are effectively screened (i.e. $\kappa \to \infty$, or alternatively $\varepsilon_{cc} = 0$). Note that changes in “apparent” protein concentration are meant to study the effect of constraining the average center-to-center distance, as gD-Crys is not well approximated as a spherical / globular protein. To properly simulate all the effects of moving to high concentration, one needs to adjust both the volume and the number of proteins in the box in order to consider other non-idealities beyond pairwise protein interactions. The focus here is on dimer-monomer equilibria; simulating association in concentrated systems is a focus of ongoing work.

After the simulations were finished, the free-energy of the system, relative to the ideal-gas, as a function of the center-of-mass distance *r* between two proteins, i.e. the potential of mean force (PMF), as well as the internal energy and entropy related with the self-association process were calculated. As mentioned in the Methods section, all the simulations were run for a temperature range between 80 and 400 K. Over this range, it is found that the simulations cover the transition between the monomeric state (high temperature) and the self-assembled state (low temperature), and such a transition will be evident from the comparison of PMFs at different temperatures. Fig. 4 illustrates the PMF (*F*), average energy (*U*), and entropy (*S*) curves for a protein concentration of 10 mg/mL and low and high ionic strength at $T = 100$, 150, 200, and 300 K. A similar figure for 5 and 10

mg/mL is provided in Fig. S4 in the Supporting Information. Here, the free-energy and average energy are calculated by combining the data collected during the simulations using WHAM[67] (details are provided in Supporting Information), whereas the entropy is calculated from the difference between those quantities (e.g., $-TS = F - U$).

The free-energy curves show that, at low temperatures, the state where both proteins are at contact is very thermodynamically favorable (i.e. for distances less or equal to the protein diameter, $r \lesssim 40$ Å). In contrast, at high temperatures, all the simulations exhibit higher thermodynamic stability for distances much larger than the protein diameter. By inspection of the free-energy curves and those of the average energy and entropy, one can see that this thermal stability at low temperatures is directly related with the range of the van der Waals interactions, since the main contribution to the PMF arises from U rather than the entropy at such conditions. For $r \approx 28$ Å, the curves show that there is a decrease in the entropy of the system as a consequence of the strong average energy.

Therefore, it is anticipated that the number of configurations leading to strong, specific protein-protein interactions (i.e., “lock-and-key” configurations) is small (see discussion below), and thus the entropy of the system is considerably reduced when these configurations are reached. In addition, the depth of the free energy minimum at these low temperature shows that there is a greater stability for the cases when the concentration is higher or when electrostatic interactions are screened. In the cases when electrostatic interactions are included, the average energy decays to zero at distances slightly larger than the protein diameter, despite the fact that the effective range for the charge–charge interactions is on the order of the protein diameter (i.e. $1/\kappa = 40$ Å). However, for distances larger than 30 Å the free-energy of self-association is completely dominated by the entropic contribution. This suggests that at high temperatures, where the contribution from U is almost negligible, it is the competition between many different configurations that have weakly interacting regions of the protein (i.e., the entropy of self-association) which determines the net B_{22} in these conditions where stable oligomers do not form. This result is in contrast to cases where it is appropriate to argue in terms of a few highly specific, “lock-and-key” interactions, where only a very small number of strongly interacting configurations determine the free-energy of protein oligomerization.

This is even more evident if one compares the thermodynamics of protein self-association in terms of the dissociation constant K_d. By defining the dimeric (monomeric) state as those configurations with $r \leq 40$ Å ($r > 40$ Å), K_d as a function of temperature was calculated for all the working conditions (see Supporting Information). The results shows that at very low temperatures K_d is on the order of nM or smaller, which is consistent for a system where strongly bound, “lock-and-key” dimers are formed. In contrast, at experimentally relevant temperatures, K_d values are much larger than the simulated protein concentrations, consistent with the experimental observation that gD-Crys is natively a monomer, with no measurable Kd value, to the best of our knowledge.[46,50,51]

The results in Fig. 4 for low and high ionic strengths also suggest that electrostatic interactions play a central role on modulating the stability of the self-assembled proteins. Figure 5 shows free-energy landscapes as contour plots of the free-energy as a function of

two selected order parameters such as center-to-center distance (r) and van der Waals energy (E_{vdw}). The color scale is shown next to each panel, with additional details in the figure caption. The two-dimensional contour free-energy curves in Fig. 5 show how both the van der Waals (E_{vdw}) and electrostatic (E_{cc}) energies contribute to the stability of the system as a function of r. Whether or not charge–charge interactions are present, the qualitative behavior of the free energy surfaces with respect to r and E_{vdw} remains unchanged. However, the quantitative behavior is affected by E_{cc}. Interestingly, despite the electrostatic interactions resulting in net repulsive forces at longer distances, they cause an increase in the thermodynamic stability of certain configurations at shorter distances, and thus they increase the propensity for dimer formation. The results illustrated here suggest that the effect of E_{cc} on the stability of the self-assembled proteins is a consequence of local charged residues rather than the overall charge distribution. This type of behavior is not completely surprising as it has been shown that single-charge point mutants which increase the net charge of γD-crystallin can ultimately favor the formation of both native and non-native aggregates.[44–46,48]

Furthermore, the curves in Figures 4 and 5 clearly depict a thermal transition between monomeric and dimeric states. By calculating the theoretical heat capacity (C_v) as function of temperature, such a transition can be observed in more detail. Here, the heat-capacity is calculated as the variance of the internal energy.[80] That is,

$$\frac{C_v}{k_B} = \beta^2 \left(\left\langle U^2 \right\rangle - \langle U \rangle^2 \right)$$

where $\beta = k_B T$, and $\langle \ldots \rangle$ indicates the ensemble average. Figure 6 illustrates the resulting C_v profiles for all the cases tested here.

In all the cases the C_v profiles show a “melting” temperature (T_m) between 135 and 150 K. Here, “melting” corresponds to shifting from a stable dimer (a more “condensed” state) to a stable monomer (less “condensed” state). This should not be confused with the T_m for protein unfolding, which occurs at much higher temperatures. The T_m values in Fig. 6 correlate with what one would expect for the thermal stability of protein self-assembly such that those conditions that stabilize protein dimer (e.g. high protein concentration and/or high ionic strength) correspond to a larger T_m value. However, in all the cases, these T_m values also occur at low temperatures compared to typical experimental conditions. As a consequence of the approximations in the coarse-grained model (e.g. implicit-solvent and rigid molecules), it is expected that these results underestimate the quantitative range of temperatures at which this transition occurs. However, given that γD-crystallin is a natively monomeric protein which is soluble to high concentrations *in vivo* and *in vitro*, a native order–disorder transition (e.g., protein self-assembly or protein crystallization) is expected to occur only at low temperatures or at much higher concentrations than those tested here.

The free-energy landscapes at low temperature also show a set of different free-energy minima for inter-protein distances smaller than the average protein diameter, with a global minimum on the PMFs for $r \approx 25$ Å. Note that gD-Crys is not spherical, so distances between centers-of-mass less than the average diameter simply indicate approach of the

proteins along their shorter axes. These features correspond to multiple possible stable configurations/orientations for the two self-assembled proteins, and the magnitude of the minima are outside the range of statistical noise in the data. Given that γD-crystallin is a two-domain protein whose structure resembles two overlapping spheres, this global minimum would correspond to configurations where both domains are simultaneously in contact laterally. However, information related to the orientation of both proteins when they are at contact, as well as the amino-acids involved on these configurations, cannot be extracted from just the free-energy surfaces plotted above. The next section illustrates how this information can be drawn from the set of order parameters $\{\Gamma_i^\alpha\}$ as well as how the charge–charge interactions modulate the configuration/orientation of the self-assembled proteins.

Note that the simulations and analysis of protein self-association thermodynamics shown above are restricted to conditions where protein unfolding is not a key intermediate step (i.e., only native-state aggregation). The highest temperature conditions are required only to assure efficient sampling of the most relevant states at intermediate and lower temperatures (e.g., to avoid sampling bottlenecks in simulating directly at low temperature).

If one considers non-native aggregation, a fully flexible CG model such as those employed elsewhere is required.[28,31,57] In the present case, the experimental unfolding transition for gD-Crys occurs near or above the boiling point of water.[45,51] Even for native-state association where flexible loops or other segments rearrange upon binding, the computational burden and advanced methodology for properly employing a hybrid rigid-flexible model requires a non-trivial mixture of constrained and unconstrained MD algorithms.[64,65] While these are not anticipated to be concerns for gD-Crys under the intermediate and low-temperature conditions of interest here, future work on less thermally stable proteins will require these additional aspects to be addressed.

