# Molecular Dynamics Simulations of Polyglutamine Aggregation Using Solvent-Free Multiscale Coarse-Grained Models

**Yanting Wang[†] and Gregory A. Voth*[,‡]**

*Key Laboratory of Frontiers in Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, 55 East Zhongguancun Road, Beijing, 100190 China, and Department of Chemistry, James Franck Institute, and Computation Institute, University of Chicago, 5735 South Ellis Avenue, Chicago, Illinois, 60637*

*Received: January 26, 2010; Revised Manuscript Received: June 2, 2010*

The multiscale coarse-graining (MS-CG) method is used to construct solvent-free CG models for polyglutamine peptides having various repeat lengths. Because the resulting CG models have fewer degrees of freedom than a corresponding all-atom simulations, they make it possible to study the self-assembly of polyglutamines at high concentrations for the first time by allowing for better equilibration and statistical sampling that is well beyond the range achievable by all-atom models. Molecular dynamics (MD) simulations performed with these models show that polyglutamine monomers with repeat lengths $\leq 28$ fluctuate between their folded and unfolded states. Monomers with 32 or more residues are stable and form $\alpha$-helix solid structures. The degree of monomer compactness increases with chain length in both cases. CG MD simulations of equilibrium polyglutamine aggregates show that even at high concentrations, the system statistically fluctuates between heterogeneous and homogeneous configurations, rather than simply aggregates. The degree of aggregation and fluctuation increases with concentration and chain length. All of these phenomena are consistent with the experimental observations and may be explained by a mechanism that the collective nonbonded interactions between polyglutamine molecules in water solution are only weakly attractive. Finally, this work demonstrates that computer simulation of polypeptides self-assembly and aggregation, which is presently beyond the reach of all-atom MD simulations, is attainable using solvent-free MS-CG models.

## 1. Introduction

The self-assembly and aggregation of polypeptides are essential for building biomolecular materials and also of medical and clinical importance (see, e.g., refs 1 and 2). Because current all-atom molecular dynamics (MD) simulations of polypeptides have a very limited spatial and temporal range (typically several nanometers and tens of nanoseconds) as compared with experimental scales (more than micrometers and microseconds), the molecular origins of many experimental results have yet to be explored via computer simulation.

The expansion of polyglutamine (pGLN) sequences within proteins is believed to be the clinical cause of several neural diseases, including Huntington's disease and Alzheimer's disease.[3] In Huntington's disease, pGLNs with 36 or fewer repeated units are almost harmless, but those with 38 or more are toxic. In addition, the onset of the disease occurs earlier as the length of the repeated sequence increases.[4] The neurotoxicity of aggregated pGLNs and possible treatments have been investigated in several experimental studies,[5−10] but the details of this aggregation process at the molecular level are still unclear.

In an X-ray diffraction study, Perutz et al.[11] observed that synthetic poly(L-glutamine) forms $\beta$-sheets by hydrogen bonds. The electron microscopy studies by Scherzinger et al.[12] show that, in addition to pGLN repeat length, the protein concentration and processing time are critical to the organization of pGLN-containing Huntington fragments into amyloid-like fibrils. Thakur and Wetzel[13] studied the aggregation of various pGLN mutations. On the basis of a number of observed macroscopic properties, Chen et al.[14] concluded that pGLN follows the common protein aggregation pathway[15] in forming amyloid fibrils.

Taken together, the above results imply that amyloid fibril growth is a nucleation-dependent process. That is, its kinetics are influenced by protein concentration and length and can be accelerated by seeding. Chen et al. also found that the aggregation is nucleated by a monomer.[16] Furthermore, Slepko et al. discovered that the lag time of pGLN aggregation decreases exponentially with increasing repeat length, and the presence of normal-length pGLN in solution can greatly accelerate the nucleation of expanded pGLN peptides.[17] Colby et al. have studied the probability of aggregate formation in cells as a function of time and concentration with a stochastic mathematical model, extending the above aggregation nucleation scheme and confirming it with experiments.[18] Crick et al.[19] performed fluorescence correlation spectroscopy measurements to show that polyglutamine monomers form collapsed structures in water, and Walters and Murphy[20] investigated the monomeric conformations and aggregation properties of pGLNS containing 8−24 glutamines with several experimental methods. They concluded that not all peptides form secondary structure, but longer monomers are more collapsed; longer peptides can aggregate more easily and generate sediments.

If the details of pGLNs nucleation and aggregation remain unclear, one reason is that it is difficult experimentally to observe molecular structures and their changes on short time scales. Computer simulations are, therefore, quite valuable to help in the understanding of the molecular-scale interactions underlying aggregation. Only a few simulations have been performed for pGLNs, however; these are described below.

* To whom correspondence should be addressed. E-mail: gavoth@uchicago.edu.
[†] Key Laboratory of Frontiers in Theoretical Physics.
[‡] University of Chicago.

10.1021/jp1007768 
Published on Web 06/15/2010

Combining discrete molecular dynamics (DMD) based on a bead-string model with all-atom MD, Khare et al.[21] showed that the transition from a random coil to a parallel $\beta$-helix occurs in single pGLNs with more than 37 repeat units. In both implicit-solvent, united-atom MD simulations and explicit-solvent, all-atom MD simulations, Zanuy et al.[22] found a common motif in the organization of the pGLN fibril: $\beta$-helices and flat $\beta$-sheet segments linked together in a superhelical arrangement. The DMD simulations of Marchut and Hall[23] with discontinuous potentials gave rise to the spontaneous formation of aggregates and $\beta$-sheet annular structures. On the basis of all-atom MD simulations with repeat lengths of 5 and 15, Wang et al.[24] concluded that these smaller peptides are disordered. In addition, the monomeric pGLNs with longer repeat chains were more compact and exhibited more conformational fluctuation. Vatalis et al.[25] found that not only is water a poor solvent for polar polyglutamines, but also the polyglutamine monomer is intrinsically disordered and collapses in water. The same group performed atomistic simulations of polyglutamine and suggested that disordered, soluble, and molten oligomers are an early intermediate stage in the process of polyglutamine aggregation,[26] whereas the $\beta$-sheet formation is a feature of many-body systems rather than monomers or oligomers.[27] Esposito et al.[28] tested the dynamics of the steric zipper motifs of monomeric and assembled pGLNs and found that hydrogen bonding between side chains plays an important stabilizing role.

As with many complex biomolecular problems, all-atom MD simulations of pGLN aggregation are still beyond the power of present-day computers. Thus, most existing MD simulations employ simplified dynamics and potentials, limit their scope to only monomers, or have very short durations as compared with the characteristic time scale of the system. A realistic coarse-grained model can greatly speed up pGLN simulations by neglecting some atomistic details of the system. Many different CG approaches have already been developed and applied to biological systems (e.g., refs 29−33). Among these, the multiscale coarse-graining (MS-CG) methodology[34−37] rigorously and systematically builds up a CG model from all-atom simulations of the targeted system. MS-CG simulations have been successfully applied to molecular liquids,[35] ionic liquids,[38−40] nanoparticles,[41] and biomolecular systems.[34,42−49]

One of the difficulties with all-atom simulations of biological systems is the enormous number of water molecules involved. In principle, the MS-CG methodology can eliminate the atomistic forces due to water molecules by constructing effective CG forces between CG sites.[48,49] In this paper, the MS-CG method has been used to generate such "solvent-free" MS-CG models for pGLNs. As a result, both the monomeric form and aggregated pGLN structures could be simulated for effectively larger times with the systems being better equilibrated and adequately sampled. The nature of the intrinsic fluctuations is revealed for both cases, and we find that longer pGLNs exhibit stronger attractions. In agreement with experimental results,[12] the aggregation of pGLNs also depends on concentration, and even at high concentrations, the system fluctuates between heterogeneous and homogeneous configurations, rather than aggregates statistically. The simulation results may be explained by a mechanism wherein the collective nonbonded interactions between polyglutamine molecules in water solution are weakly attractive. The outcome of the present work further indicates that the MS-CG methodology may be applied to study the self-assembly and aggregation of other polypeptides to overcome the difficulty of the limited spatial and temporal scales accessible by all-atom MD simulations alone.

