UNIVERSITY OF CALIFORNIA,
IRVINE

Development of multiple conformation Monte Carlo method and its application to protein aggregation in cataract formation

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Chemistry

by

Vera Dmitrievna Prytkova

Dissertation Committee:
Professor Douglas Tobias, Chair
Professor Rachel Martin
Professor Ioan Andricioaei

2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to acknowledge and thank the following important people who have supported me during my PhD program. First of all, my graduate advisor, professor Douglas J. Tobias, for his guidance in my academic research. I would also like to thank Dr. Alfredo J. Freites for many hours spent explaining concepts and technical details of anything I would struggle with. Many thanks to all the people I worked with – Dr. Eric Wong, Professor Matthias Heyden, Dr. Domarin Khago, and Professor Rachel W. Martin for all the contributions they had to the exciting scientific story that has developed.

CURRICULUM VITAE

Vera Dmitrievna Prytkova

Sep 2006 – Jun 2011	B.A. in Biochemical Physics, Saratov State University named after N.G. Chernyshevsky
Jun 2008 – Sep 2008	Researcher at Northwestern University in the group of Prof. George Schatz
Jun 2009 – Sep 2009	
Jun 2010 – Sep 2010	
Dec 2010	Researcher at Instituto di Fisica Applicata Consiglio Nazionale delle Ricerche (IFAC-CNR), Italy, in the group of Roberto Pini
Sep 2009 – Mar 2011	Researcher at Saratov State University named after N.G. Chernyshevsky, in the group of Prof. Valery Tuchin
Mar 2011 – Jun 2011	Researcher at UC Irvine, in the group of Prof. Vartkess Apkarian, Chemistry and Space-Time Limit Center
Sep 2011 – Jun 2016	Teaching Assistant at UC Irvine
Jun 2017	M.S. in Chemistry, UC Irvine
Mar 2018	PhD in Chemistry, UC Irvine

PUBLICATIONS

Cataract-associated W42R gamma D-crystallin pathway for aggregation
Vera Prytkova, Eric Wong, J. Alfredo Freites, Douglas J. Tobias
Manuscript in preparation

Influence of conformational flexibility on protein-protein interactions
Bibhab Bandhu Majumdar, Vera Prytkova, J. Alfredo Freites, Matthias Heyden, Douglas J. Tobias
Manuscript in preparation

Multi-conformation Monte Carlo: a method for introducing flexibility in efficient simulations of many-protein systems
Vera Prytkova, Matthias Heyden, Domarin Khago, J. Alfredo Freites, Carter T. Butts, Rachel W. Martin, Douglas J. Tobias
Journal of Physical Chemistry B, DOI: 10.1021/acs.jpcb.6b00827

Introducing Molecular Flexibility in Efficient Simulations of Many-Protein Systems Biophysical Society Meeting, Baltimore, MD
Vera Prytkova, Matthias Heyden, Douglas Tobias
Biophysical Journal 108 (2), 470a, 2015

The calculations of electromagnetic fields around nanoparticles
OPTO Meeting for Young Researchers & Fifth International SPIE Students' Chapter Meeting, Torun, Poland, 2010
Vera Prytkova, Valery Tuchin

The calculations of electromagnetic fields around nanoparticles embedded in biological media

SPIE Meeting in San Diego, CA
Vera Prytkova, Valery Tuchin
Plasmonics: Metallic Nanostructures and Their Optical Properties VIII, 2010

The calculations of electromagnetic fields around nanoparticles embedded in biological media
Seventh Framework Program Photonics4Life, Barcelona, Spain, 2009

Discrete dipole approximation as a method of studying electromagnetic fields around nanostructures with surface roughness
Saratov Fall Meeting, Saratov, Russia, 2009
Vera Prytkova, Valery Tuchin

ABSTRACT OF THE DISSERTATION

Development of multiple conformation Monte Carlo method and its application to protein aggregation in cataract formation

By

Vera Dmitrievna Prytkova

Doctor of Philosophy in Chemistry

University of California, Irvine, 2018

Professor Douglas Tobias, Chair

Realistic biological conditions are characterized by high concentrations of biomolecular solutes. Protein conformations and protein-protein interactions can be affected by crowding. The inclusion of a high number of proteins to model such environments necessitates the use of computationally inexpensive methods, such as rigid-body Brownian dynamics or Monte Carlo (MC) simulations. However, the rigid body representation of many protein systems gives rise to artifacts in protein-protein interactions. Presented here is the multi-conformational Monte Carlo (mcMC) method that avoids such artifacts by incorporating molecular flexibility at a low computational cost. We employ it to study the interaction of eye lens proteins, crystallins. In a healthy eye lens crystallins are the structure proteins. Their solubility at concentrations exceeding 400 mg/mL ensures lens transparency. The aggregation of crystallins leads to lens opacification, called cataract. We study how known point mutations associated with cataract formation lead to altered protein-protein interaction and creation of large aggregates.

INTRODUCTION

The living systems are characterized by highly concentrated solutions consisting of various macromolecules in solvent, such as the cytoplasm and the extracellular matrix.(*1*) The high concentration affects the process of protein folding, biochemical reactions, protein-protein interactions.(*2*) Fully atomistic simulations of highly concentrated biological environments in explicit solvent remain extremely computationally demanding. However, inexpensive computational algorithms have been developed to study such environments. We employ one such algorithm, Monte Carlo method.

Theory of Monte Carlo Simulations

In equilibrium statistical mechanics the integrals that must be evaluated generally take the form(*3*)

$$I = \int dx\, \varphi(x) f(x) \qquad (1)$$

where x is an n-dimensional vector, $\varphi(x)$ is an arbitrary function, $f(x)$ is a probability distribution function. This integral represents the ensemble average of a physical observable. Central limit theorem tells us, that if we take a set of M n-dimensional vectors $x_1, x_2, \dots, x_M$ that are sampled from $f(x)$ and approximate the average:

$$\widetilde{I_M} = \frac{1}{M} \sum_{i=1}^{M} \varphi(x_i) \qquad (2)$$

Then it can be considered to be an estimator of the integral I considering M is large.

$$\lim_{M \to \infty} \widetilde{I_M} = I \qquad (3)$$

For Monte Carlo Simulations the vectors $x_1, x_2, \dots, x_M$ are generated sequentially $x_1 \to x_2 \to \dots \to x_M$ with a rule that specifies how to generate x_{i+1} given x_i. This sequence of vectors is called a Markov chain. If x and y are accessible microstates of a physical system, and $R(x|y)$ is the probability for moving to a state x given that the system is currently in the state y. For this probability to be a valid rule for a Markov chain, it must satisfy the detailed balance:

$$R(x|y)f(y) = R(y|x)f(x) \tag{4}$$

Detailed balance ensures microscopic reversibility of the process and guarantees unbiased sampling of the state space.

Metropolis et al. have proposed the technique in 1953, which they called M(RT)^2 algorithm (it stands for the first letters of the authors' names)(*4*). The algorithm starts with a rule for generating a trial move from y to x. We call it $T(x|y)$ and it is normalized

$$\int dx\, T(x|y) = 1 \tag{5}$$

When the trial move is generated, it can be either accepted or rejected. If the move is accepted, the system moves to state x, otherwise it returns to state y. $A(x|y)$ is the probability that a move is accepted. The transition probability is

$$R(x|y) = A(x|y)T(x|y) \tag{6}$$

Considering the detailed balance,

$$A(x|y)T(x|y)f(y) = A(y|x)T(y|x)f(x) \tag{7}$$

We can find the relation of the acceptance probabilities

$$A(x|y) = \frac{T(y|x)f(x)}{T(x|y)f(y)}A(y|x) = r(x|y)A(y|x) \tag{8}$$

If the move from y to x is favorable, then $A(x|y) = 1$, then the reverse move from x to y is less favorable and $A(x|y) < 1$, implying that $r(x|y) > 1$. Then

$$A(x|y) = \min\,[1, r(x|y)] \tag{9}$$

In the case of the canonical ensemble that we employ in our simulations, for a system of N particles with coordinates $\boldsymbol{r}_1, \boldsymbol{r}_2, \dots, \boldsymbol{r}_N$ the acceptance probability from configuration $\boldsymbol{r} \equiv \boldsymbol{r_1}, \boldsymbol{r}_2, \dots, \boldsymbol{r}_N$ to $\boldsymbol{r}'$ becomes

$$A(\boldsymbol{r}' |\, \boldsymbol{r}) = \min\,[1, e^{-\beta[U(\boldsymbol{r}') - U(\boldsymbol{r})]}] \tag{10}$$

Where $U(\boldsymbol{r})$ is the potential energy, and $\beta = 1/k_B T$. Therefore, the acceptance probability is determined by the change in the potential energy that results from the move.

In our simulations, the trial moves are attempted either to displace a protein or to rotate it. When this method is used to model rigid proteins in implicit solvent, artifacts arise due to the lack of protein flexibility. This can be avoided by introducing flexibility by a MC move, where in addition to translation or rotation a protein attempts to swap the conformation from a library of rigid conformations. Each protein conformation has a weight assigned to it.

Several questions can be asked to such a scheme. The first question – does this swap move satisfy the detailed balance condition?

Let's reinstate:

$$R(x|y) = T(x|y) \min\left[1, e^{-\frac{U_y - U_x}{kT}}\right] \tag{11}$$

In our case we want to include extra information on the relative stability of the states y and x via the a priori weighting factors. Hence, we intend to change the equilibrium properties on the system, because we include additional free energy information. In mcMC, this describes the relative internal free energy of the distinct conformations. Expressed for two states y and x, with intrinsic free energies G_y^{int} and G_x^{int}, we obtain:

$$e^{[-(G_y^{int} - G_x^{int})/kT]} = \frac{w_y}{w_x} \tag{12}$$

Effectively, we want to modify the relative stability of the two states y and x to obtain as a result:

$$f^{*}(y) = w_{y} f(y) \tag{13}$$

$$f^{*}(x) = w_{x} f(x) \tag{14}$$

In the algorithm, this is achieved by modifying the a priori probabilities $T(x|y)$ and $T(y|x)$, which now become:

$$T^{*}(y|x) = w_{y} T(y|x) \tag{15}$$

$$T^{*}(x|y) = w_{x} T(x|y) \tag{16}$$

Or alternatively

$$R^{*}(x|y) = w_{x} T(x|y) \min\,[1, e^{\left[-\frac{U_{y}-U_{x}}{kT}\right]}] \tag{17}$$

$$R^{*}(y|x) = w_{y} T(y|x) \min\,[1, e^{\left[-\frac{U_{x}-U_{y}}{kT}\right]}] \tag{18}$$

We simply modify the chance of attempting a trial move that generates y or x state, irrespective of the state the system is currently in. The detailed balance condition now becomes:

$$f^{*}(y) R^{*}(y|x) = f^{*}(x) R^{*}(x|y) \tag{19}$$

Now that we put equations together, we get:

$$w_{y} w_{x} f(y) R(y|x) = w_{x} w_{y} f(x) R(x|y) \tag{20}$$

This equation is exactly equivalent to the detailed balance condition stated earlier. Therefore, detailed balance is not affected by the modified a priori probabilities, only the equilibrium between the states y and x is modified as intended.

$$\frac{f^{*}(y)}{f^{*}(x)} = \frac{w_{y} f(y)}{w_{x} f(x)} = e^{[-\frac{G_{y}^{int}-G_{x}^{int}}{kT}]} \frac{f(y)}{f(x)} \tag{21}$$

Another question of interest is – what is the size of the conformational library necessary to represent the flexibility of a protein? The answer to this question depends on the conformational flexibility of the protein. Crystallins, the eye lens proteins, that we will further consider, do not significantly change their conformation under physiological conditions – at high concentrations and neutral pH. A fairly small number of conformations will be necessary to represent a crystallin. However, proteins like intrinsically disordered proteins may have so many conformations, that their representation will require a very large conformational library. Another important parameter is aggregation propensity. The conformations are chosen with respect to the orientations of protein side chains, since they are the ones primarily responsible for protein-protein interactions. Since the conformations we were able to obtain were typically generated from Molecular Dynamics simulations of proteins at infinite dilution or a protein dimer at a higher concentration, but with limited protein-protein sampling capability, simulation of high concentration solutions is no easy task. The conformational library at high protein concentration may be different from that at lower concentration, especially if proteins interact a lot. Using conformation library from dilute MD simulations is not ideal for high concentration mcMC simulations, but it is a lot better than using one protein conformation.

We use mcMC to explore eye lens proteins, crystallins, and mutations that are known to be associated with cataract. Our hypothesis states – mutations, that lead to changes in protein-protein or protein-solvent interactions, cause proteins to aggregate without significant conformational change. The reason for this assumption is that neither MD simulations, nor experimental data at physiological conditions – high protein concentrations and neutral pH – have observed significant protein unfolding. This fact makes mcMC a great method to study potential aggregate formation in such a system due to a relatively small size of a necessary conformational library necessary to represent each protein.

Eye lens proteins – crystallins

The primary function of the eye lens is to focus light on the retina. The eye lens is composed of bundled fiber cells that loose their nuclei, ribosomes, and organelles during embryonic development. These cells contain structural proteins, called crystallins, at very high concentrations. The liquid-like and short-range order of the crystallins lead to the transparency and refractive properties required for the proper lens function. When crystallins begin to aggregate, either due to inherent or acquired mutations or due to age-related changes such as oxidation, deamidation, truncation, glycation, or methylation, the eye lens becomes cloudy and the cataract forms.(*5*)

There are two classes of crystallins: the βγ-crystallins are purely structural, while the α-crystallins have a chaperone function of binding to misfolded proteins to keep them in solution.(*5*)

γD crystallin is the third most abundant of the γ-crystallins. It is expressed primarily in the central lens nucleus. γS crystallin is expressed postnatally and is located preferentially in the lens cortex (periphery). γ-crystallins are compact, ~21 kDa, primarily β-sheet proteins consisting of two double Greek key domains. The two domains are not identical, but highly symmetric.(*6*) High resolution structures have been determined for wild type γD-crystallin and its congenital cataract-related R58H and P23T variants by x-ray crystallography and solution NMR.(*7-9*) Physicochemical and biophysical studies of several cataract-related γD-crystallin mutants have shown that, although certain mutations produce negligible or minor structural changes, they have profound effects on interprotein interactions. For example, the alteration in surface charge in the R58H and R26S mutants reduces solubility and promotes the formation of protein crystals, while the change in hydrophobicity associated with the P23T mutation produces non-crystalline aggregates with low and retrograde solubility.(*10*) A 1.7 Å resolution x-ray crystal structure of the congenital cataract-related W42R mutant of γD-crystallin has been solved.(*11*) The crystal structure reveals no significant difference between the mutant and native structure. Our MD simulations show fluctuation of domain positions with respect to one another without separate domain unfolding, but with an exposure of hydrophobic residues positioned at the interdomain interface that remained buried in the wild type γD crystallin. These residues lead to increased protein-protein interactions and large amorphous in shape aggregates.

Biophysical studies have shown that the βγ-crystallins exhibit attractive interactions despite their very low aggregation propensity(*12*), which implies a delicate balancing act between attractive and repulsive interactions among the crystallins. The solution state NMR structures of γS-crystallin and its childhood-onset cataract-related variant G18V have been solved.(*13*) G18V variant is structurally very similar to the wild type γS crystallin, but has a lower thermal stability and solubility.(*14, 15*) It exhibits strong interaction with αB-crystallin.(*13*) Despite the abundance of information on aggregation propensity and reduced solubility of the variant, the molecular mechanism of aggregation remains unknown.

Chapter 1 of this dissertation describes the introduction of conformational flexibility as a move in Monte Carlo simulations. Chapter 2 explores the question of the conformational library size and the amount of sampling necessary to generate the library. This chapter is focused on Monte Carlo method development, however simulations of γB and γD crystallins are used. Chapter 3 addresses the question of weights adjustment for a high protein concentration. Chapters 4 and 5 show how cataract-related mutations in γD and γS crystallins lead to local structural changes with subsequent enhancement in protein-protein interaction. Chapter 5 shows that new MD simulations of G18V variant started from a solution state NMR structure reveal a small structural change that may be responsible for increased protein-protein interaction.

CHAPTER 1

Multi-conformation Monte Carlo: A Method for Introducing Flexibility in Efficient Simulations of Many-protein Systems

This chapter was published by V. D. Prytkova, B. M. Heyden, D. Khago, J. A. Freites, C. Butts, R. W. Martin, D. J. Tobias, Multi-Conformational Monte Carlo: A Method for Introducing Flexibility in Efficient Simulations of Many-Protein Systems, J.Phys.Chem. B, 120 (33), 2016

INTRODUCTION

Biological environments, such as the cytoplasm and the extracellular matrix, are characterized by high concentrations of proteins and other biomacromolecular solutes (e.g., 300-400 mg/mL in the cytoplasm of *Escherichia coli*(*1*)). Under such crowded conditions, intermolecular interactions cannot be neglected and have a significant influence on the stability of folded proteins as well as their dynamics and aggregation propensities.(*16*)

To model protein-protein interactions and protein aggregation using computer simulations, multiple protein molecules must be included in the simulation system. Fully atomistic simulations including explicit representations of the aqueous solvent are currently only feasible for systems containing a limited number of biomolecular solutes (on the order of 10) and 100 ns timescales.(*17*) Brownian dynamics (BD) simulations employing an implicit representation of the solvent have emerged as a powerful approach to modeling many-protein systems on significantly longer timescales.(*18-21*) In BD simulations protein molecules are usually modeled as rigid bodies and their translational and rotational motions are generated with picosecond timesteps using the Ermak-McCammon algorithm.(*22*) Fast potential and force calculations are achieved through the use of pre-evaluated, constant potential terms on space-filling grids. This approach allows the simulation of solutions containing ~1,000 atomically detailed protein molecules for ~10-100 μs, which is long enough to obtain converged structural and thermodynamic properties for concentrated protein solutions in which the proteins are not aggregating strongly.(*18-20*) In addition to providing structural and thermodynamic information, BD simulations have been used to investigate the effects of crowding on diffusion in protein solutions and a model of the *E. coli* cytoplasm.(*19, 23*)

Monte Carlo (MC) simulations are potentially an attractive alternative approach for the modeling of aggregating systems,(*24, 25*) or to generally improve the configurational sampling efficiency when the sampling of explicit dynamics is not the primary goal. MC simulations based on highly coarse-grained colloidal sphere protein models have been employed to investigate phase behavior in protein solutions and protein crystallization.(*26-30*) MC simulations of more detailed protein models with residue level coarse-graining have been used to study the effects of solution conditions and ion binding on protein-protein interactions,(*31-33*) as well as protein self-assembly.(*34*) As we will show below, Metropolis MC simulations(*4*) of atomically detailed proteins can be used to investigate the structural and thermodynamic properties of crowded protein solutions with at least as good sampling efficiency as BD simulations. For more efficient sampling of strongly aggregating systems, MC simulations offer the possibility of specialized trial moves designed to expedite the formation and destruction of clusters, such as in the aggregation-volume-bias MC method pioneered by Siepmann, Chen, and co-workers.(*24, 25, 35*)

Here, we evaluate the feasibility of performing MC simulations using the protein-protein interaction potential developed by Wade and co-workers(*20*) for BD simulations of many-protein systems. We use new experimental static light scattering (SLS) data for the optimization of the interaction potential parameters in simulations of solutions of hen egg white lysozyme (HEWL). The optimized parameters were validated by comparing structure factors computed from the simulations to those derived from small-angle x-ray and neutron scattering measurements. In conventional rigid-body MC simulations using a single protein configuration, we find a strong dependence of both structural and thermodynamic properties on the specifics of the protein conformation. These results highlight the importance of incorporating protein conformational flexibility in the simulations. We have, therefore, implemented a new technique, which we refer to as multi-conformation Monte Carlo (mcMC); mcMC incorporates conformational flexibility by swapping protein conformations within a discrete library determined by clustering of protein configurations from an atomistic MD simulation of a single protein in explicit solvent. The approach is similar in spirit to the use of pre-evaluated libraries of molecular fragment conformations in configurational-bias Monte Carlo simulations.(*36-39*) However, in mcMC simulations sampling of intramolecular degrees of freedom is restricted to a set of discrete conformations, which allows the use of pre-evaluated potential grids for highly efficient energy calculations.(*18-20*) The HEWL solution simulations with mcMC show better agreement with experimental data compared to the results of scMC simulations using a single protein configuration, and eliminate the bias imposed by the use of a single structure.

METHODS

Protein-protein interaction potential. The overall protein-protein interaction potential we employ, which was developed by Mereghetti et al. for many-protein Brownian dynamics simulations using the SDAMM software package by Mereghetti et al.,(*20, 40*) contains four contributions. The first two account for the interactions of the charges on one protein with the electrostatic potential of a second protein and an "electrostatic desolvation" penalty when the charges on one protein enter a low dielectric cavity of a second protein. The two electrostatic contributions contain an explicit dependence on the solution ionic strength. The third contribution is a short-ranged attractive "nonpolar desolvation" potential that mimics hydrophobic interactions, and the fourth term describes soft-core repulsive interactions between atoms on different proteins.

Prior to the simulations the potentials for each simulated protein conformation were pre-computed on cubic grids. To determine the appropriate grid sizes, we examined the convergence of the radial distribution functions, $g(r)$, of the protein centers-of-mass with grid size (Figure S1 in the Supporting Information) in single-conformation MC simulations (described below). The results reported herein were generated using 200 x 200 x 200 grids with a 1 Å spacing for all of the terms in the interaction potential; Figure S1 shows that these grids are sufficient to obtain converged $g(r)$s.

The electrostatic potential grids were computed at each ionic strength considered for an atomistic representation of the proteins with charges corresponding to the OPLS force field(*41*) by finite-difference solution of the non-linear Poisson-Boltzmann equation using either the UHBD(*42*) or APBS(*43*) software packages. Dielectric constants of 78.4 and 2.0 for the solvent and protein, respectively, and ion exclusion radii of 1.5 Å were used. For increased computational efficiency, the number of charged protein sites involved in the evaluation of the electrostatic potential terms during the simulations was reduced by using the effective charge formalism of Gabdoulline and Wade(*44*).

We used the parameters in the potential function reported by Mereghetti et al.(*20*) with the following exceptions: (1) the empirical scaling coefficients of the electrostatic and nonpolar desolvation potentials were varied and optimized by comparing simulated osmotic virial coefficients with experimental data (see Figure S1.2). Unless otherwise noted, the default scaling parameters of 0.36 (unitless) and −0.0090 kcal/mol/Å^2 were used for the electrostatic and nonpolar desolvation terms, respectively; (2) we increased the parameter *σ* in the soft-core repulsion potential from 3 Å to 10 Å in order to increase the energetic penalty of overlapping protein atoms. This was necessary because MC trial moves with such unfavorable configurations need to be rejected. With the original parameter, the energetic penalty for overlapping protein molecules was too small, allowing compensation via electrostatic terms. In BD simulations, such configurations are not accessible, allowing a lower penalty and a smoother soft-core repulsion that, in turn, prevents the occurrence of large forces.

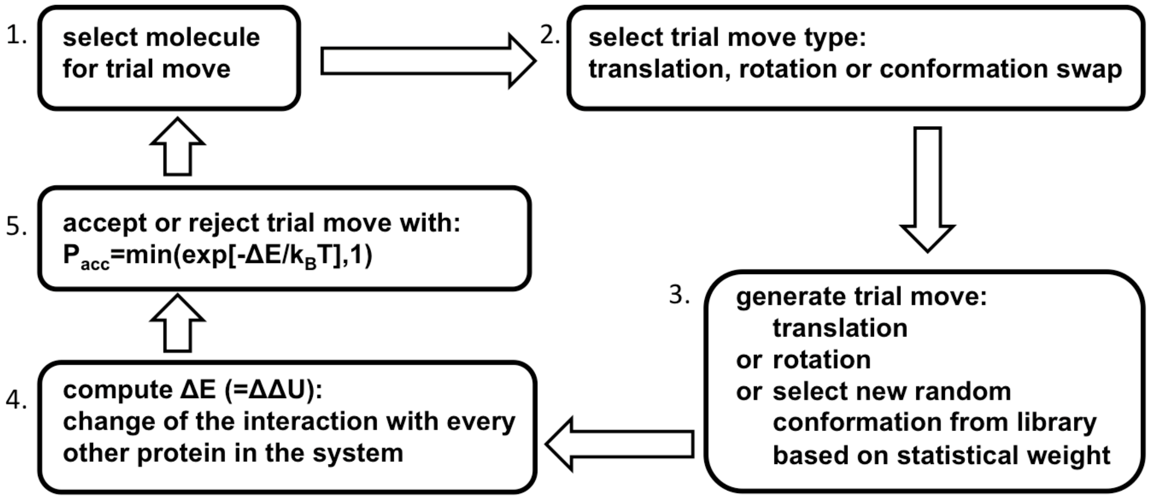
Single-conformation Monte Carlo simulations. We implemented single-conformation MC (scMC) simulations based on translational and rotational trial moves of randomly selected molecules in the SDAMM software package for BD simulations of many protein systems.(*20*) The step size of the translational and rotational trial moves adapts during the simulation to yield an acceptance ratio of roughly 50% to produce efficient sampling under all conditions. Trial moves are accepted with a probability given by the Metropolis criterion:(*4*)

$$P_{acc} = \min\left[1,\ \exp\left(\frac{-\Delta\Delta U}{k_B T}\right)\right], \qquad (1.1)$$

where *ΔΔU* is the difference in the protein-protein interaction potential between the current and trial configurations, k_B is Boltzmann's constant, and *T* is the temperature.

We performed scMC simulations of HEWL solutions over a range of protein concentrations and ionic strengths (Table S1.1) using three different structures obtained from neutron diffraction(*45, 46*) (PDB IDs 1IO5 and 1LZN) as well as solution NMR(*47*) (PDB ID 1E8L, model 1) experiments, all of which contain information on protonation states and proton coordinates in addition to the heavy atom coordinates. 1LZN has two protonated glutamates and a total charge of +11e, while 1IO5 and 1E8L carry a total charge of +9e.

Multi-conformation Monte Carlo simulations. To incorporate flexibility of the simulated proteins, we introduce a Monte Carlo trial move that, in addition to translational and rotational trial moves, attempts to swap protein conformations drawn from within a discrete library of conformations. This library is generated by clustering of protein conformations from an atomistic MD simulation of a single protein in explicit solvent. The candidate conformation is selected at random from the library with probability proportional to the population size of the corresponding cluster in the MD simulation (see below). As in the scMC simulations, a standard Metropolis acceptance criterion is used. This approach ensures that, in the dilute-solution limit (i.e., no protein-protein interactions), the resulting distribution of conformational states converges to the distribution observed in the MD simulation of a single solvated protein. In the case of interacting proteins, the distribution of protein conformations can change due to the stabilization/destabilization of conformations in bound states. Potential grids are pre-computed for each of the structures in the library. The size of the library is limited only by the memory requirements of these grids. The conformational swap moves allow sampling of the most favourable conformations of interacting proteins at a computational cost that is the same as that of the rigid-body translational or rotational moves. The algorithm is sketched in Scheme 1.1.


1.
select molecule for trial move
2.
select trial move type:
translation, rotation or conformation swap
3.
generate trial move:
translation
or rotation
or select new random conformation from library based on statistical weight
4.
compute ΔE (=ΔΔU):
change of the interaction with every other protein in the system
5.
accept or reject trial move with:
P_{acc}=min(exp[-ΔE/k_BT],1)

Scheme 1.1: Sketch of the multi-conformational Monte Carlo algorithm for the simulation of flexible proteins using a library of conformations and standard translational and rotational trial moves.

Various approaches could be employed to generate a suitable library of conformations. Here, we used a 150 ns atomistic MD simulation trajectory of a single protein in explicit solvent using a HEWL solution NMR structure(*47*) (PDB ID 1E8L, model 1) as the initial configuration. To generate the library of protein conformations, we computed the heavy (not hydrogen) atom root-mean squared deviation (RMSD) matrix between configurations saved every 10 ps, and employed a simple clustering algorithm(*48*) using a 1 Å RMSD cutoff to extract the 50 most populated clusters. By including all non-hydrogen atoms, we ensure that backbone and side chain fluctuations give rise to distinct conformations in the library. The cluster centroids constitute the library of conformations, and the cluster populations were used to assign a statistical weight to each conformation in the library.

The MD trajectory was generated using the GROMACS software package.(*49*) The OPLS all-atom force field(*41*) was used for the protein and ions, and the TIP3P model(*50*) was used for water. The charge of the protein was neutralized with chloride ions, the system was solvated by

3579 water molecules, and periodic boundary conditions were applied in three dimensions. Short-ranged interactions were truncated with a 9 Å cutoff, while long-ranged electrostatic interactions were computed with the smooth particle-mesh Ewald method(*51*) on a 1.2 Å real-space grid. Covalent bonds and the geometry of water molecules were constrained with the LINCS(*52*) and SETTLE(*53*) algorithms, respectively. After an initial energy minimization, the system was equilibrated for 100 ps using a 1 fs integration time step in the isothermal-isobaric ensemble with harmonic restraints on protein heavy atom positions using a force constant of 1000 kJ/mol/nm^2, followed by a 1 ns unrestrained equilibration using a 2 fs time step. During equilibration, a Berendsen(*54*) weak coupling thermostat and barostat was employed with time constants of 0.5 ps and 1.0 ps, respectively, and 300 K and 1 bar target values. The production simulation of 150 ns duration was generated with a 2 fs time step and the Nosé-Hoover thermostat(*55*) for temperature control and the Parrinello-Rahman barostat(*56*) for pressure control.

Calculation of osmotic second virial coefficients and structure factors. The osmotic second virial coefficient, B_2, is the second-order coefficient in the Taylor series expansion of the osmotic pressure in terms of number density;(*57*) it provides one of the very few experimental measures of pairwise interactions between protein molecules in solution: $B_2 < 0$ implies attractive interactions, while $B_2 > 0$ implies repulsive interactions, and the magnitude of B_2 quantifies the strength of the interactions. The osmotic second virial coefficient is computed from simulations according to:(*57*)

$$B_2 = -2\pi \int_0^\infty \left[g\left(r\right) - 1 \right] r^2 \, dr \ , \tag{1.2}$$

where $g(r)$ is the radial distribution function of the protein centers-of-mass. Here, the radial distribution function describes a potential of mean force between two molecules, $W(r) = -k_BT \ln g(r)$, which includes averaging over all possible orientations and conformations of both molecules.

