J|A|C|S
ARTICLES
Published on Web 08/26/2006

Atomically Detailed Simulations of Concentrated Protein Solutions: The Effects of Salt, pH, Point Mutations, and Protein Concentration in Simulations of 1000-Molecule Systems

Sean R. McGuffee and Adrian H. Elcock*

Contribution from the Department of Biochemistry, University of Iowa, Iowa City, Iowa 52242

Received February 28, 2006; E-mail: adrian-elcock@uiowa.edu

Abstract: An ability to accurately simulate the dynamic behavior of concentrated macromolecular solutions would be of considerable utility in studies of a wide range of biological systems. With this goal in mind, a Brownian dynamics (BD) simulation method is reported here that allows systems to be modeled that comprise in excess of 1000 protein molecules, all of which are treated in atomic detail. Intermolecular forces are described in the method using an energy function that incorporates electrostatic and hydrophobic interactions and that is calibrated to reproduce experimental thermodynamic information with a single adjustable parameter. Using the method, BD simulations have been performed over a wide range of pH and ionic strengths for three proteins: hen egg white lysozyme (HEWL), chymotrypsinogen, and T4 lysozyme. The simulations reproduce experimental trends in second virial coefficients (B_{22}) and translational diffusion coefficients, correctly capture changes in B_{22} values due to single amino acid substitutions, and reveal a new explanation for the difficulties reported previously in the literature in reproducing B_{22} values for protein solutions of very low ionic strength. In addition, a strong correlation is found between a residue's probability of being involved in a protein−protein contact in the simulations and its probability of being involved in an experimental crystal contact. Finally, exploratory simulations of HEWL indicate that the simulation model also gives a promising description of behavior at very high protein concentrations (~250 g/L), suggesting that it may provide a suitable computational framework for modeling the complex behavior exhibited by macromolecules in cellular conditions.

Introduction

All molecules in cellular environments are subject to nonspecific interactions with other molecules that can in principle profoundly affect their behavior.[1,2] One way to investigate the effects of nonspecific macromolecular interactions is to study the behavior of concentrated protein solutions: the measured translational diffusion coefficients,[3] second virial coefficients,[4] and scattering intensities of protein solutions can all provide important information regarding transient interactions between protein molecules. To fully understand such interactions however it is important to develop a link between the experimental observables and the protein structure, and this is often best done through the use of molecular models implemented in computer simulations. The desired characteristics of models depend of course on their intended areas of application, but to be useful in the present context a working molecular model of a protein must meet the following criteria: (1) it must be sufficiently sophisticated that it provides an accurate and predictive description of protein−protein interaction thermodynamics, (2) it must provide an easy route to calculation of intermolecular forces so that it can be incorporated into dynamic simulations, and (3) it must be sufficiently rapid to compute that it can be used in simulations of systems comprising many (hundreds of) protein molecules. The present work describes a model capable of fulfilling these three criteria.

Requirement (3) places an immediate and potentially severe limit on the form of any proposed model. Routinely available computational resources are sufficiently restricted that it is currently infeasible to simulate the dynamics of concentrated protein solutions with all atoms of the solvent treated explicitly; instead, it is necessary to employ a simplified treatment of the solvent. Although it is possible to do this and still retain a degree of explicit solvent modeling—as for example is done in dissipative particle dynamics[5] (DPD)—in the present model, a completely implicit solvent representation has been chosen: the solvent's thermodynamic effects must therefore be implicitly incorporated into the intermolecular energy functions (see Methods), and its purely dynamic effects must be accounted for in the equations of motion, which in the present case is achieved by use of a Brownian dynamics (BD) simulation algorithm.[6]

BD is already widely used in simulations of colloidal systems, where idealized structural models of the macromolecules (e.g.,

(1) Minton, A. P. *J. Biol. Chem.* **2001**, *276*, 10577.
(2) Ellis, R. J. *Curr. Opin. Struct. Biol.* **2001**, *11*, 114.
(3) Price, W. S.; Tsuchiya, F.; Arata, Y. *J. Am. Chem. Soc.* **1999**, *121*, 11503.
(4) Velev, O. D.; Kaler, E. W.; Lenhoff, A. M. *Biophys. J.* **1998**, *75*, 2682.
(5) Symeonidis, V.; Karniadakis, G. E.; Caswell, B. *Comput. Sci. Eng.* **2005**, *7*, 39.
(6) Ermak, D. L.; McCammon, J. A. *J. Chem. Phys.* **1978**, *69*, 1352.

10.1021/ja0614058 CCC: $33.50

spheres) are appropriate, and it might be imagined that similar models could also be used in simulations of concentrated protein systems; if this was the case, it would be possible to conduct simulations of extremely large protein systems (containing thousands of molecules) for very long periods of time (e.g., milliseconds). At least two observations suggest however that accurate modeling of protein solutions requires a higher degree of structural detail in the protein models. First, it has been shown that the magnitude of the excluded volume contribution to the second virial coefficient (B_{22}) can be strongly dependent on the level of structural detail employed in the protein model: an atomically detailed model of lysozyme gives a 40% larger excluded-volume contribution than a spherical model of the same overall dimensions.[7] Second, it has been shown that B_{22} values can be sensitive to mutation of a *single* amino acid in the protein and that this can be so even for mutations that cause no change in the net charge of the protein.[8] Since accounting for the latter observation is essential if the protein model is intended to meet requirement (1), it is clear that, at the least, individual amino acids must be resolved and represented in the protein model. In fact, the present model goes some way beyond this minimum level of detail and represents proteins in atomic detail, albeit with the restriction that they are considered to be rigid bodies.