**TABLE 1: The Heterogeneity Order Parameter $\hat{h}_0$ for Ideal Uniform Systems Containing $N_s$ Sites**

| $N_s$ | $\hat{h}_0$ |
|---|---|
| 1 | 1.0000 |
| 8 | 4.1464 |
| 27 | 10.8388 |
| 64 | 12.9513 |
| 125 | 15.3220 |
| 216 | 15.5285 |
| 343 | 15.7368 |
| 512 | 15.7431 |
| 729 | 15.7495 |
| 1000 | 15.7495 |
| 1728 and larger | 15.7496 |

### 2. Heterogeneity Order Parameter

The radial distribution function[50] (RDF) quantifies local structure of a condensed phase system. However, the MS-CG model developed in this study aims to simulate *global* aggregation behavior; local structures are of secondary importance. Thus, we will rely on the *heterogeneity order parameter* (HOP), originally defined as a global geometric measure of heterogeneity in ionic liquid systems,[51,52] to quantify the degree of pGLN aggregation. For a given configuration in this study, the HOP is defined for a certain type of CG site (e.g., sites A) to have a Gaussian form:

$$\hat{h} = \frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}\exp(-r_{ij}^2/2\sigma^2) \quad (1)$$

where $r_{ij}$ is the distance between sites $i$ and $j$, corrected for periodic boundary conditions, and $\sigma = L/N_s^{1/3}$. Here, $L$ is the side length of the cubic simulation volume and $N_s$ is the total number of sites. According to this definition, the shorter distances between sites contribute more to the HOP value. The HOP takes a larger value when more CG sites are topologically closer.

The form of eq 1 ensures that the HOP is topologically invariant with the absolute size of the simulation box. When the number of sites is very limited, however, the HOP exhibits a finite size effect. To demonstrate this effect, some HOP values for ideally uniform systems with $N_s = n^3$ sites ($n = 1, 2, 3, ...$) in a cubic volume are listed in Table 1. It can be seen that the HOP attains its ideal, constant value (15.7496) only for $N \geq 729$.

To obtain a measure that is close to zero when sites are distributed almost uniformly, we further define the *reduced* HOP as

$$h = \hat{h} - \hat{h}_0 \quad (2)$$

where $\hat{h}_0$ is the HOP of a perfectly uniform distribution.

### 3. Solvent-Free Multiscale Coarse-Graining of Polyglutamine

The MS-CG approach[36,37] rigorously constructs a CG model for molecular systems from atomistic trajectories to obtain a two-body decomposition of the many-body potential of mean force of the system. First, instantaneous atomic positions and forces are sampled from an equilibrium all-atom MD simulation of the targeted system. The atoms are grouped into CG sites, whose centers of mass and net atomistic forces can then be

calculated. Finally, a variational principle is employed to calculate the set of effective CG forces that best reproduces (in the least-squares sense) the averaged instantaneous atomic forces. This approach has been related to the Yvon−Born−Green theory[53] which related two- and three-body structural correlations.

Although the MS-CG theory ensures that equilibrium structures can be reasonably simulated at the CG level, diffusion is always faster for CG species than for all-atom molecular systems. Thus, CG MD simulations with a given simulated time can correspond to much longer all-atom MD simulations. It is possible to correct for the faster diffusion,[54] but in the present study, it is actually a desirable feature. The characteristic time scale of polypeptide aggregation is too long to be well sampled by an all-atom MD simulation, but approachable for the MS-CG simulation.

All CG models eliminate some atomistic degrees of freedom and therefore lose some atomistic details. The preferred resolution of the CG model depends on the nature of the problem to be solved. To investigate the global aggregation behavior of pGLNs, we construct a very simplified, solvent-free model with two kinds of CG sites: one representing a whole backbone group and one for a whole side chain group. Although the local structures within our CG model are slightly different from those that would be obtained by a fully atomistic model, the global behavior quantified by the HOP is found to be in very good agreement.

**3.1. All-Atom Simulation.** The CHARMM (version 32b2)[55,56] script was used to generate the topology of one pGLN molecule containing 8 amino acid residues (abbreviated as Q8). The N-terminal is capped with ACE ($C_2H_3O$), and the C-terminal is capped with CT3 ($CNH_4$) so that both terminals are charge-neutral. The monomeric Q8 was then duplicated to obtain 32 molecules, and the remaining volume was filled with 6927 water molecules using the Gromacs (version 3.3)[57,58] software tools. The simulation box size is 64 Å. The OPLS-AA force field[59,60] parameters were assigned to the Q8 molecules, and SPC force field[61,62] parameters were assigned to the water molecules. In the initial configuration, the monomers were all straight and placed on lattice positions evenly distributed throughout the simulation box.

Using the Gromacs MD simulation program, the initial configuration was equilibrated with the constant *NVT* ensemble at a very high temperature ($T = 2000$ K). This step ensures that the system loses its memory of the initial configuration and achieves equilibration after 1 ns. The equilibrated configuration was then subjected to an annealing process, sequentially cooling from $T = 2000$ K down to 1700, 1400, 1200, 1000, 900, 800, 700, 600, 500, 400, and 310 K. At each temperature, the system was equilibrated for 1 ns. Annealing greatly accelerated equilibration of the system at $T = 310$ K. The configuration obtained at $T = 310$ K was equilibrated for an additional 1 ns (2 ns total). Finally, a production run of 1 ns was performed under the constant *NVT* ensemble at the same temperature. A total of 1000 configurations were evenly sampled during the production run. The net force on each atom was recorded at the same time. In all of the above simulations, the MD time step was 1 fs. The Berendsen thermostat[63] was employed to keep the simulation temperature constant, and the Ewald method[50] was employed to calculate long-range electrostatic interactions. Both the short-range, nonbonded interactions and real-space contributions to the Ewald sum had a cutoff of 12 Å.

**3.2. Coarse-Graining Procedure.** The MS-CG method[36,37] was applied to the atomic coordinates and net forces obtained in the all-atom simulation to build a CG model for pGLN. The CG strategy of Q8 is shown in Figure 1. The backbone is divided into several CG units (B), and each side chain is defined as a single CG site (S). The ACE is CG site A, and the CT3 is CG site C. In a solvent-free CG model, the water molecules are not represented explicitly. However, their contributions are included in the force models for each CG site. Obviously, this CG model has many fewer degrees of freedom than the all-atom simulation. More detailed CG models are possible, but more details on local structure were not needed for the purpose of this study.