The structure factor, $S(q)$, is an interference function that arises from interparticle interactions and can be extracted from small-angle x-ray and neutron scattering measurements. The structure factor is also readily computed from the protein-protein radial distribution function according to:(*58*)

$$S(q) = 1 + 4\pi\rho \int_0^\infty \left[g(r) - 1\right] \frac{\sin(qr)}{qr} r^2 \, dr \ , \qquad (1.3)$$

where q is the modulus momentum transfer vector, and ρ is the solution density.

Sample preparation and static light scattering experiments. Lyophilized hen egg white lysozyme (Cat. No. 195303) was purchased from MP Biomedicals (Solon, OH). Lysozyme was dissolved in 10 mM sodium phosphate buffers containing 0.05% sodium azide (pH 4.7 and 6.9) with NaCl concentrations of 50, 75, 100, 125, 150, 200, 250, and 300 mM for a final protein concentration of 50 mg/mL. Serial dilutions were performed to prepare samples with protein concentrations ranging from 2.5 to 50 mg/mL for light scattering measurements. The concentrations were checked by UV absorbance measurements using ε = 2.64 mL mM^{-1} cm^{-1} at 280 nm. A Dawn HELEOS multi-angle light scattering instrument and an Opitlab rEX refractive index detector (Wyatt Technology, Santa Barbara, CA) were used to collect the data required for experimental B_2 determination. Samples were injected using the batch-mode technique from lowest to highest concentrations after filtering to ensure monodispersity.

Estimation of osmotic second virial coefficients from static light scattering data. Scattering intensity data at each concentration was processed to remove artifacts caused by sample injection, and the median of the remaining observations employed as the scattering intensity measurement for each detector. Medians were also taken for the refractive index increment at each concentration; readings at concentrations greater than 0.02 g/mL exceeded the range of the differential refractometer, and were treated as missing for purposes of analysis.

For small particles in dilute solution, Zimm's (*59*) second order expansion of light scattering intensity (in terms of the excess Rayleigh ratio, R_θ) with respect to concentration leads to the approximation:(*60*)

$$\frac{Kc}{R_\theta} \approx \frac{1}{MP(\theta)} + 2A_2 c\,, \tag{1.4}$$

where $K = 4\pi^2 (dn/dc)^2 n_0^2 / N_A \lambda_0^4$, with *dn/dc* the refractive index increment, n_0 the solvent refractive index, λ_0 the vacuum wavelength of the incident light, N_A Avogadro's number, *M* the (mass weighted) mean particle mass, *c* the protein concentration, $P(\theta)$ a size-specific factor that depends upon the detection angle θ relative to the angle of incidence, and $A_2 = B_2 N_A / M^2$ is the osmotic second virial coefficient in a power series expansion of the osmotic pressure in terms of concentration. The $P(\theta) \to 1$ limit is realized as the particle radius of gyration r_g approaches 0; for monomeric or small oligomeric particles with $r_g \ll \lambda_0$, angular dependence is negligible, and we employ this limit here. Note that no angular dependence was detected in our experiments, which is consistent with predictions for a beam wavelength of 658 nm and $r_g \approx 0.14$ nm.(*61*)

As B_2 represents a very small deviation in local effective particle density (relative to uniform mixing), it is challenging to estimate with high precision. We employ a number of techniques to address this issue. Given multiple observations of R_θ at varying concentration, it is natural to estimate A_2 by regression of Kc/R_θ on $2c$; when the scattering particles are monodisperse and of known mass, improved precision can be obtained by employing $Kc/R_\theta - 1/M$ as the response and fitting a zero-intercept model. When particles are known to be monodisperse but the oligomer size is not known, greater precision can still be obtained by fitting models to *k*-mers of orders 1, 2, … and selecting the *k* that minimizes the squared error in the predicted scattering intensity. As our samples were filtered to ensure monodispersity (with verification by dynamic light scattering) and the monomer mass is known, we employ this strategy here. This estimate also depends upon *dn*/*dc*, which must itself be estimated by regressing n_c (the measured refractive index at concentration *c*) against *c*. Because n_c is in practice far more reliably measured than *c* itself, further gains in precision can be obtained by using the refractive index data to correct the measured concentration values prior to estimation of A_2 (i.e., regressing *c* on n_c and employing the predicted $\hat{c}$ values in place of *c*). Combining the above leads to the following multi-stage procedure for estimating A_2: (1) regress n_c on *c* to obtain an estimate of *dn*/*dc* ($\widehat{dn/dc}$); (2) regress *c* on n_c to obtain corrected concentration estimates $\hat{c}$; (3) for $k \in$ 1, 2, …, regress $Kc/R_\theta - 1/(kM)$ on $2\hat{c}$ to obtain $\widehat{A_2}|k$, selecting the *k* leading to the minimum squared error in R_θ and associated $\widehat{A_2}$ for the final B_2 estimate.

Classical estimates of the precision of $\widehat{A_2}$ are problematic both because of the interdependence of $\widehat{A_2}$ on $\widehat{dn/dc}$ and because of the contribution of concentration to both sides of the regression. Here, we instead employ a non-parametric bootstrap procedure (using the boot library for R(*62*)) to estimate confidence intervals. Specifically, the above procedure was repeated for 5000 random with-replacement joint resamples of the refractive index and scattering intensity data (with sample sizes preserved for each subset), and 95% confidence intervals were estimated from the resulting bootstrap replicates using the bias-corrected/adjusted percentile (BCa) method of Efron (*63*). These are shown in Figures 1.3 and 1.10 as vertical lines. Oligomer size estimates were further inspected by examination of bootstrap standard errors for signs of instability; samples were estimated to be monomeric for all replicates in all conditions examined.

RESULTS AND DISCUSSION

Single-configuration Monte Carlo simulations. In protein solution simulations at low to medium concentrations (e.g., < 100 mg/mL), translational MC trial move steps can be significantly larger than BD time steps, increasing the efficiency of configurational sampling. This is illustrated in Figure 1.1, which displays the convergence of the radial distribution function between protein centers-of-mass in scMC and BD simulations of HEWL under identical conditions at a concentration of 10 mg/mL. The sampling efficiency in the scMC simulations is increased by roughly two orders of magnitude compared to BD, producing a converged radial distribution function after 100k MC cycles (one cycle consists of N_{prot} trial moves, where N_{prot} is the number of protein molecules in the simulation).

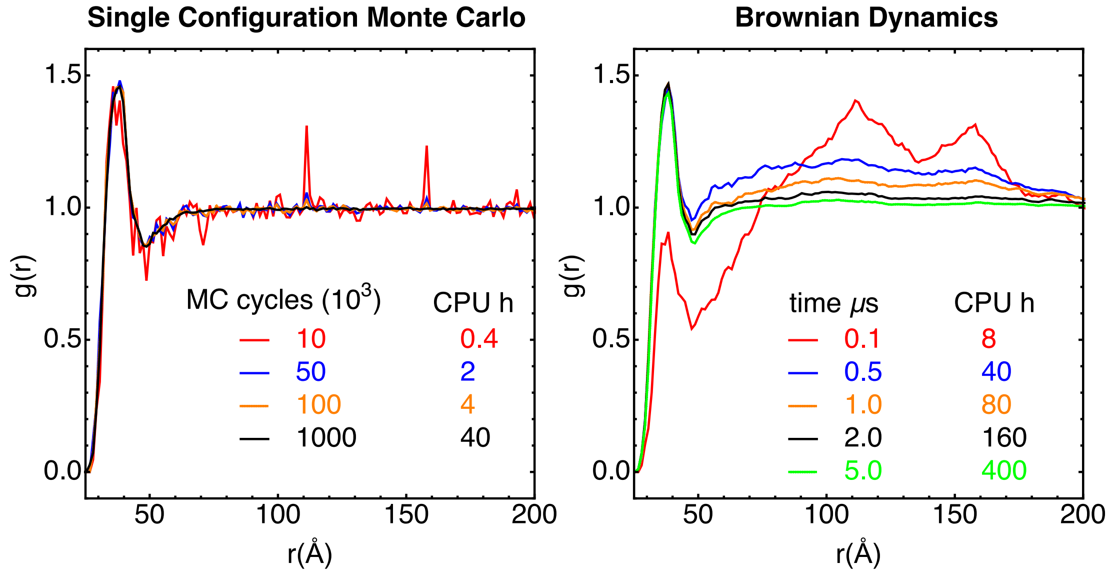


Figure 1.1: Convergence of the protein-protein radial distribution function from scMC (left panel) and BD (right panel) simulations of HEWL solutions containing 200 rigid proteins (1LZN structure) at a concentration of 10 mg/mL and ionic strength of 200 mM. The sampling efficiency, in terms of CPU hours needed to generate a converged radial distribution function, is roughly two orders of magnitude higher for scMC than BD simulations (CPU h on single Intel Xeon E5430 processor with 2.66 GHz). scMC simulation lengths are expressed in MC cycles; a single MC cycle consists of N_{prot} trial moves, where N_{prot} is the number of protein molecules in the simulation.

The scMC sampling efficiency advantage over BD vanishes at high concentrations, when the step size of translational MC trial moves need to be reduced in order to maintain acceptance ratios on the order of 50%, as shown in Figure 1.2 for a 169 mg/mL HEWL solution. However, we point out that MC simulations offer additional advantages for the simulation of slowly converging systems, e.g., in the event of protein aggregation. For example, biased sampling schemes, such as the aggregation volume biased MC technique developed by Siepmann and co-workers (*24,* *25*), can be readily implemented to improve the sampling of the formation and destruction of clusters.

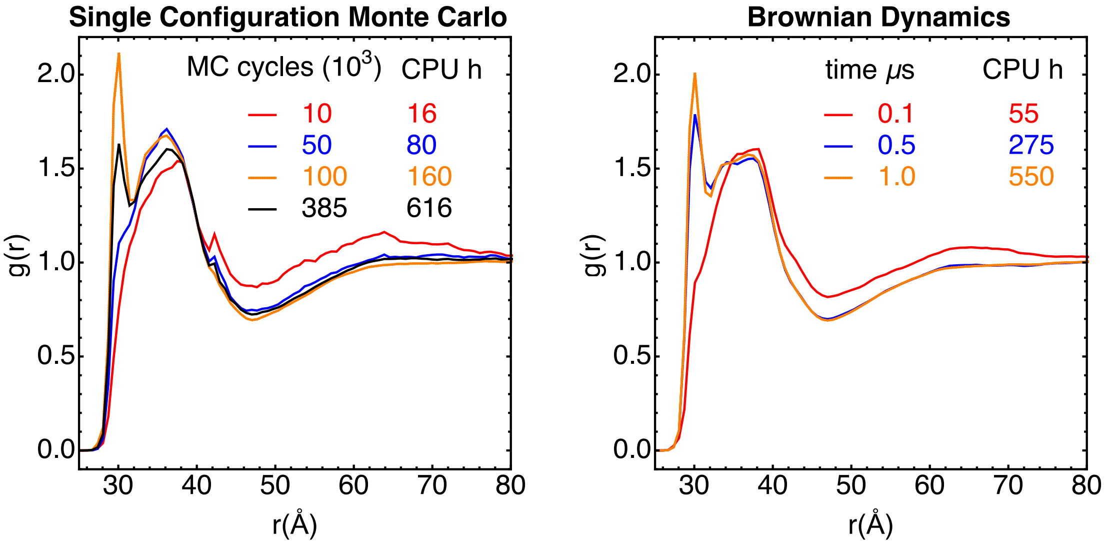


Figure 1.2: Convergence of the protein-protein radial distribution function from scMC (left panel) and BD (right panel) simulations of HEWL solutions containing 475 rigid proteins (1E8L structure) at a concentration of 169 mg/mL and ionic strength of 100 mM. The sampling efficiency is similar for both scMC and BD simulations at this concentration (CPU h on single Intel Xeon E5430 processor with 2.66 GHz); scMC simulation lengths are expressed in MC cycles; a single MC cycle consists of N_{prot} trial moves, where N_{prot} is the number of protein molecules in the simulation.

Previous BD simulations of HEWL solutions using the same (*20*) or similar (*18*) protein interaction potentials were validated using osmotic second virial coefficients as a function of solution ionic strength reported in the literature (*58*). Here, we report two new sets of B_2 values from SLS measurements on HEWL at two different pH values (see Figure 1.3). Our experimental results are in good agreement with the literature (*58, 64-67*) and are consistent with Derjaguin-Landau-Verwey-Overbeek (DLVO) theory for colloidal systems (*65, 68*).

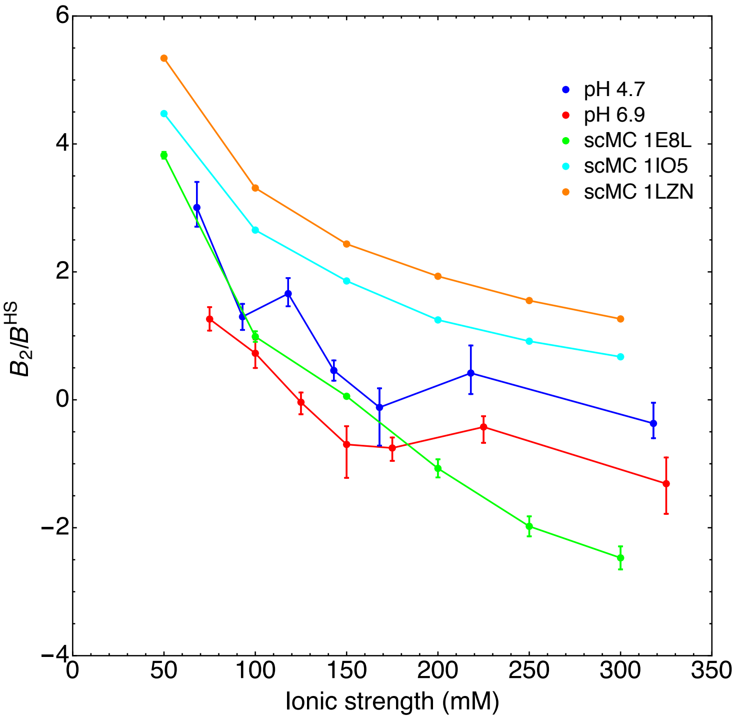


Figure 1.3: HEWL second virial coefficient (B_2; in units of B_2 for hard spheres, B^{HS} = 4 x protein volume) estimated from static light scattering experiments and scMC simulations as a function of solution ionic strength. Solution pH values of 4.7 and 6.9 reported in the experimental measurements correspond to estimated HEWL net charges of +10e and +8e, respectively (*69*). Proteins in solution simulations based on crystal structures (1IO5 and 1LZN, net charge +9e and +11e, respectively) are overall more repulsive than indicated by experiments. HEWL in simulations based on the solution NMR structure (1E8L), which has the same net charge as 1IO5, are overall more attractive than in crystal structure simulations, and appear to be consistent with experiments (see text for more details) at ionic strength values less than ~0.2 M. However, at higher ionic strengths, the proteins appear to be more attractive than indicated by experiments.

Comparison of the ionic strength dependence of the B_2 values computed from scMC simulations at a concentration of 10 mg/mL using different input structures reveals that the protein-protein interactions are strongly dependent on the choice of structure, and shows that a single structure is not able to reproduce the trend in the experimental data (Figure 1.3). Electrostatic repulsion is overestimated and, hence, B_2 is too large, in the two simulations based on crystal structures (1IO5 and 1LZN, net charge +9e and +11e, respectively). At ionic strength values below ~0.2 M, the B_2 values obtained from the solution NMR structure (1E8L, net charge +9e) follow the correct qualitative trend compared to the experimental data, but the preponderance of electrostatic interactions is evident by the lack of a plateau at ionic strength values above ~0.2 M. Notably, the protein-protein interactions are significantly more attractive in the 1E8L simulations than in the simulations based on the crystal structure with the same net charge (1IO5).

In addition to the overall increased protein-protein attractive interaction, the corresponding radial distribution functions between protein centers exhibit a spurious peak at r ~ 30 Å in simulations of the 1E8L HEWL structure, corresponding to specific protein-protein contacts that are not present in the simulation based on the 1IO5 and 1LZN crystal structures (Figure 1.4). Furthermore, the differences between the 1E8L and 1IO5 radial distribution functions are much greater than the differences between the 1IO5 and 1LZN simulations, suggesting that the specifics of protein conformations can have a much more dramatic effect than the total charge.

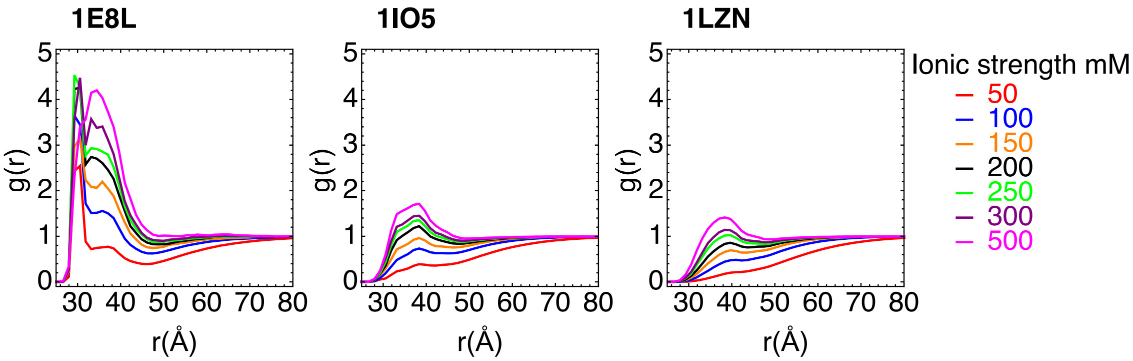


Figure 1.4: Protein-protein radial distribution functions *g*(*r*) from the scMC simulations from which the B_2 values plotted in Figure 1.3 were obtained (10 mg/mL). In addition to a spurious peak at ~30 Å in simulations based on the 1E8L structure, there is more contrast in the protein-protein interactions when comparing the provenance of the protein structure (e.g., 1E8L vs. 1IO5) than when comparing different protein net charges (1IO5 vs. 1LZN).

In Figure 1.5, we compare isosurfaces of spatial distribution functions (SDFs), specifically, distributions of the density of other protein centers-of-mass around a tagged central protein, normalized by the bulk density at a protein concentration of 10 mg/mL with an ionic strength of 100 mM. Despite the apparent similarity of the radial distribution functions for scMC simulations based on the two crystal structures 1IO5 and 1LZN, the SDF from the simulation based on the 1IO5 structure reveals interaction sites that are not observed in the SDF from the simulation based on the more repulsive 1LZN structure. Similarly, the comparison to the SDF obtained from the simulation based on the solution NMR structure (1E8L) shows not only an overall increased attraction, but also additional locations of preferred contact sites.

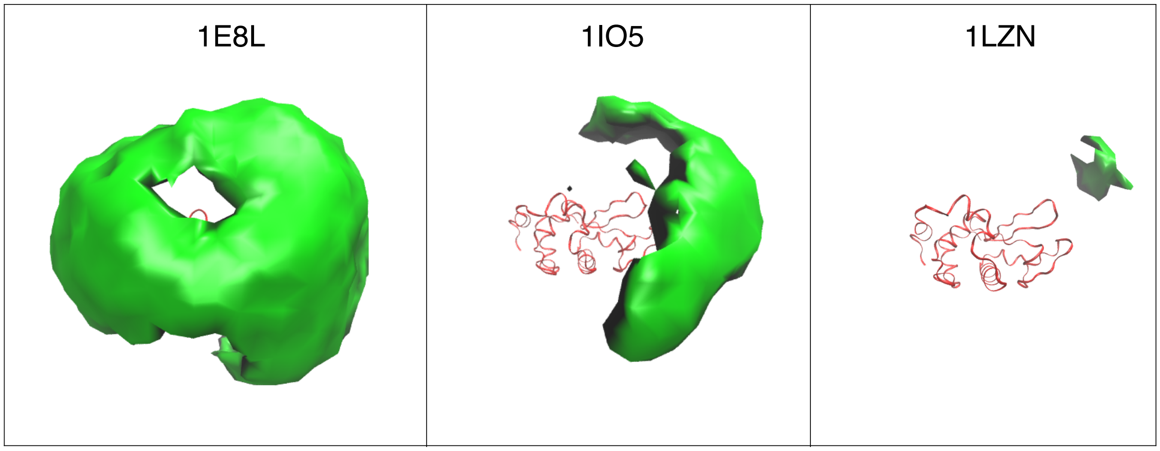


Figure 1.5: Spatial distribution functions (see text for details) for scMC simulations at 100 mM ionic strength. The green surfaces represent regions of 1.7 times the bulk density computed on a 160 Å cubic grid with a 4 Å grid spacing centered on and aligned with respect to the reference molecule shown in red. In addition to the greater protein-protein attraction in the 1E8L simulation, the simulations based on the two crystal structures exhibit interaction sites that are distinct from each other.

The SDF for the 1E8L simulation as shown in Figure 1.5 computed with an increased resolution on a 1 Å grid (Figure 1.6A) reveals two highly localized sites with a more than 1500-fold increase of the concentration relative to the bulk solution (corresponding to a stabilization in free energy of –4.3 kcal/mol). The corresponding dimer structures, extracted from the simulation trajectories, show binding motifs characterized by energetically favorable contacts of charged side chains (Figure 1.6B). Such specific binding motifs were not observed in simulations of the 1LZN and 1IO5 HEWL structures. The highest local density represents only 9-fold and 20-fold increases relative to the bulk density in scMC simulations based on the 1LZN and 1IO5 structures, respectively.

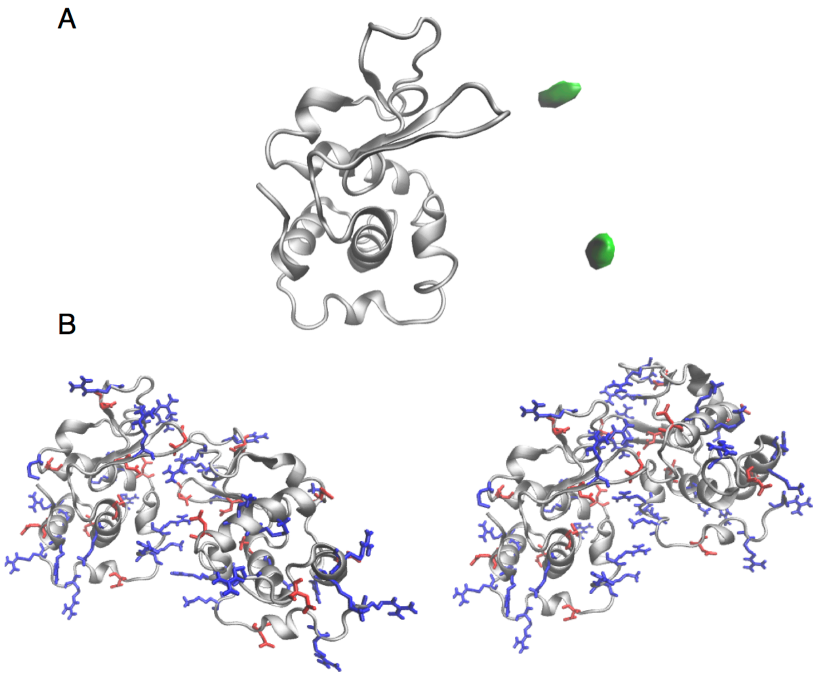


Figure 1.6: (A) Spatial distribution function for the 1E8L simulation at 100 mM ionic strength computed on a 100 Å cubic grid with a grid spacing of 1 Å. The green isosurfaces correspond to a 1500-fold increase of the local density relative to the bulk density. (B) Dimer configurations corresponding to the regions of increased concentration shown in (A). Solvent exposed basic (blue) and acidic (red) side chains are rendered in ball-and-stick representation.

The fact that 1IO5 and 1E8L structures have the same total charge and protonation state indicates that the highly specific binding observed in the 1E8L simulations originates in the specific arrangements of side chains on the protein surface, which is a consequence of the differences in structure determination methods. The 1E8L structure represents a protein in an aqueous solution environment where the charged side chains are in more extended conformations than those in the crystal environment required for the determination of the 1IO5 structure via neutron diffraction. A superposition of the 1IO5 and 1E8L structures (Figure 1.7) indicates, apart from the almost perfect alignment of the protein backbone, that the basic side chains tend to be significantly more extended and solvent exposed in the solution NMR structure than in the crystal, where they are, on average, more folded onto the protein surface. The solvent exposed basic side chains will be floppy and explore multiple conformations in solution. However, in scMC simulations as well as in rigid-body BD simulations, they are rigid, effectively losing their conformational entropy. This leads to enhanced favorable inter-protein interactions between exposed side chains of opposite charge, as observed in Figure 1.6, which in turn create highly specific binding motifs. This behavior is unrealistic, as the conformational entropy of the solvent-exposed side chains should decrease the population of such highly specific conformations. Thus, our results demonstrate that artifacts in inter-protein interactions may arise due to including only single side chain conformations in rigid-body BD or MC simulations of many-protein systems.

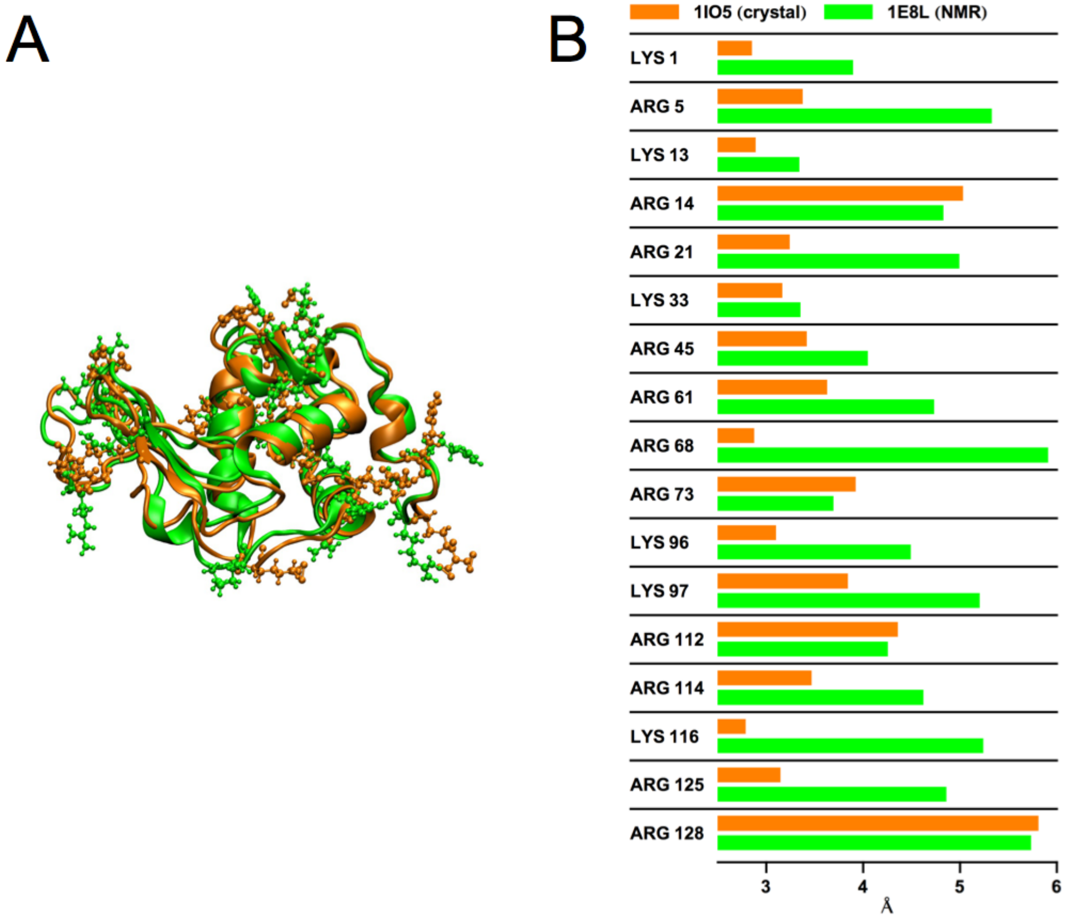


Figure 1.7: Conformations of basic side chains in the neutron diffraction structure 1IO5 (orange) and the solution NMR structure 1E8L (green) of HEWL (A), and distances between the centers of charge and the protein surface (B), indicate that they are more extended and solvent-exposed in the solution NMR structure than in the crystal structure. The behavior of these side chains in solution is poorly modeled with a single protein configuration.

Multi-conformation Monte Carlo simulations. To eliminate the dependence of the simulation results on the specifics of the protein conformation, we implemented the mcMC method, which introduces conformational flexibility in the protein molecules by allowing them to convert from one conformation to another within a structural library generated from an all-atom MD simulation (see Methods for details). While the degree of flexibility introduced this way is obviously limited and not able to describe the entire protein conformational space, we posit that any reasonable choice of the ensemble will provide a significant improvement over the modeling of proteins in solution as single-conformation rigid bodies.

Our library consists of 50 structures that collectively represent a broad sampling of the basic and acidic side chain conformations without significant structural changes to the HEWL backbone (backbone RMSD < 2 Å). The variation of the positions of the positive and negative charge centers within the ensemble is shown in Figure 1.8A. The statistical weights associated with the library of conformations, obtained from the relative population of each cluster in the MD trajectory and used in the generation of conformational swap trial moves, are depicted in Figure 1.8B.

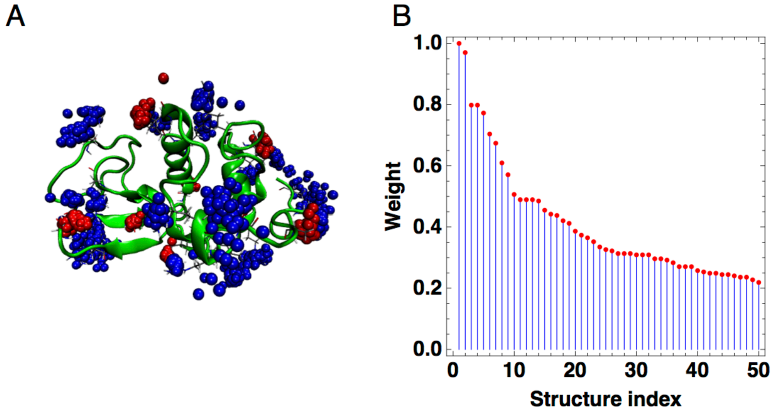


Figure 1.8: Structural library of HEWL employed in mcMC simulations obtained from a 150 ns all-atom MD simulation trajectory. (A) Superposition of the charge centers of basic (blue) and acidic (red) amino acid side chains in the 50 conformations that comprise the library. B) The statistical weights of individual structures, determined by their relative populations along the MD trajectory, are used as probabilities for the generation of conformational swap trial moves in the mcMC simulations.