Structural features of self-assembly

As mentioned in the Methods section, in order to analyze the structural features of the protein dimers, a series of 346 order parameters were sampled during the REMD simulations, with each of these parameters related to the orientation of each amino acid residue with respect to the center-to-center vector between proteins. Because of the redundancy and high dimensionality in how the order parameters were defined, PCA is applied to find a unique set of orthogonal axis (or principal components), which contains the same statistical information as the full set of values of $\{\Gamma_i^\alpha\}$. For simplicity and ease of comparison across different sample conditions (e.g., temperature and protein concentration), the set of sampled $\{\Gamma_i^\alpha\}$ for the conditions of 10 mg/mL, low ionic strength, and T = 80 K was used to calculate the reference principal components. The resulting reference principal components were used to transform the sets of order $\{\Gamma_i^\alpha\}$ from other simulations to a set or orthogonal orientational coordinates or principle components P_i; i denotes which principal component, with lower i values indicating those components that capture most of the variation in the overall data set. Details on the calculation and resulting principal components are provided on Supporting Information. Note that by applying PCA, the set of

346 different Γ_i^{α} is reduced to only 3 principal components which provides approximately 82% of the total statistical variability of the data, with P_1, P_2, and P_3 covering 48%, 18%, and 16% of this variability, respectively.

Figure 7 shows the two-dimensional (2D) distribution of the number of configurations as a function of P_1 and P_2 for the series of concentrations and ionic strength values tested here. These distributions correspond to density maps based on the number density of points for a given value of P_1 and P_2 regardless the values of the other principal component P_3 and the center-to-center distance r. For easier comparison, density maps are shown as the negative base-10-logarithm of the probability distribution function. On this scale, the most likely configurations (i.e., local maxima on the distribution function) are shown as local minima in the figure. This is akin to free energy landscapes,[80,81] where the lowest free energy "basins" of configurations are the most likely states at a given thermodynamic state point, so long as they are sufficiently deep and there are not higher minima that are very broad and therefore entropically stabilized compared to narrow, deeper minima. The color scale is shown next to each panel. Analogous 2D profiles for the distribution of configurations with respect to P_1 and P_3, and P_1 and r are provided on the Supporting Information.

The main interest for the use of PCA is to characterize the orientations of each of the proteins in the self-assembly process, as well as to identify those residues which have the greatest effect on the self-association of γD-crystallin. Therefore the following discussion focuses first on T = 80 K, where dimerization is favored. Although the distribution functions shown in Fig. 7 and those in the Supporting Information are independent of the center-to-center distance, at the temperature evaluated here the configurational space is expected to be mainly formed by structures with $r \leq 40\mathscr{L}$ (i.e., when both proteins are at contact) as it was illustrated in Fig. 4. This is not necessarily what will occur quantitatively for higher temperatures (see also below).

The density maps in Fig. 7 illustrate that the dimeric state is composed of a set of different stable configurations/orientations for both proteins, which can be identified by the local minima on the figure (or equivalently maxima for density of configurations, or density of points per unit area in the figure). Regardless of the solvent conditions or temperature, two of the orientations (labeled **A** and **B** in Fig. 7) are relatively dominant configurational states when the dimer is formed, whose populations are at least ten times larger than any of the other stable orientations. From the other possible orientations, one of them (orientation **C**) is present in a large population for all the conditions except for the case of low protein concentration. This is perhaps a consequence of the decrease on the entropy of the system with concentration (see Fig. S3 and discussion below). Nevertheless, the results in Fig. 7 suggests that charge–charge interactions play a central role in modulating the self-association of γD-crystallin. This is apparent because, while changes in protein concentration appear to have an equal effect on the distribution of protein orientations, the effective range of the electrostatic interactions seems to influence the actual likelihood of specific stable configurations (cf. panels **b** and **d** in the figure), and thus it biases the resulting populations of stable structures for the protein dimers.

Given that the choice of order parameters and the resulting principal components are only related to the orientations of the proteins in the system, this approach can be used more generally to identify the structural features of the proteins at contact. Fig. 8 illustrates the extension of the above analysis to characterize the structures that contribute to the calculation of B_{22} at different temperatures. These configurations are shown along the same set of principal component axes P_1 and P_2 used above. Figure 8 was constructed by calculating B_{22} using Mayer sampling to bias the simulation to the configurations that most greatly contribute to the integral for B_{22}. The value of P_1 and P_2 was calculated for the configurations during the Mayer sampling simulation. The resulting two-dimensional histogram of the density or frequency of (P_1, P_2) pairs is shown as Figure 8.

For comparison, the right-most panels in Fig. 8 show the cumulative distribution functions for the principal component P_1 obtained from: configurations from the calculation of B_{22}; and the distribution of P_1 obtained from REMD simulation for both c = 10 mg/mL and $1/\kappa$ = 40 L. Those configuration identified above as the predominant configurations for the dimeric state (i.e., configurations **A** and **B**) are also highlighted in the right-hand panels, to show which regions of P_1 values correspond to the “wells” in Figure 7. Interestingly, comparison of both the important configurations and the resulting cumulative distribution functions of configurations/orientations on the B_{22} calculations with those obtained from REMD (cf. Fig. 7c) shows a good qualitative agreement among them in terms of the pattern of the density of configurations across the landscape. Although the comparison cannot be done on a quantitative basis given that B_{22} is independent of concentration, the fact that this distribution of important configurations for B_{22} maps with that from a canonical ensemble at finite concentration suggests that B_{22} itself can be used as an indicator of protein self-association, at least for these low-temperature conditions where self-association is prevalent.

Additionally, the results shown in Fig. 8 argue the importance of not only the strength of the protein–protein interaction, but also the entropy of the system (i.e., the distribution of configurations) for the calculation of B_{22}. This highlights an interesting question that has been previously posed -is it only a small number of configurations that dominate B_{22} for real proteins in solution? The results in Fig. 8 show that the answer to this question depends on what one considers “small”. The fraction of configurations contributing most strongly to B_{22} is of the order of 1–10 percent of the full density of states under most conditions. This indicates that many configurations (~ 90%) do not contribute appreciably, but ~1–10 percent is still a relatively large fraction of the configuration space in statistical mechanical terms.[60,80] The “lock-and-key” approach states that only a very small number of stable configurations contribute to the integral of B_{22}, implying that the strength of the interaction energy from these configurations is very large. Based on the result here, such an argument might apply at low “effective” temperatures, where dense protein phases (e.g., protein clusters or crystals) are prominent and the interaction energy is large (e.g., see Fig. 5), or where stable dimers or higher oligomers form under dilute conditions. However, at the experimental conditions where B_{22} is typically measured (i.e., effective temperatures high enough that stable dimers do not form), the average interaction energy is small as consequence of the large number of stable configurations. For instance, the orientations **A**

and **B** (highlighted in Fig. 8, right panels) represents about 60% of the contribution to B_{22} at T = 100 K, but they contribute less than 1% at T = 250 K.

One of the goals of this work was to identify amino-acids, or "patches" of amino-acids, that have the greatest effect on the formation of the protein oligomers (i.e., "hot-spots" on the protein surface). The distributions in Fig. 7 and those in the Supporting Information provide a means to address this question. They allow one to characterize the configuration of each of the proteins in the self-assembled dimer by providing a unique set of "coordinates" in terms of positions and orientations (i.e. r, P_1, P_2, and P_3) for the thermodynamically stable states. Once the most stable configurations are identified, the position of each of the residues, relative to the center-of-mass, is recovered by transforming the set of principal components back to the original order parameters $\{\Gamma_i^\alpha\}$. The main advantage of this procedure is that the "hot-spot" residues, i.e. those amino-acids at contact in the self-assembled dimer, can be directly identified from the resulting set of $\{\Gamma_i^\alpha\}$. Fig. 9 provides an schematic representation of the primary structure of γD-crystallin, highlighting the position of those amino acids involved on the dominant orientations of the self-assembled protein (e.g. orientations **A**, **B**, and **C** on Fig. 7). For illustrative purposes, Figure S10 in Supporting Information provides representative snapshots from the REMD simulations for each of the dominant dimeric orientations.