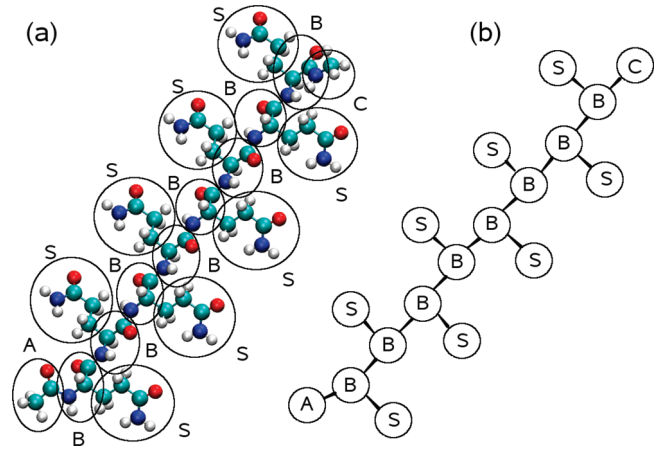


**Figure 1.** Schematic of the coarse-graining of Q8. (a) Atomic structure and coarse-graining scheme. Cyan spheres represent carbon atoms; red, oxygen; blue, nitrogen; and white, hydrogen. (b) Coarse-grained model. All coarse-grained sites have zero partial charges.

The sampled atomistic MD trajectories were converted into CG trajectories using the centers of mass of the CG sites (without water molecules) and the net forces on them (including the water contribution). The MS-CG procedure[37] with the $\delta$-function basis set was then applied to the CG trajectory to obtain effective bonded and nonbonded CG forces. Because there were totally 32 Q8 molecules, the nonbonded interactions were, in fact, calculated by averaging over all intermolecular CG sites and intramolecular CG sites that have three or more CG sites in between. Both the bin widths for nonbonded forces and bond stretching forces were set to 0.04 Å. The bonded valence angle and dihedral angle forces were divided into 100 bins. Because the distribution of Q8 monomers in the equilibrated solution is highly heterogeneous, the quality of the CG model is very sensitive to the cutoff distance for effective nonbonded CG forces. By experimenting with different values, a cutoff of 9.60 Å was found to yield the best CG model.

A few of the effective bonded potentials are plotted in Figure 2. Because it is convenient to use analytical expressions rather than discrete values, the simulated data are represented using fitted curves. The bond stretching potentials were fitted with a harmonic function, given by

$$U(r) = \frac{1}{2}K(r - r_0)^2 \tag{3}$$

The valence angle potentials were fitted with a quartic polynomial, given by

$$U(\theta) = \frac{1}{2}K_1(\theta - \theta_0)^2 + \frac{1}{3}K_2(\theta - \theta_0)^3 + \frac{1}{4}K_3(\theta - \theta_0)^4 \tag{4}$$

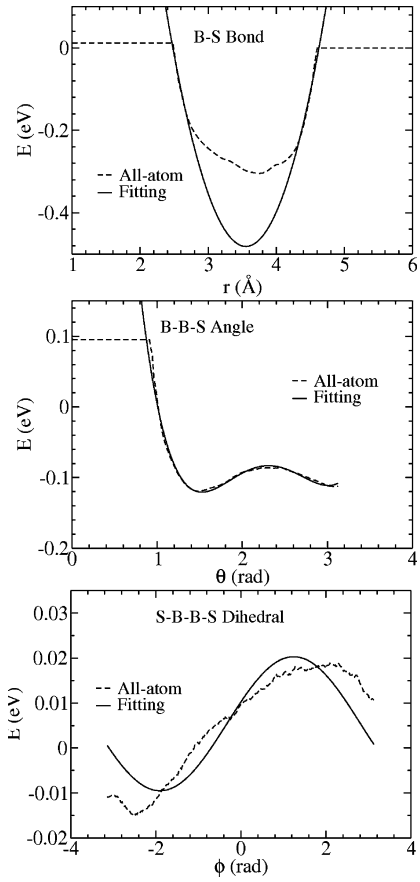


**Figure 2.** Fitting the B−S bond, B−B−S valence angle, and S−B−B-S dihedral angle potentials of Q8 with the functions given in eqs 3, 4, and 5.

The dihedral angle potentials were fitted with a cosine function, given by

$$U(\phi) = A[1 + \cos(\phi - \phi_0)] \quad (5)$$

In the above equations, $r$, $\theta$, and $\phi$ are bond length, valence angle, and dihedral angle, respectively. The corresponding equilibrium constants are $r_0$, $\theta_0$, and $\phi_0$, and $K$, $K_1$, $K_2$, $K_3$, and $A$ are force constants.

Example fitting functions for bond stretching, valence angle, and dihedral angle potentials are illustrated in Figure 2. The best-fit parameters for all bonded potentials are listed in Table 2. The valence angle and dihedral angle potentials are reasonably well described by the functions given in eqs 4 and 5, respectively, but the bond stretching potentials cannot be fit well by any simple function. The harmonic fit shown in Figure 2 was biased to better describe the upper part of the potential curve. Thus, the bond stretching interaction is slightly stronger in the CG simulation than in the collective atomistic simulation. However, the essence of the potential is maintained; because the bond stretching interaction is very rigid, a stronger force parameter has an effect on only the vibration frequency, not the structure.

The B−B, B−S, and S−S nonbonded CG forces obtained in the all-atom simulations are shown in Figure 3. Although they are in some ways similar to van der Waals (VDW) interactions, the forces cannot be fit with a simple function. Note that these effective forces include the effective contribution from water molecules. For each effective force, there is a "core region" $R < R_{core}$ that cannot be adequately sampled because the repulsive

**TABLE 2: Effective CG Parameters for Bonded Interactions**

| bond/angle | $r_0/\theta_0/\phi_0$ | $K/K_1/A$ | $K_2$, eV/rad$^3$ | $K_3$, eV/rad$^4$ |
|---|---|---|---|---|
| A−B bond | 3.13570 Å | 1.59599 eV/Å$^2$ | | |
| B−S bond | 3.55241 Å | 0.83696 eV/Å$^2$ | | |
| B−B bond | 3.40601 Å | 1.91808 eV/Å$^2$ | | |
| B−C bond | 2.67344 Å | 5.04676 eV/Å$^2$ | | |
| A−B−S valence angle | 98.837° | 0.3143 eV/rad$^2$ | −0.3675 | −0.0548 |
| A−B−B valence angle | 99.413° | 0.2546 eV/rad$^2$ | −1.0949 | 0.8515 |
| B−B−S valence angle | 87.415° | 0.4978 eV/rad$^2$ | −0.9713 | 0.4282 |
| B−B−B valence angle | 98.029° | 0.7885 eV/rad$^2$ | −2.0906 | 1.2288 |
| B−B−C valence angle | 113.289° | 0.2221 eV/rad$^2$ | −0.8650 | 0.8688 |
| S−B−C valence angle | 95.019° | 0.3826 eV/rad$^2$ | −0.6466 | 0.2015 |
| A−B−B−S dihedral angle | 116.4434° | 0.08336 eV | | |
| A−B−B−B dihedral angle | 37.2017° | 0.02846 eV | | |
| S−B−B−S dihedral angle | 71.0703° | 0.01491 eV | | |
| B−B−B−S dihedral angle | 46.9110° | 0.00844 eV | | |
| B−B−B−B dihedral angle | 24.5235° | 0.02386 eV | | |
| B−B−B−C dihedral angle | −10.6132° | 0.04640 eV | | |
| S−B−B−C dihedral angle | 144.1722° | −0.01949 eV | | |

VDW interaction prevents atoms from attaining small separations. To avoid system crash during CG MD simulations, the forces were tabulated outside of $R_{core}$ and extended into the core region using $F(0) = F(R_{core}) + 10$ eV/Å, where $R_{core}$ is independently determined for each force curve. The tabulated force data are available from the authors upon request.