To maximize agreement between experimental second virial coefficients and simulations with the mcMC algorithm, we adapted the protein-protein interaction potential by rescaling the empirical nonpolar desolvation (ND) potential term (as also described in other studies with the employed interaction potential (*20, 40*)). This term describes a short-ranged, uniform attraction between protein surfaces and mimics hydrophobic interactions. We also considered the effects of simultaneous rescaling of the empirical electrostatic desolvation, ED (repulsive, non-uniform), and nonpolar desolvation, ND (attractive, uniform), terms. We found that the two desolvation terms have compensating effects, resulting in a set of optimal ED and ND scaling factor pairs that include the default scaling value of the ED term (Figure S1.2). Therefore, we opted for leaving the ED scaling factor at its default value and explored in more detail the consequences of variations in the ND scaling factor.

Figure 1.9 shows the effects of changing the strength of the nonpolar desolvation potential on the protein-protein radial distribution function. When the default value of the scaling parameter (ND = −0.0090 kcal/mol/Å^2) is used, the resulting radial distribution functions (Figure 9, left panels) are very similar to the ones obtained from scMC simulations based on the 1IO5 structure under identical conditions (middle panels of Figures 1.4 and 1.5), suggesting a comparable radially averaged interaction potential between the HEWL proteins in both simulations. However, the SDFs at 100 mM ionic strength again show significant differences between the two simulations (compare the left panel of Figure 1.9B with the middle panel of Figure 1.5), indicating the importance of side chain conformation and flexibility in determining the preferred binding geometries. The radial distribution functions from mcMC simulations with increased strengths of nonpolar desolvation interactions (ND = −0.0098 kcal/mol/Å^2 and ND = −0.0100 kcal/mol/Å^2) and the 1E8L scMC simulation (Figure 1.4, left panel) exhibit comparable main peaks at ~35 Å separation distance. However, the spurious peak at ~30 Å is absent in the mcMC simulations, as expected for simulations with flexible side chains.

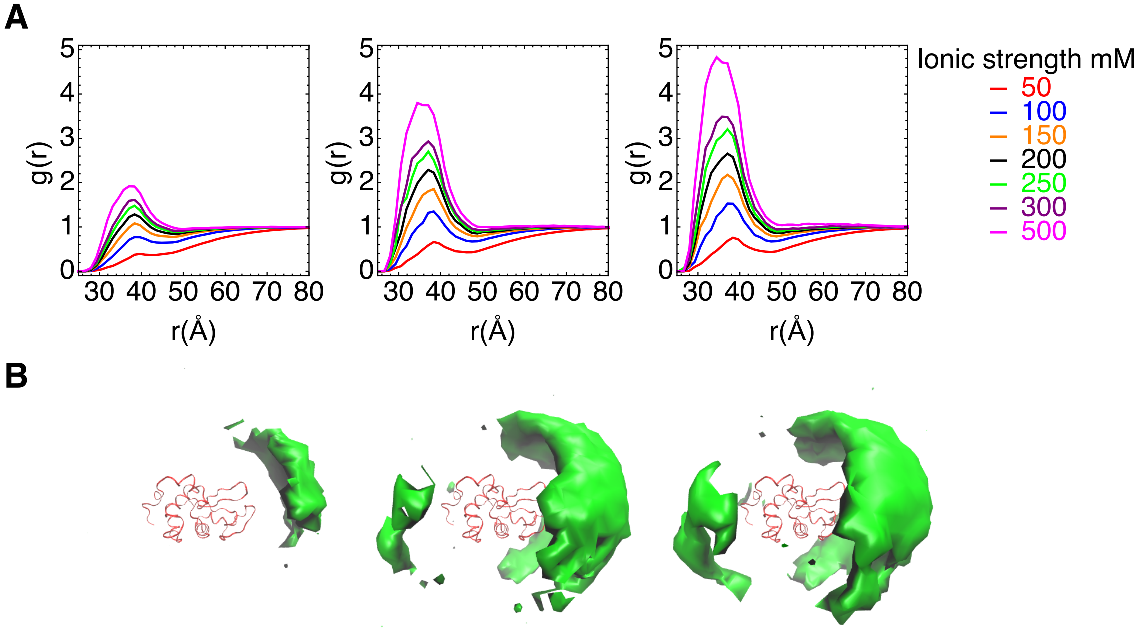


Figure 1.9: (A) Protein-protein radial distribution functions from HEWL mcMC simulations at three different values of the scaling parameter of the nonpolar desolvation potential, ND (from left to right, ND = −0.0090 kcal/mol/Å^2, −0.0098 kcal/mol/Å^2, −0.0100 kcal/mol/Å^2). (B) Isosurface contours (green) at 1.7 times the bulk density in the corresponding spatial distribution functions at 100 mM ionic strength.

The experimental osmotic second virial coefficients are reproduced well by the mcMC simulations with stronger nonpolar desolvation interactions (Figure 1.10). Notably, however, the SDFs, which define the preferred protein-protein binding interfaces, apart from the overall interaction strength, do not depend on the scaling factor of the nonpolar desolvation term (Figure 1.9B) as expected given the uniform attraction relative to the solvent accessible surface area described by the scaled ND potential term. In the scMC simulation that employed the 1E8L solution NMR structure, the agreement with the experimental B_2 values at ionic strength below 200 mM was fortuitously achieved via increased exposure of the polar and charged side chains with zero conformational entropy. The exposure of polar and charged side chains is comparable for the structures in the ensemble used for the flexible mcMC simulations, as they are obtained from MD simulations in an explicit solvent environment. Thus, in contrast to the scMC case, in the mcMC simulation energetically highly favorable protein-protein interactions between individual structures lead to a compensating decrease of the conformational entropy, thus weakening the total binding affinity.

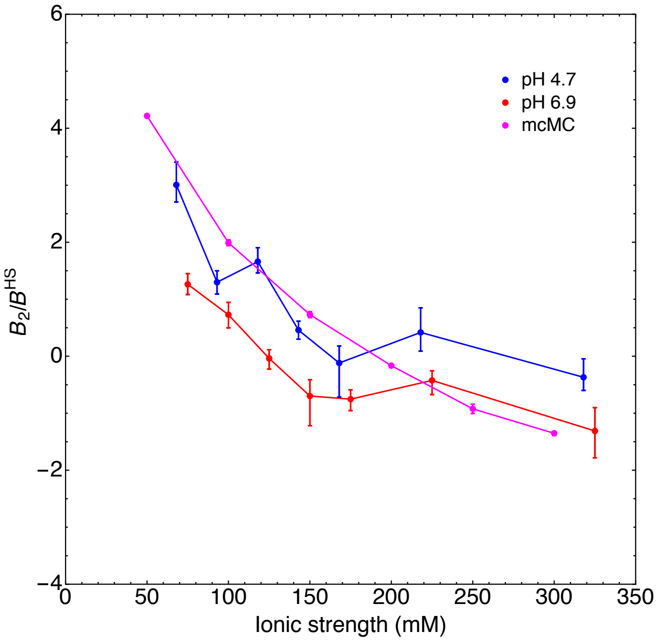


Figure 1.10: HEWL second virial coefficient (B_2; in units of B_2 for hard spheres, B^{HS} = 4 x protein volume) estimated from static light scattering experiments and mcMC simulations as a function of solution ionic strength at a protein concentration of 10 mg/mL. The mcMC simulations were performed using the optimal value of the nonpolar desolvation strength determined from the experimental estimates (ND = −0.0098 kcal/mol/Å^2; see Supporting Information for more detail on the potential parameter optimization).

While osmotic second virial coefficients are determined at low concentrations (i.e., 2.5–50 mg/mL), structure factors allow us to validate our simulations against experimental data obtained at high protein concentration. Structure factors are interference functions that arise from protein-protein interactions in small-angle x-ray and neutron scattering measurements on protein solutions; peaks in the structure factors occur at values of the wave-vector transfer, $q \sim 2\pi/d$, corresponding to preferred interactions on length scales d. Results from mcMC simulations with variable scaling of the nonpolar desolvation potentials are shown in Figure 1.11 in comparison to experimental results (*70, 71*) and a previous BD simulation study by McGuffee *et al* (*18*). The concentration used in all cases is 169 mg/mL. In order to compare our simulations with experiments at low ionic strength, we employed an ionic strength of 50 mM to ensure a sufficient decay of the electrostatic interactions within the employed potential grids. We note that the experimental studies have been carried out at a neutral pH with a slightly different protonation state (charge of +8e), while the HEWL proteins in our mcMC simulations carry a charge of +9e. Thus, minor differences between the simulation and experimental data are expected based on slight differences in solution conditions.

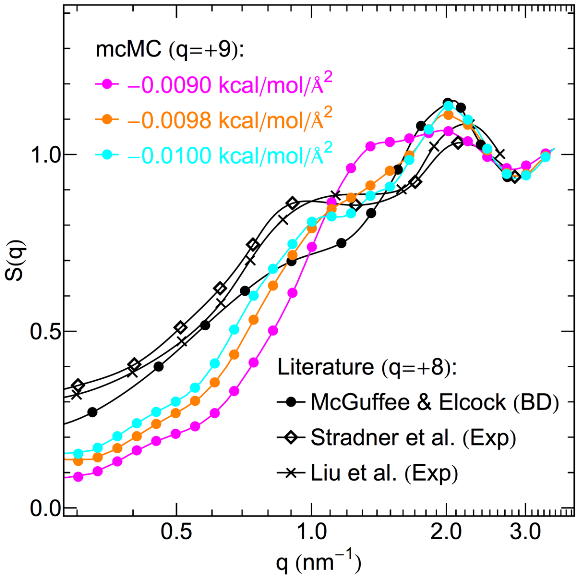


Figure 1.11: Structure factors from mcMC simulations (with varying nonpolar desolvation strength) of a HEWL solution at 169 mg/mL and 50 mM ionic strength in comparison to experimental data and a previous BD simulation, both at neutral pH. The choice made for the nonpolar desolvation strength parameter after optimization using B_2 estimates from low-concentration simulations (ND = −0.0098 kcal/mol/Å^2) also produces the best match to experimental structure factors at high concentration.

The main peak at q = 2.0 nm^{-1} and the shoulder at q = 0.9–1.0 nm^{-1} are qualitatively reproduced by the mcMC simulations with modified nonpolar desolvation potentials (ND = −0.0098 kcal/mol/Å^2 and −0.0100 kcal/mol/Å^2) (Figure 1.11). In the simulation with the default scaling factor (ND = −0.0090 kcal/mol/Å^2), the low q-shoulder is shifted towards higher q-values and is increased in intensity. At q-values below 0.7 nm^{-1}, the structure factors obtained from mcMC simulations are lower than the experimental structure factors, indicating some discrepancy in the long-range order. This discrepancy could be due to the increased long-ranged repulsion caused by the additional charge, and/or to the lack of convergence of the values of the radial distribution functions at large distances that are necessary to compute accurate values of $S(q)$ at very low q.(*18*) Nonetheless, the simultaneous overall good agreement of the mcMC simulation results (for a particular choice of the nonpolar desolvation scaling factor, ND = −0.0098 kcal/mol/Å^2) with experimentally determined osmotic second virial coefficients and structure factors shows that the mcMC simulations can provide a realistic description of protein-protein interactions. Moreover, the conformational sampling provided by mcMC alleviates the unwanted dependence of the simulated protein-protein interactions on the choice of input structure, which is an issue with conventional rigid-body simulations.

High concentrations and the resulting prevalence of protein-protein interactions also affect the population of HEWL conformations in the mcMC simulations. Conformations that are able to form dimers or oligomers with low intermolecular potential energies are stabilized, while conformations that interact less favorably with other proteins are destabilized. Figure 1.12 shows the distribution of conformations sampled by the MD simulation represented by the 50 individual conformations in the library (gray bars) together with interaction-induced changes observed in the mcMC simulations at 169 mg/mL for the three values of the nonpolar desolvation potential considered here (colored bars). Both, stabilization and destabilization effects are monotonic and approximately proportional to the scaling factor of the nonpolar desolvation potential. This result shows how mcMC simulations allow the system to adapt the distribution of protein conformations from dilute-limit conditions, in which the structural library was generated, to the high concentration regime, in which inter-protein interactions become relevant. This adaptation is limited to a change in the population of discrete conformations represented in the employed library of structures. Conformations that are unfavorable under dilute conditions and only become stable due to interactions with other proteins will have low statistical weights in a library generated by a MD simulation of a single solvated protein and, hence, require many trial moves to be sampled, even if they are energetically favored in high concentration conditions. Alternative procedures for generating the library of protein conformations, such as a more computationally demanding MD simulation of multiple protein molecules at high concentration, might be considered in such a case.

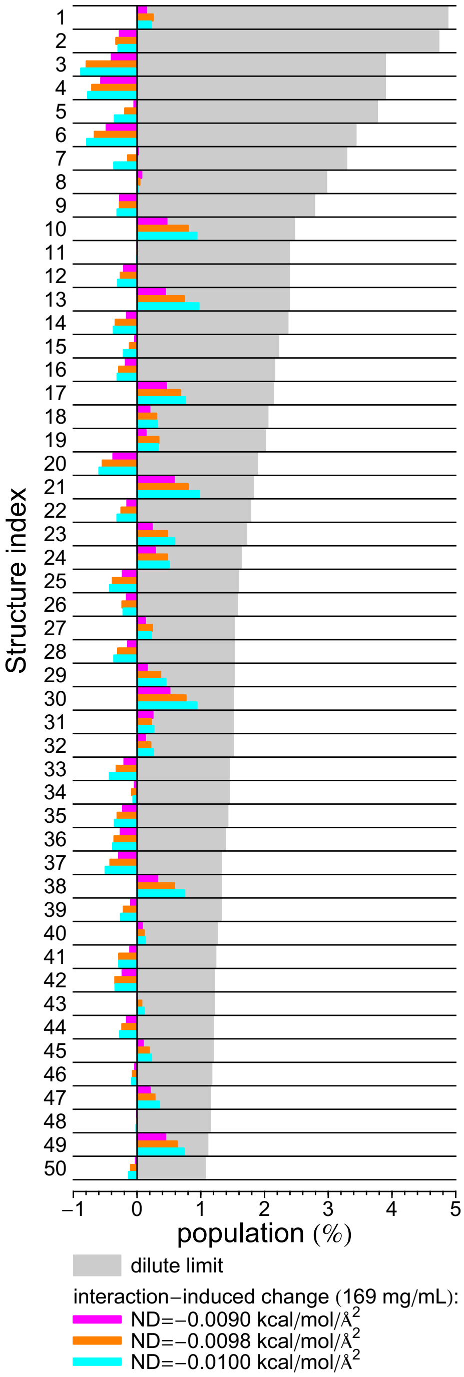


Figure 1.12: Distribution of protein conformations sampled by the MD simulation (gray bars, proportional to the statistical weights shown in Figure 1.8B) together with changes induced by inter-protein interactions in mcMC simulations at high protein concentration (169 mg/mL) for varying strength of the nonpolar desolvation potential term (colored bars).

SUMMARY

We have shown that MC simulations of many-protein systems can be performed using protein-protein interaction models developed for BD simulations. We used scMC simulations to analyze the effects of fixed side chain conformations in rigid-body simulations of systems containing many interacting proteins. Differences in the exposure of charged basic side chains, in particular between solution NMR structures and structures obtained from crystallography, can significantly modify preferential protein-protein interaction sites and the overall attraction of the proteins. Within the framework of MC simulations of protein solutions, we introduced a simple approach, mcMC, to account for molecular flexibility by allowing molecules to switch between multiple conformations within a discrete library of conformations, e.g., as obtained herein from a MD simulation of a single, fully-flexible protein in aqueous solution. Statistical weights used in the generation of mcMC trial moves for conformational changes allow us to describe distinct thermodynamic stabilities of individual conformations in the infinite dilution limit. Our approach removes potential artifacts observed in simulations of rigid protein structures, such as highly specific binding motifs involving fixed conformations of long, charged side chains, whose flexibility needs to be accounted for. In particular, after a minor reparameterization of empirical scaling parameters in the protein-protein interaction potential, we demonstrated improved agreement of simulated osmotic second virial coefficients with light scattering experiments at low protein concentrations and various salt concentrations. In addition, we could demonstrate improved agreement with experimental structure factors obtained at high protein concentration.

The mcMC approach can be employed for simulations of aggregating protein systems and crowded biomolecular solutions. Introducing flexibility for the simulated proteins, even if only within a limited set of discrete, Boltzmann-weighted conformations, will improve predictions of specific binding modes as well as overall aggregation propensities. Furthermore, the use of MC simulations allows implementation of enhanced sampling procedures for strongly aggregating systems (such as aggregation volume biased Monte Carlo(*24, 25, 27, 35*)), which are otherwise challenging to sample in conventional BD or MC simulations.

SUPPLEMENTAL INFORMATION

Table S1.1. Single-conformation Monte Carlo (scMC) simulations of hen egg white lysozyme solutions

Protein Structure (PDB ID)	Net Charge (*e*)	Number of Proteins	Concentration (mg/ml)	Ionic Strength (mM)	MC cycles[a] (x 10^3)
1E8L	+9	200	10	50	1500
1E8L	+9	200	10	100	1500
1E8L	+9	200	10	150	1500
1E8L	+9	200	10	200	1500
1E8L	+9	200	10	250	1500
1E8L	+9	200	10	300	1500
1E8L	+9	200	10	500	1500
1E8L	+9	475	169	100	385
1IO5	+9	200	10	50	1500
1IO5	+9	200	10	100	1500
1IO5	+9	200	10	150	1500
1IO5	+9	200	10	200	1500
1IO5	+9	200	10	250	1500
1IO5	+9	200	10	300	1500
1IO5	+9	200	10	500	1500
1LZN	+11	200	10	50	1500
1LZN	+11	200	10	100	1500
1LZN	+11	200	10	150	1500
1LZN	+11	200	10	200	1500
1LZN	+11	200	10	250	1500
1LZN	+11	200	10	300	1500
1LZN	+11	200	10	500	1500

[a] One MC cycle consists of N_{prot} trial moves, where N_{prot} is the number of protein molecules in the simulation.

Table S1.2. Multi-conformation Monte Carlo (mcMC) simulations of hen egg white lysozyme solutions[a,b]

Number of Proteins	Concentration (mg/ml)	Ionic Strength (mM)	MC cycles[c] (x 10^3)
200	10	50	1500
200	10	100	1500
200	10	150	1500
200	10	200	1500
200	10	250	1500
200	10	300	1500
200	10	500	1500
475	169	50	300

[a]The structure library used in the mcMC simulations was generated from an all-atom MD simulation with the 1E8L structure (net charge +9e) as the initial configuration.

[b]Each simulation indicated in this table was performed three times, each with a different value of the nonpolar desolvation parameter.

[b]One MC cycle consists of N_{prot} trial moves, where N_{prot} is the number of protein molecules in the simulation.

Dependence of protein-protein radial distribution functions on the size of the interaction potential grids. All of the terms in the protein-protein interaction potential are mapped onto cubic grids for computational efficiency. We investigated the optimal grid size for each potential term in turn by varying the grid size between 60^3 and 200^3 Å^3, while using a 200^3 Å^3 grid for all the other potential terms. Figure S1.1 shows the corresponding radial distribution functions for each test set. Figure S1.1 shows radial distribution functions obtained from scMC simulations using the 1E8L NMR solution structure at 10 mg/mL concentration and 100 mM ionic strength. The results suggest that convergence is achieved for the electrostatic potential at a minimum grid size of 100^3 Å^3 (at this ionic strength) and at 80^3 Å^3 for all other potential terms. The conservative use of 200^3 Å^3 potential grids in the remainder of this study therefore ensures minimal influence of this effective interaction potential cutoff on the reported results.

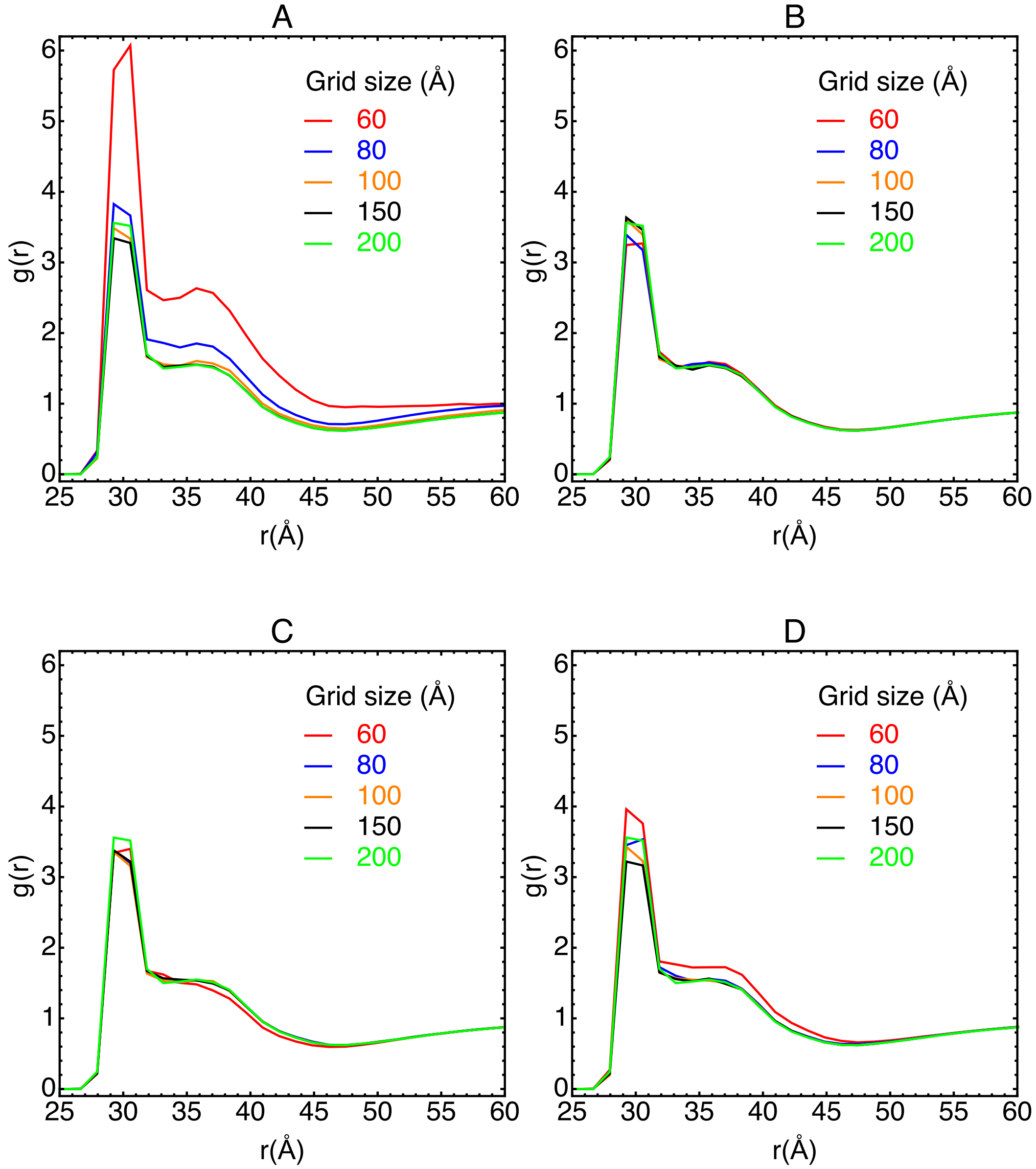
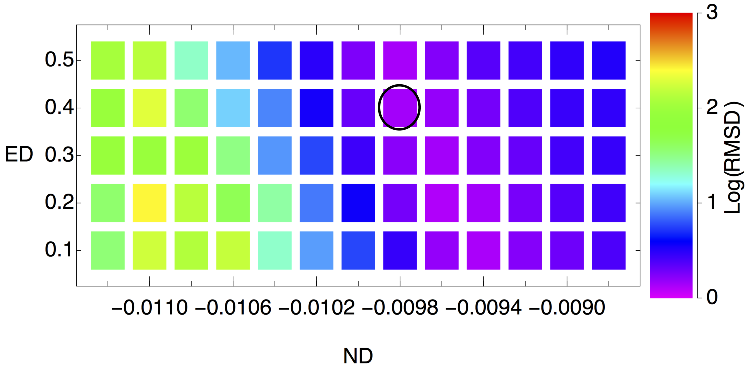


Figure. S1.1. Protein-protein radial distribution functions from HEWL scMC simulations performed with the 1E8L structure at a concentration of 10 mg/mL and an ionic strength of 100 mM. Each panel shows the effect of varying the grid size of a specific interaction potential term while using a grid size of at 200^3

Å^3 for all others: (A) Electrostatic potential term. (B) Electrostatic desolvation potential. (C) Nonpolar desolvation potential. (D) Soft-core repulsion potential.

Optimization of the parameters in the desolvation potentials. In the SDAMM potential,(*72*) the strength of the electrostatic and nonpolar desolvation potential terms are specified by empirical scaling paramters (ED and ND, respectively).(*40, 72*) As the ionic strength is increased, the Coulomb interactions are screened and the short-ranged desolvation potentials dominate the protein-protein interactions. In order to optimize the values of ED and ND, we performed a systematic search of the ED/ND parameter space by computing B_2 values in HEWL mcMC simulations at a concentration of 10 mg/ml and ionic strengths of 100 and 300 mM. We calculated the root-mean squared differences (RMSD) between the simulation values and our two sets of experimental B_2 estimates (pH 4.7 and 6.9) in the same range of ionic strength. The ND parameter was varied from –0.0114 kcal/mol/Å^2 to –0.0088 kcal/mol/Å^2 and the ED parameter from 0.1 to 0.5. The logarithm of the RMSD is shown as a function of the ND and ED parameters in Figure S1.2. It is evident that the attractive and repulsive contributions to the overall protein-protein interaction potential due to the two desolvation potentials can compensate each other to some degree, hence the corresponding scaling parameters are not entirely independent.

Because the resulting set of optimal (ED,ND) pairs includes the scaling value of the electrostatic desolvation term employed in previous HEWL BD simulations(*72*) (ED = 0.36), we opted for leaving ED at this default value and explored the effect of changing ND in more detail. In addition to the ND value used in HEWL BD simulations(*72*) (ND = -0.090 kcal/mol/Å^2), we report in the main text mcMC simulation results for the two ND values corresponding to the minimum B_2 RMSDs for ED = 0.4 (ND = –0.098 kcal/mol/Å^2 when comparing to B_2 experimental estimates at pH 4.7; ND = –0.0100 kcal/mol/Å^2 RMSD when comparing to B_2 experimental estimates at pH 6.9).



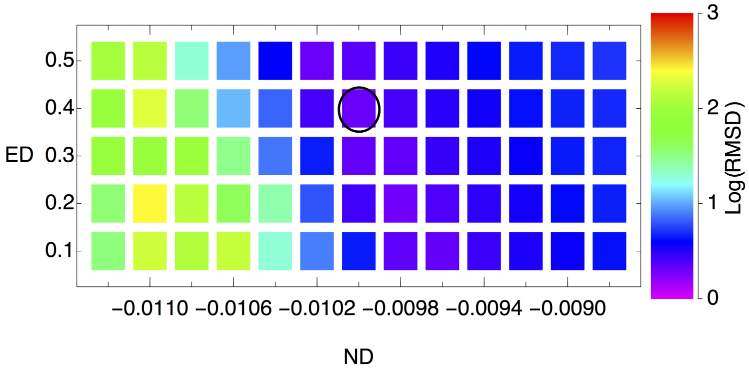


Figure. S1.2: Logarithm of the root-mean squared differences (RMSD) between B_2 estimates from mcMC HEWL simulations (using the indicated desolvation potential parameters) and experimental estimates at pH 4.7 (top) and pH 6.9 (bottom) and ionic strengths of 100 mM and 300 mM. The optimal ND values (at ED = 0.4) are encircled.

CHAPTER 2

The role of conformational flexibility in the simulation of many-protein systems

This is a manuscript in preparation by B.Majumdar, V.D. Prytkova, J.Freites, M.Heyden, D.J.Tobias.

INTRODUCTION

The crowded conditions encountered by biomacromolecules in cellular environments dominate their mobility and stability as well as the nature of their interactions (*1*). Experimental studies involving biomolecules are mostly carried out *in vitro* using buffer solutions, often neglecting the effects of macromolecular crowding. Several experimental techniques (*73-79*) have recently become available to study biomolecular interactions in crowded media both *in vitro* and *in vivo*. Atomistic simulations of such crowded systems in explicit solvent are still restricted to small numbers of biomolecules due to often prohibitive computational costs.(*17, 80*) Computationally inexpensive modeling approaches, such as rigid-body Brownian Dynamics and Monte Carlo (MC) simulations in implicit solvent, can provide a powerful alternative to fully atomistic simulations.(*19, 31, 72*) However, the accuracy of rigid-body simulations is limited due to the neglect of conformational fluctuations. Recently, Prytkova et.al. (*81*) described an efficient method to introduce conformational flexibility in MC simulations of many-protein systems. In this approach, termed multi-conformation Monte Carlo (mcMC), protein flexibility is incorporated via a conformational swap trial move from a pre-determined library of structures. In principle, the library of structures used in mcMC simulations can be generated by any suitable experimental or computational method that samples protein conformations. The probability of a conformational swap trial move can be used to describe free energy differences between protein conformations in the infinite dilution limit. The mcMC sampling then takes into account changes in the relative stability of conformations due to interprotein interactions at increased concentration.