In the case of the structures **A** and **C**, most of the "hot-spots" are located on the C-terminal domain, with a large portion of them being shared by both configurations. Since both states differ slightly in terms of the orientation of both proteins (data not show), it is not surprising that there is a similarity in terms of the contacts residues between these two orientations. Nonetheless, the common regions among the "hot-spots" suggest particular regions of the protein surface with a high propensity for self-assembly. On the other hand, the relevant amino-acids on the configuration **B** correspond to completely different regions of the protein, covering both the N- and C-terminal domains. Interestingly, all of these "hot-spot" regions cover residues which are experimentally known to affect native or non-native aggregation of γD-crystallin, such as S130,[49] P23,[46] and I90.[47] Moreover, the present coarse-grained model also identifies the region between V126–Y134 as a key region for the self-association process. Based on a consensus among different aggregation predictors (e.g., AGGRESCAN, TANGO, and PASTA),[41,82–84] this region was also found in a previous report as an important sequence-fragment for the intrinsic aggregation propensity of gD-Crys.[49]

Summary and Conclusions

A coarse-grained model to assess protein–protein interactions and protein colloidal stability was developed and applied to protein dimerization for γD-crystallin. The resulting model contains three adjustable parameters which were calculated from experimental data of the osmotic second virial coefficient of gD-Crys, and these were reasonably similar for another test protein when comparing prediction vs. experimental B_{22} values. Using REMD simulations in combination with PCA, the thermodynamics and the structural features of gD-Crys dimers were evaluated. The results show that at experimentally relevant conditions for gDCrys the competition between many possible dimer configurations (i.e., the entropy of

the system) dominates over the direct energetic contributions to the the free-energy of protein self-association unless one is at conditions where stable dimers form in solution. This argues that one must be careful in interpreting B_{22} as being indicative of a small number of dominant pair-wise orientational configurations that are analogous to those in a protein crystal. Rather, there appear to be many configurations (in a statistical mechanical sense) that contribute to B_{22}. Some of these likely include those “lock-and-key” configurations, but these will not be the only significant contributions to B_{22} if one is working at conditions where the monomer state is that for natively monomeric proteins.[24,26,50,73]

Additionally, it was found that the short-range attractive interactions determine which are the most stable configurations for protein dimers when they are stable, but charge-charge interactions help dictate the relative populations or the free-energy of the different orientational states that make up the dimer ensemble. Finally, the results shown here allow identification of surface residues or regions in gD-Crys that are involved in protein self-association (i.e., “hot-spots”) by extending the analysis to low temperatures where protein oligomers are the most stable state. The resulting “hot-spots” agree with those previously obtained from different aggregation predictor profilers and experimental results, and also postulate new aggregation-prone sites.

In summary, the present coarse-grained model, combined with PCA, provides a powerful tool for the analysis of the thermodynamics and the structural aspects of the self-assembly of natively monomeric proteins. Although it is computationally more expensive than simple protein aggregation predictors, it provides a more detailed description of the process since it explicitly considers the contribution of the electrostatic interactions, the short-ranged attractions and repulsions, as well as the entropy of the system, and does so at the resolution of individual amino acids on the protein surface. Furthermore, the results illustrate the capability of the model to capture the important amino-acids in the self-association process of γD-crystallin. It will be extended in future work to predict and control the effect of possible point mutations on the aggregation of this and other proteins.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors gratefully acknowledge funding from the National Institutes of Health (R01 EB006006), the National Institute of Standards and Technology (NIST 70NANB12H239), and the National Science Foundation (CBET 0853639) in support of this work.

References

1. Stradner A, Sedgwick H, Cardinaux F, Poon WCK, Egelhaaf SU, Schurtenberger P. Equilibrium Cluster Formation in Concentrated Protein Solutions and Colloids. Nature. 2004; 432:492–495. [PubMed: 15565151]
2. Wang Y, Lomakin A, Latypov RF, Benedek GB. Phase Separation in Solutions of Monoclonal Antibodies and the Effect of Human Serum Albumin. Proc. Natl. Acad. Sci. U.S.A. 2011; 108:16606–16611. [PubMed: 21921237]

3. Li Y, Ogunnaike BA, Roberts CJ. Multi-Variate Approach to Global Protein Aggregation Behavior and Kinetics: Effects of pH, Nacl, and Temperature for Alpha-Chymotrypsinogen A. J. Pharm. Sci. 2010; 99:645–662. [PubMed: 19653264]
4. Zhang J, Liu XY. Effect of Protein-Protein Interactions on Protein Aggregation Kinetics. J. Chem. Phys. 2003; 119:10972–10976.
5. Shire SJ, Shahrokh Z, Liu J. Challenges in the Development of High Protein Concentration Formulations. J. Pharm. Sci. 2004; 93:1390–1402. [PubMed: 15124199]
6. Saluja A, Kalonia DS. Nature and Consequences of Protein-Protein Interactions in High Protein Concentration Solutions. Inter. J. Pharm. 2008; 358:1–15.
7. Keskin O, Gursoy A, Ma B, Nussinov R. Principles of Protein-Protein Interactions: What Are the Preferred Ways for Proteins to Interact? Chem. Rev. 2008; 108:1225–1244. [PubMed: 18355092]
8. Zhou H-X. Influence of Crowded Cellular Environments on Protein Folding, Binding, and Oligomerization: Biological Consequences and Potentials of Atomistic Modeling. FEBS Lett. 2013; 587:1053–1061. [PubMed: 23395796]
9. Nooren IMA, Thornton JM. Diversity of Protein-Protein Interactions. EMBO J. 2003; 22:3486–3492. [PubMed: 12853464]
10. Huang P-S, Love JJ, Mayo SL. A de novo Designed Protein-Protein Interface. Prot. Sci. 2007; 16:2770–2774.
11. Chi EY, Krishnan S, Randolph TW, Carpenter JF. Physical Stability of Proteins in Aqueous Solution: Mechanism and Driving Forces in Nonnative Protein Aggregation. Pharm. Res. 2003; 20:1325–1336. [PubMed: 14567625]
12. Valente JJ, Payne RW, Manning MC, Wilson WW, Henry CS. Colloidal Behavior of Proteins: Effects of the Second Virial Coefficient on Solubility, Crystallization and Aggregation of Proteins in Aqueous Solution. Curr. Opn. Pharm. Biotechnol. 2005; 6:427–436.
13. Xu D, Lin SL, Nussinov R. Protein Binding versus Protein Folding: The Role of Hydrophilic Bridges in Protein Associations. J. Mol. Biol. 1997; 265:68–84. [PubMed: 8995525]
14. Ford DM. Enthalpy-Entropy Compensation is not a General Feature of Weak Association. J. Am. Chem. Soc. 2005; 127:16167–16170. [PubMed: 16287305]
15. Kern N, Frenkel D. Fluid-Fluid Coexistence in Colloidal Systems with Short-Ranged Strongly Directional Attraction. J. Chem. Phys. 2003; 118:9882–9889.
16. Minton AP. Static Light Scattering from Concentrated Protein Solutions, I: General Theory for Protein Mixtures and Application to Self-Associating Proteins. Biophys. J. 2007; 93:1321–1328. [PubMed: 17526566]
17. Young TM, Roberts CJ. A Quasichemical Approach for Protein-Cluster Free Energies in Dilute Solution. J. Chem. Phys. 2007; 127:165101. [PubMed: 17979394]
18. Rosenbaum DF, Zukoski CF. Protein Interactions and Crystallization. J. Cryst. Growth. 1996; 169:752–758.
19. Neal BL, Asthagiri D, Velev OD, Lenhoff AM, Kaler EW. Why is the Osmotic Second Virial Coefficient Related to Protein Crystallization? J. Cryst. Growth. 1999; 196:377–387.
20. Dumetz AC, Chockla AM, Kaler EW, Lenhoff AM. Effects of pH on Protein-Protein Interactions and Implications for Protein Phase Behavior. BBA-Proteins Proteomics. 2008; 1784:600–610. [PubMed: 18258214]
21. Schaink HM, Smit JAM. Determination of the Osmotic Second Virial Coefficient and the Dimerization of Beta-Lactoglobulin in Aqueous Solutions with Added Salt at the Isoelectric Point. Phys. Chem. Chem. Phys. 2000; 2:1537–1541.
22. Weiss WF IV, Young TM, Roberts CJ. Principles, Approaches, and Challenges for Predicting Protein Aggregation Rates and Shelf Life. J. Pharm. Sci. 2009; 98:1246–1277. [PubMed: 18683878]
23. Roberts CJ, Das TK, Sahin E. Predicting Solution Aggregation Rates for Therapeutic Proteins: Approaches and Challenges. Inter. J. Pharm. 2011; 418:318–333.
24. Gliko O, Pan W, Katsonis P, Neumaier N, Galkin O, Weinkauf S, Vekilov PG. Metastable Liquid Clusters in Super- and Undersaturated Protein Solutions. J. Phys. Chem. B. 2007; 111:3106–3114. [PubMed: 17388477]