A number of BD simulation studies have already been described in which atomically detailed, rigid models of proteins have been employed (for a review see ref 9). These previous studies have for the most part been aimed at reproducing the kinetics of diffusion-limited association reactions,[9] and the simulations have therefore been used to model the mutual diffusion of only two protein molecules up to the moment at which they form a reactive encounter complex. It is obviously not possible to model the behavior of concentrated protein solutions with only two protein molecules however, and in the simulations described in the present work therefore the number of simulated molecules is increased by almost 3 orders of magnitude to 1000. Some of the algorithmic developments allowing such simulations to be performed over 10-μs timescales on single CPUs have been described in previous work conducted by our group;[10] all of our work has been based on the sophisticated two-molecule BD model originally developed by Gabdoulline and Wade for modeling protein−protein association rate constants.[11,12] Previous applications of our group's extended methodology have considered the effects of solute competition on substrate channeling in an enzyme[13] and the effects of macromolecular crowding on release of protein from the GroEL chaperonin.[14]

Although the computational framework that has been established makes it technically feasible to simulate the dynamics of concentrated protein systems, it does not guarantee that the resulting simulations will be realistic. In order to do this, it is essential that the energy functions used to model the intermolecular interactions be properly calibrated, and this in turn requires that good quality experimental data describing the thermodynamics of protein solutions be available. The latter need can be conveniently met by measurements of the second virial coefficient B_{22} of protein solutions: B_{22} describes, in principle, the deviations from ideal behavior due to interactions between pairs of molecules and can be measured with a number of experimental techniques, most usually static light scattering (SLS) measurements.[4] Importantly, B_{22} is sensitive to pH, ionic strength, and, as noted above, amino acid point mutations,[8,15] and it therefore can be used to test computational models of intermolecular interactions quite extensively. A number of attempts have been made previously to compute B_{22} values with structurally detailed models of proteins, starting with the pioneering work of the Lenhoff group.[7,16−20] As far as we are aware however all of these previous studies have computed B_{22} from calculations of the interaction between only two protein molecules. In contrast, in the present study, B_{22} is computed from 1000-molecule BD simulations of protein solutions performed at concentrations identical to those used in the experiments; as is discussed in some detail, this ability to perform simulations that closely mimic the experimental conditions is shown to be important for rationalizing the experimental B_{22} data obtained in low salt concentrations.

Table 1. Physical Properties Assigned to the Simulated Proteins and Experimental Conditions Used to Parametrize the Simulation Model's Energy Function

protein	M_w (Da)	D_{trans} (Å²/ns)	D_{rot} (/ns)	pH	pI	[salt] (mM)	ϵ_{LJ} (kcal/mol)
HEWL	14 296	10.96	0.019 69	9.0	10.5	100	0.28
T4 lysozyme	18 551	9.860	0.014 29	7.0	9.6	55	0.22
chymotrypsinogen	25 651	9.101	0.011 38	6.8	8.8	100	0.23

We report simulations of solutions of three different proteins: hen egg white lysozyme (HEWL), chymotrypsinogen, and T4 lysozyme. The energy model for each protein has first been parametrized to reproduce B_{22} data obtained in one set of experimental conditions and has then been used to predict B_{22} in other conditions: the overall good agreement that is obtained between these predictions and available experimental results indicates that the parametrized models have utility for describing the behavior of concentrated protein solutions. This utility is enhanced by the fact that the same simulations also provide a host of additional structural and dynamic information. In particular, the simulations give unusually detailed views of (a) the way translational and rotational diffusion of molecules is affected by intermolecular interactions, (b) the kinetics and thermodynamics of formation of oligomeric clusters, and (c) the surface residues that drive close interactions between neighboring proteins. Since a number of these aspects can be experimentally tested, the parameters of the simulation method can in the future be further refined, thus making it a viable framework for developing models of the more complex and concentrated macromolecular mixtures typically encountered in biological systems. As a first step in this direction, we also report simulations of highly concentrated HEWL solutions (up to 254

(7) Neal, B. L.; Lenhoff, A. M. *AIChE J.* **1995**, *41*, 1010.
(8) Chang, R. C.; Asthagiri, D.; Lenhoff, A. M. *Proteins: Struct., Funct., Genet.* **2000**, *41*, 123.
(9) Gabdoulline, R. R.; Wade, R. C. *Curr. Opin. Struct. Biol.* **2002**, *12*, 204.
(10) Elcock, A. H. *Methods Enzymol.* **2004**, *383*, 166.
(11) Gabdoulline, R. R.; Wade, R. C. *Biophys. J.* **1997**, *72*, 1917.
(12) Gabdoulline, R. R.; Wade, R. C. *J. Mol. Biol.* **2001**, *306*, 1139.
(13) Elcock, A. H. *Biophys. J.* **2002**, *82*, 2326.
(14) Elcock, A. H. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2340.
(15) Curtis, R. A.; Steinbrecher, C.; Heinemann, M.; Blanch, H. W.; Prausnitz, J. M. *Biophys. Chem.* **2002**, *98*, 249.
(16) Neal, B. L.; Asthagiri, D.; Lenhoff, A. M. *Biophys. J.* **1998**, *75*, 2469.
(17) Elcock, A. H.; McCammon, J. A. *Biophys. J.* **2001**, *80*, 613.
(18) Lund, M.; Jönsson, B. *Biophys. J.* **2003**, *85*, 2940.
(19) Asthagiri, D.; Paliwal, A.; Ábras, D.; Lenhoff, A. M.; Paulaitis, M. E. *Biophys. J.* **2005**, *88*, 3300.
(20) Stradner, A.; Sedgwick, H.; Cardinaux, F.; Poon, W. C. K.; Egelhaaf, S. U.; Schurtenberger, P. *Nature.* **2004**, *432*, 492.

g/L) and show from comparisons of computed structure factors, $S(Q)$, that the simulated behavior is in good qualitative agreement with recently reported experimental data.[20,21]

Methods

Protein Structures. Coordinate files for the three proteins studied here were downloaded from the Protein Data Bank[22] (http://www.rcsb.org), with PDB file 1HEL[23] being used for HEWL, 1L87[24] for T4 lysozyme, and 2CGA[25] for chymotrypsinogen. Hydrogens and any missing side chain atoms were added to each structure using the molecular modeling program WHATIF;[26] the same program was used to perform the side chain replacements necessary to construct single-residue mutants of HEWL and T4 lysozyme. As necessary input for the BD simulations, rotational and translational diffusion coefficients for the proteins at *infinite dilution* (Table 1) were determined by inputting the protein structures to the hydrodynamics program HYDROPRO.[27]

Brownian Dynamics (BD). The multiple-macromolecule BD method used in this work extends the previously reported methodology[10] with modifications to ensure complete conservation of forces and an effort to model hydrophobic interactions between proteins (see below). The method models proteins as rigid bodies and simulates their translational and rotational motion with the BD algorithm due to Ermak and McCammon.[6] Interactions between proteins are modeled as a sum of electrostatic and van der Waals/hydrophobic interactions, with calculation of the latter terms being accelerated by modeling only non-hydrogen surface atoms with at least 2 Å^2 of solvent-exposed surface area.