While the mcMC approach represents an improvement compared to simulations based on a single conformation rigid-body simulation (*81*), a limited-size, discrete library of protein structures remains an approximate representation of the configurational space that describes the structure and dynamics of polypeptide chains. A natural question to ask in this context is how many protein structure independent samples are needed to effectively describe molecular flexibility using a finite library of protein conformations? An associated query is how much conformational sampling is required to adequately capture the portions of the protein configurational landscape that may affect protein-protein interactions. Here, we address these questions in the case of library of structures generated using all-atom molecular dynamics (MD) simulation of a single protein in explicit solvent under infinite dilution conditions. We perform mcMC simulations of 10 mg/ml solutions of hen egg white lysozyme (HEWL), with a library of conformations generated using enhanced-sampling methods; and human γD-crystallin, using long-timescale trajectories to generate the library of conformations. In addition, we consider the case of bovine γB-crystallin, which is homologous to γD-crystallin but with no net charge. We find that structural observables computed from the mcMC simulations converge rapidly for weakly interacting systems such as the γ-crystallin solutions or for solutions of HEWL under conditions where electrostatic repulsion dominates protein-protein interactions. In contrast, the specifics of the construction of the library of protein conformations become key in the structural characterization of strongly attractive systems such as HEWL solutions a high ionic strength.

METHODS

Protein-protein interaction potential. We employ the grid-based protein-protein interaction potential introduced by Mereghetti et al.(*72*) for many-protein BD simulations in the SDAMM software package(*82*), which consists of four contributions: the electrostatic interaction energy due to the electrostatic potential of one protein with the charges of a second protein;(*83*) an electrostatic desolvation penalty due to bringing solvated polar groups of one protein into a low dielectric environment upon binding another protein with the simultaneous loss of their solvation shell;(*84*) a short-range attractive non-polar desolvation interaction due to the burial of solvent exposed hydrophobic surface atoms of one protein by a second protein;(*82*) and a soft-core repulsive interaction energy between atoms from different proteins. The two electrostatic contributions depend explicitly on the solution ionic strength. The non-polar desolvation interaction can be modified by varying a scaling factor in the potential, which is used to convert the protein buried area into an energy. Proteins are modeled using an atomistic representation. All the terms of the interaction potentials were computed on 200 × 200 × 200 grids with a grid constant of 1Å.

The electrostatic potential grids were computed at different ionic strengths (see Supplemental Information) with atomic charges according to the OPLS force field(*85*) by finite difference solution of the linearized Poisson-Boltzmann equation using the UHBD(*86*) software package. Dielectric constants of 78.4 and 2.0 were used for solvent and protein, respectively. A reduced number of charged sites in each protein was used during the evaluation of the electrostatic potential energy with charge magnitudes calculated using the effective charge approximation (*83*) implemented in the SDAMM software package. We used the SDAMM potential function parameters reported by Prytkova et al. (*81*), except for the simulations of γB-crystallin, where a nonpolar desolvation parameter of −0.0082 kcal/mol/Å^2 was used in order to match a second virial coefficient experimental datum (*87*).

Multi-conformation Monte Carlo simulation. The multi-conformation Monte Carlo (mcMC) algorithm extends the conventional rigid-body simulation scheme by introducing protein conformational flexibility through a trial move that attempts to swap conformations between a protein in the simulation system and a discrete library of protein structures. In this context, under the canonical ensemble, the probability of finding the system in the state i, p_i, is proportional to

$$p_i \sim w_i \exp\left[-\frac{E_i}{k_B T}\right] \tag{2.1}$$

where E_i is the corresponding protein-protein interaction energy and w_i is a statistical weight, independent of the Monte Carlo sampling, given by

$$w_i \sim \exp\left[-\frac{G_i}{k_B T}\right] \tag{2.2}$$

where G_i is the *intrinsic* free energy of all the protein conformations present in the state i, which includes all intramolecular and protein-solvent interactions in the infinite dilution limit.

The underlying transition matrix element corresponding to the conformational swap move from state i to state j, α_{ij}, is given by

$$\alpha_{ij} = w_j \tag{2.3}$$

In other words, in contrast to rigid-body translational and rotational trial moves, the Markov chain underlying the conformational swap trial move consists of independent events.

Imposing detailed balance leads to following acceptance criterion

$$\min\left(1, \frac{\alpha_{ji}}{\alpha_{ij}}\frac{p_j}{p_i}\right) = \min\left(1, \exp\left[-\frac{\Delta E}{k_B T}\right]\right) \tag{2.4}$$

where $\Delta E = E_j - E_i$, which is identical to the acceptance criterion associated with the rigid-body motion trial moves. Thus, only the ordinary Metropolis[21] acceptance criterion is needed in mcMC simulations.

The mcMC algorithm has been implemented (*81*) into a customized version of the SDAMM software package (*82*). In this implementation, a Monte Carlo cycle comprises as many trial moves as the total number of proteins in the system. In a trial move, a rigid-body translation, rigid-body rotation, or conformational swap is chosen with equal probability, and applied to a single protein in the simulation system also selected with equal probability. Rigid-body translational and rotational trial moves are adjusted to yield an acceptance ratio of approximately 50%. In the conformational swap trial move an alternative protein conformation is chosen from the protein structure library with probability given by (2.3). The library of protein conformations and the corresponding probability distribution are sampled from extensive all-atom MD simulations of a single protein in explicit solvent, as described in the next section.

All mcMC simulations were carried out at a concentration of 10 mg/ml using 200 protein molecules. The HEWL and γB-crystallin simulations were run at 300 K, and the γD-crystallin simulations were run at 310 K. See Supporting Information for more details.

Generation of protein structure libraries. The library of protein structures employed in the mcMC simulation can be generated using several methods. Here, we used a 400-ns replica exchange molecular dynamics (REMD) simulation for HEWL, and conventional equilibrium MD simulation trajectories of 40 μs and 500 ns for γD- and γB-crystallin, respectively, all under infinite dilution conditions. To generate the library of protein structures, we computed the heavy atom RMSD matrix between protein configurations over selected portions of each trajectory (see Table S2.1), and performed a cluster analysis using the algorithm of Daura et al. (*48*), implemented in the GROMACS software package(*49*), with a RMSD cutoff of 1 Å for HEWL and γB-crystallin, and 1.7 Å for γD-crystallin. The configurations corresponding to the cluster centroids constitute the library of conformations, and the respective cluster population sizes were used to construct the underlying probability distribution used in the conformational swap trial move (*cf.* equations 2.2 and 2.3).

A 400-ns REMD simulation was performed in explicit solvent starting from a solution NMR structure of HEWL (PDB ID 1E8L).(*47*) The protonation states of the protein at pH 7 were set using the H++ webserver.(*88*) A temperature range of 300-375K was chosen and a REMD temperature generator webserver (http://folding.bmc.uu.se/remd/) was used to determine the optimum number of replicas.(*89*) REMD trajectories for 22 replicas were generated using the GROMACS 4.5.6 software package.(*49*) The OPLS all-atom force field(*85*) was used for the protein and ions, and the TIP3P model was used for water(*90*). The system was neutralized with eight chloride ions and solvated by 8819 water molecules. Covalent bonds and the geometry of water molecules were constrained with the LINCS (*53*) and SETTLE (*53*) algorithms, respectively. Short-range interactions were truncated with a 9Å cut-off and long-range electrostatics were computed using the fast smooth particle-mesh Ewald method (*51*) on a 1.2 Å grid. After energy minimization, the system was equilibrated for 100 ps with a 2 fs time step at constant temperature and pressure applying harmonic restraints on the protein heavy atom positions using a force constant of 1000 kJ mol^{-1} nm^{-2}. This was followed by a consecutive 1-ns unrestrained equilibration. During the equilibration, a Berendsen weak coupling thermostat and barostat (*54*) were used with a target temperature of 300K and a pressure of 1 bar respectively with a 1 ps time constant. The REMD simulation was performed with a 2 fs time step with an exchange attempt every 200 fs. Therein, the Nosé-Hoover thermostat (*55*) was used for temperature coupling and the Parrinello-Rahman barostat (*56*) was used for pressure coupling.

The γD-crystallin simulation system was built from the crystal structure reported by Basak et. al. (*9*) (PDB ID code 1HK0). The solution state NMR structure of the γD-crystallin P23T variant (PDB ID code 2KFB) (*91*) was used to determine the histidine protonation states for wild-type γD-crystallin. The γB-crystallin simulation system was built from the crystal structure reported by Kumaraswamy et al. (*92*) (PDB ID code 1AMM) with protonation states consistent with neutral pH. The CHARMM36 (*93*) force field was used for protein and ions, and the TIP3P model was used for water (*27*). The proteins were solvated in a cubic 80 Å x 80 Å x 80 Å box with the nearest box boundary at least 12 Å from the protein. Crystal structure waters were preserved and the γD-crystallin system was neutralized with chloride counterions. The resulting systems contain 48,309 and 48,222 atoms for γD-crystallin and γB-crystallin, respectively. The VMD 1.9.1 software package (*94*) was used to assemble the systems.

The initial part of the γD-crystallin simulation was first carried out for 20 ns using the NAMD 2.9 software package (*95*) following 10000 steps of conjugate-gradient energy minimization and a 200 ps MD simulation during which harmonic positional restraints placed on each protein heavy atom were gradually relaxed. The γB-crystallin simulation was run for 500 ns using the NAMD 2.11 software package, with a pre-equilibration consisting of 500 steps of conjugate-gradient energy minimization and a 500 ps harmonically restrained run. Both simulations were run at constant temperature (300 K for γB-crystallin and 310 K for γD-crystallin) and pressure (1 atm). The smooth particle mesh Ewald method (*51, 96*) was used to calculate electrostatic interactions. Short-range, real-space interactions were cut off at 11 Å by means of a switching function. A reversible, multiple time-step algorithm (*97*) was used to integrate the equations of motion with a time step of 4 fs for electrostatic forces, 2 fs for short-range nonbonded forces, and 1 fs for bonded forces. All bond lengths involving hydrogen atoms were held fixed using the SHAKE (*98*) and SETTLE (*53*) algorithms. A Langevin dynamics scheme was used for temperature control, and a Nosé-Hoover-Langevin piston was used for pressure control (*99, 100*).

The γD-crystallin simulation was extended for another 46-µs on the Anton 2 supercomputer, a special-purpose computer for molecular dynamics simulations of biomolecules (*101*)**.** The CHARMM36 force field (*93*) was used for protein and ions, and the TIP3P model was used for water (*90*). The reversible, multiple time-step algorithm (RESPA) (*102*) was used to integrate the equations of motion every three timesteps for long-range non-bonded forces, and every timestep for short-range nonbonded and bonded forces. The k-Gaussian split Ewald method(*103*) was used for long-range electrostatic interactions. All bonds lengths involving hydrogen atoms were fixed using the SHAKE (*98*) algorithm. The simulation was performed at constant temperature (310 K) and pressure (1 atm) using Nose-Hoover chains (*104*) and the Martyna-Tobias-Klein barostat (*104*). The RESPA algorithm and temperature and pressure controls were implemented using the multigrator scheme, allowing the simulation to run with a 2.5 fs time step (*105*). The last 40 µs of this trajectory were used in the generation of the protein structure libraries for mcMC (see Fig. S2.1).

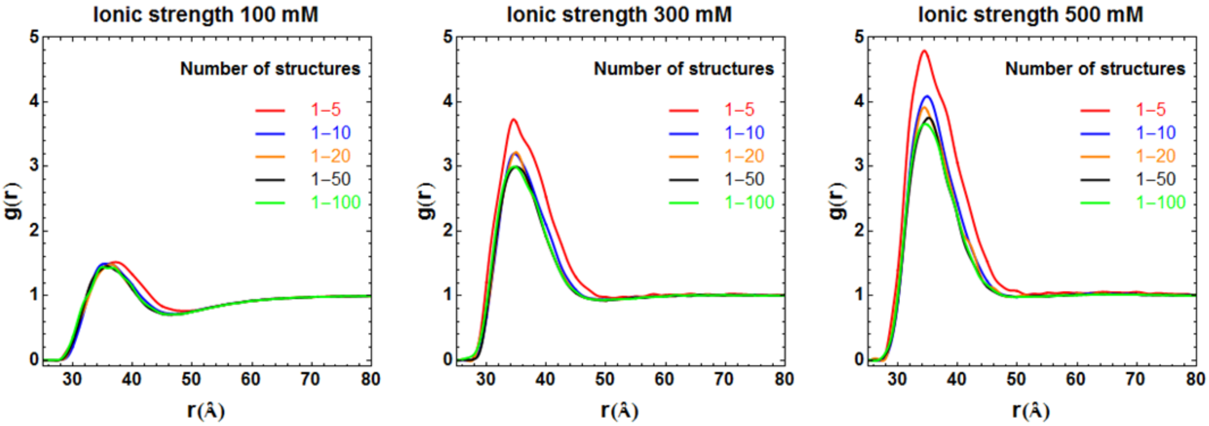


Figure 2.1: Protein-protein radial distribution functions from mcMC simulations of HEWL solutions with varying library sizes formed by the (5, 10, 20, 50, or 100) most populated conformations from a 400-ns REMD trajectory.

RESULTS AND DISCUSSION

Protein conformational flexibility is introduced in mcMC simulations by allowing molecules to swap conformations drawn from a library of structures generated, in this particular case, from all-atom MD simulations under infinite dilution conditions. To assess how the specifics of the library generation protocol affect the convergence of structural quantities computed from mcMC simulations of protein solutions, we performed simulations of 200 rigid molecules at 10 mg/ml of either HEWL, with libraries of conformations generated from a 400-ns REMD simulation, or γB- and γD-crystallin, with libraries of conformations generated from 40-μs and 500-ns MD simulation trajectories, respectively. Figure 2.1 illustrates the influence of the library size on the protein-protein interactions in HEWL solutions at various ionic strengths. Overall, the largest differences occur upon increasing the library size from five structures to ten structures. The results appear to be strongly dependent on the ionic strength, which is to be expected given that HEWL at neutral pH carries a net charge of +8e. At low ionic strength (100 mM), where the electrostatic repulsion between the proteins dominates, all of the libraries produce essentially identical results except for the one with five structures, which exhibits a small shift of the maximum of the radial distribution function. For higher ionic strengths, the structural details of the interacting proteins become more relevant as the electrostatic repulsion is increasingly screened and close encounters become more important. At both 300 and 500 mM ionic strength, simulations based on a library with only five structures result in an increased overall attraction between the proteins, as evidenced by peak shape a maximum value of the radial distribution function. Larger library sizes yield radial distribution functions that are more uniform and with lower maximum values as the library size is increased. The strongest dependence of the protein-protein radial distribution function with the library size occurs at 500 mM ionic strength where protein interactions are expected to be most attractive.

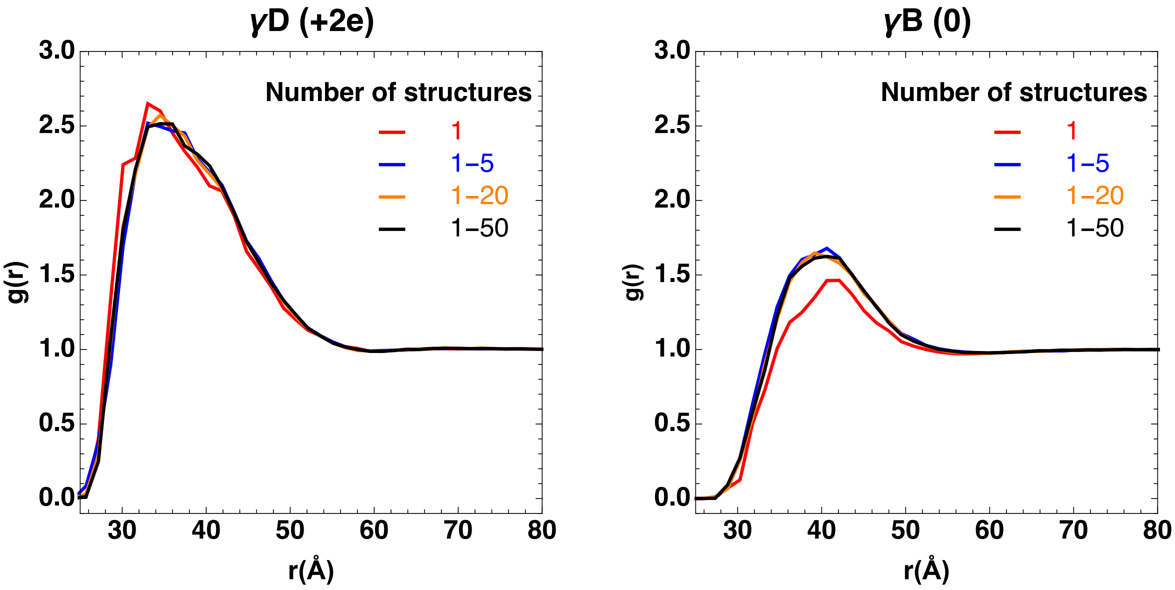


Figure 2.2: Protein-protein radial distribution functions from mcMC simulations of γD- and γB-crystallin solutions with varying library sizes formed by the (1, 5, 10, 20, or 50) most populated conformations from a 40-μs and 500-ns MD trajectory, respectively.

Because γD- and γB-crystallin have a much lower net charge at neutral pH than HEWL (+2e and 0, respectively) protein-protein interactions are already dominated by attractive dispersion forces at low ionic strength. Consistent with the results of single-conformation rigid-body MC simulations of HEWL reported by Prytkova et.al.,(*81*) use of a single protein structure in either crystallin simulation system results in a nonuniform protein-protein radial distribution function that suggest the presence of artifacts in the underlying interprotein interactions (Fig. 2.2). Notably, in contrast to HEWL, the size of the protein structure library has no effect on the corresponding radial distributions (Fig. 2.2).

As described in detail in the Methods section, the transition matrix of the conformational swap trial move is generated from the relative population sizes of the MD simulation trajectory protein conformation clusters whose centroids constitute the mcMC protein structure library.

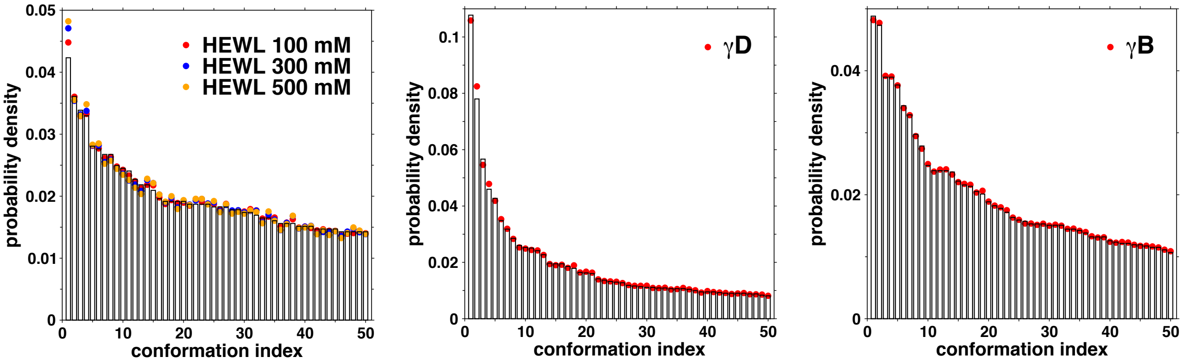


Figure 2.3: Prior probability distribution function for the 50-structure libraries (shown as bars) compared to the observed probability distribution function in mcMC simulations at 10 mg/ml (shown as dots).

Within the timescale of each MD simulation trajectory, these cluster population sizes essentially encode the relative stability of each protein conformation included in the library under infinite dilution conditions, *i.e.* in absence of protein-protein interactions. In the analyses reported in Figs. 2.1 and 2.2, the protein structures were added

to the library in a ranked fashion, starting with the centroid of the most frequently observed conformational cluster in the corresponding MD simulation. Thus, increasing the size of the structure library only adds protein configurations that are less likely to be encountered, at least in the absence of protein-protein interactions. As shown in Fig. 2.3, the sampling frequency of each protein structure in the mcMC simulation system coincides with the corresponding prior probability in the library, irrespective of their specific statistical weight. Chi-square goodness-of-fit tests (results not shown) indicate that each observed per protein probability distribution constitutes a statistical sample of the corresponding library of conformations prior distribution. Nevertheless, the similarity between the library prior probability distribution and the distribution sampled during the mcMC simulations, as measured by the Euclidean distance between them (see Fig. 2.4), are positively correlated with the strength of the protein-protein interactions. Specifically, in the HEWL simulations, the mcMC sample distributions deviations from the library prior probability distribution increases with ionic strength in the HEWL simulations. Similarly, the γB-crystallin system, which exhibits the weakest protein-protein interactions generates the closest mcMC samples to the library probability distribution.

These results suggest that under conditions where few protein pairs are formed, radially resolved interactions between proteins do not necessarily require a detailed description of the protein conformational flexibility. In the case of HEWL, the number of protein-protein pairs increases with ionic strength, and the size of the protein library becomes increasingly relevant (see Fig. 2.1).

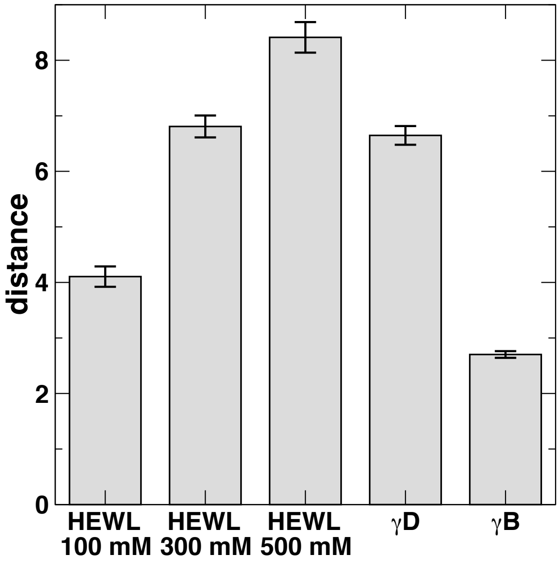


Figure 2.4: Euclidean distance between the probability distribution of the 50-structure library and the corresponding sampled distribution from mcMC simulations. In all cases, the mean value and standard error were computed from samples of 1000 Monte Carlo cycles over the whole simulation.

Simulations of protein solutions with the mcMC algorithm implicitly include stabilization effects of specific protein conformations due to protein-protein interactions (*81*). Hence, even protein conformations that are highly unlikely to be observed in dilute conditions, might become relevant in the formation of multiprotein configurations, *e.g.* protein dimers, if they are sufficiently stabilized energetically by interprotein interactions. When these interactions become more attractive, the inclusion of conformations that are less populated in infinite dilution conditions becomes more important, as they might be stabilized in complexes even at the relatively low protein concentration of 10 mg/ml employed in our simulations. Correspondingly, libraries containing 50 protein structures are required to obtain a converged protein-protein radial distribution at an ionic strength for HEWL at 300 mM ionic strength, while at 500 mM ionic strength, small changes in the protein-protein radial distribution function are still observed upon increasing the size of the structure library from 50 to 100.

The length of the MD simulation trajectory used to generate the library of structures is a complementary aspect to the library size that may also play key role in the description of protein conformational flexibility in mcMC simulations. Figure 2.5 shows the protein-protein radial

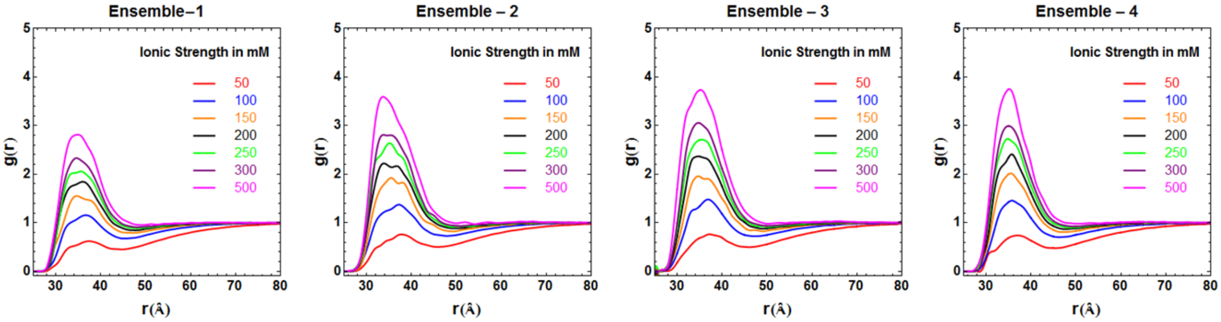


Figure 2.5: Protein-protein radial distribution functions from mcMC simulations of HEWL solutions with a library of 50 protein structures sampled using four different REMD trajectory lengths (from left to right: 100, 200, 300 and 400 ns).

distribution function from simulations of HEWL at varying ionic strengths generated with a library of 50 protein structures sampled from 100 ns, 200 ns, 300 ns and 400 ns of REMD trajectory. As indicated before, the monotonic increase of the maximum value of the radial distribution function with increasing ionic strength can be attributed to electrostatic screening. The results indicate qualitative changes in the average protein-protein interactions as the REMD sampling for the generation of the structural library is increased, most notably at higher ionic strengths. In particular, at 500 mM ionic strength there is an increase in the strength of the protein-protein interaction as the REMD sampling is increased from 100 ns to 200 ns. This indicates the occurrence of conformations within 200 ns of REMD sampling, which enhance protein-protein interaction and are not sufficiently explored within the first 100 ns, and is consistent with an analysis of the protein root mean squared fluctuations in the REMD trajectory (see Fig. S2.2). No major change in the coordination number is observed for longer REMD trajectory lengths but changes in the radial distribution peak shape suggest the formation of different protein dimer configurations with protein libraries of equal size generated from different REMD trajectory lengths. The results for HEWL (Figs. 2.1 and 2.5) indicate that dimers are the dominant protein complexes at 10 mg/ml, in agreement with previous simulations and experimental studies.[11,49]

Protein-protein radial distribution functions from mcMC simulations of γD-crystallin generated with a library of 50 protein structures sampled from MD trajectory lengths of 500 ns through 40 µs are essentially independent on the extent of sampling by MD (Fig. 2.6). This is due in part to the limited internal conformational dynamics of this protein fold (see Fig. S2.1). Taken together the

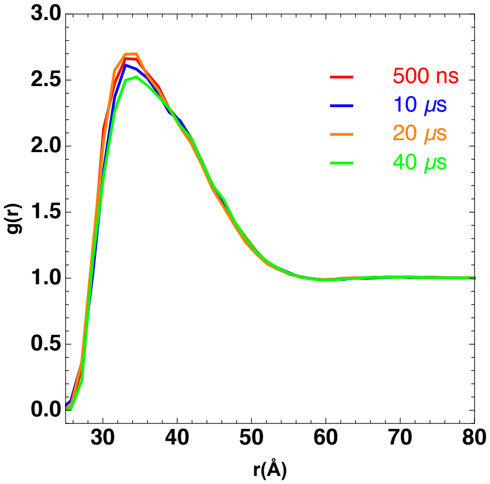


Figure 2.6: Protein-protein radial distribution functions from mcMC simulations of γD solutions with libraries of 50 protein structures sampled using four different MD trajectory lengths.

mcMC results for γD-crystallin (Figs. 2.2 and 2.6), suggest that in weakly-interacting systems mcMC simulations do not require extensive conformation libraries. Overall, the radial distribution function main peak shape for all the systems considered here suggest that protein pairwise configurations develop through specific relative orientations and/or interaction sites. To assess the effect of the mcMC protein structure library on the sampling of specific modes of protein-protein interaction, we computed spatial distribution functions (SDF) from the HEWL system at 100 mM ionic strength and the γD-crystallin crystalline system as a function of the protein structure library underlying conformational sampling (Figs. 2.7 and 2.8). SDFs represent the spatial density distribution of protein centers-of-mass around a tagged protein at the origin, normalized by the bulk density (*81*). In Figure 2.7, we compare HEWL SDF isosurfaces from mcMC simulations with protein structure libraries sampled from REMD simulations of increasing trajectory length, the observed preferred interaction sites tend to be more populated in the simulations with libraries sampled from 200 ns or longer REMD trajectories, which is consistent with the corresponding radial distribution functions (Fig. 2.5). Although there are minor differences among the SDFs, the main modes of interaction appear to be preserved after structures from the first 200 ns of the REMD conformational sampling are included into the structural library used to describe intramolecular flexibility.

The SDF analysis for γD-crystallin reveals a dominant interaction surface on the C-terminal domain that is essentially independent of the extent of conformational sampling used to generate the protein structure library (see Fig. 2.8). There are two secondary interaction regions, one on either protein domain, whose extent and specific location vary among the simulations. This variability may be attributed to the accumulated effect of rare conformational excursions that occur at intervals on the order several μs (see Fig. S2.1).