25. Young TM, Roberts CJ. Structure and Thermodynamics of Colloidal Protein Cluster Formation: Comparison of Square-Well and Simple Dipolar Models. J. Chem. Phys. 2009; 131:125104. [PubMed: 19791922]
26. Le Brun V, Friess W, Bassarab S, Garidel P. Correlation of Protein-Protein Interactions as Assessed by Affnity Chromatography with Colloidal Protein Stability: A Case Study with Lysozyme. Pharm. Develop. Technol. 2010; 15:421–430.
27. Tozzini V. Coarse-Grained Models for Proteins. Curr. Opn. Struct. Biol. 2005; 15:144–150.
28. Riniker S, Allison JR, van Gunsteren WF. On Developing Coarse-Grained Models for Biomolecular Simulation: A Review. Phys. Chem. Chem. Phys. 2012; 14:12423–12430. [PubMed: 22678152]
29. Asthagiri D, Paliwal A, Abras D, Lenhoff AM, Paulaitis ME. A Consistent Experimental and Modeling Approach to Light-Scattering Studies of Protein-Protein Interactions in Solution. Biophys. J. 2005; 88:3300–3309. [PubMed: 15792969]
30. Paliwal A, Asthagiri D, Abras D, Lenhoff AM, Paulaitis ME. Light-Scattering Studies of Protein Solutions: Role of Hydration in Weak Protein-Protein Interactions. Biophys. J. 2005; 89:1564–1573. [PubMed: 15980182]
31. Gohlke H, Thorpey MF. A Natural Coarse Graining for Simulating Large Biomolecular Motion. Biophys. J. 2006; 91:2115–2120. [PubMed: 16815893]
32. Kim YC, Hummer G. Coarse-Grained Models for Simulations of Multiprotein Complexes: Application to Ubiquitin Binding. J. Mol. Biol. 2008; 375:1416–1433. [PubMed: 18083189]
33. Hyeon C, Thirumalai D. Capturing the Essence of Folding and Functions of Biomolecules using Coarse-Grained Models. Nat. Comm. 2011; 2:487.
34. Clementi C. Coarse-Grained Models of Protein Folding: Toy Models or Predictive Tools? Curr. Opin. Struct. Biol. 2008; 18:10–15. [PubMed: 18160277]
35. Chan HS, Zhang Z, Wallin S, Liu Z. Cooperativity, Local-Nonlocal Coupling, and Nonnative Interactions: Principles of Protein Folding from Coarse-Grained Models. Ann. Rev. Phys. Chem. 2011; 62:301–326. [PubMed: 21453060]
36. Pellarin R, Caflisch A. Interpreting the Aggregation Kinetics of Amyloid Peptides. J. Mol. Biol. 2006; 360:882–892. [PubMed: 16797587]
37. Chebaro Y, Mousseau N, Derreumaux P. Structures and Thermodynamics of Alzheimer's Amyloid-beta A beta(16-35) Monomer and Dimer by Replica Exchange Molecular Dynamics Simulations: Implication for Full-Length A beta Fibrillation. J. Phys. Chem. B. 2009; 113:7668–7675. [PubMed: 19415895]
38. Hagan MF, Dinner AR, Chandler D, Chakraborty AK. Atomistic Understanding of Kinetic Pathways for Single Base-Pair Binding and Unbinding in DNA. Proc. Natl. Acad. Sci. U.S.A. 2003; 100:13922–13927. [PubMed: 14617777]
39. Leguebe M, Nguyen C, Capece L, Hoang Z, Giorgetti A, Carloni P. Hybrid Molecular Mechanics/Coarse-Grained Simulations for Structural Prediction of G-Protein Coupled Receptor/Ligand Complexes. Plos One. 2012; 7:e47332. [PubMed: 23094046]
40. Tozzini V. Minimalist Models for Proteins: A Comparative Analysis. Quart. Rev. Biophys. 2010; 43:333–371.
41. Agrawal NJ, Kumar S, Wang X, Helk B, Singh SK, Trout BL. Aggregation in Protein-Based Biotherapeutics: Computational Studies and Tools to Identify Aggregation-Prone Regions. J. Pharm. Sci. 2011; 100:5081–5095. [PubMed: 21789769]
42. Andley UP. Crystallins in the Eye: Function and Pathology. Prog. Retin. Eye Res. 2007; 26:78–98. [PubMed: 17166758]
43. Moreau KL, King JA. Protein Misfolding and Aggregation in Cataract Disease and Prospects for Prevention. Trends Mol. Med. 2012; 18:273–282. [PubMed: 22520268]
44. Pande A, Pande J, Asherie N, Lomakin A, Ogun O, King J, Benedek GB. Crystal Cataracts: Human Genetic Cataract Caused by Protein Crystallization. Proc. Natl. Acad. Sci. U.S.A. 2001; 98:6116–6120. [PubMed: 11371638]
45. Flaugh SL, Kosinski-Collins MS, King J. Interdomain Side-Chain Interactions in Human Gamma D Crystallin Influencing Folding and Stability. Protein Sci. 2005; 14:2030–2043. [PubMed: 16046626]

46. McManus JJ, Lomakin A, Ogun O, Pande A, Basan M, Pande J, Benedek GB. Altered Phase Diagram Due to a Single Point Mutation in Human Gamma D Crystallin. Proc. Natl. Acad. Sci. U.S.A. 2007; 104:16856–16861. [PubMed: 17923670]
47. Moreau KL, King J. Hydrophobic Core Mutations Associated with Cataract Development in Mice Destabilize Human gamma D-Crystallin. J. Biol. Chem. 2009; 284:33285–33295. [PubMed: 19758984]
48. Zhang L-Y, Gong B, Tong J-P, Fan DS-P, Chiang SW-Y, Lou D, Lam DS-C, Yam GH-F, Pang C-P. A Novel Gamma D Crystallin Mutation Causes Mild Changes in Protein Properties but Leads to Congenital Coralliform Cataract. Mol. Vision. 2009; 15:1521–1529.
49. Sahin E, Jordan JL, Spatara ML, Naranjo A, Costanzo JA, Weiss WF, Robinson AS, Fernandez EJ, Roberts CJ. Computational Design and Biophysical Characterization of Aggregation-Resistant Point Mutations for Gamma D Crystallin Illustrate a Balance of Conformational Stability and Intrinsic Aggregation Propensity. Biochemistry. 2011; 50:628–639. [PubMed: 21184609]
50. Pellicane G, Costa D, Caccamo C. Microscopic Determination of the Phase Diagrams of Lysozyme and Gamma-Crystallin Solutions. J. Phys. Chem. B. 2004; 108:7538–7541.
51. Goulet DR, Knee KM, King JA. Inhibiton of Unfolding and Aggregation of Lens Protein Human Gamma D Crystallin by Sodium Citrate. Exp. Eye Res. 2011; 93:371–381. [PubMed: 21600897]
52. Kramer RM, Shende VR, Motl N, Pace CN, Scholtz JM. Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. Biophys. J. 2012; 102:1907–1915. [PubMed: 22768947]
53. Grüenberger A, Lai P-K, Blanco MA, Roberts CJ. Coarse-Grained Modeling of Protein Second Osmotic Virial Coefficients: Sterics and Short-Ranged Attractions. J. Phys. Chem. B. 2013; 117:763–770. [PubMed: 23245189]
54. Wang Z-H, Lee HC. Origin of the Native Driving Force for Protein Folding. Phys. Rev. Lett. 2000; 84:574–577. [PubMed: 11015967]
55. Fitzpatrick AW, Knowles TPJ, Waudby CA, Vendruscolo M, Dobson CM. Inversion of the Balance Between Hydrophobic and Hydrogen Bonding Interactions in Protein Folding and Aggregation. Plos Compt. Biol. 2011; 7:e1002169.
56. Thirumalai D, Reddy G, Straub JE. Role of Water in Protein Aggregation and Amyloid Polymorphism. Acc. Chem. Res. 2012; 45:83–92. [PubMed: 21761818]
57. Bereau T, Deserno M. Generic Ccoarse-Grained Model for Protein Folding and Aggregation. J. Chem. Phys. 2009; 130:235106. [PubMed: 19548767]
58. Miyazawa S, Jernigan RL. Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. J. Mol. Biol. 1996; 256:623–644. [PubMed: 8604144]
59. Pacios LF. Distinct Molecular Surfaces and Hydrophobicity of Amino Acid Residues in Proteins. J. Chem. Inf. Comp. Sci. 2001; 41:1427–1435.
60. Ben-Naim, A. Statistical Thermodynamics for Chemists and Biochemists. Plenum Press; New York, New York: 1992.
61. Singh JK, Kofke DA. Mayer Sampling: Calculation of Cluster Integrals Using Free-Energy Perturbation Methods. Phys. Rev. Lett. 2004; 92:220601. [PubMed: 15245206]
62. Benjamin KM, Singh JK, Schultz AJ, Kofke DA. Higher-Order Virial Coefficients of Water Models. J. Phys. Chem. B. 2007; 111:11463–11473. [PubMed: 17850128]
63. Sugita Y, Okamoto Y. Replica-Exchange Multicanonical Algorithm and Multicanonical Replica-Exchange Method for Simulating Systems with Rough Energy Landscape. Chem. Phys. Lett. 2000; 329:261–270.
64. Miller TF, Eleftheriou M, Pattnaik P, Ndirango A, Newns D, Martyna GJ. Symplectic Quaternion Scheme for Biophysical Molecular Dynamics. J. Chem. Phys. 2002; 116:8649–8659.
65. Ikeguchi M. Partial Rigid-Body Dynamics in NPT, NPAT and NP Gamma T Ensembles for Proteins and Membranes. J. Comput. Chem. 2004; 25:529–541. [PubMed: 14735571]
66. Ferrenberg AM, Swendsen RH. Optimized Monte-Carlo Data-Analysis. Phys. Rev. Lett. 1989; 63:1195–1198. [PubMed: 10040500]