**3.3. Validation of the Coarse-Grained Force Field.** Two MS-CG MD simulations were performed for a system including 32 Q8 monomers using the force fields constructed in the previous section, and their results were compared with the all-atom MD simulations. The only difference between the two CG MD simulations was their initial configuration. The first started from an equilibrated configuration identical to that used by the all-atom MD simulation to generate a directly comparable case. The second started from the unequilibrated initial configuration of the all-atom MD simulation, with the Q8 monomers evenly distributed on lattice positions, to determine whether the CG model is capable of producing pGLN aggregation.

The DL_POLY (version 2.14)[64] MD simulation program was used to run both simulations. First, the initial configurations

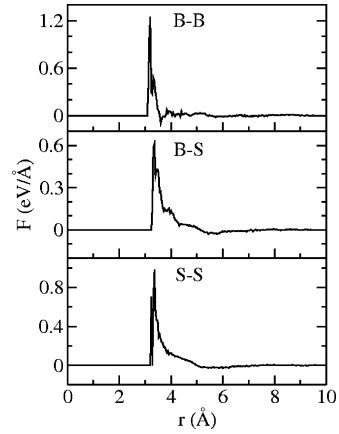


**Figure 3.** Coarse-grained forces between CG sites B-B, B-S, and S-S of polyglutamines, effectively including the water contribution.

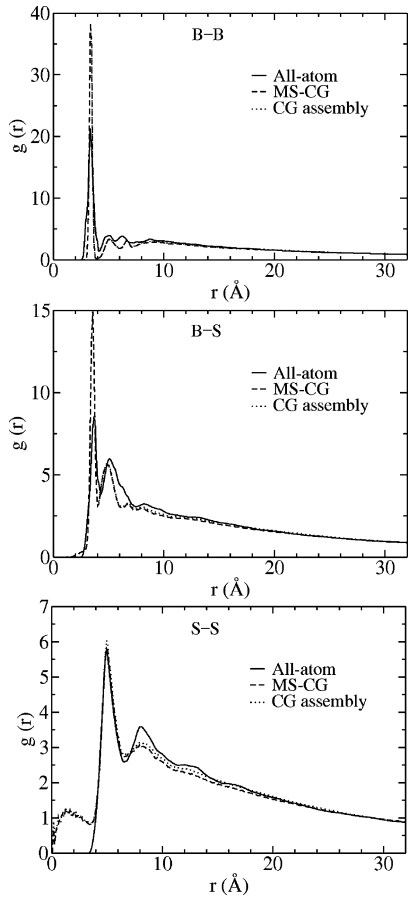


**Figure 4.** The radial distribution functions of polyglutamines from all-atom and coarse-grained molecular dynamics simulations. The "MS-CG" simulation starts from the last equilibrium configuration of the corresponding all-atom simulation, whereas the "CG assembly" simulation starts from a lattice configuration.

were run through a constant *NVT* MD simulation for 1 ns with a time step of 1 fs. The Hoover thermostat[65] was employed to hold the temperature at $T = 310$ K. The resulting RDFs of B−B, B−S, and S−S pairs are compared with those from the all-atom MD simulation in Figure 4. In these pictures, "MS-CG" denotes the CG run corresponding to the all-atom simulation initial conditions, and "CG assembly" denotes the run starting from a uniform Q8 distribution. The RDFs of the two CG simulations are almost the same. For B−B and B−S pairs, the first peaks (corresponding to bonded interactions) are about twice as high in the CG simulations as they are in the all-atom simulation. This difference is directly related to the fact that the fitted harmonic CG potential shown in Figure 2 is stronger at small separations than the atomistic potential. All other features of the RDFs are in reasonable agreement. For S−S pairs, except for a small peak appearing inside the core region, the CG RDFs are in very good agreement with the atomistic RDF. The satisfactory agreement between all RDFs at middle and long distances indicates that the CG model correctly reproduces the global behavior of the system. This claim is also demonstrated by the following comparison of HOP values.

The average HOP values, $\hat{h}$, calculated for all three simulations are compared in Table 3. It is clear that they agree within statistical error. However, the fluctuations in $\hat{h}$ observed in the all-atom MD simulation are several times smaller than those observed in the two CG MD simulations. This difference might be attributed to the fact that the all-atom MD simulation evolves much more slowly, so the 1 ns simulation covers only a very small portion of one fluctuation period. One nanosecond in the CG simulation, on the other hand, may correspond to several fluctuations of the aggregated system.

**TABLE 3: Comparison of Heterogeneity Order Parameters for Q8 from the All-Atom and CG Simulations[a]**

| site | all-atom | MS-CG | CG assembly |
|---|---|---|---|
| S | $33.47 \pm 0.40$ | $32.88 \pm 1.99$ | $32.89 \pm 2.09$ |
| B | $35.13 \pm 0.39$ | $33.18 \pm 1.89$ | $33.19 \pm 2.00$ |
| A | $15.62 \pm 0.13$ | $15.90 \pm 0.81$ | $15.77 \pm 0.91$ |
| C | $14.26 \pm 0.17$ | $13.96 \pm 0.60$ | $14.00 \pm 0.63$ |

[a] The "MS-CG" simulation starts from the last equilibrium configuration of the corresponding all-atom simulation; the "CG assembly" simulation starts from a lattice configuration.

**3.4. Extension to Longer pGLN Chains.** In the following section, the MS-CG force field developed for Q8 is extended to enable CG simulations of pGLN molecules with more amino acid residues. This was done by increasing the number of B and S sites, multiplying the number of bonded interactions required, and directly applying the effective CG bonded and nonbonded interactions derived in the previous section to the new interactions. Note that the validity of this simple strategy is not guaranteed by the MS-CG theory, because the effective CG interactions are fitted to the set of atomistic forces for a given system at a given thermodynamic condition. However, the simulation temperature is still $T = 310$ K, and the way the CG MD simulations are performed ensures that the new CG sites experience local environments that are very similar to those around the original sites. Thus, it is reasonable to expect that applying the CG force fields derived for Q8 directly to CG models with longer pGLN molecules will introduce relatively small errors. It should also be noted that this strategy has been successfully implemented for ionic liquids with increasing cation alkyl tail lengths.[66]

## 4. Monomer Simulations

CG MD simulations were performed for monomeric pGLNs with repeat lengths ranging from Q8 to Q56, with an interval of four amino acid residues. Each simulation contained a single monomer, and all monomers were initially configured as linear peptides. The simulations were run for 400 ns in a constant *NVT* ensemble at $T = 310$ K, with a time step of 4 fs (for $10^8$ steps). In such long simulations, unless its parameter is chosen exactly correctly, the Hoover thermostat allows the total energy to drift, with unphysical results. To avoid the complexity of finely tuning the Hoover parameter, we instead used the Evans thermostat.[67] The box lengths of the simulations were 8 Å times the number of residues, large enough that each monomer was free of interactions from its images due to the periodic boundary conditions. As previously noted, the elimination of so many degrees of freedom (especially for the water molecules) means that the time scale of statistical sampling in CG simulations is effectively several orders of magnitude longer than it would be in a corresponding atomistic simulation.