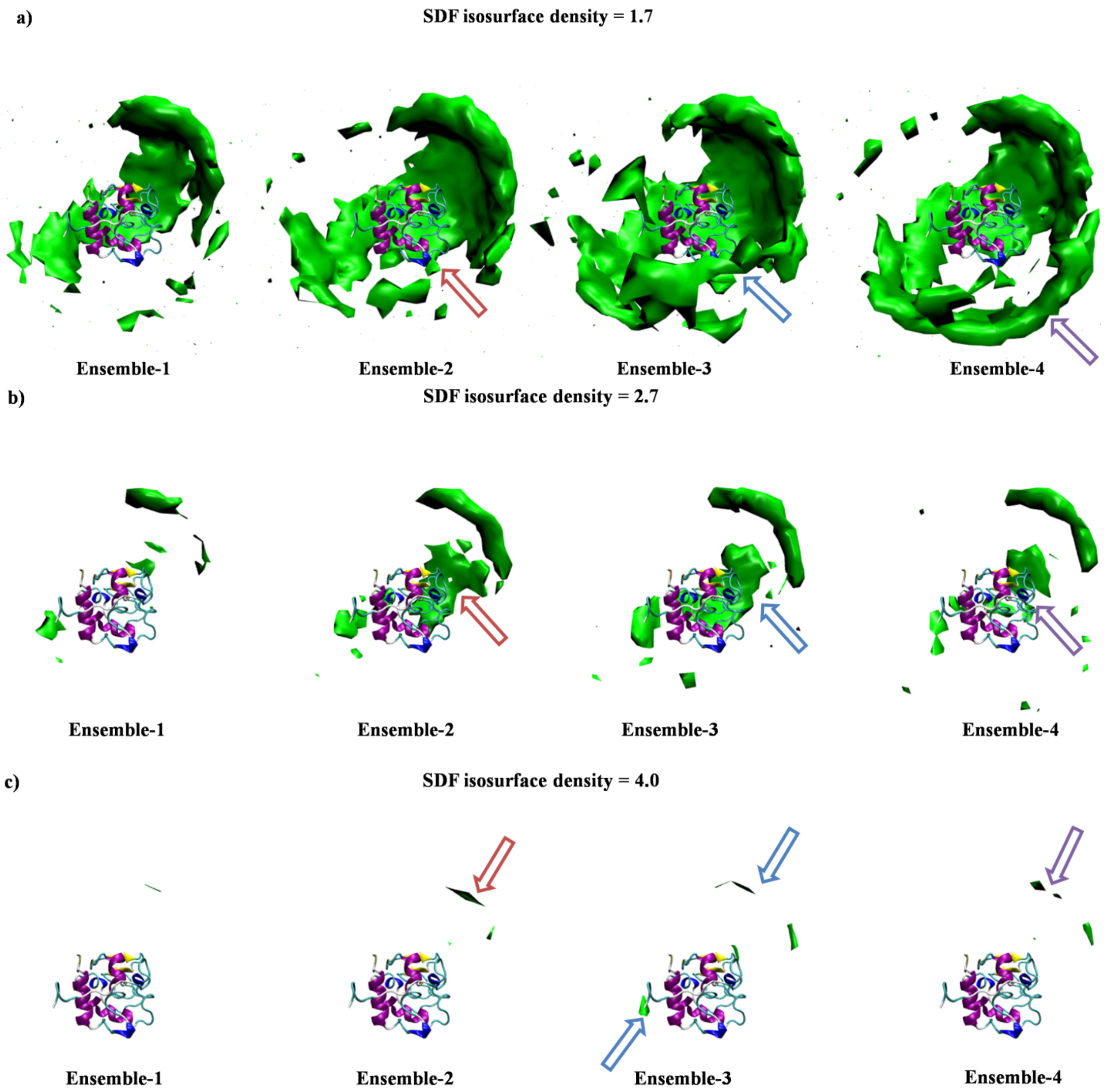


Figure 2.7: Spatial distribution functions from mcMC simulations of HEWL solutions generated with a library of 50 protein structures sampled using four different REMD trajectory lengths (from left to right: 100, 200, 300 and 400 ns). a) SDF isosurface at 1.7 times the bulk density, b) SDF isosurface at 2.7 times the bulk density, c) SDF isosurface at 4.0 times the bulk density. The arrows point to the dominant interaction sites in each simulation**.** The protein structure with the highest statistical weight sampled from the 400-ns REMD trajectory is shown in all panels in the same orientation.

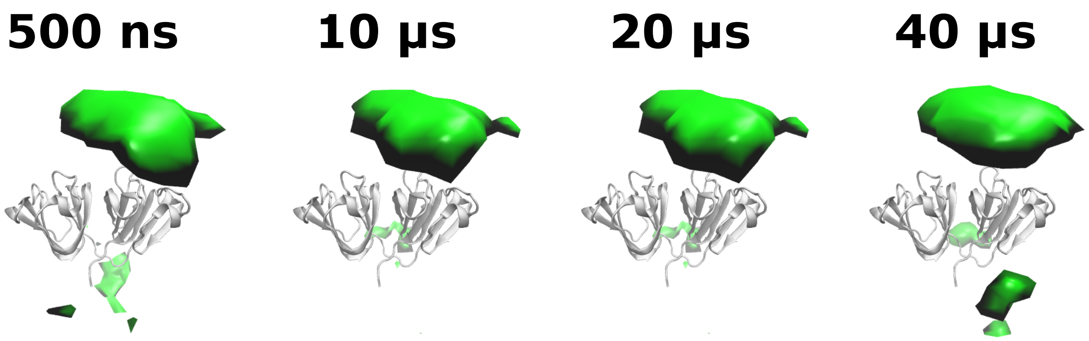


Figure 2.8: Spatial distribution functions from mcMC simulations of γD-crystallin solutions with libraries of 50 protein structures sampled using four different MD trajectory lengths. The isosurface contours (shown in green) correspond to 7.5 times the bulk density. The protein structure with the highest statistical weight sampled from the 40-μs MD trajectory is shown in all panels in the same orientation.

SUMMARY

In the characterization of protein-protein interactions using Monte Carlo simulations of many-protein systems, mcMC augments the computational efficiency of rigid-body modeling approaches with increased realism through the incorporation of a novel MC trial move, which approximates the intramolecular degrees and interactions between a single protein and its solvation shell via a finite-size library of protein conformations. Moreover, mcMC eliminates the undesirable dependence, observed in rigid body MC simulations, on the initial structure chosen to initiate the simulation (*81*). The conformational library is generated using atomistic MD simulations under infinite dilution conditions. We demonstrated here that, in the case of weakly-interacting systems or when electrostatic repulsion dominates interprotein interactions, the library size and the extent of conformational sampling involved in its generation do not play a crucial role in the description of protein-protein interactions. These factors do become relevant under conditions in which protein-protein interactions are increasingly attractive (*e.g.* high ionic strength HEWL solutions). We find that insufficient conformational sampling in the generation of the library of structures results in a pronounced underestimation of attractive protein-protein interactions. However, we also find that specific high-affinity pairwise configurations do not change significantly with prolonged sampling. This result is encouraging because it demonstrates that information on the most relevant modes of interactions between the proteins, which may lead to dimer and oligomer formation and eventual aggregation, can already be obtained with a limited description of the protein conformational flexibility. Predicted dimer structures can then be used as starting points for additional simulations with fully flexible proteins, *e.g.* using atomistic MD, which can then be used to study the interactions in more detail.

SUPPLEMENTAL INFORMATION

Table S2.1. Multi-conformation Monte Carlo (mcMC) simulations of hen egg white lysozyme 10 mg/ml solutions[a]

REMD[b] trajectory length (ns)	Library Size (number of protein conformations)
400	5[c]
400	10[c]
400	20[c]
400	50[d]
400	100[c]
300	50[d]
200	50[d]
100	50[d]

[a] All simulations systems consisted of 200 proteins and were run for 0.6 x 10^6 MC cycles. Every MC cycles comprised 200 trial moves. [b]The structure libraries used in the mcMC simulations were generated from

an all-atom REMD simulation of the specified trajectory length, and with the PDB ID 1E8L structure (net charge +8e) as initial configuration. [c]Separate simulations were performed at 100, 300, and 500 mM ionic strength. [d]Separate simulations were performed at 50, 100, 150, 200, 250, 300, and 500 mM ionic strength.

Table S2.2. Multi-conformation Monte Carlo (mcMC) simulations of gD-crystallin 10 mg/ml solutions[a]

MD[b] trajectory length (μs)	Library Size (number of protein conformations)
40	1
40	5
40	20
20	50
10	50
0.5	50

[a] All simulations systems consisted of 200 proteins and were run for 4.0 x 10^6 MC cycles. Every MC cycles comprised 200 trial moves. [b]The structure libraries used in the mcMC simulations was generated from an all-atom MD simulation of the specified trajectory length, and with the PDB ID 1HK0 structure (net charge +3e) as initial configuration.

Table S2.3. Multi-conformation Monte Carlo (mcMC) simulations of gB-crystallin 10 mg/ml solutions[a]

MD[b] trajectory length (ns)	Library Size (number of protein conformations)	MC cycles[c] (x10^6)
500	1	2
500	5	2
500	20	2
500	50	3

[a] All simulations systems consisted of 200 proteins. [b]The structure libraries used in the mcMC simulations was generated from an all-atom MD simulation of the specified trajectory length, and with the PDB ID 1AMM structure (no net charge) as initial configuration. [c]Every MC cycles comprised 200 trial moves.

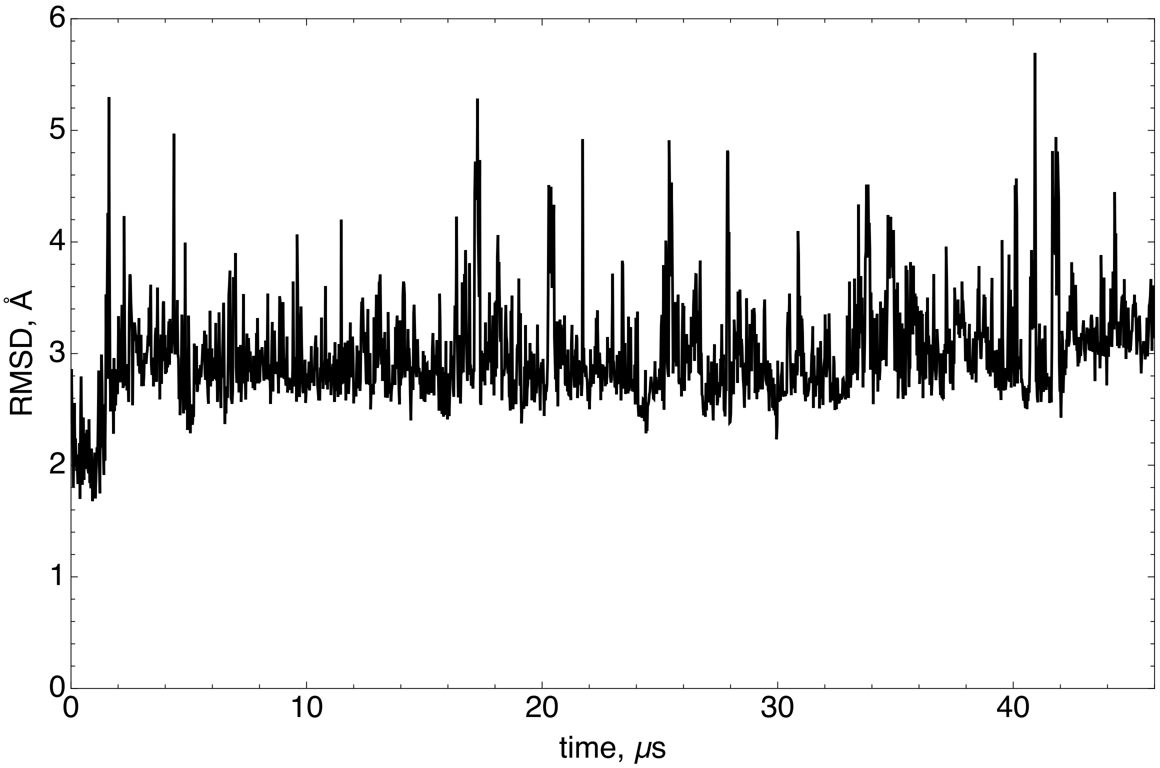


Figure S2.1: Evolution of the γD-crystallin heavy atom root mean squared deviation (RMSD) from the initial configuration in a 46-μs all-atom MD trajectory under infinite dilution conditions. The first 6 μs of the trajectory exhibit a transient and were not included in the generation of mcMC conformation libraries. The portions of the trajectory used in the RMSD cluster analysis correspond to 500 ns, 10 μs, 20 μs, and 40 μs after the first 6 μs of the trajectory.

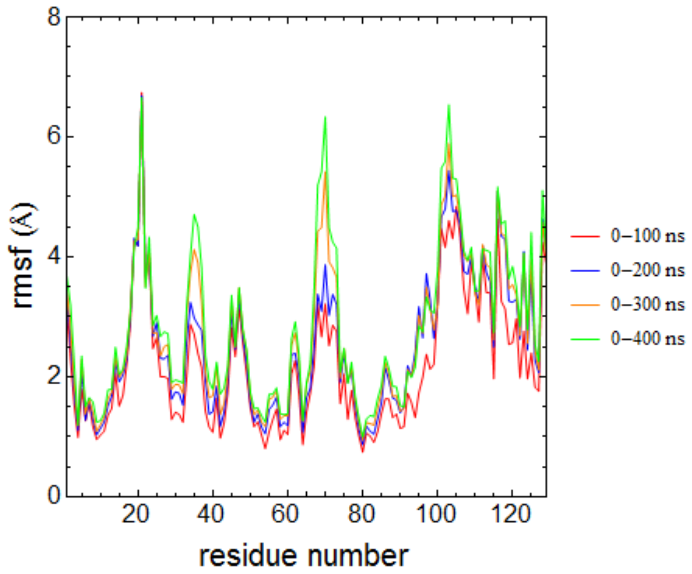


Figure S2.2: HEWL heavy atom root mean squared fluctuations (RMSF) for four portions of a 400-ns REMD trajectory. The RMSF converges for most residues after the first 100 ns.

CHAPTER 3

Multi-conformational ensemble refinement for high concentration protein simulations

An ensemble of input conformations for mcMC simulations is chosen by running an MD simulation of a protein at infinite dilution and selecting protein conformations most likely appearing in that simulation based on sidechain fluctuations. However, at higher protein concentrations, the protein may take different conformations due to crowding and protein-protein interactions that never appear in simulations at infinite dilution. Ideally, we'd need to run an MD simulation at a concentration similar to an intended concentration in mcMC simulation, however such calculations are usually computationally expensive at a time range necessary to sufficiently sample the conformational space of a protein.

As was mentioned in chapter 1, each conformation sampled by MD simulation is assigned a weight proportional to the frequency of occurrence of this conformation in a simulation. This weight is used in mcMC simulations to limit how frequently the conformation shows up. After running mcMC simulation, the agreement between input weights and the frequency of conformational occurrence can be checked.

The convergence of distribution of input weights and output conformational frequency will be checked for γS crystallin. In case of a low concentration mcMC simulation, such as 10 g/L, the agreement is exact, as presented in Fig. 3.1. This result is expected since infinite dilution MD simulation, from which the conformations of γS crystallin were extracted are representative for mcMC simulation at 10 g/L.

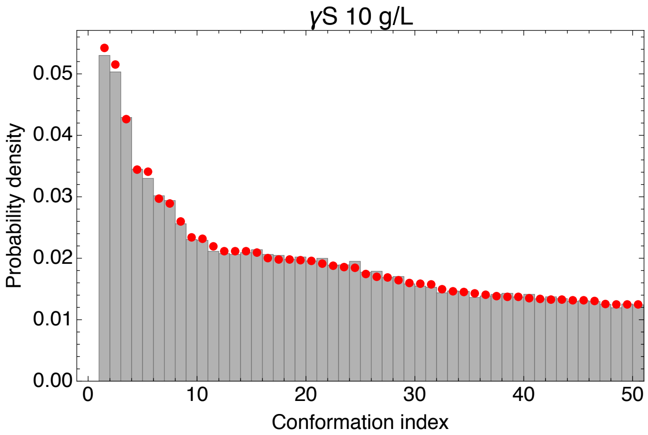


Figure 3.1: In red – the input weights of every conformation in mcMC library. In gray – the probability of each conformation showing up as the result of mcMC simulation of wild-type gS protein at 10 g/L.

However, in case of a much higher concentration, such as 200 g/L, differences between distribution of input weights and output conformational frequency become apparent (Fig.3.2).

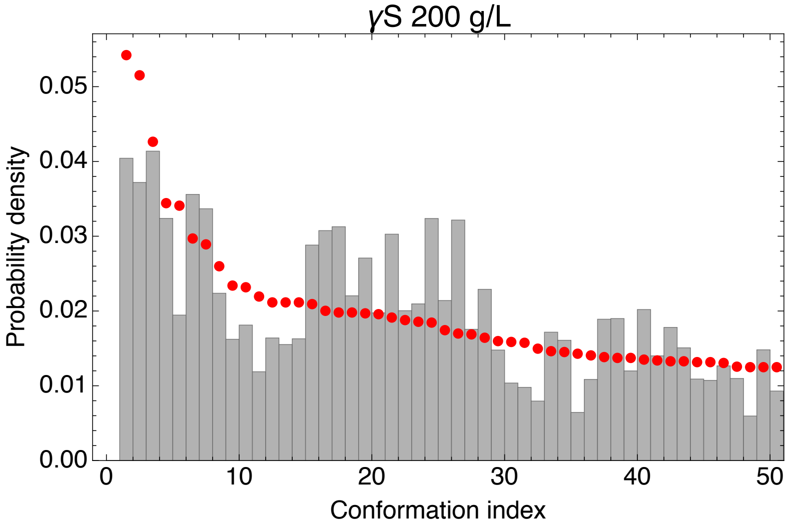


Figure 3.2: In red – the input weights of every conformation in mcMC library. In gray – the probability of each conformation showing up as the result of mcMC simulation of wild-type γS protein at 200 g/L.

We set to test the following hypothesis. If at higher protein concentration, the conformations of proteins are the same as the ones at lower concentrations, but the probability of them appearing is different, we can update the distribution of input weights in mcMC simulation until it matches the output conformational frequency. The input ensemble was updated every 2x10^4 MC cycles. The output conformational frequency of one run would become the set of input weights for the next run. Each time the two were compared.

In addition, we wanted to find if this way we can assign the weights for NMR conformations of wild-type γS protein (*13*). In this case the initial input weights were all set equal since no information about weights could be obtained.

The results of these computer simulations are presented in figures 3.3 and 3.4. In figure 3.3 γS crystallins MD conformational library was used. After 2x10^5 mcMC cycles, we can observe that only a small number of conformation stays, while the weights of some others appear to be zero or very close to zero. This situation is even more extreme in a simulation of NMR ensemble. Conformation number 17 has the highest weight in this ensemble.

If we consider the energy per protein in figures 3.5 and 3.6, we can see that particular conformations are chosen to minimize the energy of the system. Unfortunately, instead of reweighting the entire ensemble to create new weights more representative of a higher concentration, this method selects only a few conformations defeating the purpose of mcMC, that is to make proteins flexible by introducing multiple conformations. In fact, it is unclear, that selecting a conformation of a protein that leads to maximum aggregation and therefore to lowest energy of the protein system is representative of a realistic protein behavior at high concentrations.

Reflecting on this experience, it rather makes sense to establish the particular selection of conformations for the library and weights assigned to each conformation not through the mcMC simulation, but by running an MD simulation at an appropriate concentration.

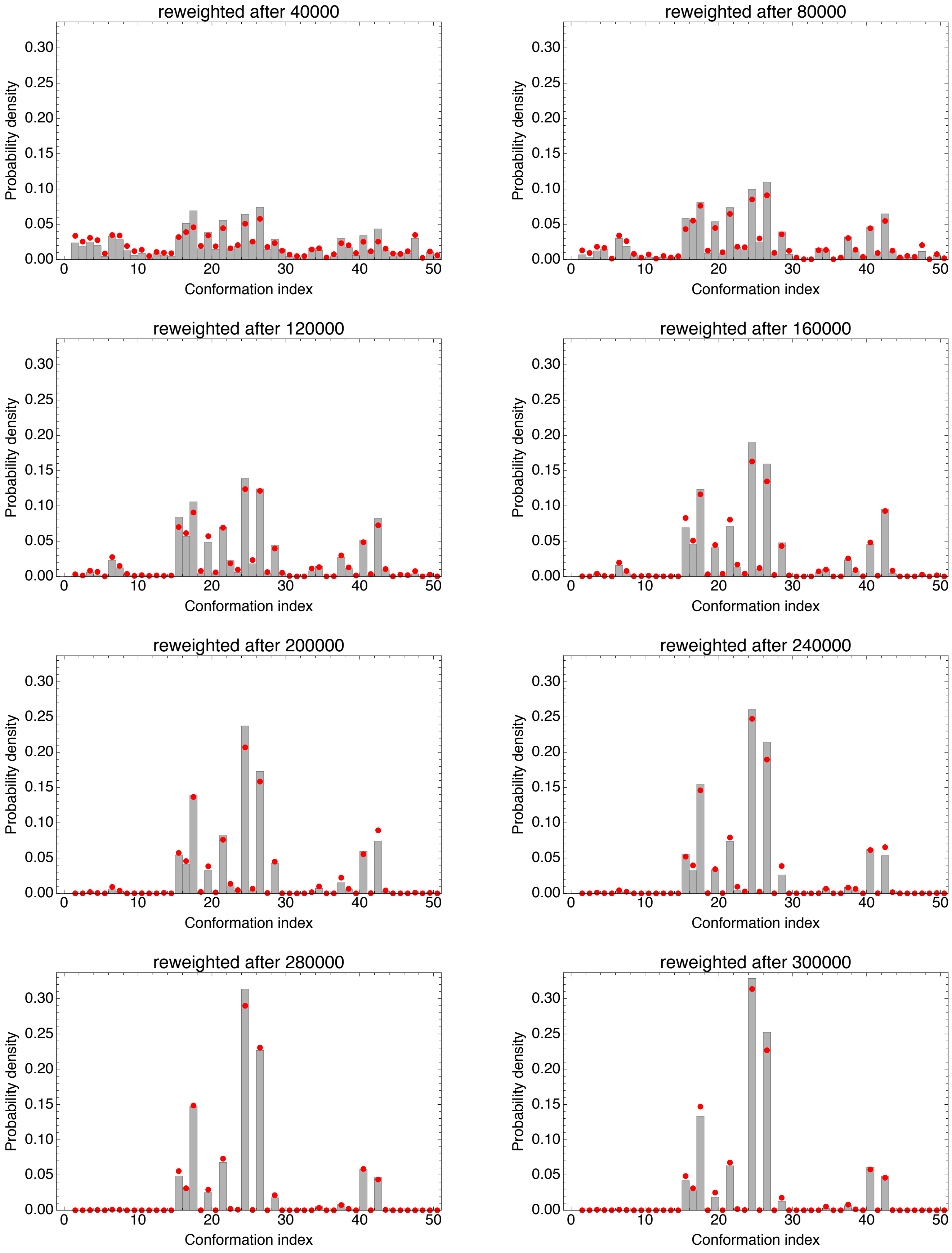

reweighted after 40000
reweighted after 80000
reweighted after 120000
reweighted after 160000
reweighted after 200000
reweighted after 240000
reweighted after 280000
reweighted after 300000
Probability density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Conformation index
0
10
20
30
40
50

Figure 3.3: In red – the input weights of every conformation in mcMC library for a previous run of 20000 mcMC cycles. In gray – the probability of each conformation showing up as the result of mcMC simulation of wild-type γS protein at 200 g/L for a conformational library obtained from MD simulations.

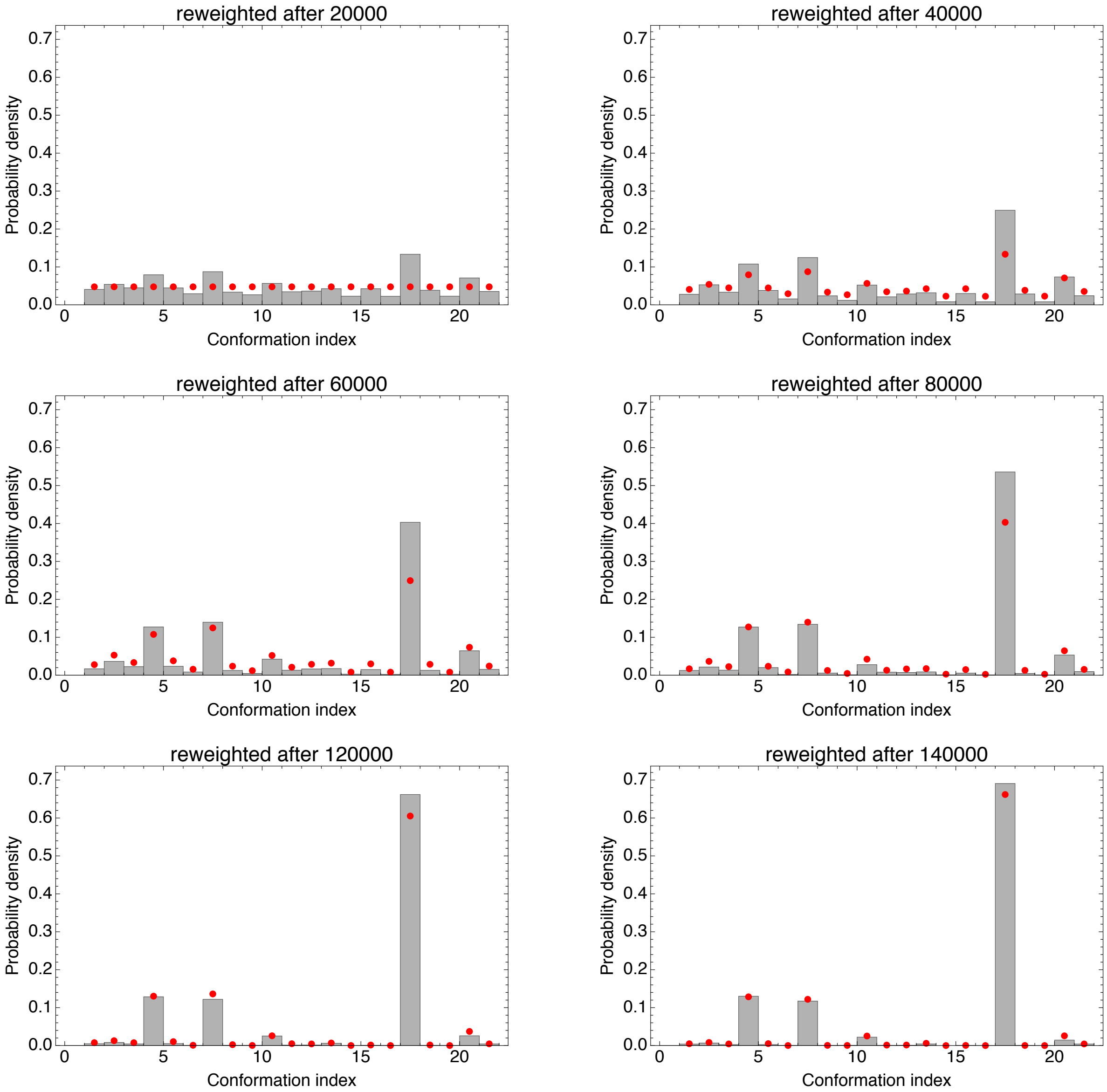


Figure 3.4: In red – the input weights of every conformation in mcMC library for a previous run of 20000 mcMC cycles. In gray – the probability of each conformation showing up as the result of mcMC simulation of wild-type γS protein at 200 g/L for a conformational library obtained from NMR experiments.

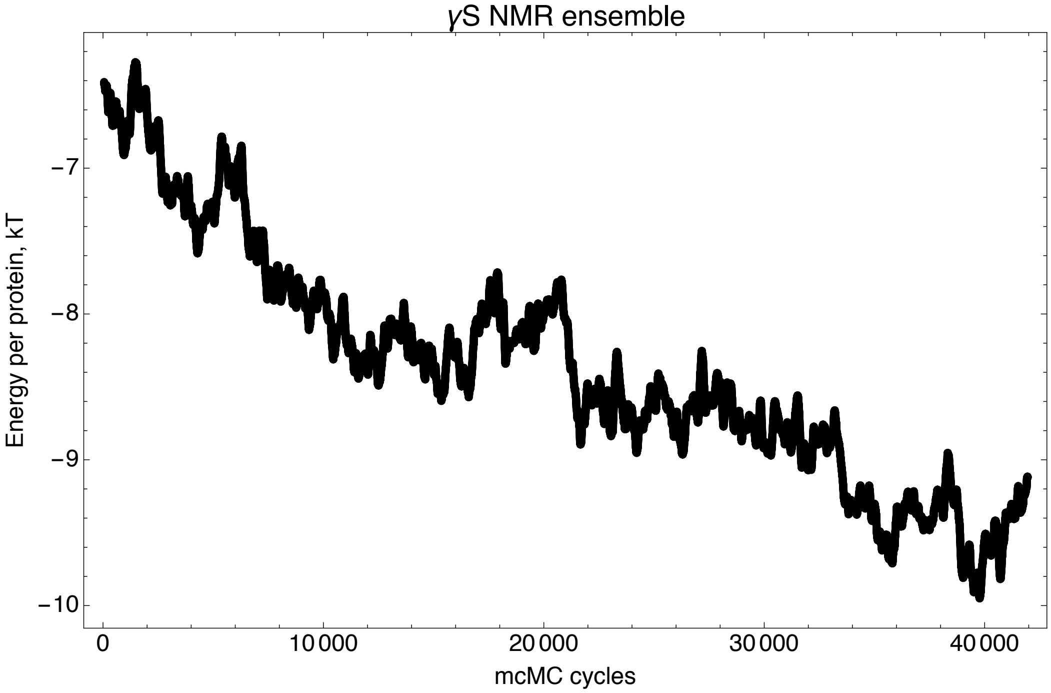


Figure 3.5: The total energy per protein in a simulation of wild-type γS protein at 200 g/L for a conformational library obtained from MD simulations.

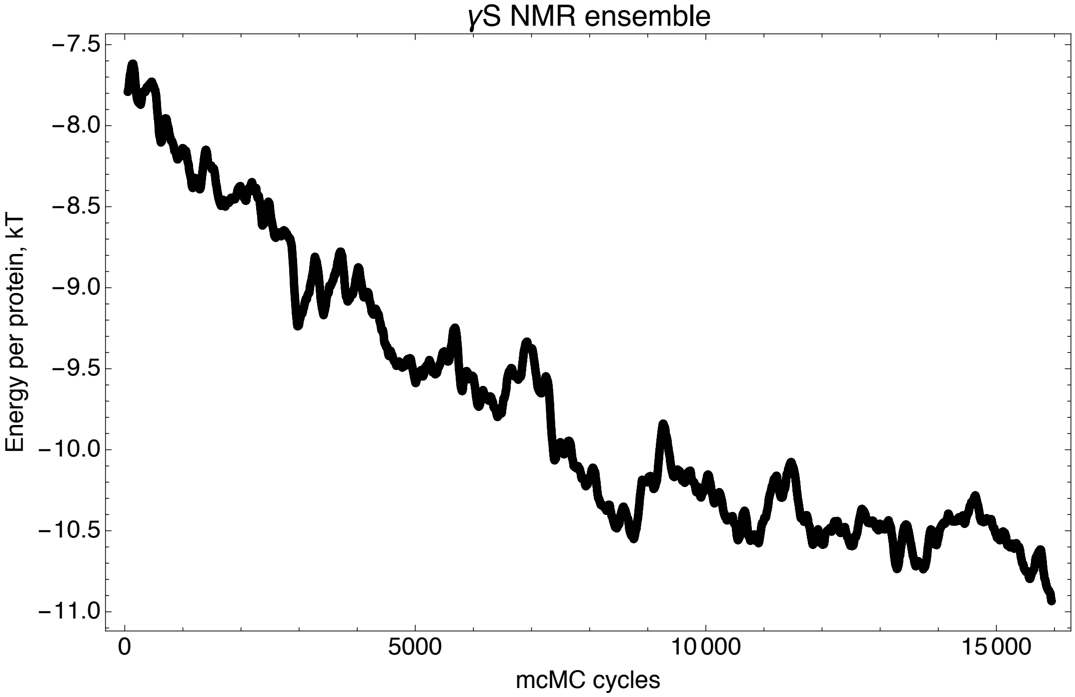


Figure 3.6: The total energy per protein in a simulation of wild-type γS protein at 200 g/L for a conformational library obtained from NMR experiments.