67. Chodera JD, Swope WC, Pitera JW, Seok C, Dill KA. Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations. J. Chem. Theory Comput. 2007; 3:26–41.
68. Ogunnaike, BA. Random Phenomena: Fundamentals of Probability and Statistics for Engineers. CRC Press; Boca Raton, Florida: 2010.
69. Toofanny RD, Jonsson AL, Daggett V. A Comprehensive Multidimensional-Embedded, One-Dimensional Reaction Coordinate for Protein Unfolding/Folding. Biophys. J. 2010; 98:2671–2681. [PubMed: 20513412]
70. Anderson VL, Ramlall TF, Rospigliosi CC, Webb WW, Eliezer D. Identification of a Helical Intermediate in Trifluoroethanol-Induced Alpha-Synuclein Aggregation. Proc. Natl. Acad. Sci. U.S.A. 2010; 107:18850–18855. [PubMed: 20947801]
71. Fasolo M, Sollich P. Fractionation Effects in Phase Equilibria of Polydisperse Hard-Sphere Colloids. Phys. Rev. E. 2004; 70:041410.
72. Blanco MA, Sahin E, Li Y, Roberts CJ. Reexamining Protein-Protein and Protein-Solvent Interactions from Kirkwood-Buff Analysis of Light Scattering in Multi-Component Solutions. J. Chem. Phys. 2011; 134:225103. [PubMed: 21682538]
73. Velev OD, Kaler EW, Lenhoff AM. Protein Interactions in Solution Characterized by Light and Neutron Scattering: Comparison of Lysozyme and Chymotrypsinogen. Biophys. J. 1998; 75:2682–2697. [PubMed: 9826592]
74. Zhang Y, Cremer PS. Interactions Between Macromolecules and Ions: The Hofmeister Sseries. Curr. Opn. Chem. Biol. 2006; 10:658–663.
75. Marek PJ, Patsalo V, Green DF, Raleigh DP. Ionic Strength Effects on Amyloid Formation by Amylin Are a Complicated Interplay among Debye Screening, Ion Selectivity, and Hofmeister Effects. Biochemistry. 2012; 51:8478–8490. [PubMed: 23016872]
76. Box, GEP.; Draper, NR. Response Surfaces, Mixtures, and Ridge Analyses. 2nd ed.. Wiley-Interscience; Hoboken, New Jersey: 2007.
77. Siderius DW, Krekelberg WP, Roberts CJ, Shen VK. Osmotic Virial Coefficients for Model Protein and Colloidal Solutions: Importance of Ensemble Constraints in the Analysis of Light Scattering Data. J. Chem. Phys. 2012; 136:175102. [PubMed: 22583267]
78. Sandler, SI. Chemical, Biochemical, and Engineering Thermodynamics. John Wiley & Sons, Inc.; Hoboken, New Jersey: 2006.
79. Blanco MA, Perevozchikova T, Martorana V, Manno M, Roberts CJ. Protein– Protein Interactions in Dilute to Concentrated Solutions: 1. α-Chymotrypsinogen in Acidic Conditions. 2013 Unpublished work.
80. McQuarrie, DA. Statistical Mechanics. University Science Books; Sausalito, California: 2000.
81. Wales DJ, Bogdan TV. Potential Energy and Free Energy Landscapes. J. Phys. Chem. B. 2006; 110:20765–20776. [PubMed: 17048885]
82. Conchillo-Sole O, de Groot NS, Aviles FX, Vendrell J, Daura X, Ventura S. AGGRESCAN: A Server for the Prediction and Evaluation of “Hot Spots” of Aggregation in Polypeptides. BMC Bioinformatics. 2007:8. [PubMed: 17212835]
83. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L. Prediction of Sequence-Dependent and Mutational Effects on the Aggregation of Peptides and Proteins. Nat. Biotech. 2004; 22:1302–1306.
84. Trovato A, Chiti F, Maritan A, Seno F. Insight Into the Structure of Amyloid Fibrils from the Analysis of Globular Proteins. PLOS Comput. Biol. 2006; 2:1608–1618.

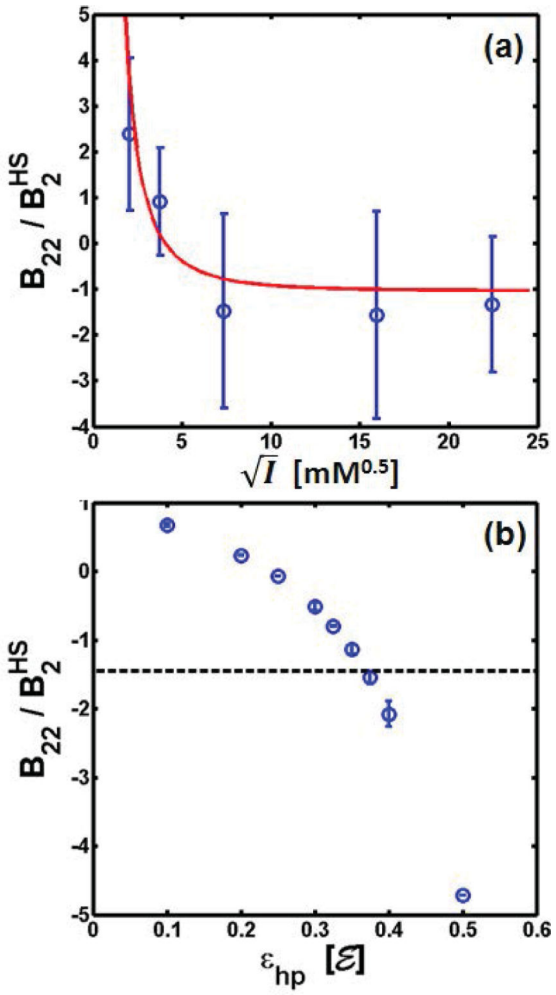

Figure 1.

Comparison of experimental and calculated B_{22} values for γD-crystallin at T = 300 K: (a) Theoretical (solid line) and experimental (symbols) B_{22} as a function of the square root of the ionic strength. Theoretical B_{22} curve is calculated for a value of $\varepsilon_{hp} = 0.375\mathscr{E}$ and $\varepsilon_{cc} = 0.125\mathscr{E}$. κ is assumed proportional to $\sqrt{I}$, with a proportionality constant a = 3.5 ± 0.5 $M^{-1/2}nm^{-1}$ obtained from fitting the resulting response surface of B_{22} to the experimental data. See main text for additional details. (b) B_{22} calculated via Mayer Sampling (symbols) as a function of ε_{hp} at high ionic strength (i.e., $\kappa \to 0$, and/or $\varepsilon_{cc} = 0$). The horizontal dashed line corresponds to the experimental result determined via light scattering for an ionic strength of 504.2 mM. Values of B_{22} are relative to the hard-sphere second virial coefficient B_2^{HS} for a protein diameter of 4 nm.

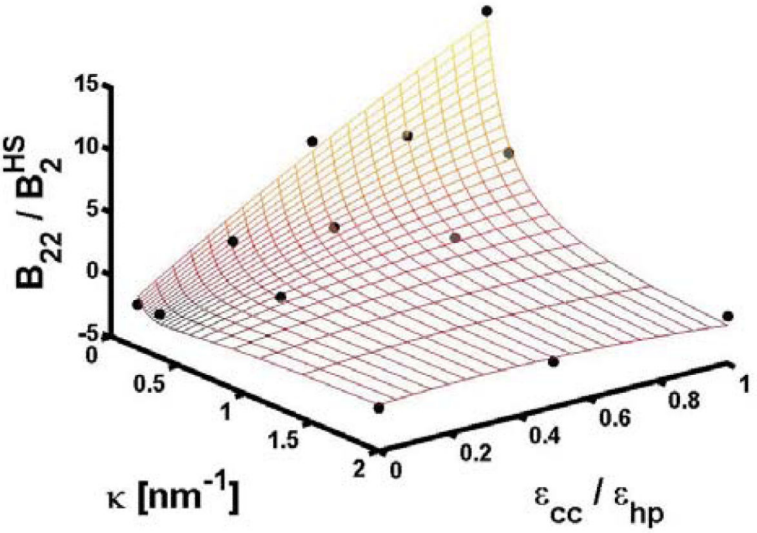

Figure 2.