A total of $10^4$ configurations were sampled for each CG run. The compactness of the pGLN monomers was roughly quantified as $L = d/n$, where $d$ is the distance from CG site A to CG site C and $n$ is the repeat length. For each monomer, the instantaneous $L$ values were averaged over all sampled configurations to determine their ensemble average value $\langle L \rangle$. The results are plotted in Figure 5. From Q8 to Q28, $\langle L \rangle$ decreases monotonically with repeat length. The fluctuations in $\langle L \rangle$, depicted as error bars of one standard deviation, are rather large.

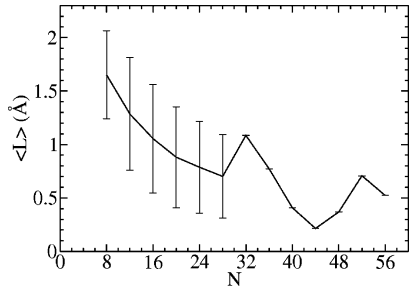


**Figure 5.** Average length per residue for monomeric polyglutamines with various repeat lengths. The error bars are one standard deviation.

This observation agrees with the all-atom MD simulations by Wang and co-workers,[24] who found that Q5 and Q15 monomers are disordered and give unstable configurations. They also observed that Q15 experienced more fluctuation than Q5. However, our own standard deviations do not show much difference for fluctuations from Q8 to Q28.

In simulations of Q32 through Q56, $\langle L \rangle$ does not fluctuate at all. Visual examination of the trajectories showed that these monomers quickly formed stable structures. Two random configurations of Q28 and the solid configurations of Q32−Q56 are shown in Figure 6. The two configurations of Q28 demonstrate that shorter pGLN monomers fluctuate between compact and extended shapes. Q32 stabilizes in a hook-shaped configuration. As for Q36 and longer monomers, the greater part of the structure forms an α-helix. Because the C-terminal (CG site C) points out of the coil in some of the solid configurations, $\langle L \rangle$ does not monotonically decrease with repeat length. However, the coiled parts of the monomers do become more compact with increasing repeat length. Note that the solid structures obtained in these simulations might not be the only ones possible. A complete investigation of all possible solid structures is left to future investigations.

The above simulation results suggest that the collective attraction between GLN amino acid residues in water increases with increasing repeat length. Our simulation results that Q8−Q28 do not form a secondary structure and longer monomers are more collapsed is also consistent with the experimental observation by Walters and Murphy[20] for Q8−Q24. When the repeat length is ≥32, the collective attraction and bonded interactions are strong enough to stabilize the monomer into a solid structure. In shorter monomers, the free energy difference between the folded and unfolded structures is so small that the solid state cannot stabilize at $T = 310$ K.

**5. Simulations of Aggregation.** The MS-CG models developed in previous sections were also used to simulate the aggregation of pGLNs. Although it is possible to study the nonequilibrium process of early self-assembly, in this work, we concentrate on the structures and properties of equilibrium states. The final monomer configurations from the simulations of section 4 were duplicated to obtain 27 identical monomers evenly distributed on lattice positions in a cubic simulation volume. The number of monomers was chosen to be small so that very long simulations were easy to perform. DL_POLY[64] was used to perform constant *NVT* CG MD simulations. The time step was 4 fs, and the Evans thermostat[67] was used to maintain the system at 310 K. The monomers aggregated and equilibrated in these simulations after about 10 ns, as was monitored by the HOP. The equilibrated system then went through a 400 ns CG production run, with $10^4$ evenly spaced outputs.

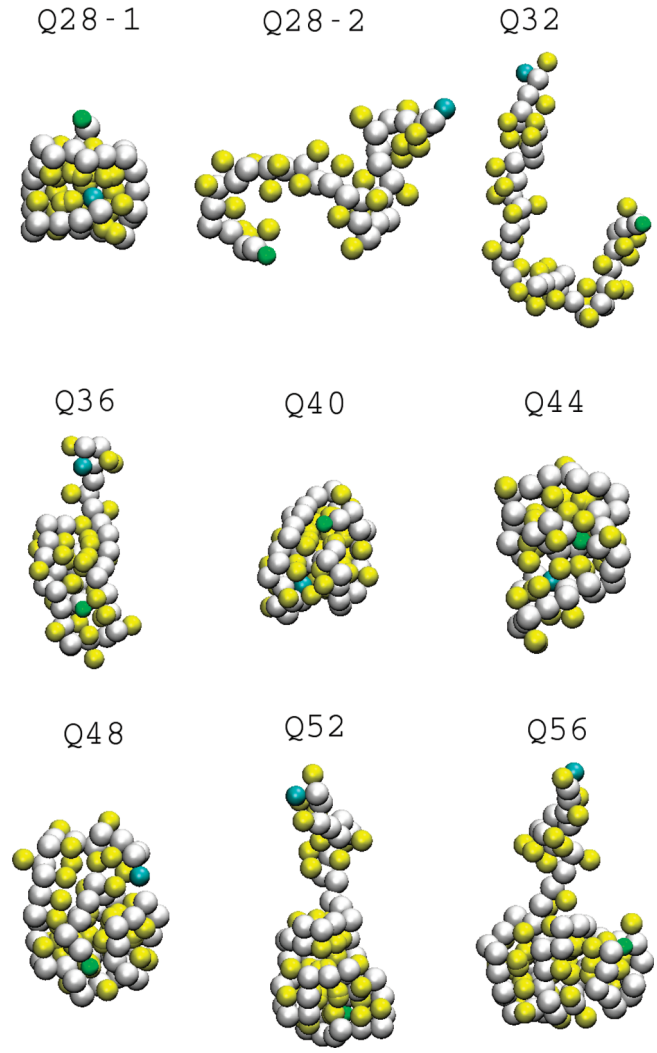


**Figure 6.** Pictures of monomeric polyglutamines. The first two pictures are random snapshots of Q28, a molecule in fluctuation. The others represent solid structures in equilibrium obtained for longer polyglutamines. White spheres represent the B CG sites; yellow, the S sites; green, the A sites; and cyan, the C sites.

The Q36 system with 27 monomers was simulated in a cubic volume of size 15.6 nm, corresponding to a concentration of about 11.8 mM. This concentration was intentionally chosen to be orders of magnitude higher than typical experimental values (usually on the order of micromolar) for two reasons: (1) the equilibration time is much shorter for higher-concentration systems, and (2) we wished to find out whether fluctuations exist in solutions with extremely high concentrations.

The instantaneous reduced HOPs for CG site A of the equilibrated 27-Q36 system are plotted in Figure 7. CG sites C and B give almost identical reduced HOP values, indicating that this measure of the global aggregation does not alter with the specific choice of CG site. The degree of aggregation fluctuates over a wide range of values, from very compact ($h > 9$) to very homogeneous ($h \sim 0$), even at high concentration. Two random configurations from the sampled trajectory are shown in Figure 8. In the left panel, the 27 monomers are grouped into several clusters. In the right panel, they have aggregated into a single structure. The large fluctuation in HOP demonstrates that the collective intermolecular attractions among pGLNs in solvent are relatively weak, comparable to the thermal energy at $T = 310$ K.