In all simulations reported here, electrostatic interactions between proteins were modeled with the "effective charge" method developed by Gabdoulline and Wade.[28] In this approach, the electrostatic forces on protein atoms are determined by the interaction of their effective charges with the electrostatic potentials generated by other nearby proteins. The requisite electrostatic potentials are obtained by solving the linearized Poisson−Boltzmann (PB) equation[29] with the finite-difference program UHBD[30] and stored in memory as three-dimensional grids that translate and rotate during the BD simulations with the protein from which they are generated. Partial charges and atomic radii for the PB calculations were taken from the PARSE parameter set;[31] partial charges for the atoms of ionizable residues were obtained by linearly interpolating between those of the protonated and unprotonated forms of the residue so that the net charge was equal to that computed from the residue's pK_a in the revised "null model" described by Antosiewicz et al.[32] For both HEWL and chymotrypsinogen, the net protein charges obtained with this approach were found to be in good agreement with those measured experimentally[33,34] (Figure S1). In line with the only modest changes observed in crystallographic structures of HEWL with pH,[35] an identical protein structure was used for simulations in all pH conditions. The solvent dielectric was set to 78.4 to match the dielectric of water (at 25 °C), and the dielectric within the protein interior was set to 12.0 as a simple compromise between the lower dielectric of protein interiors and the higher dielectric of protein exteriors.[36]

To properly account for the possibility of very long-range electrostatic interactions at low salt concentrations (5 mM), a modification to the simulation code was made allowing each protein to be assigned two electrostatic potential grids. For very long-range interactions, a coarse 200 × 200 × 200 potential grid of spacing 1.5 Å was used, thus allowing interactions between proteins separated by as much as 150 Å to be computed. For accurate representation of electrostatic interactions at short range (where substantial potential gradients can be encountered), a more detailed potential grid of spacing 0.5 Å was computed with dimensions sufficient to encompass a 20 Å shell around the protein surface.

To provide a simple combined model of van der Waals and hydrophobic interactions between the carbon and sulfur atoms of neighboring proteins, a Lennard−Jones potential was used:

$$U(r) = 4\epsilon_{LJ}\left[\left(\frac{\sigma_{LJ}}{r}\right)^{12} - \left(\frac{\sigma_{LJ}}{r}\right)^{6}\right]$$

where the potential energy, $U(r)$, depends on the distance, r, between atoms, σ_{LJ} is the distance at which $U(r)$ changes from being favorable to unfavorable, and ϵ_{LJ} is the well depth of the energy minimum. For interactions involving all other combinations of atom types, a purely repulsive potential was used, it being assumed that they make no significant net contribution to interactions other than those modeled by the electrostatic term:

$$U(r) = 4\epsilon_{LJ}\left[\left(\frac{\sigma_{LJ}}{r}\right)^{12}\right]$$

This model of atomic interactions, in which only interactions between hydrophobic atoms are energetically rewarded, was used by us recently to model ligand−receptor interactions.[37] In all simulations, σ_{LJ} was set to 4 Å, and ϵ_{LJ} was treated as a free parameter that was adjusted separately for each protein so that the computed B_{22} reproduced the experimental value in a single chosen condition of pH and salt concentration (listed in Table 1).

It is obviously a considerable simplification to assume that hydrophobic interactions can be described with a Lennard−Jones potential. One drawback is that it overlooks the fact that the free energy surface for association of hydrophobic groups has separate contact and solvent-separated minima; it should be remembered however that the continuum electrostatic model that we use also introduces the same simplification into the treatment of charge−charge interactions. A second limitation is that, as pointed out by a reviewer, it assumes that interactions between hydrophobic groups are pairwise-additive, even though there is evidence from molecular dynamics simulations that such interactions may have many-body characteristics.[38−40] In future developments of the present simulation model it may be possible to use more elegant hydrophobic models that attempt to incorporate both desolvation barriers and many-body effects (e.g., ref 41).

In order to provide the best opportunity for properly parametrizing the van der Waals/hydrophobic interactions, the solution conditions for each protein were chosen such that electrostatic interactions were at least partially suppressed by the presence of salt in substantial concentrations (55−100 mM) and by the pH being at or near the