METHODS

mcMC simulations of wild type γS crystallin were created from two input ensembles of conformations. One ensemble was generated from MD simulation and 50 input conformations were used.

The MD simulations were run by Eric Wong. The MD simulation system of γS-crystallin was built from the lowest energy solute-state NMR conformations deposited in the Protein Data Bank by Kingsley et al. (*13*) (PDB ID code 2M3T). Protein atoms were parameterized with the CHARMM36(*93*) force field. The protein was solvated in an orthorhombic box of TIP3P(*90*) waters, 70 Å x70 Å x70 Å in size. The net charge of the system was neutralized with chloride counterions. The VMD 1.9.1 software package (*94*) was used to assemble the system.

Pre-production simulation was carried out using NAMD 2.9. The protein-solvent system was minimized for 10,000 steps at constant temperature (300 K) and pressure (1 atm). Harmonic positional restraints placed on each protein heavy atom and were gradually relaxed over 1 ns. The system was then equilibrated at constant temperature and pressure for a total of 434 ns. The smooth particle mesh Ewald method (*51, 96*) was used to calculate electrostatic interactions. Short-range, real-space interactions were cut off at 11 Å by means of a switching function. A reversible, multiple time-step algorithm (*97*) was used to integrate the equations of motion with a time step of 4fs for electrostatic forces, 2 fs for short-range nonbonded forces, and 1 fs for bonded forces. All bond lengths involving hydrogen atoms were held fixed using the SHAKE (*98*) and SETTLE (*53*) algorithms. A Langevin dynamics scheme was used for temperature control, and a Nosé-Hoover-Langevin piston was used for pressure control (*99*).

The initial weights were generated by calculating the RMSD based clustering of side chain fluctuations. Another ensemble was generated by NMR and published by Kingsley et al (*13*). The weights were assigned to be equal due to a lack of a better method. The four potential terms were computed in exactly the same way as described in chapters 1 and 2, apart from the nonpolar term, where a nonpolar desolvation parameter of −0.0078 kcal/mol/Å^2 was used to set the second virial coefficient to zero. Typically, an experimental value of the second virial coefficient would be used, however no such value was found in up to date literature. For the present study, however, the value of nonpolar desolvation parameter has no effect. The grids were calculated at 300 K and 50 mM ionic strength. Simulation, result of which is presented in figure 3.1, is run at 10 mg/mL protein concentration using 200 proteins. All other simulations are run at 200 mg/mL protein concentration using 375 proteins.

Input weights were updated every 2x10^4 mcMC cycles. MD ensemble was run for 3.2x10^5 mcMC cycles, while NMR ensemble was run for 1.6x10^5 mcMC cycles until little change between the distribution of input weights and output conformational frequency is observed.

CHAPTER 4

Cataract-related W42R γD-crystallins show spontaneous domain separation and increased propensity for interprotein interaction at high concentration

This is a manuscript in preparation by E.Wong, V.Prytkova, J.Freites, D.Tobias.

INTRODUCTION

Cataract, the opacification of the eye lens, is the leading cause of blindness in many developing countries (*106*)**.** This opacification is caused by the aggregation of a family of proteins in the eye lens called crystallins (*107*). These proteins make up 90% of the protein content in the eye lens fiber cells. Upon cell differentiation, these cells are denucleated, resulting in the loss of protein turnover for all eye lens proteins. Thus, in order for the eye lens to maintain its function, crystallin proteins must remain soluble at concentrations exceeding 400 g/L for an entire lifetime (*5*). However, congenital defects and post-translational modifications, such as UV photo-oxidation, deamidation, and truncation, can result in the formation of large protein aggregates that diffract light (*108-110*). These light diffracting aggregates make up the cloudiness that is characteristic of nuclear cataract.

Crystallin proteins are divided into 3 families: α-, β-, and γ-crystallins. α-Crystallins are heat shock proteins that serve as holdase chaperones. These chaperones bind misfolded β/γ-crystallins, but do not refold them, preventing further aggregation (*13*). β- and γ-crystallins are dimeric and monomeric structural proteins, respectively. These structural proteins allow the eye lens to modulate the index of refraction while maintaining its lens transparency. Many congenital (*8, 11, 111, 112*) and post-translational modifications (*113, 114*) have been linked to cataract formation, often through different mechanisms of structural change. In this paper, we focus on the aggregation-related properties of the monomeric human γD-crystallin (WT HγD) and its cataract-related W42R variant (W42R HγD).

γD-crystallin is a 173 residue, ~21 kDa, monomeric protein comprised of primarily β-sheets organized into two Greek key domains. The congenital cataract-related W42R variant was reported to have changes in tertiary structure and a reduced thermal stability (*115*). However, the reported crystal structure of W42R HγD is near identical to that of wild-type (*11*). Recent experiments have identified the formation of internal disulfide cross linkages upon denaturation of the N-terminal domain (*116*), suggesting the existence of a small population of partially unfolded intermediates. However, structural details of this intermediate state and its aggregation pathway at physiological conditions remain uncertain.

Several mechanisms of aggregation have been reported, often involving different degrees of unfolding. Large-scale denaturation of the Greek key domain structure can lead to the formation of amyloid fibrils, characterized by fibrillar aggregates containing intermolecular β-sheets (*10, 117*). Moderate unfolding, resulting in the separation of the N- and C-terminal domains, can lead to domain-swapped aggregates, where the separated domains reform new inter-protein domain interfaces (*118*). Lastly, minimal unfolding can alter interprotein interactions such that the proteins aggregate in a native-like state (*119, 120*)**.** It is important to note, that the aggregation pathway is dependent on the type of stress (chemical, pH, thermal, etc.) put on the protein. pH stress on γ-crystallins has been shown to form amyloid fibrillar aggregates that differ from aggregates formed at physiological conditions (*10, 112*). Furthermore, some cataract-forming crystallin variants have been shown to be more thermally stable than non-cataract forming variants (*15*). Therefore, it is important to characterize γ-crystallins variants at physiological conditions to understand the pathway to cataract formation.

To computationally investigate the cataract-related conformations and interactions of W42R HγD and its aggregates, we use a combination of microsecond-scale molecular dynamics (MD) simulations and Multi-Conformation Monte-Carlo (mcMC) simulations to model the single protein dynamics and interactions at high concentration, respectively. Additionally, we use network analysis to investigate the morphologies of the aggregates. We show that the N- and C-terminal domains of W42R HγD separate in absence of thermal or chemical denaturation. The resulting domain-separated conformations contain patches of exposed hydrophobic residues that, in turn, become the primary sites of interprotein interaction in Monte Carlo simulations of W42R HγD at high concentrations. These domain-separated conformations of W42R HγD show a propensity to form higher order aggregates than WT HγD in mcMC simulations. Based on our overall results, we provide atomistic computational evidence of significant conformational changes in W42R HγD that result in large-scale aggregation.

METHODS

Single Protein Molecular Dynamics Simulation System Preparation and Equilibration

The initial protein coordinates of WT HγD and W42R HγD were built from the crystal structures deposited into the Protein Data Bank (PDB ID code 1HK0 for WT HγD and 4GR7 for W42R HγD) (*9, 11*). Histidine protonation states were set to be the same as those published in the solution state NMR structure of the P23T variant of γD-crystallin (PDB ID code 2KFB) (*91*). Protein atoms were parameterized with the CHARMM36 force field (*93*). The crystal structure waters were kept and the proteins were solvated in a cubic TIP3P (*90*) water box measuring 80 Å on a side. The system was neutralized with chloride counterions. The single protein systems contained 48,309 atoms and 48,367 atoms for WT HγD and W42R HγD, respectively. All system preparation was performed using the VMD 1.9.1 software package (*94*).

A 20 ns pre-production simulation equilibration was performed with NAMD 2.9 (*95*). The prepared systems were minimized for 10,000 steps in the NPT ensemble at 310 K and 1 atm. Protein heavy atoms were restrained with harmonic positional restraints and were gradually relaxed over 200 ps. NAMD was parameterized with the smooth particle mesh Ewald method (*96, 121*) for long-range electrostatic interactions, a real space interaction cutoff at 11 Å, and an integration time step of 2 fs/timestep. The RESPA algorithm (*97*) was used with a timestep of 4 fs for electrostatic forces, 2 fs for nonbonded forces, and 1 fs for bonded forces. Hydrogen covalent bonds were held fixed using the SHAKE (*98*) and SETTLE (*53*) algorithms. Constant temperature and pressure was maintained using a Langevin thermostat and a Nosé-Hoover-Langevin piston (*99, 100*).

Microsecond Time Scale Molecular Dynamics Simulations

Production simulations were performed on the Anton 2 supercomputer, a special-purpose computer for molecular dynamics simulations of biomolecules (*101*). Protein and solvent atoms were parameterized with the CHARMM36 (*93*) and TIP3P (*90*) force fields, respectively. The multigrator scheme (*105*) was used to integrate Newton's equation of motion at 2.5 fs/time step. Using the RESPA algorithm (*97*), long-range nonbonded, short-range nonbonded, and bonded forces were calculated at a timestep of 7.5 fs, 2.5 fs, and 2.5 fs, respectively. Long-range electrostatic forces were calculated using the k-Gaussian split Ewald method (*103*). Hydrogen covalent bonds were held fixed using the SHAKE (*98*) algorithm. Constant temperature and pressure was maintained using Nose-Hoover chains (*104*) and the Martyna-Tobias-Klein barostat (*99*), respectively. Single protein simulations of WT HγD and W42R HγD were run for a total of 50 µs and 17 µs of production simulation, respectively.

W42R HγD Dimer System Preparation and Simulation

To prepare the two protein MD simulation of W42R HγD, the protein and its solvation shell (waters within 6 Å of the protein) were extracted from the last frame of the 17 µs single protein MD simulation. Two copies of the proteins were placed such that the proteins are separated by at least 9 Å and the solvating waters do not overlap. The cubic water box is parameterized, solvated, neutralized, and equilibrated as previously described. The resulting system contains a total of 46,655 atoms contained in an 80 Å x 80 Å x 80 Å periodic cell, the same dimensions as the single protein simulation. The two-protein simulation of γD-W42R was run for a total of 7 µs on the Anton 2 supercomputer using the same parameters as the single protein simulation.

mcMC simulation protein-protein interaction potential

Our mcMC simulation employ protein-protein interaction potential developed by Mereghetti et al. (*122*). Two proteins interact through the following potential function:

$$\Delta U = \frac{1}{2}\sum_{i_2} \Phi_{el_1}(r_{i_2}) \cdot q_{i_2} + \frac{1}{2}\sum_{j_1} \Phi_{el_2}(r_{j_1}) \cdot q_{j_1} + \sum_{i_2} \Phi_{el_1}(r_{i_2}) \cdot q_{i_2}^2 + \sum_{j_1} \Phi_{el_1}(r_{j_1}) \cdot q_{j_1}^2 + \sum_{m_2} \Phi_{ND_1}(r_{m_2}) \cdot SASA_{m_2} + \sum_{n_1} \Phi_{ND_2}(r_{n_1}) \cdot SASA_{n_1} + \sum_{m_2} E_{Softcore_1}(r_{m_2}) + \sum_{n_1} E_{Softcore_2}(r_{n_1}) \qquad (4.1)$$

The first two terms denote the interaction of electrostatic potential of one of the proteins with the charges of another protein (*83*). The charges are computed through the effective charge approximation implemented in SDAMM software package. The second two terms refer to electrostatic desolvation penalty that appears due to location of solvated polar groups of one protein in proximity of the low dielectric environment of another protein and consequential simultaneous loss of solvation shell (*123*). Terms five and six correspond to an attractive short-range non-polar desolvation interaction between two proteins that appears when solvent exposed hydrophobic atoms of one protein are buried by another protein. This interaction can be scaled by modifying a prefactor β used to convert the buried area of the protein surface into a desolvation energy. The value used in our simulation is -9 cal mol^{-1} $Å^{-2}$. Seventh and eighth terms denote the softcore repulsive interaction energy terms.

The interaction potential terms were computed prior to simulations on 200 Å x 200 Å x 200 Å grids with the grid spacing of 1 Å. The electrostatic potential grids were computed at 50 mM ionic strength according to OPLS force field (*85*) by finite difference solution of the linearized Poisson-Boltzmann equation using the UHBD (*86*) software package.

Multiple conformation Monte Carlo simulations

A multiple conformation Monte Carlo (mcMC) algorithm (*81*) employs translational and rotational moves on randomly selected proteins in combination with a conformational swap from a finite size library of structures. For rotational and translational moves a basic Metropolis scheme (*124*) is used and the size of moves is adjusted to provide a 50% acceptance ratio. The appearance of each conformation in the simulation is proportional to its probability of appearance in MD simulation from which it was extracted. The entire MD simulation is used for clustering based on sidechain orientation. Top fifty structures are selected for the mcMC conformational swap library.

Each trial move is accepted according to the Metropolis criterion with acceptance probability:

$$P_{acc} = \min\left(1, \exp\left[\frac{-\Delta E}{k_B T}\right]\right) \tag{4.2}$$

Where ΔE is the difference between the energy of the system before and after the trial move, k_B is the Boltzmann constant and T is the temperature. All simulations were performed with 375 proteins at 200 mg/mL protein concentration, which is approximately half the density of the eye lens. A total of $2x10^5$ MC cycles at 310 K are performed for each protein type.

RESULTS

W42R HγD domains separate after salt-bridge interaction

To investigate the conformational dynamics of wild-type γD-crystallin and its cataract-related W42R variant, single protein MD simulations of WT HγD and W42R HγD were run for 50 μs and 17 μs, respectively. Although the crystal structure of W42R HγD contains the W42R point mutation, the mutant structure is strikingly similar to that of the wild-type protein (backbone RMSD of 0.795 Å).(*11*) In the MD simulation of W42R HγD, the protein structure remains similar to its initial structure for several microseconds. Over the course of 6 μs, the residue R42 gradually becomes solvent exposed and ultimately forms a salt-bridge with the C-terminal carboxyl group (Figure 4.1A). Upon salt-bridge formation, the N- and C-terminal domains separate, resulting in a > 10 Å increase in the αC-RMSD for W42R HγD (Figure 4.1B). As the domains separate, the C-terminal domain rotates such that the set of hydrophobic interdomain residues become solvent exposed for both domains (Figure 4.2B). Interestingly, the internal fluctuations of the N- and C-terminal domains of the domain-separated W42R HγD is similar to that of WT HγD (Figure 4.2C). This gives no indication for further unfolding of the Greek key domains after domain separation in W42R HγD. In the case of the wild-type protein, the protein maintains its native conformation over the course of 50 μs of simulation (Figure 4.2B).

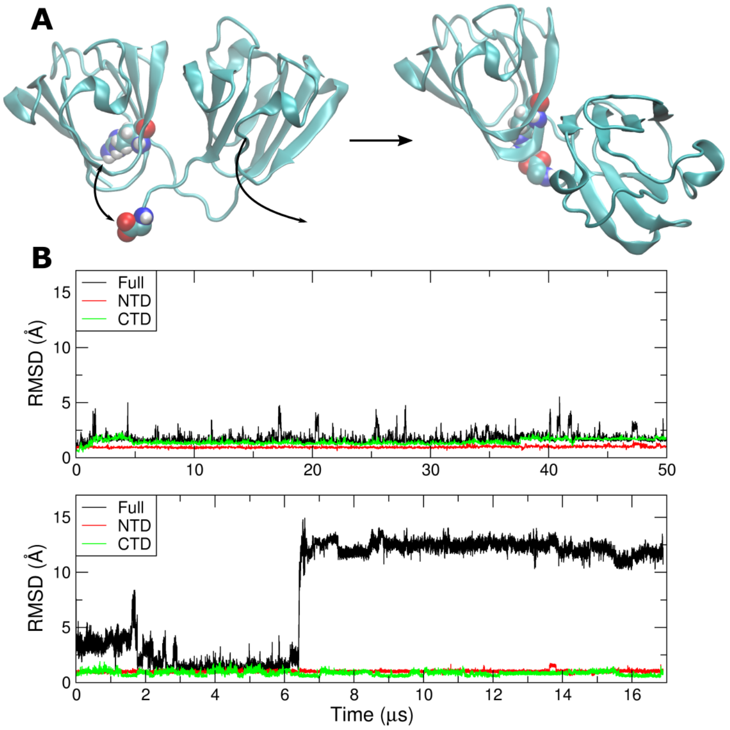


Figure 4.1: (A) Left: Initial conformation of W42R HγD. The black arrows indicate the R42-CTD carboxyl salt bridge and the separation of the CTD. Right: Snapshot of the conformation of W42R HγD after domain separation. (B) Root mean square deviation (RMSD) of the backbone alpha carbons plotted vs. time. WT-HγD and W42R HγD is plotted on the top and bottom, respectively.

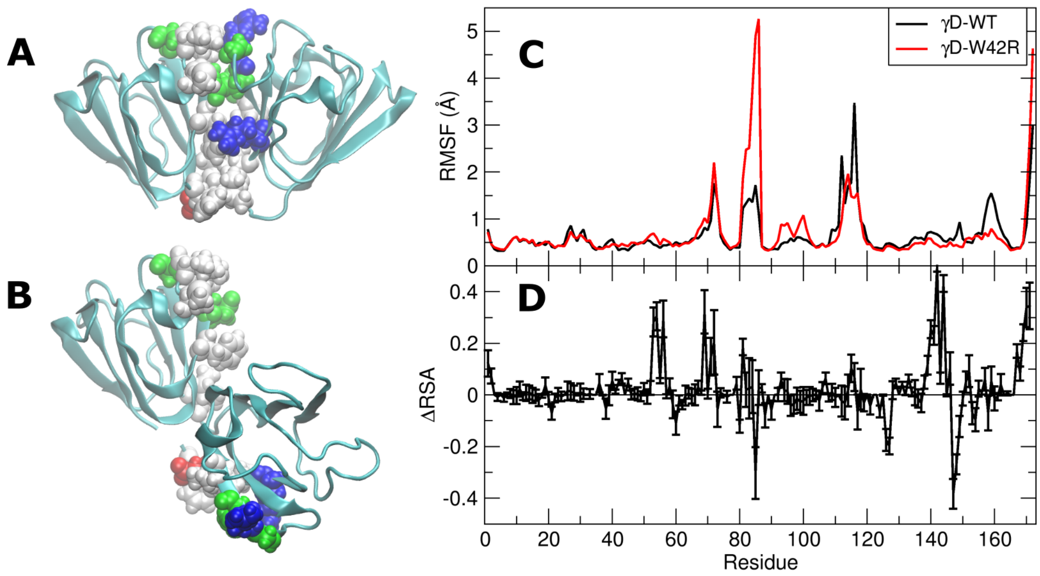


Figure 4.2: Conformational snapshots of (A) WT HγD and (B) W42R HγD taken after microsecond-scale MD simulations. Residues with a > 0.13 increase in relative solvent exposure (ΔRSA) from the W42R point mutation are represented in VDW spheres on both wild-type and mutant proteins. Residues are colored by residue type. Non-polar residues are colored white, polar residues are colored green, basic residues are colored blue, and acidic residues are colored red. (C) Intra-domain root mean square fluctuations of WT HγD and W42R HγD shown in black and red, respectively. (D) Difference in RSA between W42R HγD and WT HγD. Positive ΔRSA values correspond to an increase in residue exposure for W42R HγD.

To identify the exposure of new potential interprotein contacts, the relative solvent accessibilities (RSA) of WT HγD and W42R HγD were calculated for each clustered conformation prepared for the mcMC simulations (Figure 4.2D). The RSA represent the residue solvent accessibility normalized over the theoretical maximum solvent accessibility for each type of residue (*125*). Residues with a change in RSA > 0.13 are all located at the interdomain interface in WT HγD (shown as VDW spheres in Figure 4.2A&B). These are primarily patches of hydrophobic residues that form the hydrophobic core of the N- and C-terminal domains. These exposed hydrophobic patches give opportunities for the formation of new interprotein contacts. To test the role of these residues in interprotein interaction, we performed multi-conformation Monte-Carlo simulations of WT HγD and W42R HγD at high concentrations. Clustered conformations were obtained for WT HγD and the open conformation of W42R HγD. The top 50 most populated clusters were used as input for mcMC simulations of WT HγD and W42R HγD at 200 g/L.

Monomer mcMC simulation results

After performing Monte Carlo simulations of wild-type and W42R variant at 200 g/L for 2x10^5 MC cycles, domain center of mass based radial distribution functions were computed. The domain-based radial distribution functions were considered instead of the center of mass based radial distribution function due to the shape of the protein. Since γD-crystallin is composed of two highly homologous domains – N-terminal and C-terminal, the entire protein has an elongated structure. Considering the radial distribution function of the center of mass of each domain allows us to see which domain has the strongest preference for interaction.

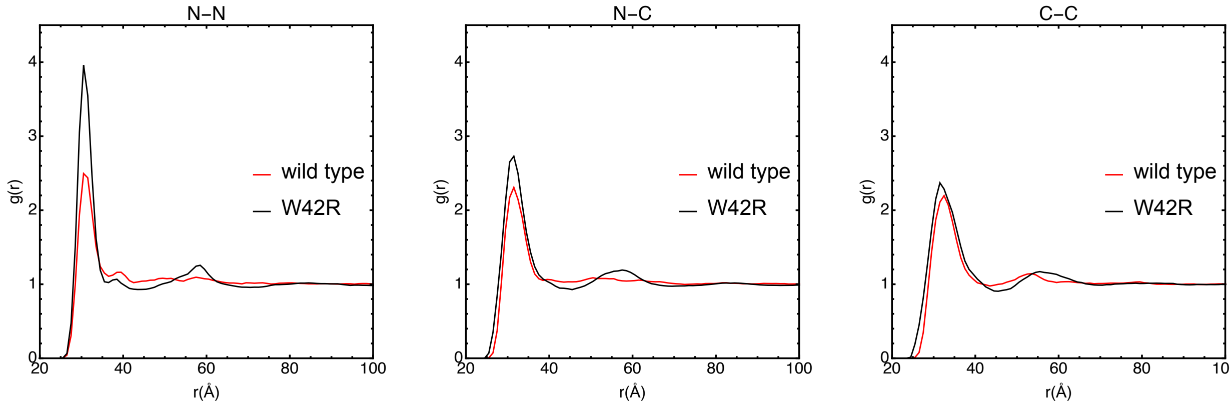


Figure 4.3: Radial distribution functions of domain centers of mass of wild type and W42R variant (structures obtained in monomer MD simulation).

According to domain-based radial distribution functions, N-terminal of W42R variant has the strongest preference for interaction with N-terminal of other proteins comparing to that of the wild-type. N-terminal to C-terminal interactions are less strong, but also more prevalent in the W42R variant, while C-terminal to C-terminal interactions are very similar in W42R variant and in the wild-type crystallin.

To examine the origin of this preferential interaction in W42R variant all protein pairs contributing to the radial distribution function were selected and the contacts between them were analyzed. The protein pairs were selected based on the distance between the centers of mass of protein domains. Two residues are said to be in contact if any two heavy atoms are within 3.5 Å distance of each other. Based on these criteria, the total number of contacts found in the wild type simulation is 24991, and for W42R simulation – 47887, which is almost twice as many as that for the wild type. The plot below shows the total number of contact found for each residue. The contacts are displayed if their number exceeds twenty. W42R has slightly more contacts between C terminals than the wild type. Specifically, residue 144, Leucine is responsible for the increased interaction between C terminals. N terminal interacts more with both N and C terminal of other proteins through residues 53, Leucine, 69, Methionine, and 71, Leucine. As can be observed in the contact analysis plot, these hydrophobic residues don't make a single specific contact, but instead contact many other residues. If we compare these residues to Figure 28D showing RSA of fifty mcMC configurations extracted from the MD simulations, we can see that these residues indeed belong to those patches of the protein that are more exposed in W42R than in the wild type.

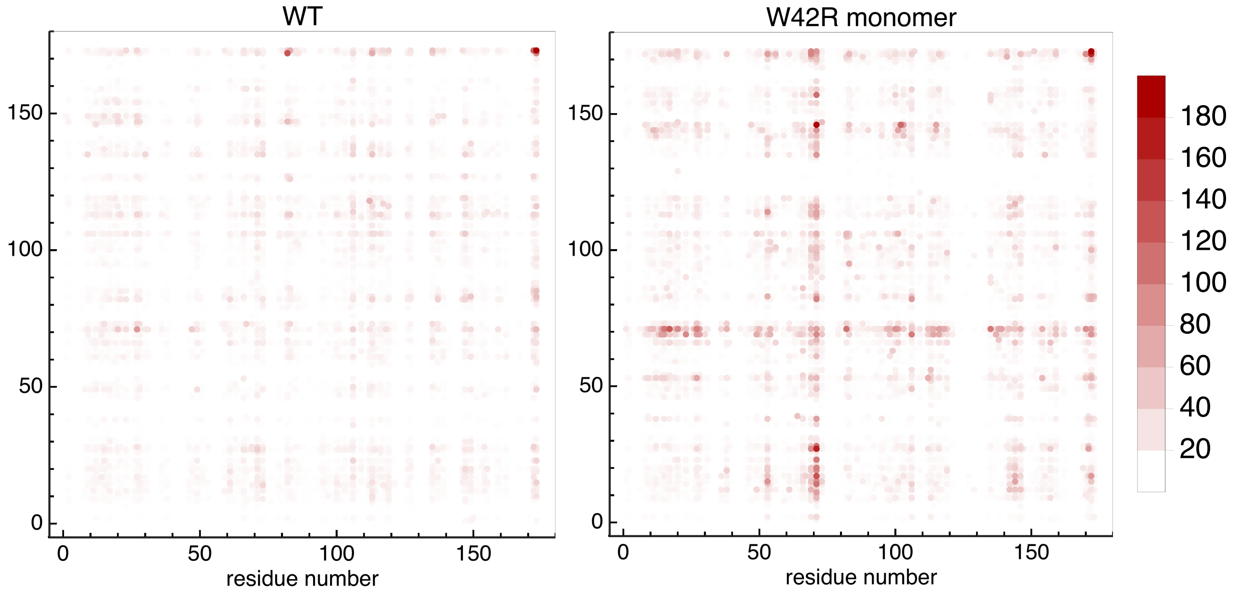


Figure 4.4: The total number of contacts of HγD and its W42R variant protein residues with other protein residues in mcMC simulations. No preferential contacts are found between wild-type proteins. The darkest points relate to contacts of the flexible end on C-terminal. However, W42R variant contacts multiple residues of other proteins with residues 53, 69, 71.

Two protein MD Simulations of W42R HγD show increased domain separation

As the increased interaction of the W42R variant is apparent both from the radial distribution functions and contact analysis, we decided to perform another MD simulation of a W42R dimer starting from the final structure of a monomer MD simulation. This allows us to see whether there is any further structural change, as well as to expand the ensemble of input structures for mcMC simulations. An MD simulation of W42R HγD was prepared with two copies of W42R HγD and their solvation shells. The two protein copies were placed such that no protein heavy atoms are within 9 Å of each other. Over the course of the 7 μs MD simulation, the two proteins diffuse together and interact such that the exposed hydrophobic interdomain residues (specifically, L53, F56, M69, and L71) from the N-terminal domain of both proteins form a new interprotein hydrophobic core. Apart from L53, these are the same residues with strong contacts observed in the monomer mcMC simulations. The interprotein contacts for the two protein W42R HγD simulation are rather sparse. However, these few contacts are sufficient to keep the two proteins associated for the duration of the simulations. When the two proteins interact, the C-terminal domain separates further from the N-terminal domain resulting in a fully exposed interdomain interface for both proteins (Figure 4.6). This results in residue F56, located deeper in the interdomain interface, becoming exposed to protein interaction.

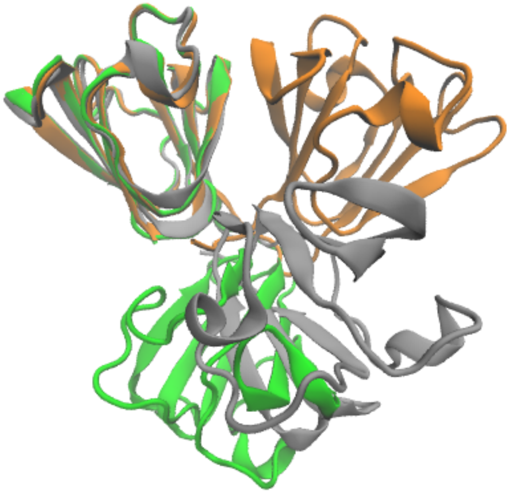


Figure 4.5: Snapshots of the highest probability structures used in mcMC simulations. In orange is the wild type structure, in grey – the W42R variant structure from monomer MD simulation. In green – W42R variant structure from dimer MD simulation. C-terminal domain is tilted by 45.2 and 65.5 degrees in monomer and dimer structures respectively comparing to the wild type.

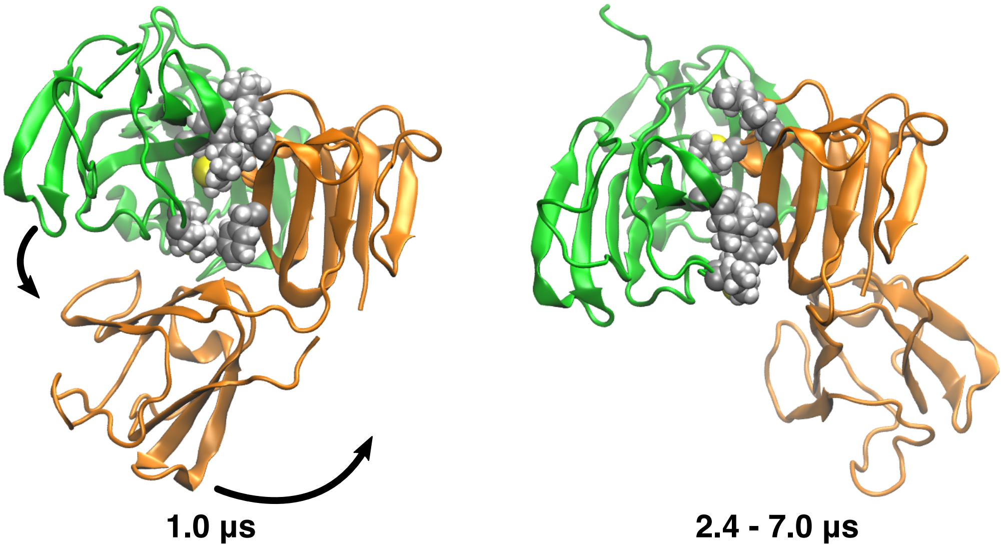


Figure 4.6: Snapshots of the two-protein MD simulation of W42R HγD after (left) initial binding and (right) after binding-induced conformational change. The change in protein conformation results in increased domain separation and increased exposure of interfacial residues. This co-conformation persists from the 2.4 µs mark to the end of the 7 µs simulation. Binding residues identified from the mcMC simulations are shown as VDW spheres on the NTD of both proteins.