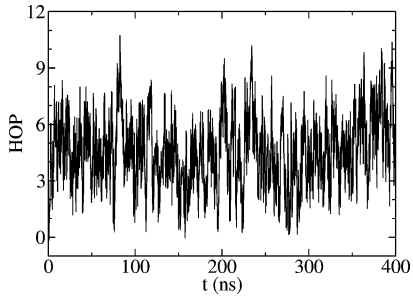


**Figure 7.** The instantaneous reduced heterogeneity order parameter of one A site in Q36. The side length of the simulation cube is 15.6 nm, corresponding to a concentration of 11.8 mM.

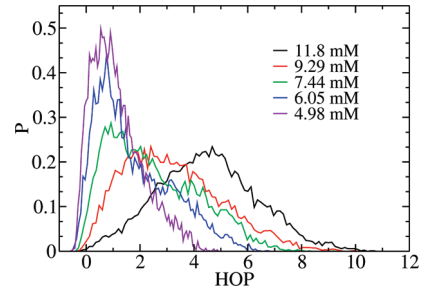


**Figure 9.** Distributions of the reduced heterogeneity order parameter obtained in CG MD simulations of Q36 at different concentrations.

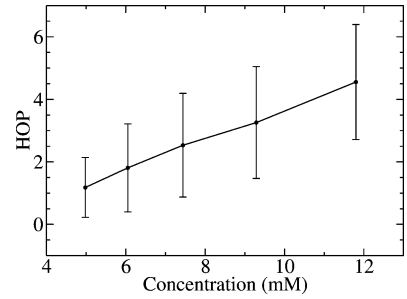


**Figure 10.** Average reduced heterogeneity order parameters from CG MD simulations of Q36 at different concentrations. The error bars represent one standard deviation.

The 27-Q36 system was also simulated in four larger volumes: 16.9 nm (concentration of 9.29 mM), 18.2 nm (7.44 mM), 19.5 nm (6.05 mM), and 20.8 nm (4.98 mM). Their reduced HOP distributions are compared in Figure 9. All five curves (including the 15.6 nm volume) are close to Γ-distributions. With increasing concentration, the distribution becomes broader, and the peak position shifts to larger values. The average value of the reduced HOP is plotted against simulation box size in Figure 10. The error bars denote the standard deviations of the reduced HOP distributions. Figures 9 and 10 show that the pGLNs are more likely to aggregate at high concentrations, but the fluctuation in HOP is also larger. This is consistent with experimental observations[12,18] that the protein concentration is critical to the amyloid-like fibril formation of pGLN-containing Huntington fragments.

The simulation box size was then fixed at 15.6 nm (11.8 mM), and the same simulations were performed for pGLNs with repeat lengths from Q8 to Q44. Simulations for Q48 and longer chains could not be performed at this volume, which is too small to accommodate so many CG sites. The average reduced HOPs and their standard deviations are plotted against the number of residues in Figure 11, and the corresponding distributions are shown in Figure 12. The average reduced HOP increases from Q8 to Q32, then is roughly constant from Q32 to Q44. These results demonstrate that the degree of aggregation increases for longer chains until the density saturates and qualitatively agrees with the experimental observation for Q8 to Q24.[20] It is interesting to note that, as shown in Figure 12, all distributions have a minimum reduced HOP less than zero. This fact implies that the systems always have certain probability to be homogeneous, no matter how long the pGLNs and how high the concentration.

The simulations were then repeated with a simulation box size of 20.8 nm (4.98 mM) for models Q8−Q56. The average reduced HOPs and their standard deviations are also plotted in Figure 11, so they might be compared with the previous set (15.6 nm volume).

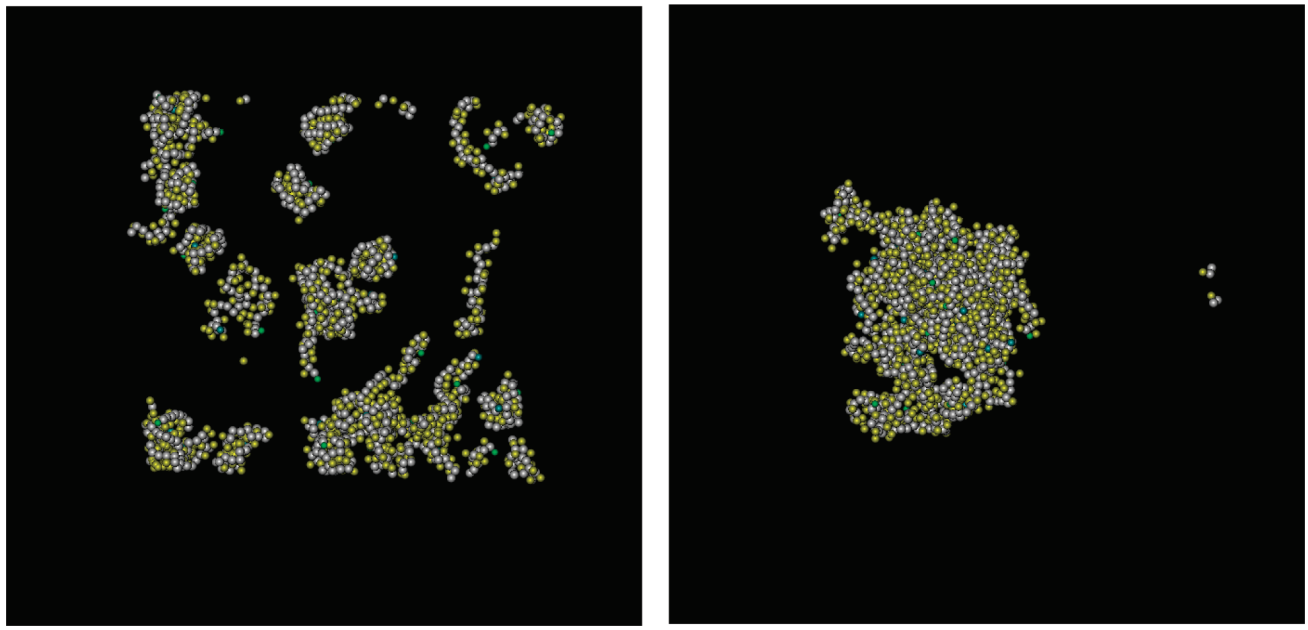

**Figure 8.** Two random snapshots from the 27-Q36 MD simulations, showing very different degrees of aggregation in the same system. The side length of the simulation cube is 15.6 nm, corresponding to a concentration of 11.8 mM. White spheres represent the B CG sites; yellow, the S sites; green, the A sites; and cyan, the C sites.