(21) Liu, Y.; Fratini, E.; Baglioni, P.; Chen, W.-R.; Chen, S.-W. *Phys. Rev. Lett.* **2005**, *95*, 118102.
(22) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235.
(23) Wilson, K. P.; Malcolm, B. A.; Matthews, B. W. *J. Biol. Chem.* **1992**, *267*, 10842.
(24) Eriksson, A. E.; Baase, W. A.; Matthews, B. W. *J. Mol. Biol.* **1993**, *229*, 747.
(25) Wang, D. C.; Bode, W.; Huber, R. *J. Mol. Biol.* **1985**, *185*, 595.
(26) Vriend, G. *J. Mol. Graph.* **1990**, *8*, 52.
(27) de la Torre, J.; Huertas, M. L.; Carrasco, B. *Biophys. J.* **2000**, *78*, 719.
(28) Gabdoulline, R. R.; Wade, R. C. *J. Phys. Chem.* **1996**, *100*, 3868.
(29) Fogolari, F.; Brigo, A.; Molinari, H. *J. Mol. Recogn.* **2002**, *15*, 377.
(30) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; McCammon, J. A. *Comput. Phys. Commun.* **1995**, *91*, 57.
(31) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978.
(32) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. *Biochemistry.* **1996**, *35*, 7819.
(33) Kuehner, D. E.; Engmann, J.; Fergg, F.; Wernick, M.; Blanch, H. W.; Prausnitz, J. M. *J. Phys. Chem. B* **1999**, *103*, 1368.
(34) Marini, M. A.; Martin, C. J. *Eur. J. Biochem.* **1970**, *19*, 162.
(35) Sukumar, N.; Biswal, B. K.; Vijayan, M. *Acta Crystallogr.* **1999**, *D55*, 934.
(36) Sept, D.; McCammon, J. A. *Biophys. J.* **2001**, *81*, 667.
(37) Rockey, W. M.; Elcock, A. H. *J. Med. Chem.* **2005**, *48*, 4138.
(38) Ghosh, T.; Garcia, A. E.; Garde S. *J. Phys. Chem. B* **2003**, *107*, 612.
(39) Moghaddam, M. S.; Shimizu, S.; Chan, H. S. *J. Am. Chem. Soc.* **2005**, *127*, 313.
(40) Czaplewski, C.; Liwo, A.; Ripoll, D. R.; Scheraga, H. A. *J. Phys. Chem. B* **2005**, *109*, 8108.
(41) Hummer, G. *J. Am. Chem. Soc.* **1999**, *121*, 6299.

protein's isoelectric point. Once ϵ_{LJ} was parametrized for each protein in this single condition, the same ϵ_{LJ} value was then used for simulations of the same protein in *all* other solution conditions, thus testing the ability of the PB electrostatic model to account directly for the effects of both pH and salt on B_{22}. It is to be noted that this involves the implicit assumption that the hydrophobic interactions are independent of salt in the range of salt concentrations studied. Of course, it is known that such interactions are actually strengthened by the addition of high (~molar) concentrations of salts such as NaCl,[42] and this effect is reproduced in potentials of mean force computed for the association of hydrophobic molecules by Monte Carlo and/or molecular dynamics methods.[43,44] However, since the highest salt concentration investigated here is 0.5 M, neglecting the salt-dependence of the hydrophobic interactions is unlikely to introduce significant errors.

Simulation Details. All simulations were performed with 1000 identical protein molecules contained within a cubic simulation box, with dimensions set such that the simulated protein concentration was identical to that studied experimentally (1.25−10 g/L); as an example, for HEWL at 10 g/L, a simulation cube of length 1339 Å was employed. Protein molecules were initially placed within the box by random rotation and translation while ensuring at least a 10 Å separation from neighboring molecules. Brownian motion of the molecules was then simulated using the Ermak−McCammon BD algorithm[6] with a time step of 2.5 ps; rotational and translational diffusion of all proteins was assumed to be isotropic, and hydrodynamic interactions between proteins were neglected. Because of the use of a comparatively large time step, it is occasionally possible for significant steric clashes to develop between atoms of neighboring proteins following a single simulation step. To alleviate any such clashes, an iterative adjustment of protein positions was performed immediately following each simulation step until no interacting pair of atoms was separated by less than 4.5 Å. Details of this adjustment algorithm, which conserves linear and angular momentum and is similar in spirit (though not in details) to the SHAKE constraint algorithm[45] commonly used in MD simulations, are provided in the Supporting Information. Examination of the total system energy in preliminary simulations using a range of different time steps indicated that 2.5 ps was the largest value that could be safely used.

All simulations were performed under constant volume conditions, and periodic boundary conditions were applied so that edge effects were avoided and the systems behaved like bulk solutions.[46] For speed, van der Waals/hydrophobic interactions were computed only between atoms separated by less than 12 Å; a list of atom pairs meeting this criterion was constructed every 20 simulation steps. Simulations were continued for periods of 10, 15, and 10 μs for systems modeled at 10 g/L, 5 g/L, and 1.25 g/L, respectively. For HEWL, a series of simulations was also performed at the much higher protein concentrations of 36, 72, 125, 169, and 254 g/L: because of the increased computational expense involved in such simulations, their total lengths were each 1 μs, respectively. For subsequent analysis of the dynamic behavior of the proteins during the simulations, all coordinates necessary for uniquely specifying the location of each protein molecule (a three-dimensional translational vector and a 3×3 rotational matrix for each molecule) were recorded every 1 ns. Based on examinations of the total system energy as a function of simulation time, the first 1 μs of each simulation was treated as an equilibration period (100 ns in the case of the very concentrated HEWL solutions) and was therefore not used for final computation of any dynamic or structural properties.

Calculation of B_{22}. A convenient route to calculating B_{22} directly from dynamic simulations is via the radial distribution function, $g(r)$. In order to compute the latter with as much statistical confidence as possible, a histogram of all protein−protein pairwise distances (defined as the distance between the proteins' centers of geometry) was updated at *every* time step of the simulation. The B_{22} was then calculated from $g(r)$ as described in Velev et al.[4] using

$$B_{22} = -\frac{2\pi}{M_W^2 N_A}\int_0^\infty (g(r) - 1) r^2 \, dr$$

where r is the protein−protein distance, M_W is the molecular weight of the protein, and N_A is Avogadro's number. Although nominally involving an integration of $g(r)$ to infinite distance, in practice the computations of B_{22} were subject to finite upper limits when statistical uncertainties in $g(r)$ at longer values of r were encountered (owing to the r^2 dr dependence, even tiny deviations in $g(r)$ from 1.0 at long distances can make significant contributions to a computed B_{22}). For simulations performed at 5−10 g/L, these uncertainties limit the precision of the calculated B_{22} values to perhaps $\pm 1 \times 10^4$ mol mL/g^2. For the lower protein concentration of 1.25 g/L, where sampling of interaction events is less thorough, the precision is somewhat lower (e.g., $\pm 5 \times 10^4$ mol mL/g^2); however, none of the key conclusions drawn here regarding the effects of protein concentration on measured B_{22} values are affected by this lower precision.