Dimer MC Simulation results

The two protein MD trajectory was clustered once again based on the sidechain RMSD to create a new ensemble of fifty conformations for mcMC simulations. We further denote this ensemble as "dimer" simulation ensemble. Below we consider the domain-based radial distribution function of the dimer ensemble of W42R variant. It can be observed that the N-terminal domain has an even higher propensity to interact in the dimer ensemble of the W42R than in the monomer ensemble.

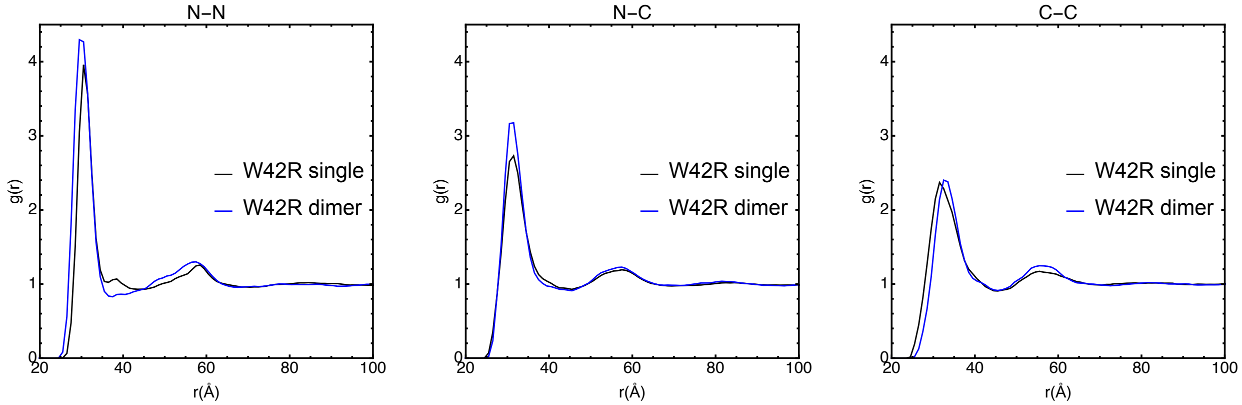


Figure 4.7: Radial distribution functions of domain centers of mass of W42R variant from structures obtained in monomer and dimer MD simulations.

The total number of protein pairs found in the contact analysis for the dimer ensemble is even higher - 67415. The comparison of the contact analysis between two ensembles of the W42R variant – the one generated from the monomer MD simulation and from the dimer MD simulation – is presented on the plot below. The main features of interaction are preserved, the same residues are participating in contact, however N terminal interacts even more in the dimer ensemble.

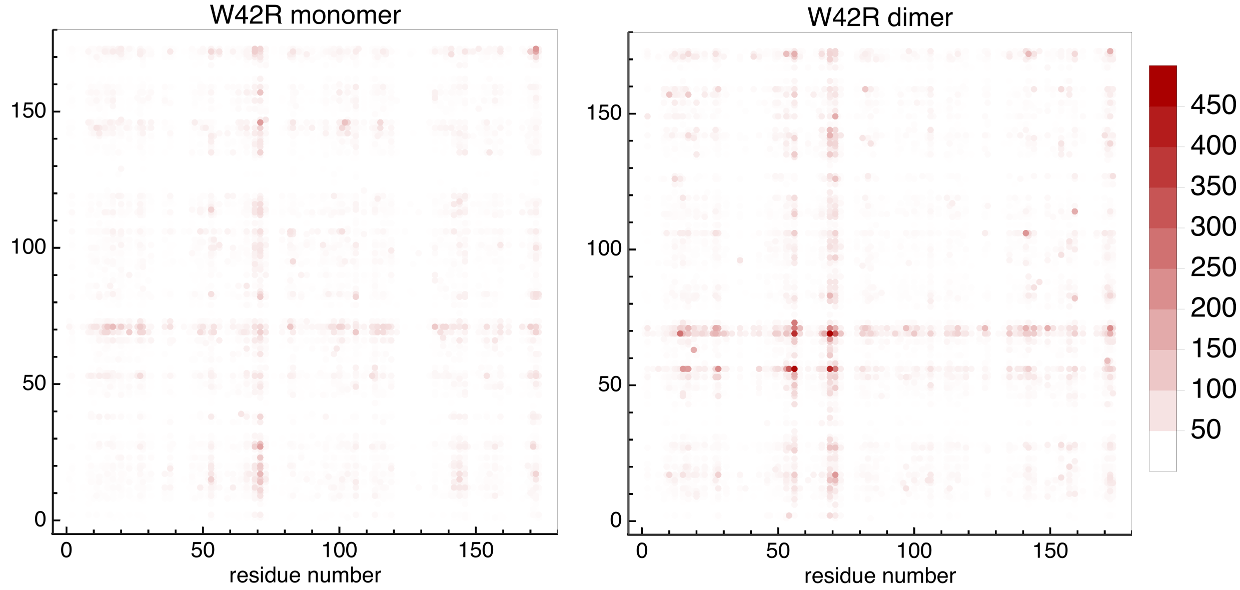


Figure 4.8: The total number of contacts for each residue of W42R variant from monomer and dimer conformations with other protein residues in mcMC simulations. Dimer conformation of W42R variant have a higher number of contacts than the monomer conformation for residues 53, 56, 69, 71.

The open conformation of W42R HγD forms larger sized aggregates relative to wild-type

We further investigate the effect of the W42R point mutation on the size and morphologies of γD-crystallin aggregates. Monte Carlo simulations indicate that increased exposure of the inter-domain interface leads to a larger number of contacts between proteins in a solution of W42R γD-crystallins. We analyze how this increased propensity to make contacts leads to cluster formation. Two proteins are said to be in contact with one another if the distance between the centers of mass of their domains is within 31 Å. Such distance is chosen due to the position of the maximum of the domain-based radial distribution function. Clusters of proteins chosen by the above criteria are now analyzed. In the plot below the probability density of cluster size distribution is shown for the wild type γD simulations as well as for the two ensembles of the W42R variant. The cluster size distributions of the monomer and dimer simulations of W42R HγD show a significant increase in cluster size over wild-type. Though WT HγD shows some propensity to form small clusters, the W42R HγD clusters can be composed of more than half of the proteins present in the simulation (375 proteins total). When the domains are further separated (represented in the dimer simulation of W42R HγD), the distribution of clusters becomes much broader, signifying a large proportion of higher order aggregates.

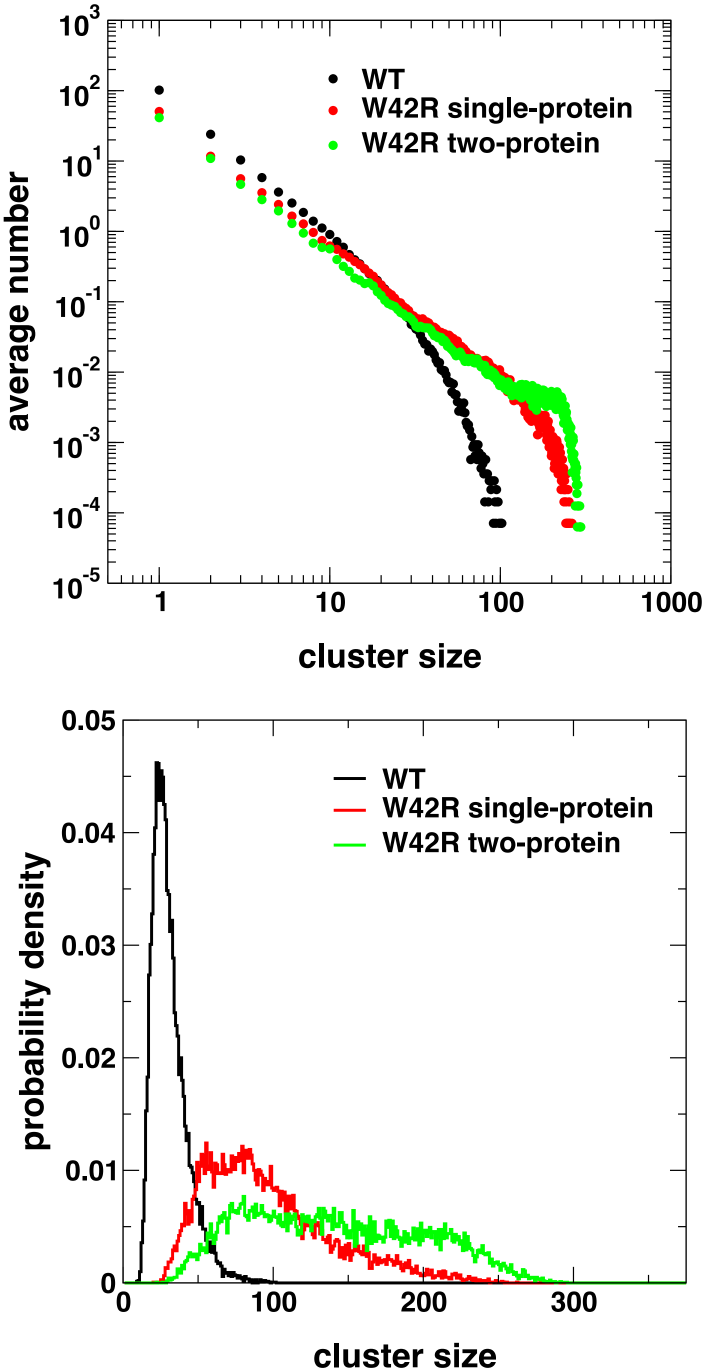


Figure 4.9: (Top) Probability density of the largest cluster sizes from Monte-Carlo simulations of 375 proteins (200 g/L). (Bottom) Distribution of the size of all clusters present in Monte-Carlo simulations of 375 proteins (200 g/L)

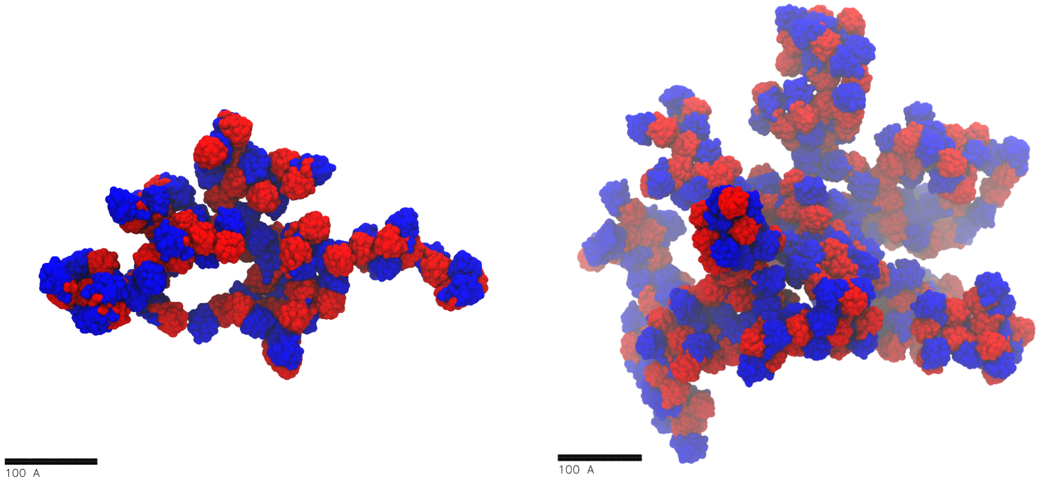


Figure 4.10: Isolated cluster conformations taken from the 95th percentile of the cluster size distribution. Clusters were formed in mcMC simulations using conformations of WT HγD (left) and the W42R HγD monomer (right) at 200 g/L. The N- and C- terminal domains of the proteins are colored in red and blue, respectively. Cluster sizes are 54 and 181 proteins for WT HγD and W42R HγD, respectively.

Visually inspecting the morphologies of the isolated clusters (Figure 4.10), there is an amorphous structure in both mutant and wild-type aggregates. However, the clusters of the mutant protein have a much larger apparent size. This is a result of the increased propensity for interprotein interaction. Additionally, the newly exposed hydrophobic interfacial residues in W42R HγD introduce a new interprotein interaction surface, allowing for a single protein to interact with several more proteins. By representing the clusters as a network of interprotein interactions, the W42R HγD conformation is shown to be capable of a higher degree of interactions (Figure 4.11). Most of these high order interactions come from interactions with the N-terminal domain.

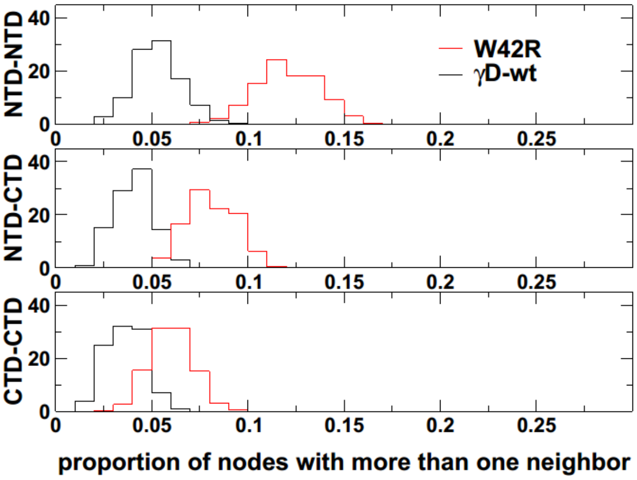


Figure 4.11: Histogram of the interprotein domain interactions involving more than one neighbor.

DISCUSSION

We compare the conformational change of W42R HγD in a long MD simulation to the wild type protein. The simulation of HγD was conducted for 50 μs and no structural change was observed. However, after 6 μs of W42R simulation the protein undergoes structural change due to solvent exposure of residue R42 and formation of a salt bridge between this residue and C-terminal carboxyl group. This leads to separation of two domains while the structure of separate domains stays intact. This structural change exposes several hydrophobic residues on the interdomain interface of both domains that were buried in the wild type HγD protein. No other significant conformational changes occur for the remaining 11 μs of simulation, suggesting that this conformation of the protein is stable in dilute protein concentrations. Simulation of W42R variant dimer reveals that hydrophobic interdomain residues participate in protein-protein interaction, and two domains separate even further as a result of protein-protein interaction, fully exposing the interdomain interface.

Early experiments by Wang et al. (*115*) report an initial comparison of the hexahistidine-tagged HγD and W42R. They report similar secondary structure between W42R and wild-type through far UV CD spectroscopy. However, ANS fluorescence experiments show a significant increase in hydrophobic exposure in W42R compared to wild-type. They suggest that this hydrophobic exposure is a result of a change in tertiary structure in W42R, consistent with the separation of intact Greek key domains observed in our MD simulations.

F. Ji et. al. report shows the crystal structure of W42R variant has two domains tilted only by 9 degrees comparing to the wild type while structurally each domain remains intact (*11*). They find that wild type protein and W42R variant possess almost identical solvent accessible surface areas (8894.32 and 8546.23 Å^2, respectively, as calculated by VMD). However, the unfolding curve of W42R variant under close to physiological conditions (37 ˚C and 7 pH) shows two-step unfolding, indicating that an unfolding intermediate exists and that W42R variant has a lower chemical stability than HγD. A domain-separated conformation has been observed in annealing simulations of wild-type γD-crystallin (*118*). They proceed to observe domain swapping interactions with the CTD. It is worth noting, however, that the unfolding simulations involved simultaneous thermal and chemical denaturation with urea, resulting in an unfolded NTD. We observe a separation of intact domains in absence of thermal and chemical denaturation.

Our MD simulations under physiological conditions, and therefore close to eye lens conditions, reveal that without unfolding the structure of separate domains, just by further tilting the angle between them, solvent accessible surface area increases from 8991.08 Å^2 in HγD to 9615.64 Å^2 in W42R HγD monomer conformation to 9784.23 Å^2 in W42R dimer conformation.

Serebryany et al. also report that tryptophan fluorescence spectra – an evidence of conformational change in HγD (references 18, 20, 39, 49, 50 in that paper) – show no difference between WT and W42R variant (*126*). They observe how in oxidizing solution, W42R spectra becomes red-shifted indicating the process of unfolding. The spectral change was observed over the course of 60 minutes – the time yet inaccessible to simulations. However, the presence of unfolded proteins was either observed or indicated to exist in conditions that were far from physiological – in the presence of oxidizing agents and at very low protein concentration. It is possible that if we had the time scale of minutes or hours, more conformational changes would take place. However, HγD proteins exist in the eye lens at concentrations exceeding 400 mg/mL, and even small structural changes associated with point mutations may lead to enhanced local interaction and aggregation while full protein unfolding may become impossible under physiological conditions.

High concentration (200 g/L) mcMC simulation of 375 proteins with conformations extracted from the MD monomer simulations show that W42R variant has a higher propensity to aggregate than the wild type HγD protein. Dimer ensemble of W42R is even more likely to aggregate than the monomer ensemble since further conformational change of the protein occurs due to protein-protein contact. The radial distribution function indicates that N-terminal of W42R variant interacts with both N-terminal and C-terminal of other proteins, while C-terminal to C-terminal interaction in W42R variant are very similar to those in wild type.

Hydrophobic residues, specifically residues 53, Leucine, 69, Methionine, and 71, Leucine, located on the interdomain interface of N-terminal, come in contact with many other residues of other proteins. Therefore, many orientations of proteins with respect to one another are possible during aggregation, as long as the residues mentioned above are participating in contacts. Those contacts are successful in creating clusters of W42R variant much larger than clusters of WT protein. The isolated aggregates are amorphous and stringy in shape since each protein may have only a few neighbors. Aggregates of similar shape were recently detected by Boatz et al. in negative-stain TEM images for another HγD variant – P23T (*10*). In fact, the shape of the aggregates is sensitive to aggregation conditions – when the pH of solution is decreased to 3, the proteins form amyloid fibrils containing interprotein β-sheets. ssNMR spectroscopy did not detect any structural change of the Greek key domains in the P23T variant at neutral pH. This indicates that at high concentrations of HγD, small changes of protein surface charge of structure disrupt the careful balance of protein solubility and aggregates form.

CHAPTER 5

Cataract related G18V variant of γS-crystallin shows increased protein-protein interaction

INTRODUCTION

γS crystallin belongs to a family of eye lens structural proteins. Its solubility at concentrations exceeding 400 mg/mL is crucial for keeping the eye lens transparent. Like other structural eye lens proteins, γS crystallin is β-sheet protein consisting of two Greek key domains. G18V variant of γS crystallin is linked to hereditary childhood-onset cortical cataract (*14*). This chapter will investigate the possible molecular mechanism for increased propensity for protein association between G18V variant of γS crystallins.

METHODS

Molecular Dynamics Simulation System Preparation and Equilibration

MD simulations were run by Eric Wong. The simulation system of wild type γS-crystallin and its G18V variant were built from the lowest energy solute-state NMR conformations deposited in the Protein Data Bank by Kingsley et al. (*13*) (PDB ID codes 2M3T and 2M3U). Protein atoms were parameterized with the CHARMM36(*93*) force field. The proteins were solvated in an orthorhombic box of TIP3P (*90*) waters with the nearest box boundary at least 12 Å from any protein atom. The net charge of the system was neutralized with chloride counterions. The VMD 1.9.1 software package (*94*) was used to assemble the system.

Pre-production simulations were carried out using NAMD 2.9 (*95*). Each protein-solvent system was minimized for 10,000 steps at constant temperature (300 K) and pressure (1 atm). Harmonic positional restraints placed on each protein heavy atom and were gradually relaxed over 1 ns. The system was then equilibrated at constant temperature and pressure for a total of 434 ns. The smooth particle mesh Ewald method (*51, 96*) was used to calculate electrostatic interactions. Short-range, real-space interactions were cut off at 11 Å by means of a switching function. A reversible, multiple time-step algorithm (*97*) was used to integrate the equations of motion with a time step of 4 fs for electrostatic forces, 2 fs for short-range nonbonded forces, and 1 fs for bonded forces. All bond lengths involving hydrogen atoms were held fixed using the SHAKE (*98*) and SETTLE(*53*) algorithms. A Langevin dynamics scheme was used for temperature control, and a Nosé-Hoover-Langevin piston was used for pressure control (*99*). 2 μs production run was performed for wild type γS-crystallin.

Molecular dynamics simulations of G18V variant performed on Anton

Additional trajectory was obtained on the Anton supercomputer, a special-purpose computer for molecular dynamics simulations of biomolecules (*127*). Protein and water atoms were parameterized with the CHARMM36(*93*) and TIP3P (*90*) force fields, respectively. The multiple time-step algorithm (*97*) was integrate the equations of motion with a time step of 7.5 fs for long-range non-bonded forces, 2.5 fs for short-range non-bonded and bonded forces. The k-Gaussian split Ewald method (*103*) was used for long-range electrostatic interactions. All bonds lengths involving hydrogen atoms were fixed using the SHAKE (*98*) algorithm. The simulation was performed at constant temperature (310 K) and pressure (1 atm) using Nose-Hoover chains (*104*) and the Martyna-Tobias-Klein barostat (*99*). The RESPA algorithm and temperature and pressure controls were implemented using the multigrator scheme, allowing the simulation to run at 2.5

fs/timestep(*105*). The simulation was run for 960 ns, leading to a total trajectory length of 1.394 µs from the combined trajectories from NAMD 2.9 and Anton.

mcMC simulations

Top 50 structures were selected as the input conformational library for both wild type γS-crystallin and its G18V variant after clustering the entire available trajectory with respect to side chain orientations. The method for computing potential terms can be found in chapters 1 and 2. The nonpolar desolvation parameter was set to −0.0078 kcal/mol/Å^2. The grids were calculated at 300 K and 50 mM ionic strength. All simulations are run at 200 mg/mL protein concentration using 375 proteins at 300 K. Each simulation is 2x10^5 mcMC cycles long.

RESULTS

The conformational dynamics of wild type γS crystallin and its cataract-related G18V variant were investigated in 2 microseconds MD simulations. The structure of C terminal domain is very similar for the two proteins. The largest structural difference comes from deviation in N-terminal domain.

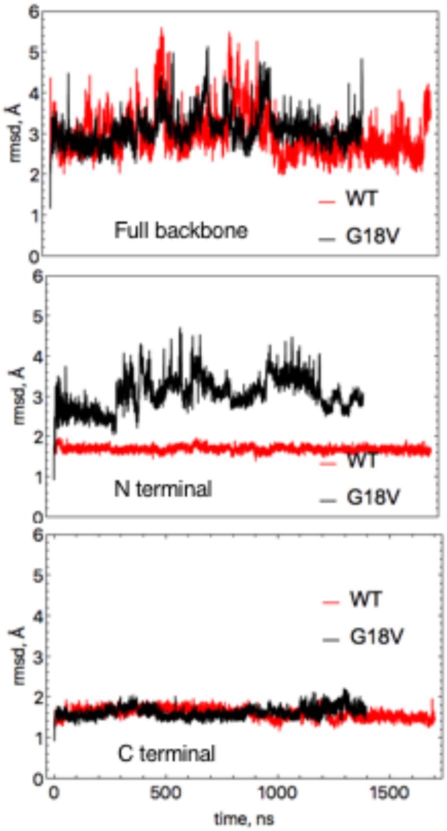


Figure 5.1:Root mean square deviation (rmsd) of the backbone in wild type and G18V variant of γS crystallin. The largest difference appears in N terminal domain RMSD.

To find the origin of this difference arising in N-terminal of G18V variant, we can consider the structure difference between the two proteins. G18V variant has the loop near residue 30 that fluctuates a lot more than that in wild type. This leads to the exposure of a hydrophobic residue, phenylalanine, in position 30, that previously seemed to be buried in the wild type γS crystallin.

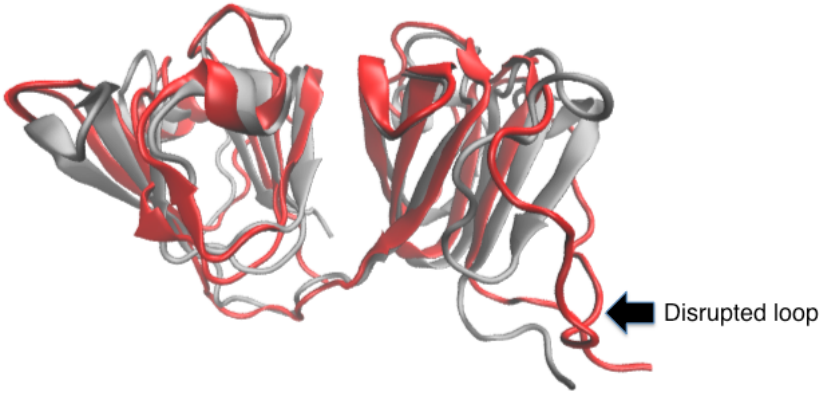


Figure 5.2: Aligned conformations of wild type and G18V variant. Position of a flexible loop is marked.

The MD trajectories were clustered to obtain 50 conformations for each protein as the input for mcMC simulations.

The domain-based radial distribution function – based on the center of mass of each domain – was computed from mcMC simulations. Figure 5.3 shows that in each case G18V variant has a higher propensity for interaction than the wild type γS crystallin.

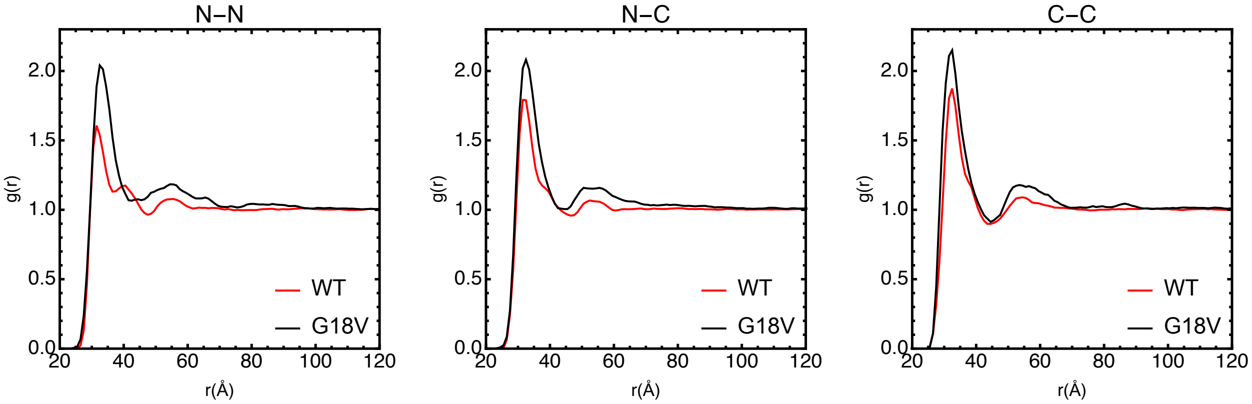


Figure 5.3: Domain-based radial distribution function. On the left it is the N-terminal to N-terminal radial distribution function. In the center – N-terminal to C-terminal, on the right – C-terminal to C-terminal. In each case the peak of the G18V variant radial distribution function is higher than that of the wild type.

The distance of the peak of the radial distribution function is used for distance cutoff to separate protein pairs from the entire trajectory. This new trajectory of pairs is then used to analyze contacts between proteins. Figure 5.4 displays the probability of making a contact mapped on a protein. In wild type crystallin the highest probability belongs to isoleucine in position 161. In the case of G18V phenylalanine in position 30 makes most contacts. In MD simulation G18V variant gets a temporarily disrupted loop that contains residue 30 that starts to stick out and make a lot of contacts. The second highest contact is phenylalanine in position 104.

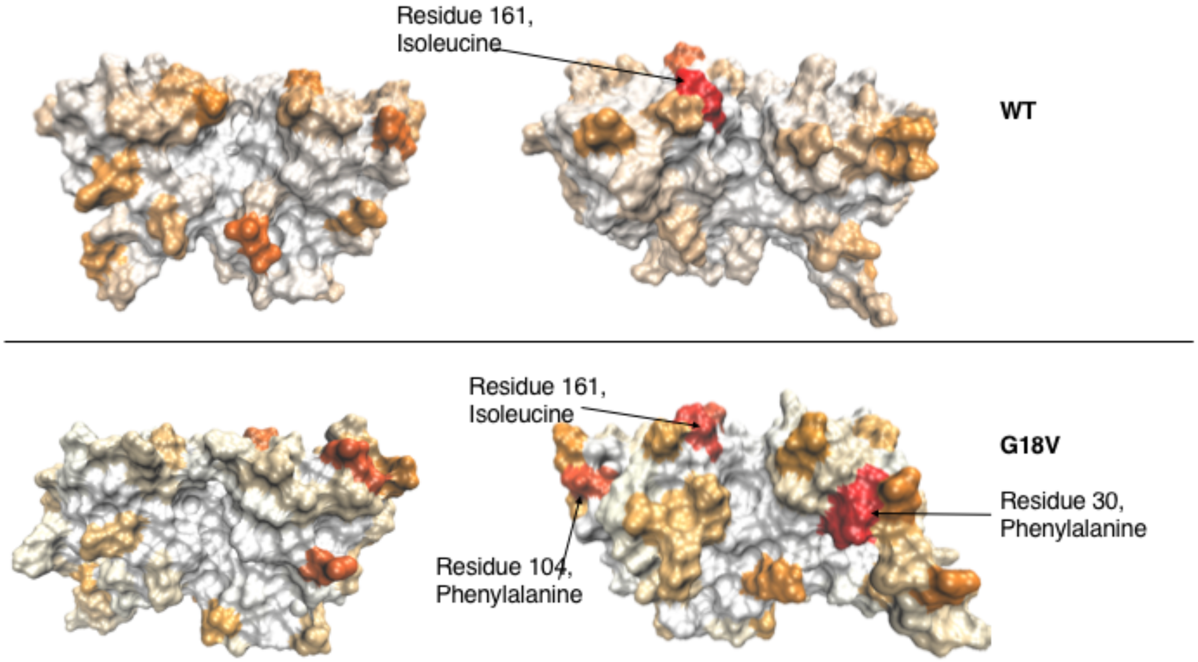


Figure 5.4: The probability of making a contact is mapped on to protein. The highest probability is shown in red, while the lowest is in white. G18V has residues 161 and 104 making highest number of contacts.