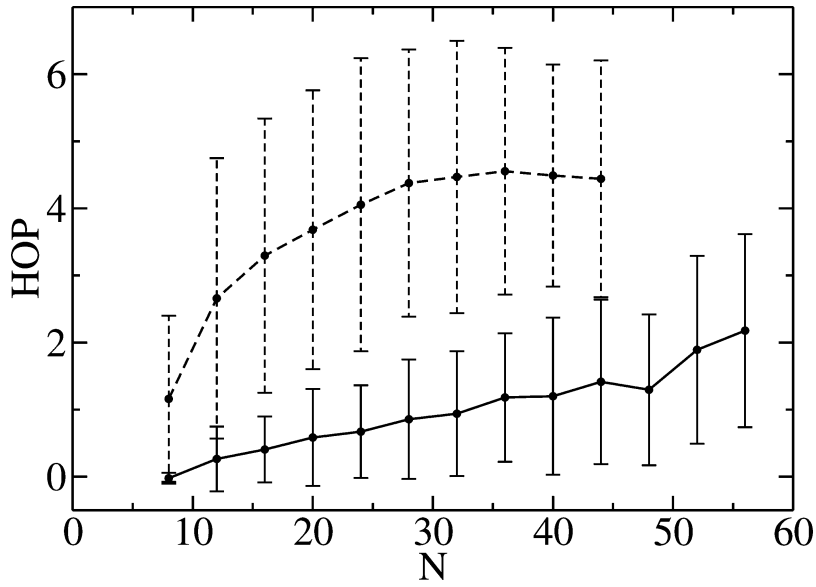


**Figure 11.** Average reduced heterogeneity order parameters from CG MD simulations of polyglutamines with various lengths. Two sets of simulations were performed, with box sizes of 15.6 nm (dashed lines) and 20.8 nm (solid lines), corresponding to concentrations of 11.8 and 4.98 mM, respectively. The error bars represent one standard deviation.

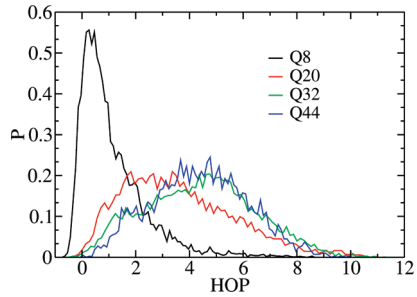


**Figure 12.** Distributions of reduced heterogeneity order parameters obtained in CG MD simulations of polyglutamines with various lengths. The simulation box size is fixed at 15.6 nm, concentration of 11.8 mM.

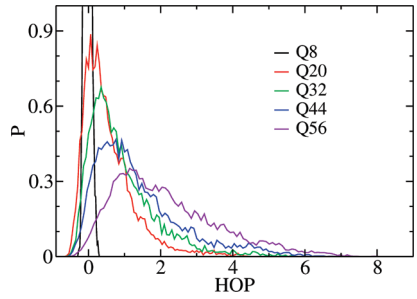


**Figure 13.** Distributions of reduced heterogeneity order parameters obtained in CG MD simulations of polyglutamines with various lengths. The simulation box size is fixed at 20.8 nm, concentration of 4.98 mM.

The corresponding distributions are shown in Figure 13. At this more dilute concentration, the average reduced HOP and its standard deviation continue increasing for longer chains. All systems show smaller HOPs and standard deviations than the corresponding 15.6 nm simulations, demonstrating that higher concentrations result in higher degrees of aggregation.

Finally, unlike the Γ-distributions observed in other systems, it is interesting that the HOP distribution of Q8 in a 20.8 nm volume has a sharp peak around zero. Visual examination of the trajectory showed that such configurations correspond to the rapid equilibration of Q8 monomers (initially evenly distributed in the simulation box) after some turbulent movement.

## 6. Conclusions and Discussion

The MS-CG method was applied to construct solvent-free CG models suitable for studying the aggregation behavior of pGLNs. These resulting MD simulations with these CG models accomplish a high degree of equilibration and statistical sampling that is well beyond the capability of all-atom MD simulations. The heterogeneity order parameter (HOP) was then used to quantify the degree of aggregation.

The single-monomer simulations indicate that Q8−Q28 monomers fluctuate between the unfolded and folded states. In contrast, Q32−Q56 are basically stable with α-helix solid structures. In both cases, the compactness of the solid structure increases with repeat length. Simulations of equilibrated pGLN aggregations show that the HOP and its variance increase with both concentration and repeat length. Nevertheless, no matter how high the concentration and how long the peptide, these systems always have a certain probability of being very homogeneous. These observations are consistent with experimental and all-atom simulations results. A unified perspective can thus be achieved from the CG simulations that the collective effective interactions among pGLN residues in water are only weakly attractive, comparable to the thermal energy at $T = 310$ K. In addition, with respect to the aggregation of pGLNs, the stronger attraction between long chains leads to denser aggregation due to the more attractive contributions from the larger numbers of residues. The S−S and B−S interactions above 5 Å are mainly responsible for the aggregation when the molecules are relatively far from each other, whereas the B−B interaction around 3.5 Å is more important when the molecules move closer. On the other hand, the weakness of the intermolecular interactions is responsible for the concentration-dependence of aggregation probability and large fluctuations in the HOP.

This work has studied only the equilibrium properties of pGLN aggregates. Nonequilibrium MD simulations investigating the details of the nucleation process and self-assembly will be necessary to complete our understanding of pure pGLN self-assembly and aggregation in vitro.

The CG MD aggregation simulations in this work contained only 27 pGLN molecules. Such small systems permit us to investigate the aggregation phenomenon on very long time

scales. Because the solvent-free CG models for pGLNs have very few degrees of freedom (one site for each backbone and one for each side chain), similar studies of larger systems are both feasible and desirable. Future work in this direction will serve to reduce the possible finite size effect and improve the physical accuracy of our results.

The concentrations simulated in this work were chosen near the saturation point to verify that fluctuation is an intrinsic property of aggregated pGLN systems. Future simulations will be more directly comparable to in vitro experiments. In addition, living cells are much more complicated than the pure pGLN aggregates simulated here. For example, sodium and other ions are also present in the solvent, and the proteins may not be pure pGLNs. More detailed CG simulations will be required to clarify how the aggregation of pGLNs is manifested in such environments.

**Acknowledgment.** This research was supported by a Collaborative Research in Chemistry grant from the National Science Foundation (CHE-0628257). The authors thank Drs. Sergei Izvekov, Gary Ayton, Pu Liu, and Qiang Shi for useful discussions. Allocations of computer time at the Lonestar supercomputer, granted by the Texas Advanced Computing Center, are gratefully acknowledged. The early stages of this research were completed at the Center for Biophysical Modeling and Simulation, University of Utah.