Calculation of Translational Diffusion Coefficients. The effective translational diffusion coefficients D_{trans} of protein molecules were computed from their center of mass trajectories using the Einstein formula:[46]

$$D_{trans} = \frac{\langle \delta x^2 \rangle}{2\,\delta t}$$

where δx is the distance traveled in one of the Cartesian directions during a time interval δt, and the brackets indicate an ensemble average. The choice of δt is a compromise between the need to have a value large enough that the effects of anomalous diffusion[47] are overcome but small enough that the statistical uncertainties in the measurements are reasonable. In the present study, two values of δt were used. For calculations aimed at best estimating the average D_{trans} of the entire population of molecules, δt was set to 100 ns; error estimates for these calculations were obtained from the standard deviation of the 1000 D_{trans} values obtained for each individual molecule. For calculations aimed at investigating the relation between a single molecule's diffusive behavior and its interaction with its immediate environment (see Results), a smaller δt of 1 ns was used to reduce statistical errors.

Calculation of Rotational Diffusion Coefficients. The effective rotational diffusion coefficients (D_{rot}) of protein molecules were obtained from the average of the autocorrelation functions of the three unit vectors describing the rotational orientation of the molecule. The autocorrelation functions were each fit to a single-exponential decay function to extract the rotational relaxation time, τ_{rot}, from which D_{rot} was obtained via the relationship $D_{rot} = 1/(2\tau_{rot})$. All single-exponential fits were sufficiently accurate ($r^2 > 0.99$) that higher-exponential fits were not considered.

Analysis of Oligomeric Clusters. The formation of oligomers of protein molecules was investigated using geometric criteria in the following way. Structural snapshots saved every 1 ns were examined for protein pairs with surface atoms within 6 Å of one another; oligomeric species were then identified by grouping together all contacting protein pairs that had a molecule in common. This 6 Å distance was chosen because it was sufficient to include the bulk of the first peak in the histogram of closest intermolecular atomic distances without extending so far into space that cases where noninteracting proteins happen to drift into contact with each other were included; any cutoff distance within the range ~5.5 Å to ~9 Å could be chosen

(42) Baldwin, R. L. *Biophys. J.* **1996**, *71*, 2056.
(43) Ghosh, T.; Kalra, A.; Garde, S. *J. Phys. Chem. B.* **2005**, *109*, 642.
(44) Thomas, A. S.; Elcock, A. H. *J. Am. Chem. Soc.* **2006**, *128*, 7796.
(45) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.
(46) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, U.K., 1987.
(47) Saxton, M. J. *Biophys. J.* **1996**, *70*, 1250.

without changing any of the qualitative conclusions drawn here. Following Carlsson et al.[48] the association constant ($K_{a,i}$) of an oligomer of "i" monomers was expressed in terms of the average concentrations of oligomers and monomers observed during the production stage of the simulation using: $K_{a,i} = [P_i]/([P_{i-1}][P_1])$, where $[P_i]$ is the concentration of an oligomer of size i.

Kinetics of Monomer Dissociation. The kinetics of monomer dissociation from each oligomeric species (defined by the same geometric criteria outlined above) was computed as follows. First, all occurrences of the oligomeric species (dimer, trimer, etc.) during the production stage of the simulation were examined in order to identify those cases where the oligomer was eventually destroyed by dissociation of a *single* monomer: all cases where the oligomer decayed by some other process (e.g., by loss of a dimer, loss of multiple monomers simultaneously, or addition of a monomer to form a higher-order oligomer) were ignored in order to simplify interpretation. The lifetimes of all oligomers satisfying this single-monomer-loss criterion were then used to construct a time-dependent decay plot for the population of that type of oligomer. This decay was in all cases found to fit well to a double-exponential function, the faster component of which was due to rapid recrossing of the 6 Å threshold distance used to designate proteins as being in contact, and was not considered to be representative of a true dissociation event. The time constant of the slower component (τ_{slow}), which was considered to be representative of a genuine dissociation event, was used to define a unimolecular dissociation rate constant, k_{off}, through the relation $k_{off} = 1/\tau_{slow}$. Sampling of dissociation events was sufficient to allow k_{off} values to be determined in this way for dimeric, trimeric, and tetrameric species only; although dissociation events were also observed for pentamers and certain higher-order oligomers, sampling was insufficient to produce reasonable rate estimates.

Surface Atom Contact Probabilities. The propensity of each surface atom in a protein to be involved in interactions with neighboring molecules was calculated from the frequency (f_i) with which the atom was found within 6 Å of an atom on a neighboring protein during the production stage of the simulation. For each protein studied, these frequencies were converted into effective contact probabilities by dividing each atom's frequency f_i by f_{max}, the maximum contact frequency found for any of the atoms of the protein. These effective contact probabilities could then be compared with the probability of the atom being involved in an experimental crystal contact in the following way. For each protein studied, a survey of crystal structures solved in different space groups was conducted. For HEWL, wild-type structures were taken from the space groups listed in ref 49 (pdb codes 1HEL; 1LYS; 1LZT; 132L); for chymotrypsinogen, structures in the three space groups were taken from ref 50 (pdb codes 2CGA; 1EX3; 1CHG); for T4 lysozyme, since true wild-type structures are not available, near-wild-type structures were selected instead: proteins were included in this list only if they had two or fewer mutations and if the sites of the mutations themselves were not solvent-exposed, in order to minimize any influence of the mutations on the surface interactions (pdb codes 1L87; 175L; 180L; 148L; 1P7S; 1QTH). The atoms involved in contacts with neighboring molecules in each of these crystal structures were then identified using the "Crystal Symmetry" module on the WHATIF webserver (http://swift.cmbi.kun.nl/WIWWWI). Then, for each surface atom in the particular protein studied, the probability of it being involved in a crystal contact was obtained by dividing the number of structures in which it was found to be engaged in a contact by the total number of structures in the sample.