In the Figure 5.5 residues 30 and 104 of G18V variant make contacts with many other residues. Both wild type γS crystallin and its G18V variant make some contacts with residues 161.

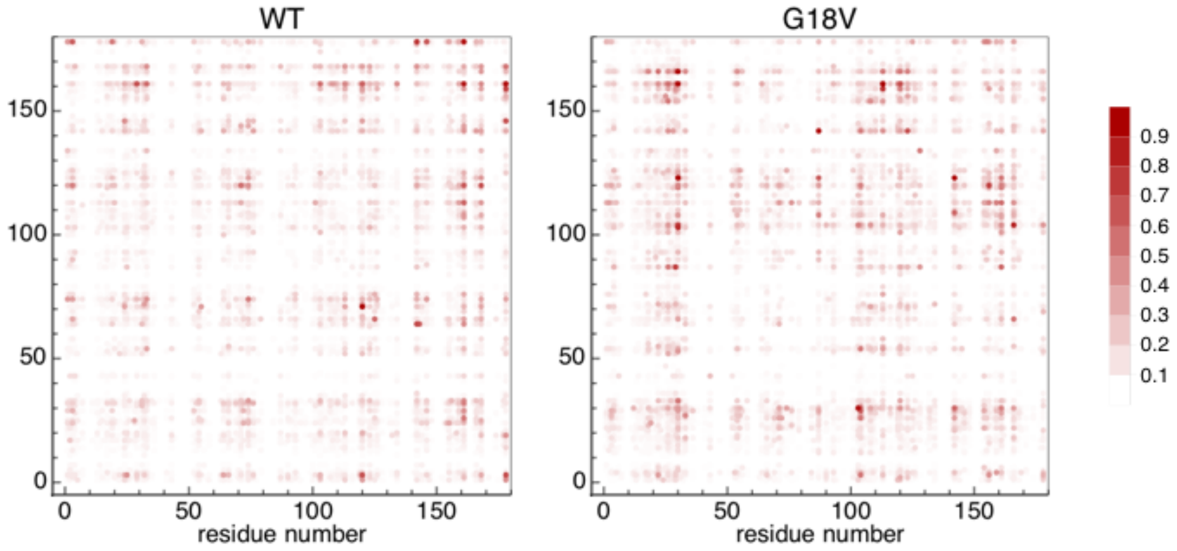


Figure 5.5: Contact probability map for wild type γS crystallin (on the left) and its G18V variant (on the right).

When the above criterion for making clusters is used (based on the peak of the radial distribution function), the distribution of cluster sizes can be plotted. G18V variant tends to make clusters of a larger size than the wild type crystallin.

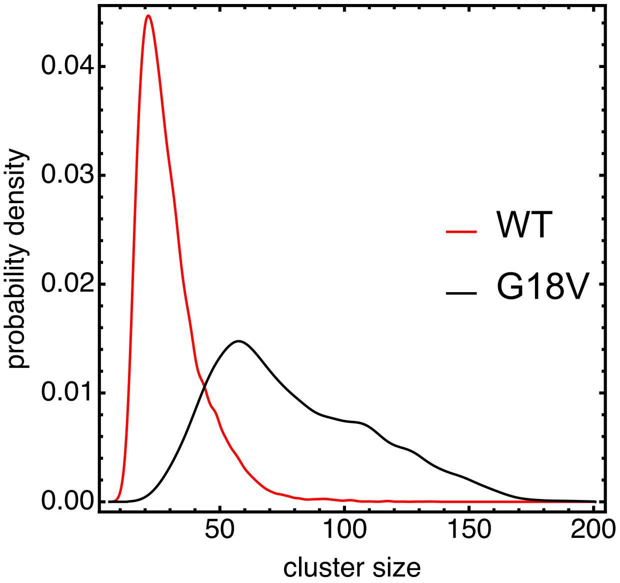


Figure 5.6: Distribution of cluster sizes in 200 g/L wild type γS crystallin simulation and its G18V variant.

From the results presented above, we can conclude that the partial unfold around the mutation site creates a flexible loop in G18V variant that exposes residue 30, which starts to highly interact with other proteins. Residue 161 also becomes exposed and creates a second location for protein-protein interaction. These interaction locations increase the propensity for interprotein interaction and lead to increased aggregation.

REFERENCES

1. R. J. Ellis, Macromolecular crowding: obvious but underappreciated. *Trends Biochem. Sci.* **26**, 597-604 (2001).
2. H. X. Zhou, G. N. Rivas, A. P. Minton, Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annu. Rev. Biophys.* **37**, 375-397 (2008).
3. M.E. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation*. (Oxford University Press, New York, 2010), pp. 700.
4. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-1092 (1953).
5. H. Bloemendal *et al.*, Ageing and vision: structure, stability and function of lens crystallins. *Prog. Biophys. Mol. Biol.* **86**, 407-485 (2004).
6. K. K. Sharma, P. Santhoshkumar, Lens aging: effects of crystallins. *Biochim. Biophys. Acta, Gen. Subj.* **1790**, 1095-1108 (2009).
7. P.B. Banerjee, S. Puttamadappa, A. Pande, A. Shekhtman, J. Pande, Structure, dynamics and surface hydrophobicity of the cataract-associated mutant, Pro23Thr of human gamma D-crystallin: molecular basis of cataract formation. *Biophys. J.* **100**, 539-539 (2011).
8. F. Ji, L. Koharudin, J. Jung, A. Gronenborn, Crystal structure of the cataract-causing P23T D-crystallin mutant. *Proteins: Struct., Funct., Bioinf.* **81**, 1493-1498 (2013).
9. A. Basak*,* O. Bateman, C. Slingsby, A. Pande, N. Asherie, O. Ogun, G. B. Benedek, J. Pande, High-resolution X-ray crystal structures of human gamma D crystallin (1.25 angstrom) and the R58H mutant (1.15 angstrom) associated with aculeiform cataract. *J. Mol. Biol.* **328**, 1137-1147 (2003).
10. J. Boatz, M. Whitley, M. Li, A. Gronenborn, P. van der Wel, Cataract-associated P23T gamma D-crystallin retains a native-like fold in amorphous-looking aggregates formed at physiological pH. *Nat. Commun.* **8**, (2017).
11. F. Ji, J. Jung, L. Koharudin, A. Gronenborn, The human W42R gamma D-crystallin mutant structure provides a link between congenital and age-related cataracts. *J. Biol. Chem.* **288**, 99-109 (2013).
12. A. Tardieu, F. Veretout, B. Krop, C. Slingsby, Protein interactions in the calf eye lens - interactions between beta-crystallins are repulsive whereas in gamma-crystallins they are attractive. *Eur. Biophys. J. Biophy.* **21**, 1-12 (1992).
13. C. N. Kingsley, W. D. Brubaker, S. Markovic, A. Diehl, A. J. Brindley, H. Oschkinat, R. W. Martin, Preferential and specific binding of human alpha B-crystallin to a cataract-related variant of gamma S-crystallin. *Structure* **21**, 2221-2227 (2013).
14. Z. Ma, G. Piszczek, P. Wingfield, Y. Sergeev, J. Hejtmancik, The G18V CRYGS mutation associated with human cataracts increases gamma S-crystallin sensitivity to thermal and chemical stress. *Biochemistry* **48**, 7334-7341 (2009).
15. W. D. Brubaker, J. A. Freites, K. J. Golchert, R. A. Shapiro, V. Morikis, D. J. Tobias, R. W. Martin, Separating instability from aggregation propensity in gamma S-crystallin variants. *Biophys. J.* **100**, 498-506 (2011).

16. R. J. Ellis, A. P. Minton, Protein aggregation in crowded environments. *Biol. Chem.* **387**, 485-497 (2006).
17. M. Feig, Y. Sugita, Variable interactions between protein crowders and biomolecular solutes are important in understanding cellular crowding. *J. Phys. Chem. B* **116**, 599-605 (2012).
18. S. R. McGuffee, A. H. Elcock, Atomically detailed simulations of concentrated protein solutions: the effects of salt, pH, point mutations, and protein concentration in simulations of 1000-molecule systems. *J. Am. Chem. Soc.* **128**, 12098-12110 (2006).
19. S. R. McGuffee, A. H. Elcock, Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *Plos. Comput. Biol.* **6**, e1000694 (2010).
20. P. Mereghetti, R. R. Gabdoulline, R. C. Wade, Brownian dynamics simulation of protein solutions: structural and dynamical properties. *Biophys. J.* **99**, 3782-3791 (2010).
21. M. Dlugosz, J. Trylska, Diffusion in crowded biological environments: applications of brownian dynamics. *Bmc. Biophys.* **4**, 3 (2011).
22. D. L. Ermak, J. A. Mccammon, Brownian dynamics with hydrodynamic interactions. *J. Chem. Phys.* **69**, 1352-1360 (1978).
23. E. Marcos, P. Mestres, R. Crehuet, Crowding induces differences in the diffusion of thermophilic and mesophilic proteins: a new look at neutron scattering results. *Biophys. J.* **101**, 2782-2789 (2011).
24. B. Chen, J. I. Siepmann, A novel Monte Carlo algorithm for simulating strongly associating fluids: applications to water, hydrogen fluoride, and acetic acid. *J. Phys. Chem. B* **104**, 8725-8734 (2000).
25. B. Chen, J. I. Siepmann, Improving the efficiency of the aggregation-volume-bias Monte Carlo algorithm. *J. Phys. Chem. B* **105**, 11275-11282 (2001).
26. A. Lomakin, N. Asherie, G. B. Benedek, Monte-Carlo study of phase separation in aqueous protein solutions. *J. Chem. Phys.* **104**, 1646-1656 (1996).
27. B. Chen, R. B. Nellas, S. J. Keasler, Fractal aggregates in protein crystal nucleation. *J. Phys. Chem. B* **112**, 4725-4730 (2008).
28. I. Staneva, D. Frenkel, The role of non-specific interactions in a patchy model of protein crystallization. *J. Chem. Phys.* **143**, 194511 (2015).
29. H. Liu, S. K. Kumar, Vapor-liquid coexistence of patchy models: relevance to protein phase behavior. *J. Chem. Phys.* **127**, 084902 (2007).
30. D. Fusco, P. Charbonneau, Crystallization of asymmetric patchy models for globular proteins in solution. *Phys. Rev. E* **88**, 012721 (2013).
31. M. Lund, B. Jönsson, A mesoscopic model for protein-protein interactions in solution. *Biophys. J.* **85**, 2940-2947 (2003).
32. M. Lund, Anisotropic protein-protein interactions due to ion binding. *Coll. Surf. B: Biointerfaces* **137**, 17-21 (2015).
33. W. Li*,* B. A. Persson, M. Morin, M. A. Behrens, M. Lund, M. Zackrisson Oskolkova, Charge-induced patchy attractions between proteins. *J. Phys. Chem. B* **119**, 503-508.
34. A. Kurut, B. A. Persson, T. Åkesson, J. Forsman, M. Lund, Anisotropic interactions in protein mixtures: self assembly and phase behavior in aqueous solution. *J. Phys. Chem. Lett.* **3**, 731-734 (2012).

35. B. Chen, J. I. Siepmann, K. J. Oh, M. L. Klein, Aggregation-volume-bias Monte Carlo simulations of vapor-liquid nucleation barriers for Lennard-Jonesium. *J. Chem. Phys.* **115**, 10903-10913 (2001).

36. J. R. Errington, A. Z. Panagiotopoulos, New intermolecular potential models for benzene and cyclohexane. *J. Chem. Phys.* **111**, 9731-9738 (1999).

37. M. D. Macedonia, E. J. Maginn, A biased grand canonical Monte Carlo method for simulating adsorption using all-atom and branched united atom models. *Mol. Phys.* **96**, 1375-1390 (1999).

38. A. Sepehri, T. D. Loeffler, B. Chen, Improving the efficiency of configurational-bias Monte Carlo: a density-guided method for generating bending angle trials for linear and branched molecules. *J. Chem. Phys.* **141**, (2014).

39. J. K. Shah, E. J. Maginn, A general and efficient Monte Carlo method for sampling intramolecular degrees of freedom of branched and cyclic molecules. *J. Chem. Phys.* **135**, (2011).

40. R. R. Gabdoulline, R. C. Wade, On the contributions of diffusion and thermal activation to electron transfer between phormidium laminosum plastocyanin and cytochrome f: Brownian dynamics simulations with explicit modeling of nonpolar desolvation interactions and electron transfer events. *J. Am. Chem. Soc.* **131**, 9230-9238 (2009).

41. W. L. Jorgensen, J. Tirado-Rives, The OPLS potential functions for proteins - energy minimizations for crystals of cyclic-peptides and crambin. *J. Am. Chem. Soc.* **110**, 1657-1666 (1988).

42. J. D. Madura, J. M. Briggs, R. C. Wade, M. E. Davis, B. A. Luty, A. Ilin, J. Antosiewicz, M. K. Gilson, B. Bagheri, L. R. Scott, J. A. McCammon , Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian dynamics program. *Comp. Phys. Commun.* **91**, 57-95 (1995).

43. N. A. Baker, D. Sept, S. Joseph, M. J. Holst, J. A. McCammon, Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA* **98**, 10037-10041 (2001).

44. R. R. Gabdoulline, R. C. Wade, Effective charges for macromolecules in solvent. *J. Phys. Chem.* **100**, 3868-3878 (1996).

45. N. Niimura, Y. Minezaki, T. Nonaka, J. Castagna, F. Cipriani, P. Hoghoj, M. S. Lehmann, C. Wilkinson, Neutron Laue diffractometry with an imaging plate provides an effective data collection regime for neutron protein crystallography. *Nat. Struct. Biol.* **4**, 909-914 (1997).

46. C. Bon, M. S. Lehmann, C. Wilkinson, Quasi-Laue neutron-diffraction study of the water arrangement in crystals of triclinic hen egg-white lysozyme. *Acta Crystallogr. D* **55**, 978-987 (1999).

47. H. Schwalbe, S. B. Grimshaw, A. Spencer, M. Buck, J. Boyd, C. M. Dobson, C. Redfield, L. J. Smith, A refined solution structure of hen lysozyme determined using residual dipolar coupling data. *Protein. Sci.* **10**, 677-688 (2001).

48. X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, A. E. Mark, Peptide folding: when simulation meets experiment. *Angew. Chem.-Intl. Ed.* **38**, 236-240 (1999).

49. B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl, GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory and Comp.* **4**, 435-447 (2008).

50. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926-935 (1983).

51. U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, L. G. Pedersen, A smooth particle mesh ewald method. *J. Chem. Phys.* **103**, 8577-8593 (1995).

52. B. Hess, H. Bekker, H. J. C. Berendsen, J. G. E. M. Fraaije, LINCS: a linear constraint solver for molecular simulations. *J. Comp. Chem.* **18**, 1463-1472 (1997).

53. S. Miyamoto, P. A. Kollman, Settle - an analytical version of the Shake and Rattle algorithm for rigid water models. *J. Comp. Chem.* **13**, 952-962 (1992).

54. H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola, J. R. Haak, Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684-3690 (1984).

55. S. Nosé, A molecular-dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **52**, 255-268 (1984).

56. M. Parrinello, A. Rahman, Polymorphic transitions in single-crystals - a new molecular-dynamics method. *J. Appl. Phys.* **52**, 7182-7190 (1981).

57. T. L. Hill, *An Introduction to Statistical Thermodynamics*. (Dover Publications, New York, 1986), pp. xiv, 508 p.

58. O. D. Velev, E. W. Kaler, A. M. Lenhoff, Protein interactions in solution characterized by light and neutron scattering: comparison of lysozyme and chymotrypsinogen. *Biophys. J.* **75**, 2682-2697 (1998).

59. B. H. Zimm, The scattering of light and the radial distribution function of high polymer solutions. *J. Chem. Phys.* **16**, 1093-1099 (1948).

60. P. J. Wyatt, Light scattering and the absolute characterization of macromolecules. *Anal. Chem. Acta* **272**, 1-40 (1993).

61. W. R. Krigbaum, F. R. Kuegler, Molecular conformation of egg-white lysozyme and bovine alpha-lactalbumin in solution. *Biochemistry* **9**, 1216--1223 (1970).

62. R Development Core Team, Vienna, Austria, (2008).

63. B. Efron, R. Tibshirani, *An Introduction to the Bootstrap*. (Chapman and Hall, London, 1993).

64. R. Piazza, M. Pierno, Protein interactions near crystallization: a microscopic approach to the Hofmeister series. *J. Phys.-Cond. Matter* **12**, A443-A449 (2000).

65. M. Muschol, F. Rosenberger, Interactions in undersaturated and supersaturated lysozyme solutions: static and dynamic light scattering results. *J. Chem. Phys.* **103**, 10424-10432 (1995).

66. B. Guo, S. Kao, H. McDonald, A. Asanov, L. L. Combs, W. William Wilson, Correlation of second virial coefficients and solubilities useful in protein crystal growth. *J. Cryst. Growth* **196**, 424-433 (1999).

67. D. F. Rosenbaum, A. Kulkarni, S. Ramakrishnan, C. F. Zukoski, Protein interactions and phase behavior: sensitivity to the form of the pair potential. *J. Chem. Phys.* **111**, 9882-9890 (1999).

68. G. Pellicane, D. Costa, C. Caccamo, Microscopic determination of the phase diagrams of lysozyme and gamma-crystallin solutions. *J. Phys. Chem. B* **108**, 7538-7541 (2004).

69. D. E. Kuehner, J. Engmann, F. Fergg, M. Wernick, H. W. Blanch, J. M. Prausnitz, Lysozyme net charge and ion binding in concentrated aqueous electrolyte solutions. *J. Phys. Chem. B* **103**, 1368-1374 (1999).
70. A. Stradner, F. Cardinaux, P. Schurtenberger, A small-angle scattering study on equilibrium clusters in lysozyme solutions. *J. Phys. Chem. B* **110**, 21222-21231 (2006).
71. Y. Liu, E. Fratini, P. Baglioni, W. R. Chen, S. H. Chen, Effective long-range attraction between protein molecules in solutions studied by small angle neutron scattering. *Phys. Rev. Lett.* **95**, 118402 (2005).
72. P. Mereghetti, R. Gabdoulline, R. Wade, Brownian dynamics simulation of protein solutions structural and dynamical properties. *Biophys. J.* **99**, 3782-3791 (2010).
73. A. Miklos, C. Li, N. Sharaf, G. Pielak, Volume exclusion and soft interaction effects on protein stability under crowded conditions. *Biochemistry* **49**, 6984-6991 (2010).
74. X. Xie, P. Choi, G. Li, N. Lee, G. Lia, Single-molecule approach to molecular biology in living bacterial cells. *Annu. Rev. Biophys.* **37**, 417-444 (2008).
75. I. Pastor, E. Vilaseca, S. Madurga, J. L. Garces, M. Cascante, F. Mas, Diffusion of alpha-chymotrypsin in solution-crowded media. A fluorescence recovery after photobleaching study. *J. Phys. Chem. B* **114**, 12182-12182 (2010).
76. M. Weiss, O. Wolfbeis, *Fluorescence Methods and Applications: Spectroscopy, Imaging, and Probes* **1130**, 21-27 (2008).
77. C. Li, G. Wang, Y. Wang, R. Creager-Allen, E. A. Lutz, H. Scronce, K. M. Slade, R. A. S. Ruf, R. A. Mehl, G. J. Pielak, Protein F-19 NMR in Escherichia coli. *Journal of the American Chemical Society* **132**, 321-327 (2010).
78. M. Senske *et al.*, Protein Stabilization by Macromolecular Crowding through Enthalpy Rather Than Entropy. *J. Am. Chem. Soc.* **136**, 9036-9041 (2014).
79. D. Gnutt, M. Gao, O. Brylski, M. Heyden, S. Ebbinghaus, Excluded-volume effects in living cells. *Angew. Chem. Int. Ed.* **54**, 2548-2551 (2015).
80. M. Feig, Y. Sugita, Reaching new levels of realism in modeling biological macromolecules in cellular environments. *J. Mol. Graphics Modell.* **45**, 144-156 (2013).
81. V. D. Prytkova, M. Heyden, D. Khago, J. A. Freites, C. T. Butts, R. W. Martin, D. J. Tobias, Multi-conformation Monte Carlo: a method for introducing flexibility in efficient simulations of many-protein systems. *J. Phys. Chem. B* **120**, 8115-8126 (2016).
82. M. Martinez, N. J. Bruce, J. Romanowska, D. B. Kokh, M. Ozboyaci, X. Yu, M. A. Ozturk, S. Richter, R. C. Wade, SDA 7: a modular and parallel implementation of the simulation of diffusional association software. *J. Comp. Chem.* **36**, 1631-1645 (2015).
83. R. Gabdoulline, R. Wade, Effective charges for macromolecules in solvent. *J. Phys. Chem.* **100**, 3868-3878 (1996).
84. A. Elcock, R. Gabdoulline, R. Wade, J. McCammon, Computer simulation of protein-protein association kinetics: Acetylcholinesterase-fasciculin. *J. Mol. Biol.* **291**, 149-162 (1999).
85. W. Jorgensen, J. Tiradorives, The OPLS potential function for proteins - energy minimizations for crystals of cyclic-peptides and crambin. *J. Am. Chem. Soc.* **110**, 1657-1666 (1988).
86. J. D. Madura, J. M. Briggs, R. C. Wade, M. E. Davis, B. A. Luty, A. Ilin, J. Antosiewicz, M. K. Gilson, B. Bagheri, L. R. Scott, J. A. McCammon, Electrostatics and diffusion of molecules

in solution - simulations with the University of Houston Brownian dynamics program. *Comp. Phys. Commun.* **91**, 57-95 (1995).

87. S. Bucciarelli, N. Mahmoudi, L. Casal-Dujat, M. Jehannin, C. Jud, A. Stradner, Extended law of corresponding states applied to solvent isotope effect on a globular protein. *J. Phys. Chem. Lett.* **7**, 1610-1615 (2016).

88. J. C. Gordon, J. B. Myers, T. Folta, V. Shoja, L. S. Heath, A. Onufriev, H++: a server for estimating pK(a)s and adding missing hydrogens to macromolecules. *Nucleic Acids Res.* **33**, W368-W371 (2005).

89. A. Patriksson, D. van der Spoel, A temperature predictor for parallel tempering simulations. *Phys. Chem. Chem. Phys.* **10**, 2073-2077 (2008).

90. W. Jorgensen, J. Chandrasekhar, J. Madura, R. Impey, M. Klein, Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926-935 (1983).

91. J. Jung, I. Byeon, Y. Wang, J. King, A. Gronenborn, The structure of the cataract-causing P23T mutant of human gamma D-crystallin exhibits distinctive local conformational and dynamic changes. *Biochemistry* **48**, 2597-2609 (2009).

92. V. Kumaraswamy, P. Lindley, C. Slingsby, I. Glover, An eye lens protein-water structure: 1.2 angstrom resolution structure of gamma B-crystallin at 150K. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **52**, 611-622 (1996).

93. R. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, A. D. MacKerell, Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *J. Chem. Theory Comput.* **8**, 3257-3273 (2012).

94. W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics. *J. Mol. Graphics Modell.* **14**, 33-38 (1996).

95. J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, K. Schulten, Scalable molecular dynamics with NAMD. *J. Comp. Chem.* **26**, 1781-1802 (2005).

96. T. Darden, D. York, L. Pedersen, Particle mesh Ewald - an N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089-10092 (1993).

97. H. Grubmuller, H. Heller, A. Windemuth, K. Schulten, Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions. *Mol. Simul.* **6**, 121-142 (1991).

98. J. Ryckaert, G. Ciccotti, H. Berendsen, Numerical-integration of cartesian equations of motion of a system with constraints – molecular dynamics of N-alkanes. *J. Comput. Phys.* **23**, 327-341 (1977).

99. G. Martyna, D. Tobias, M. Klein, Constant-pressure molecular dynamics algorithms. *J. Chem. Phys.* **101**, 4177-4189 (1994).

100. S. Feller, Y. Zhang, R. Pastor, B. Brooks, Constant-pressure molecular dynamics simulation – the Langevin piston method. *J. Chem. Phys.* **103**, 4613-4621 (1995).

101. D. E. Shaw, J. P. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, et al., Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. *Sc14: International Conference For High Performance Computing, Networking, Storage and Analysis*, 41-53 (2014).

102. M. Tuckerman, B. Berne, G. Martyna, Reversible multiple time scale molecular dynamics. *J. Chem. Phys.* **97**, 1990-2001 (1992).

103. Y. Shan, J. Klepeis, M. Eastwood, R. Dror, D. Shaw, Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. *J. Chem. Phys.* **122**, (2005).

104. G. Martyna, M. Klein, M. Tuckerman, Nose-Hoover chains – the canonical ensemble via continuous dynamics. *J. Chem. Phys.* **97**, 2635-2643 (1992).

105. R. A. Lippert, C. Predescu, D. J. Ierardi, K. M. Mackenzie, M. P. Eastwood, R. O. Dror, D. E. Shaw, Accurate and efficient integration for molecular dynamics simulations at constant temperature and pressure. *J. Chem. Phys.* **139**, (2013).

106. World Health Organization. (2014).

107. U. Andley, Crystallins in the eye: function and pathology. *Prog. Retinal Eye Res.* **26**, 78-98 (2007).

108. P. Hains, R. Truscott, Post-translational modifications in the nuclear region of young, aged, and cataract human lenses. *J. Proteome Res.* **6**, 3935-3943 (2007).

109. J. Hejtmancik, Congenital cataracts and their molecular genetics. *Sem. Cell Dev. Biol.* **19**, 134-149 (2008).

110. K. Moreau, J. King, Protein misfolding and aggregation in cataract disease and prospects for prevention. *Trends Mol. Med.* **18**, 273-282 (2012).

111. W. Brubaker, R. Martin, H-1, C-13, and N-15 assignments of wild-type human gamma S-crystallin and its cataract-related variant gamma S-G18V. *Biomol. NMR Assign.* **6**, 63-67 (2012).

112. K. W. Roskamp, D. M Montelongo, C. D. Anorma, D. N. Bandak, J. A. Chua, K. T. Malecha, R. W. Martin, Multiple aggregation pathways in human gamma S-crystallin and its aggregation-prone G18V variant. *Invest. Ophthalmol. Vis. Sci.* **58**, 2397-2405 (2017).

113. Z. Xia, Z. Yang, T. Huynh, J. King, R. Zhou, UV-radiation induced disruption of dry-cavities in human gamma D-crystallin results in decreased stability and faster unfolding. *Sci. Rep.* **3**, (2013).

114. V. Lapko, A. Purkiss, D. Smith, J. Smith, Deamidation in human gamma S-crystallin from cataractous lenses is influenced by surface exposure. *Biochemistry* **41**, 8638-8648 (2002).

115. B. B. Wang, C. Yu, Y. B. Xi, H. C. Cai, J. Wang, S. Zhou, S. Zhou, Y. Wu, Y. B. Yan, X. Ma, L. Xie, A novel CRYGD mutation (p.Trp43Arg) causing autosomal dominant congenital cataract in a chinese family. *Hum. Mutat.* **32**, E1939-E1947 (2011).

116. E. Serebryany, J. C. Woodard, B. V. Adkar, M. Shabab, J. A. King, E. I. Shakhnovich, An internal disulfide locks a misfolded aggregation-prone intermediate in cataract-linked mutants of human gamma D-crystallin. *J. Biol. Chem.* **291**, 19172-19183 (2016).

117. K. Papanikolopoulou, I. Mills-Henry, S. L. Thol, Y. Wang, A. A. R. Gross, D. A. Kirschner, S. M. Decatur, J. King, Formation of amyloid fibrils in vitro by human gamma D-crystallin and its isolated domains. *Mol. Vis.* **14**, 81-89 (2008).

118. P. Das, J. King, R. Zhou, Aggregation of gamma-crystallins associated with human cataracts via domain swapping at the C-terminal beta-strands. *Proc. Nat. Acad. Sci. U. S. A.* **108**, 10514-10519 (2011).

119. G. Benedek, Cataract as a protein condensation disease - The Proctor Lecture. *Invest. Ophthalmol. Vis. Sci.* **38**, 1911-1921 (1997).

120. A. Pande, K. Ghosh, P. Banerjee, J. Pande, Increase in surface hydrophobicity of the cataract-associated P23T mutant of human gamma D-crystallin is responsible for its dramatically lower, retrograde solubility. *Biochemistry* **49**, 6122-6129 (2010).

121. U. Essmann, L. Perera, M. L. Berkowitz, A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577-8593 (1995).

122. P. Mereghetti, R. R. Gabdoulline, R. C. Wade, Brownian dynamics simulation of protein solutions structural and dynamical properties. *Biophys. J.* **99**, 3782-3791 (2010).

123. A. H. Elcock, R. R. Gabdoulline, R. C. Wade, J. A. McCammon, Computer simulation of protein-protein association kinetics: Acetylcholinesterase-fasciculin. *J. Mol. Biol.* **291**, 149-162 (1999).

124. N. Metropolis, S. Ulam, The Monte Carlo method. *J. Am. Stat. Assoc.* **44**, 335-341 (1949).

125. M. Tien, A. Meyer, D. Sydykova, S. Spielman, C. Wilke, Maximum allowed solvent accessibilites of residues in proteins. *Plos One* **8**, (2013).

126. E. Serebryany, J. A. King, WT human gamma D crystallin promotes aggregation of its oxidation-mimicking mutants. *Biophys. J.* **110**, 387A-388A (2016).

127. D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Ierardi, Anton, a special-purpose machine for molecular dynamics simulation. *Commun. Acm* **51**, 91-97 (2008).