Response surface for B_{22} as function of screening length parameter κ and the electrostatic interactions parameter ε_{cc}. Symbols correspond to the values calculated via Mayer sampling for a value of ε_{hp} = 0.375 $\mathscr{E}$. The response surface is obtained by fitting the calculated B_{22} values to a full-quadratic polynomial in $1/\kappa$ and $\varepsilon_{cc}/\varepsilon_{hp}$ -i.e., including cross-terms. Details about the response surface and its fitted parameters are provided in the Supporting Information.

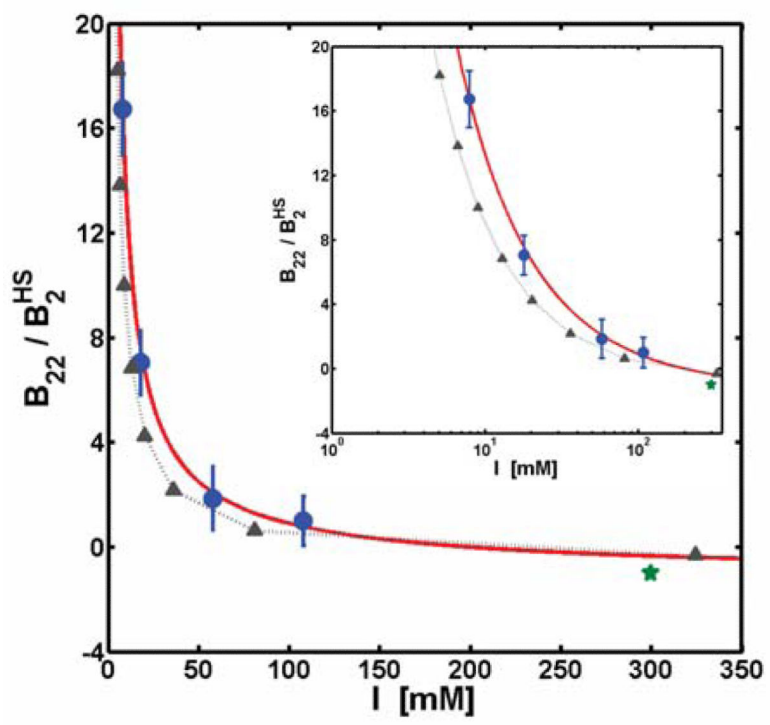

Figure 3.

Comparison of theoretical and experimental osmotic second virial coefficient for α-chymotrypsinogen as a function of ionic strength I at T = 300 K. Theoretical B_{22} curves correspond to: (solid line) B_{22} response surface and (triangles) Mayer Sampling simulations. The response surface was calculated as a function of κ and $\varepsilon_{hp}/\varepsilon_{cc}$ for $\varepsilon_{hp}/\varepsilon_{cc}$ = 3, ε_{hp} = 4.08 $\mathscr{E}$, and $\kappa=\alpha_{aCgn}\sqrt{I}$, with α_{aCgn} = 3.08 ± 0.1 $M^{-1/2}nm^{-1}$ obtained from fitting the response surface to the experimental data (see Supporting Information). B_{22} values from Mayer Sampling were calculated using the same values of the adjustable parameters obtained for gD-Crys (i.e., ε_{hp} = 0.375$\mathscr{E}$, ε_{cc} = 0.125$\mathscr{E}$, and α = 3.5 $M^{-1/2}nm^{-1}$). Other symbols correspond to experimental B_{22} values published elsewhere for: (circles) pH=3.5 in 10mM citrate buffer and NaCl concentration between 0 and 100 mM;[79] and (stars) pH=3 and I = 300 mM.[73] Although the experimental value of B_{22} at I = 300 mM was measured at pH=3, Velev et al.[73] showed that B_{22} is pH-independent for aCgn at that ionic strength. Inset provides the same data with I in a logarithmic scale.

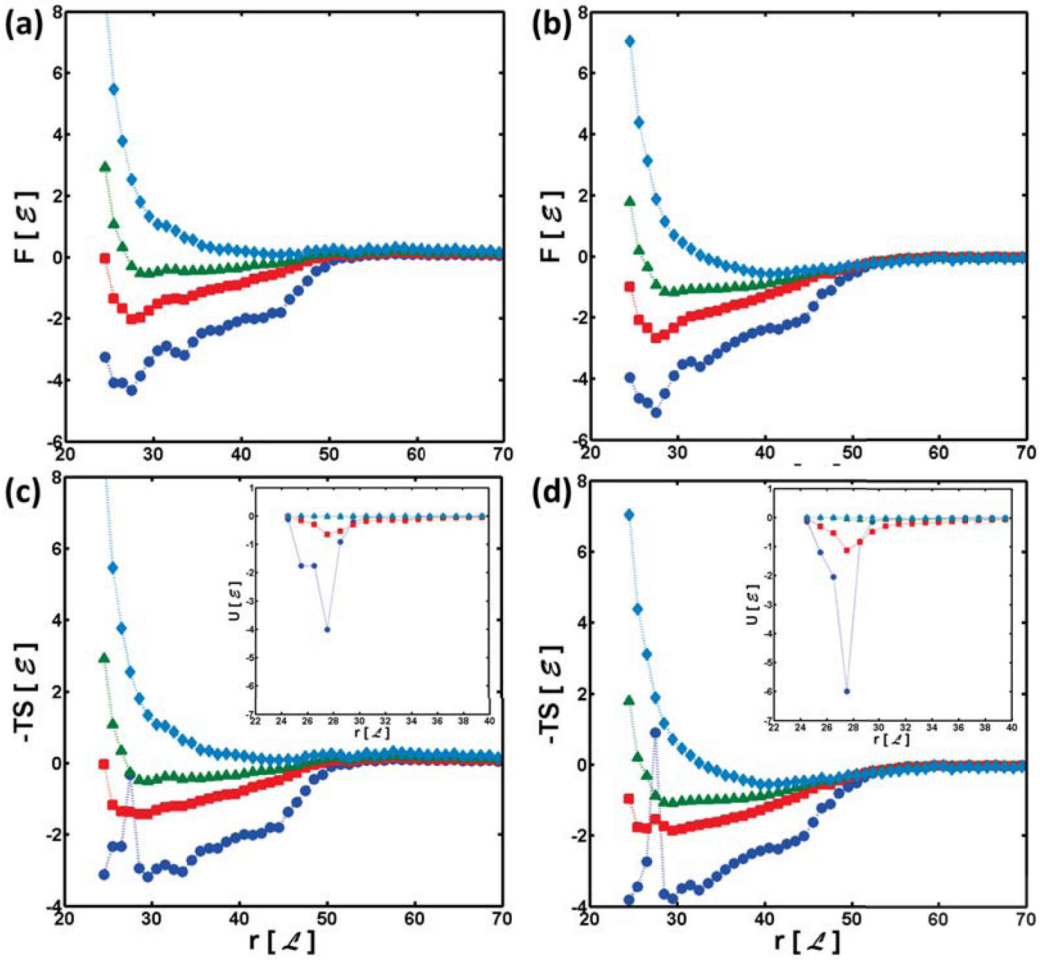

Figure 4.

Potential of mean force *F* (panel **a** and **b**), entropy *S* (panel **c** and **d**), and average interaction energy *U* (insets) curves as a function of the protein center-to-center distance *r*. Curves are shown at *T* = 100 (circles), 150 (squares), 200 (triangles), and 300 K (diamonds) for two different working cases: (first column) low ionic strength ($1/\kappa = 40$ Å); and (second column) high ionic strength ($1/\kappa = 0$ Å). 95% confidence intervals for all the values shown here are smaller than 0.05 ℰ, and are not visible at the scale of the plots.