An alternative way to describe the relative propensity of an atom to be involved in an intermolecular contact is to convert the contact frequencies to free energy form using: $\Delta G^\circ_{contact} = -RT \ln(f/f_{max})$, where f_{max} is the maximum contact frequency found for any of the atoms of the protein. Using this definition, the $\Delta G^\circ_{contact}$ is zero for the atom most frequently involved in contacts with neighboring proteins and positive for all other atoms. Since the $\Delta G^\circ_{contact}$ value of an atom depends not only on its effective energetic interaction with other molecules but also on its physical accessibility to the atoms of other proteins, it was of interest to see if these two effects could be separated. To do this, a 1000-molecule simulation of each protein studied was conducted in which all electrostatic and hydrophobic interactions were switched off, and all atoms were in effect treated as hard spheres. These simulations, which were conducted using exactly the same protocol as the simulations used to predict B_{22} values, allow $\Delta G^\circ_{contact}$ values to be computed where the only determining factor is the effective accessibility of the atoms. Subtracting these control ΔG° values from those measured during more "realistic" simulations can in principle give a more direct measure of how energetic interactions determine an atom's involvement in interprotein contacts.

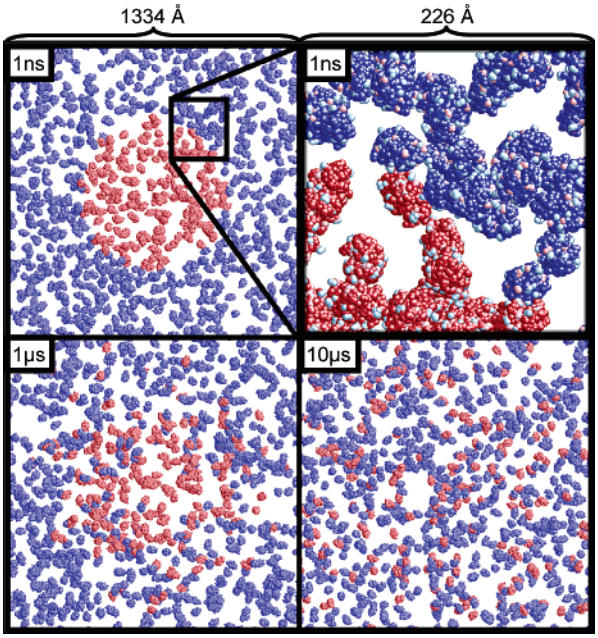

Figure 1. Snapshots of the 1000-molecule BD simulation of 10 g/L HEWL at pH 9, 100 mM salt taken at points 1 ns, 1 μs, and 10 μs into the simulation. Proteins located in the center of the box during the first snapshot are arbitrarily colored red for visualization purposes only: in the actual simulations all molecules were modeled as identical. The expansion in the upper right demonstrates the atomic level of detail of the simulation model; positive and negative "effective" charges on individual molecules are colored in light blue and red, respectively. This figure was prepared with RasMol.[51]

Scattering Data. Following Velev et al.[4] the structure factor, $S(Q)$, was calculated from the simulations via $g(r)$:

$$S(Q) = 1 + 4\pi\rho \int_0^\infty (g(r) - 1) \frac{\sin(Qr)}{Qr} r^2 \, dr$$

with ρ as the protein concentration, Q the wavevector (nm^{-1}), and the formally infinite upper limit of integration replaced by the distance at which it could be safely assumed that $g(r) = 1$.

Results

Structural snapshots taken from a typical BD simulation (HEWL at a concentration of 10 g/L at pH 9 and in 100 mM salt) are shown in Figure 1. The progressive mixing of the 1000 molecules that occurs over the time scale of the simulations can be seen simply by tracking the diffusion of a subpopulation of the molecules, arbitrarily colored red at the beginning of the

(48) Carlsson, F.; Malmsten, M.; Linse, P. *J. Phys. Chem. B* **2001**, *105*, 12189.
(49) Vaney, M. C.; Maignan, S.; Riès-Kautt, M.; Ducruix, A. *Acta Crystallogr.* **1996**, *D52*, 505.
(50) Pjura, P. E.; Lenhoff, A. M.; Leonard, S. A.; Gittis, A. G. *J. Mol. Biol.* **2000**, *300*, 235.

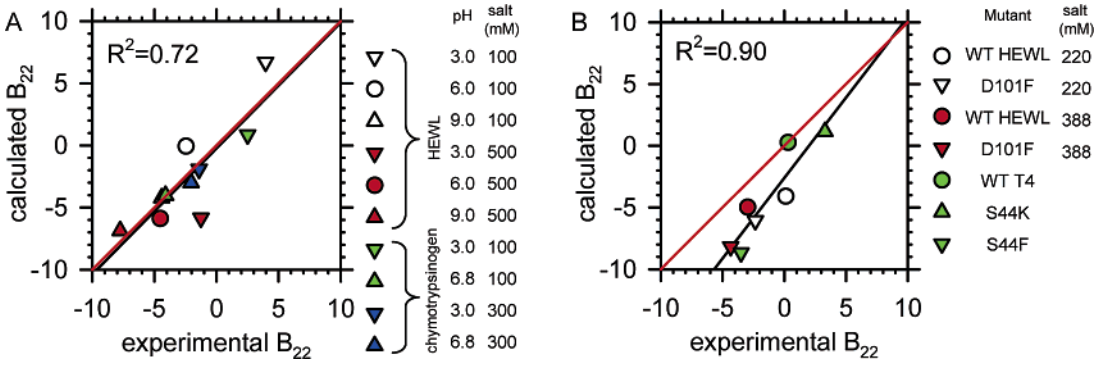

Figure 2. (A) Comparison of computed B_{22} values ($10^4 \times$ mol mL/g^2) with experimental values for wild-type proteins in different conditions of pH and salt concentration. (B) Comparison for wild-type and single-residue mutants for HEWL and T4 lysozyme.

simulation. The extent of assimilation achieved by the end of the 10 μs simulation provides a straightforward but important indication that the simulated time scale is likely to be sufficient for a relatively thorough sampling of the system's behavior.