## References and Notes

(1) Zhang, S. *Nat. Biotechnol.* **2003**, *21*, 1171.
(2) Okazawa, H. *Cell. Mol. Life. Sci.* **2003**, *60*, 1427.
(3) Cummings, C. J.; Zoghbi, H. Y. *Hum. Mol. Genet.* **2000**, *9*, 909.
(4) Myers, R. H.; Marans, K. S.; MacDonald, M. E. In *Genetic Instabilities and Hereditary Neurological Diseases*; Wells, R. D., Warren, S. T., Eds.; Academic: San Diego, 1998; pp 301−323.
(5) Sánchez, I.; Mahlke, C.; Yuan, J. *Nature* **2003**, *421*, 373.
(6) Muchowski, P. J.; Schaffar, G.; Sittler, A.; Wanker, E. E.; Hayer-Hartl, M. K.; Hartl, F. U. *Proc. Natl. Acad. Sci., U.S.A.* **2000**, *97*, 7841–7846.
(7) Heiser, V.; Engernann, S.; Brocker, W.; Dunkel, I.; Boeddrich, A.; Waelter, S.; Nordhoff, E.; Lurz, R.; Schugardt, N.; Rautenberg, S.; Herhaus, C.; Barnickel, G.; Bottcher, H.; Lehrach, H.; Wanker, E. E. *Proc. Natl. Acad. Sci., U.S.A.* **2002**, *99*, 16400–16406.
(8) Chen, S.; Berthelier, V.; Yang, W.; Wetzel, R. *J. Mol. Biol.* **2001**, *311*, 173–182.
(9) Yang, W.; Dunlap, J. R.; Andrews, R. B.; Wetzel, R. *Hum. Mol. Genet.* **2002**, *11*, 2905–2917.
(10) Hughes, R. E.; Olson, J. M. *Nat. Med.* **2001**, *7*, 419–423.
(11) Perutz, M. F.; Johnson, T.; Suzuki, M.; Finch, J. T. *Proc. Natl. Acad. Sci., U.S.A.* **1994**, *91*, 5355–5358.
(12) Scherzinger, E.; Sittler, A.; Schweiger, K.; Heiser, V.; Lurz, R.; Hasenbank, R.; Bates, G. P.; Lehrach, H.; Wanker, E. E. *Proc. Natl. Acad. Sci., U.S.A.* **1999**, *96*, 4604–4609.
(13) Thakur, A. K.; Wetzel, R. *Proc. Natl. Acad. Sci., U.S.A.* **2002**, *99*, 17014–17019.
(14) Chen, S.; Berthelier, V.; Hamilton, J. B.; O'Nuallain, B.; Wetzel, R. *Biochemistry* **2002**, *41*, 7391–7399.
(15) Ferrone, F. *Methods Enzymol.* **1999**, *309*, 256–274.
(16) Chen, S.; Ferrone, F. A.; Wetzel, R. *Proc. Natl. Acad. Sci., U.S.A.* **2002**, *99*, 11884–11889.
(17) Slepko, N.; Bhattacharyya, A. M.; Jackson, G. R.; Steffan, J. S.; Marsh, J. L.; Thompson, L. M.; Wetzel, R. *Proc. Natl. Acad. Sci., U.S.A.* **2006**, *103*, 14367–14372.
(18) Colby, D. W.; Cassady, J. P.; Lin, G. C.; Ingram, V. M.; Wittrup, K. D. *Nat. Chem. Biol.* **2006**, *2*, 319–323.
(19) Crick, S. L.; Jayaraman, M.; Frieden, C.; Wetzel, R.; Pappu, R. V. *Proc. Natl. Acad. Sci., U.S.A.* **2006**, *103*, 16764–16769.
(20) Walters, R. H.; Murphy, R. M. *J. Mol. Biol.* **2009**, *393*, 978–992.
(21) Khare, S. D.; Ding, F.; Gwanmesia, K. N.; Dokholyan, N. V. *PLoS Comput. Biol.* **2005**, *1*, 230–235.
(22) Zanuy, D.; Gunasekaran, K.; Lesk, A. M.; Nussinov, R. *J. Mol. Biol.* **2006**, *358*, 330–345.
(23) Marchut, A. J.; Hall, C. K. *Biophys. J.* **2006**, *90*, 4574–4584.
(24) Wang, X.; Vitalis, A.; Wyczalkowski, M. A.; Pappu, R. V. *Proteins: Struct. Funct. Bioinf.* **2006**, *63*, 297–311.
(25) Vitalis, A.; Wang, X.; Pappu, R. V. *Biophys. J.* **2007**, *93*, 1923–1937.
(26) Vitalis, A.; Wang, X.; Pappu, R. V. *J. Mol. Biol.* **2008**, *384*, 279–297.
(27) Vitalis, A.; Lyle, N.; Pappu, R. V. *Biophys. J.* **2009**, *97*, 303–311.
(28) Esposito, L.; Paladino, A.; Pedone, C.; Vitagliano, L. *Biophys. J.* **2008**, *94*, 4031–4040.
(29) *Coarse-Graining of Condensed Phase and Biomolecular Systems*; Voth, G. A., Ed.; CRC Press: Boca Raton, FL, 2009.
(30) Shih, A. Y.; Freddolino, P. L.; Arkhipov, A.; Schulten, K. *J. Struct. Biol.* **2007**, *157*, 579.
(31) May, E. R.; Kopelevich, D. I.; Narang, A. *Biophys. J.* **2008**, *94*, 878.
(32) Kim, Y. C.; Hummer, G. *J. Mol. Biol.* **2008**, *375*, 1416.
(33) Chen, N.-Y.; Su, Z.-Y.; Mou, C.-Y. *Phys. Rev. Lett.* **2006**, *96*, 078103.
(34) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 6573.
(35) Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2005**, *123*, 134105.
(36) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. *J. Chem. Phys.* **2008**, *128*, 244114.
(37) Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. *J. Chem. Phys.* **2008**, *128*, 244115.
(38) Wang, Y.; Izvekov, S.; Yan, T.; Voth, G. A. *J. Phys. Chem. B* **2006**, *110*, 3564.
(39) Jiang, W.; Wang, Y.; Yan, T.; Voth, G. A. *J. Phys. Chem. C* **2008**, *112*, 1132.
(40) Wang, Y.; Jiang, W.; Yan, T.; Voth, G. A. *Acc. Chem. Res.* **2007**, *40*, 1193.
(41) Izvekov, S.; Violi, A.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 17019.
(42) Izvekov, S.; Voth, G. A. *J. Chem. Theory Comput.* **2006**, *2*, 637.
(43) Shi, Q.; Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2006**, *110*, 15045.
(44) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 2469.
(45) Liu, P.; Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 11566.
(46) Zhou, J.; Thorpe, I. F.; Izvekov, S.; Voth, G. A. *Biophys. J.* **2007**, *92*, 4289.
(47) Thorpe, I. F.; Zhou, J.; Voth, G. A. *J. Phys. Chem. B* **2008**, *112*, 13079–13090.
(48) Lu, L.; Voth, G. A. *J. Phys. Chem. B* **2009**, *113*, 1501–1510.
(49) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2009**, *113*, 4443–4455.
(50) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, 1987.
(51) Wang, Y.; Jiang, W.; Voth, G. A. Spatial Heterogeneity in Ionic Liquids. In *Ionic Liquids IV: Not Just Solvents Anymore*; Brennecke, J. F., Rogers, R. D., Seddon, K. R., Eds.; American Chemical Society: Washington DC, 2007; pp 272−307.
(52) Wang, Y.; Voth, G. A. *J. Phys. Chem. B* **2006**, *110*, 18601–18608.
(53) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 4116–4127.
(54) Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2006**, *125*, 151101.
(55) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J. *J. Comput. Chem.* **1983**, *4*, 187.
(56) MacKerel, A. D., Jr.; Brooks, B.; Brooks, C. L., III; Nisson, L.; Roux, B.; Won, Y.; Karplus, M. CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. In *The Encyclopedia of Computational Chemistry*; John Wiley & Sons: Chichester, 1998; Vol. 1; pp 271−277.
(57) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43.
(58) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Mod.* **2001**, *7*, 306–317.
(59) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
(60) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.
(61) Ferguson, D. M. *J. Comput. Chem.* **1995**, *16*, 501–511.
(62) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction Models for Water in Relation to Protein Hydration. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel Publishing Company: Dordrecht, 1981; pp 331.
(63) Berendsen, H. J. C.; Postma, J. P. M.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
(64) Forester, T. R.; Smith, W. *Dl_Poly User Manual*; CCLRC, Daresbury Laboratory: Daresbury, Warrington, UK, 1995.
(65) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.
(66) Wang, Y.; Feng, S.; Voth, G. A. *J. Chem. Theory Comput.* **2009**, *5*, 1091–1098.
(67) Evans, D. J.; Morriss, G. P. *Comput. Phys. Rep.* **1984**, *1*, 297.

JP1007768