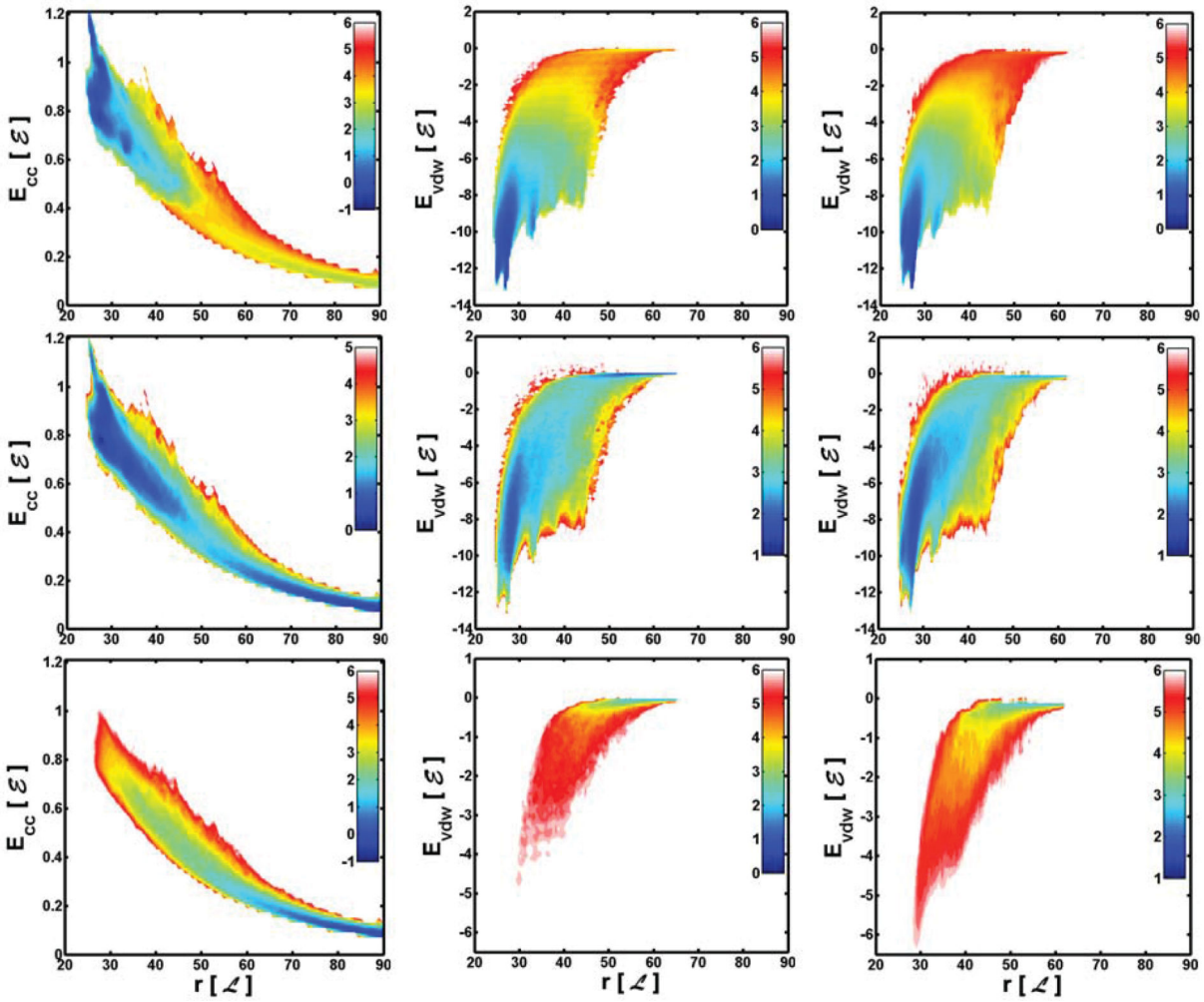

Figure 5.

Two-dimensional free energy surfaces as a function of the electrostatic (E_{cc}) or van der Waals (E_{vdw}) energies and the center-to-center distance r for γD-crystallin at a protein concentration of 10 mg/mL and T = 100 (top row), 150 (middle row), and 300 K (bottom row). First and second columns correspond to simulation at large screening length ($1/\kappa = 40$ $\mathscr{L}$); third column provides the free energy surface for the case when electrostatic interactions are completely screened ($1/\kappa = 0$). Colors indicates free-energy values in units of $\mathscr{E}$, and its scale is shown next to each panel. With the exception of the left-most panel in the middle row, all color scales are the same for each panel.

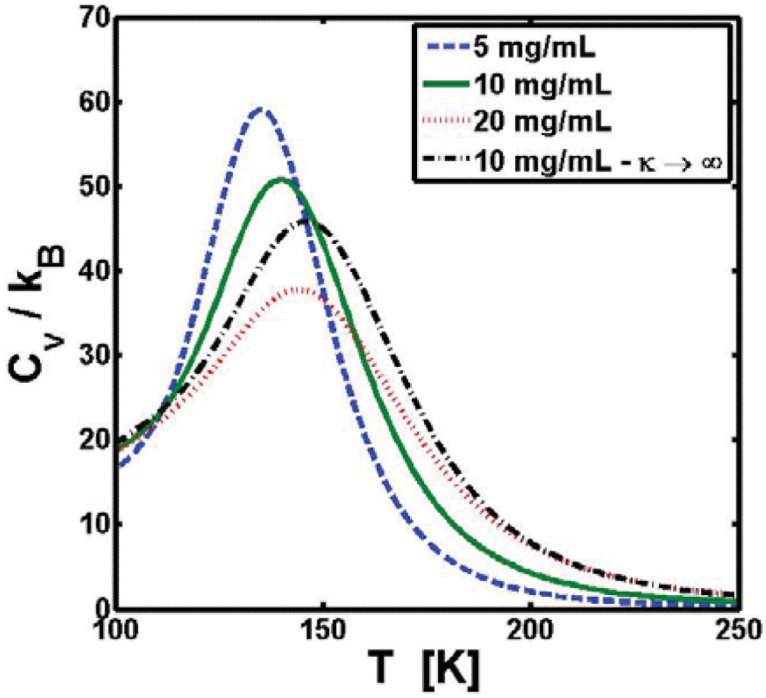

Figure 6.

Heat capacity C_v vs. T for all the simulations tested here: (dashed line) c = 5 mg/mL and $1/\kappa$ = 40 $\mathscr{L}$; (solid line) c = 10 mg/mL and $1/\kappa$ = 40 $\mathscr{L}$; (dotted line) c = 20 mg/mL and $1/\kappa$ = 40 $\mathscr{L}$; and (dash-dot line) c = 10 mg/mL and $1/\kappa$ = 0 $\mathscr{L}$.

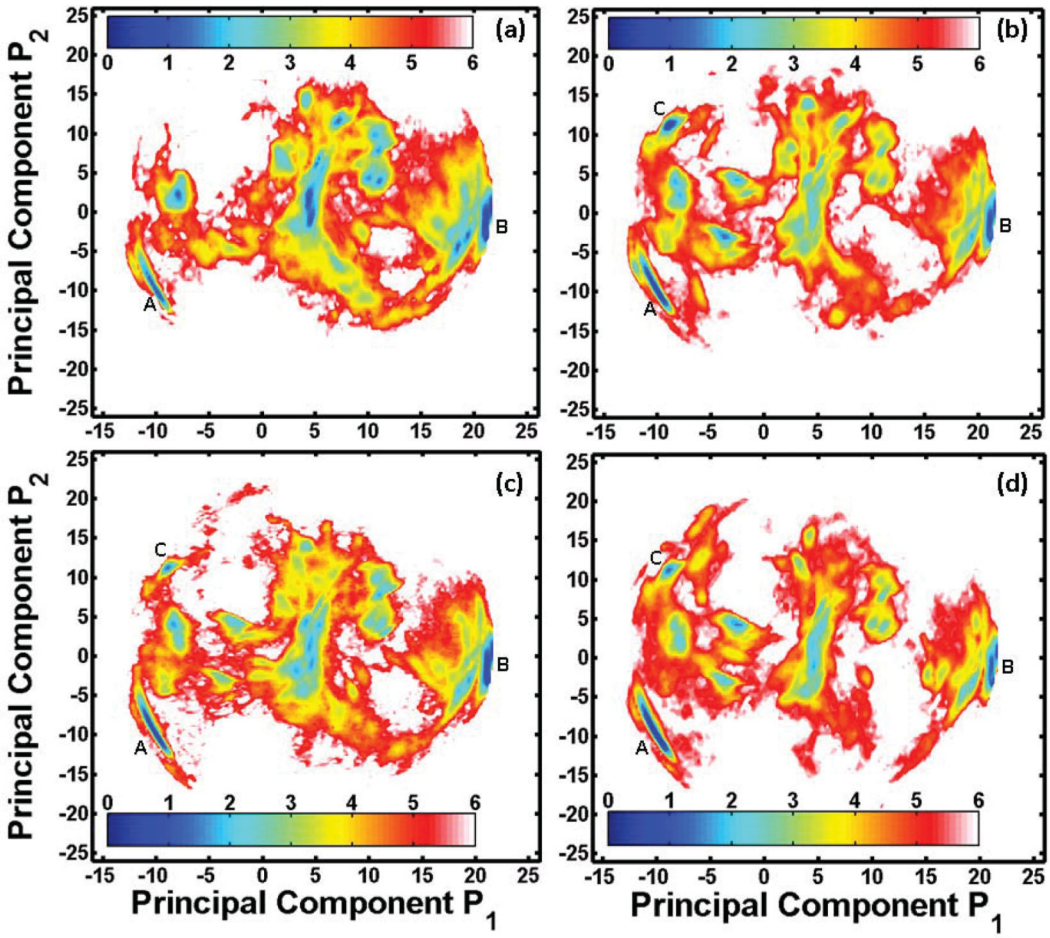

Figure 7.