***B*$_{22}$ Computations.** The energy function used in the simulations has a single adjustable parameter (ϵ_{LJ}) that has been altered separately for wild-type HEWL, T4 lysozyme, and chymotrypsinogen to optimize agreement between the computed and experimental B_{22} values in a single solution condition (Figure S2). Encouragingly, the optimal values of ϵ_{LJ} obtained for the three proteins are all within ~25% of each other (Table 1), which suggests, in line with previous work,[17] that it may eventually be possible to develop a transferable energetic model that can routinely be used in a predictive setting. Since one of the main purposes of the present paper is to reproduce B_{22} values however, no attempt was made here to use a single compromise value of ϵ_{LJ} in simulations of all three proteins. Instead, what was investigated was (a) whether the independently parametrized values for HEWL and chymotrypsinogen would accurately describe B_{22} for the wild-type proteins in other conditions of pH and salt concentration and (b) whether the parametrized values for HEWL and T4 lysozyme would allow accurate prediction of experimentally measured B_{22} values of site-directed mutants.

A comparison of computed and experimental B_{22} values for wild-type HEWL and chymotrypsinogen in salt concentrations ranging from 100 mM to 500 mM and at pH's from 3 to 9 is shown in Figure 2A; these data were all obtained from simulations performed at protein concentrations of 10 g/L. A linear fit of the data (omitting the two points that represent the parametrization conditions) gives an R^2 value of 0.72 with a gradient of 1.04. The former is comparable to, though not better than, the correlation obtained from a previous simpler model due to Velev et al.[4] for the same data points (R^2 of 0.81); it should be noted however that the latter was parametrized via a global fit and so is *a priori* expected to perform better over a wide range of conditions. For chymotrypsinogen, the simulations successfully capture the nontrivial result[4,16] that at pH 3 B_{22} is lower (more favorable) in 300 mM salt than in 100 mM salt, but at pH 6.8 it is higher in 300 mM than in 100 mM salt (due to a salt suppression of favorable short-range electrostatic interactions). For HEWL, the pH dependence of B_{22} is nicely reproduced at 100 mM but, interestingly, is markedly underestimated at 500 mM salt; this suggests the possibility that the Poisson−Boltzmann electrostatic model implemented here may overestimate the screening of electrostatic interactions at higher salt concentrations.

A comparison of computed and experimental B_{22} values for wild-type and site-directed mutant proteins is shown in Figure 2B. Since the mutant proteins were simulated using the exact same values of ϵ_{LJ} as those developed for the corresponding wild-type proteins and since none of these specific simulations were directly parametrized to match experimental data, the plot shown in Figure 2B represents a *bona fide* test of the simulation model's predictive abilities; a linear fit of the data gives an R^2 value of 0.90 with a gradient of 1.19. For T4 lysozyme, the simulations qualitatively capture the fact that B_{22} increases with the S44K mutation but decreases with the S44F mutation;[8] the latter mutation is of particular interest because it causes no change in the protein's net charge and is therefore beyond description by more simplified physical models. For HEWL, the qualitative effects of the D101F mutant studied by the Blanch and Prausnitz groups[15] are also correctly reproduced. This is notable because the mutation, in principle, introduces two opposing effects which must be properly balanced for the correct result to be obtained: on the one hand, the loss of the negative charge of the aspartate residue might, on purely electrostatic grounds, be expected to increase B_{22} somewhat (since it increases the net charge on the protein); on the other hand, the addition of the phenylalanine side chain would be expected to decrease B_{22} (since it introduces a new hydrophobic "patch"[15] on the protein surface that could promote interactions with other molecules). The two effects can be decoupled in simulations by calculating B_{22} for a wild-type model of HEWL in which the aspartate side chain charges have been set to zero; interestingly, when these simulations are performed, the B_{22} in this artificial mutant is found to be more or less identical to the wild-type value (-4.1 vs -4.0) $\times$ 10^{-4}mol mL/g^2.

Low Salt Behavior. As noted in the Introduction, several computational studies have already addressed the modeling of B_{22} data with structurally detailed protein models, and a number studies[4,16−18,48] have specifically attempted to reproduce the experimental data reported by Velev et al.[4] There have been two features common to these previous studies: (1) B_{22} was obtained by computing the interaction of only two protein molecules, and (2) the resulting calculated B_{22} values in 5 mM salt were significantly more positive than the corresponding experimental values. A key finding that emerges from the present study is that the first of these features is almost certainly responsible, at least in part, for the second feature and that

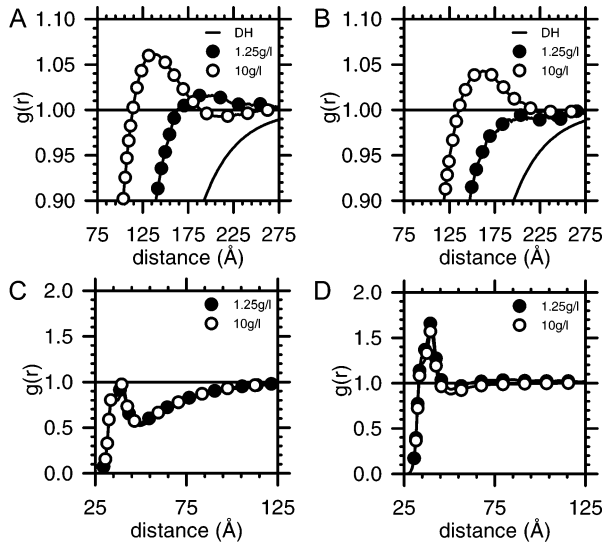

Figure 3. Comparison of radial distribution functions, $g(r)$, obtained from simulations showing the dependence on protein concentration. (A) HEWL in 5 mM salt, pH 3 conditions; solid line indicates the prediction of the Debye−Hückel equation (see text for details). (B) Chymotrypsinogen in 5 mM salt, pH 3. (C) HEWL in 5 mM salt, pH 9. (D) HEWL in 100 mM salt, pH 3.

higher-order interactions between protein molecules, which are captured naturally in multimolecule simulations of the kind reported here, are necessary in order to properly describe the low-salt experimental data.