Two-dimensional contour maps for the negative base-10-logarithm of the probability distribution function of the number of configurations as a function of the principal components P_1 and P_2 for γD-crystallin at T = 80 K. Each of the panels corresponds to one of the working cases: (a) c = 5mg/mL and $1/\kappa = 40\mathscr{L}$; (b) c = 10mg/mL and $1/\kappa = 40\mathscr{L}$; (c) c = 20mg/mL and $1/\kappa = 40\mathscr{L}$; and (d) c = 10mg/mL and $1/\kappa = 0\mathscr{L}$. Colors indicate the order of magnitude of the probability distribution function. The selected configurations for further analysis (**A**, **B**, and **C**) are labeled in each of the panels.

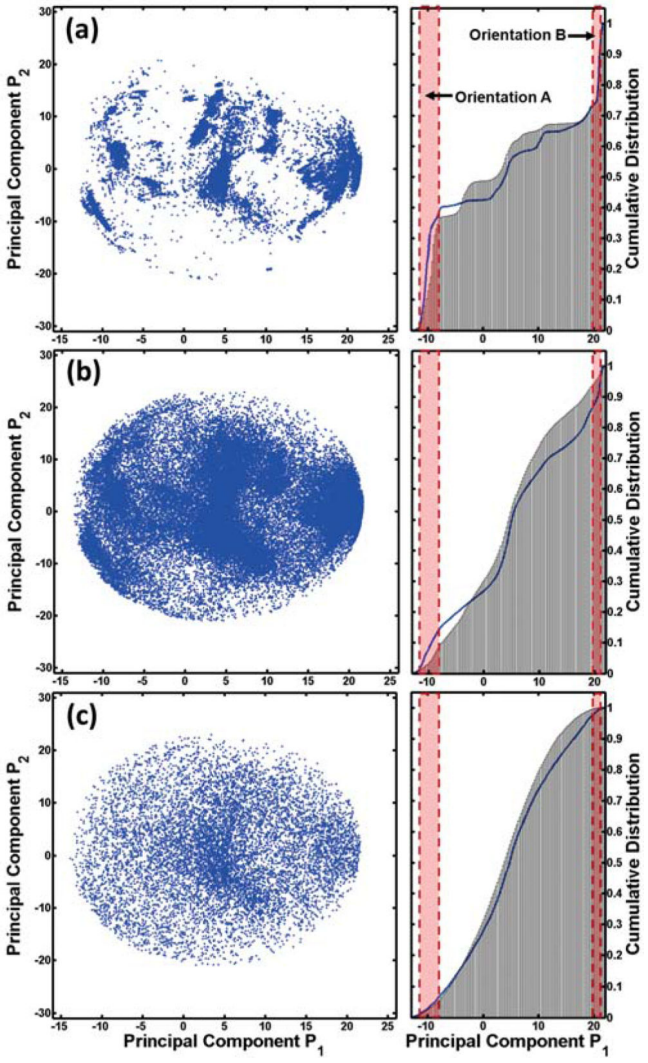

Figure 8.

P_1 *versus* P_2 for the relevant configurations for the calculation of B_{22} for γD-crystallin at $1/\kappa = 40$ $\mathscr{L}$ and $T = 100$ (a), 150 (b), and 250 K (c). Rightmost panels next to each of (a), (b) and (c) compare the cumulative distribution function of P_1 obtained from the B_{22} calculations (gray vertical bars) with that obtained from applying WHAM to the REMD simulation for $c = 10$ mg/mL and $1/\kappa = 40$ $\mathscr{L}$ at the same temperatures (solid blue curve). The range of P_1 values corresponding to the two most populated configurations (orientations **A** and **B**) are also highlighted with red shaded regions on the cumulative distribution function in the righthand panels.

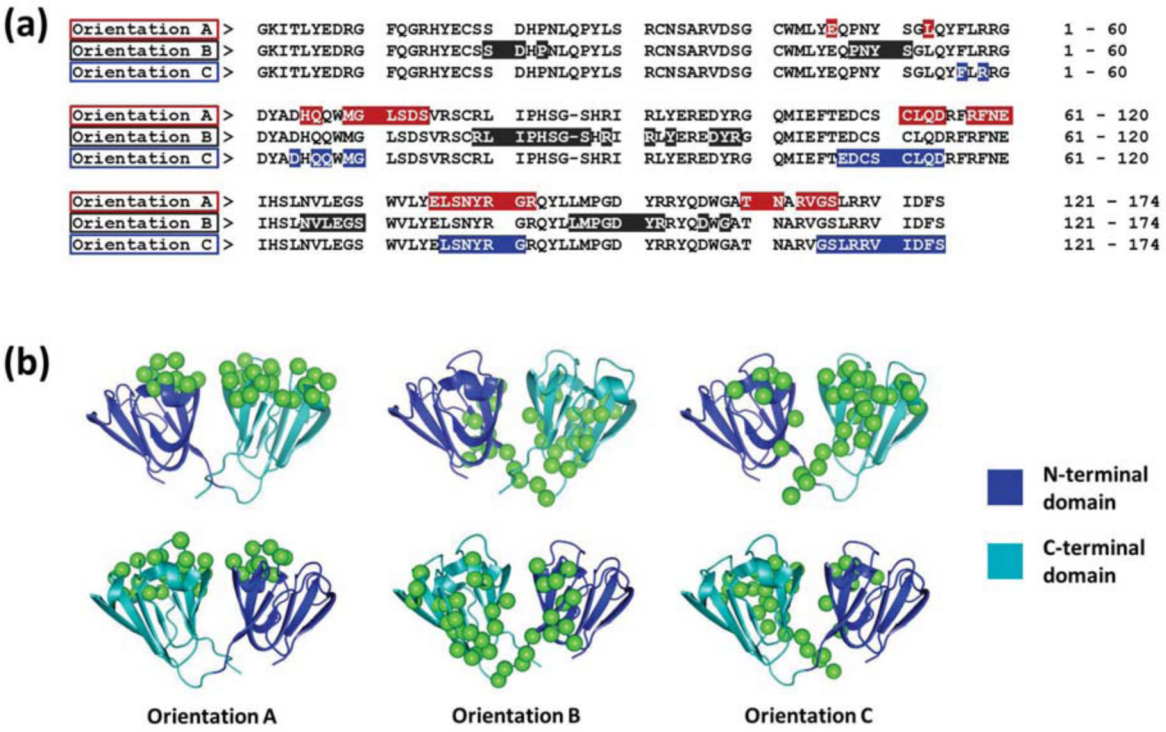

Figure 9.

Self-association-prone sites (i.e., "hot-spot" residues) on γD-crystallin found from REMD simulations for the selected protein relative orientations: **A**, **B**, and **C**. Identity of the "hot-spot" residues are shown in panel (a), while panel (b) provides an schematic representation of these sites along the protein crystal structure. Note that the identified "hot-spot" residues agree with those obtained in a previous work[49] based on a consensus among different aggregation predictors. Self-association-prone sites are shown as green beads in the figure. An alternative representation of the same information contained in panels (a) and (b) is provided in the form of illustrative snapshots of the low-temperature structures for the dimeric configurations **A**, **B**, and **C** in Figure S10 in the Supporting Information.

Table 1

Van der Waals diameter[59] (σ_i) in units of Å, and relative hydrophobic scores[57] (ε_i) for each type of natural occurring amino acid.

Residue	σ_i	ε_i	Residue	σ_i	ε_i
Lys	7.03	0.00	His	6.29	0.25
Glu	6.40	0.05	Ala	5.02	0.26
Asp	5.83	0.06	Tyr	7.11	0.49
Asn	5.95	0.10	Cys	4.92	0.54
Ser	5.28	0.11	Trp	6.70	0.64
Arg	7.32	0.13	Val	6.05	0.65
Gln	6.35	0.13	Met	6.32	0.67
Pro	5.62	0.14	Ile	6.36	0.84
Thr	5.81	0.16	Phe	6.95	0.97
Gly	4.31	0.17	Leu	6.55	1.00