The road to this conclusion emerges from a comparison of the behavior observed in 1000-molecule BD simulations performed at different protein concentrations. Figure 3A shows the protein−protein radial distribution functions, $g(r)$, obtained from BD simulations of HEWL at pH 3, 5mM salt, performed with protein concentrations of 1.25 g/L and 10 g/L; these two concentrations span the range used by Velev et al.[4] to experimentally determine B_{22}. Also shown in the same figure is the $g(r)$ calculated from the Debye−Hückel equation for the interaction of two proteins with the same diameter (38.2 Å) and net charge (+13.8e) as HEWL (see the solid line in Figure 3A). This latter plot should provide a reasonable approximation to the long-range interaction expected at infinite dilution of the protein (i.e., 0 g/L), so the three plots together should allow the behavior expected at three different protein concentrations (0, 1.25, and 10 g/L) to be examined. Not surprisingly, the Debye−Hückel (0 g/L) result predicts that close approach of the molecules will be strongly disfavored ($g(r) \approx 0$) and that the electrostatic repulsion will be sufficiently long-ranged that the bulk solution value ($g(r) \approx 1$) will only be reached when the center−center separation distance between the two molecules is 200−300 Å. Since this is the behavior expected when two isolated molecules interact, it is also almost certainly the behavior that will have been present in the previous calculations of B_{22} at 5 mM salt reported in the literature. As shown in Figure 3A however, this behavior is very different from that observed in BD simulations performed at the experimentally studied protein concentrations. The $g(r)$'s obtained from the BD simulations at 1.25 and 10 g/L protein concentrations indicate that close approach of the protein molecules is considerably less repulsive than at 0 g/L, and in fact, a modest but clear *peak* value ($g(r) \approx 1.06$) is obtained at a separation of ∼135 Å at 10 g/L, with a more minor peak ($g(r) \approx 1.02$) being obtained at a separation of ∼185 Å at 1.25 g/L (additional simulations were performed to demonstrate that the positions and heights of these peaks were not dependent on the cutoff distance assigned to electrostatic interactions). These results are important for two reasons. First, the fact that $g(r) > 1$ is obtained indicates the presence of a weak, effective long-range attraction between the molecules, despite the fact that all of the direct pairwise interactions between proteins are purely repulsive at these distances. Second, the observation of clear differences in the $g(r)$'s computed at 1.25 g/L and 10 g/L indicates that the effective pairwise interaction between HEWL molecules is likely to change over the range of protein concentrations studied experimentally by Velev et al. at 5 mM salt.

Before considering the consequences of these results for the computed and experimentally measured B_{22} values, it is worth examining the $g(r)$'s obtained with other proteins and/or conditions. Figure 3B shows the corresponding results obtained with chymotrypsinogen at pH 3, 5 mM salt. Overall, the behavior is very similar to that obtained with HEWL in the same conditions: an effective long-range attraction between molecules is again obtained at a protein concentration of 10 g/L, and although an attractive ($g(r) > 1$) peak does not actually appear at a concentration of 1.25 g/L, it is still apparent that the effective interaction is significantly less repulsive than that predicted at 0 g/L from the Debye−Hückel equation (solid line in Figure 3B). That somewhat weaker effects are obtained with chymotrypsinogen compared to those obtained with HEWL is consistent with the former protein's lower charge density: although the net charges on the two proteins are essentially identical at pH 3, chymotrypsinogen is a considerably larger molecule (46.2 Å[52] diameter vs 38.2 Å[19]). Again, the differences between the $g(r)$'s obtained from simulations at 1.25 and 10 g/L indicate that the effective pairwise interaction of chymotrypsinogen molecules is likely to be changing over the range of protein concentrations studied experimentally.

The presence of attractive long-range peaks in $g(r)$'s from simulations in which all protein molecules are like-charged, although perhaps counterintuitive at first sight, is not uncommon and has already been observed previously in simulations of highly charged colloidal systems (see for example refs 53 and 54). The effective attraction is essentially a consequence of the fact that when the protein concentration is comparatively high and the salt concentration is low, the length scale over which the repulsive net-charge interaction acts is similar to the average distance between neighboring protein molecules. Since molecules are surrounded on all sides by neighbors with which they are engaged in repulsive interactions, increasing the separation between any one pair of protein molecules in an attempt to relieve their electrostatic repulsion only tends to result in increasing the electrostatic repulsion experienced by both molecules from *other* nearby molecules: as a result, there is a preferred separation distance that manifests itself as a local maximum in $g(r)$.

Support for the idea that very long-range electrostatic interactions cause the differences in behavior at different protein

(51) Sayle, R.; Milner-White, E. J. *Trends Biochem. Sci.* **1995**, *20*, 374.
(52) Paliwal, A.; Asthagiri, D.; Abras, D.; Lenhoff, A. M.; Paulaitis, M. E. *Biophys. J.* **2005**, *89*, 1564.
(53) Vlachy, V.; Marshall, C. H.; Haymet, A. D. J. *J. Am. Chem. Soc.* **1989**, *111*, 4160.
(54) Giacometti, A.; Gazzillo, D.; Pastore, G.; Das, T. K. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2005**, *71*, 031108.

concentrations comes from examining simulations performed in conditions where the long-range electrostatic interactions are weakened. One such set of conditions is found at higher pH's where the net charge on the protein molecules is reduced. In Figure 3C the $g(r)$'s from simulations of HEWL at pH 9, 5 mM salt are shown for 1.25 and 10 g/L. In these conditions favorable short-range ("hydrophobic") interactions almost cancel the repulsive electrostatic interaction of the protein net charges such that $g(r)$ close in almost rises above 1; the presence of significant nonelectrostatic contributions means that for these conditions the Debye−Hückel equation no longer provides a useful description of $g(r)$. More important, however, is the result that, throughout the entire distance range examined, the $g(r)$'s for 1.25 and 10 g/L are very similar to one another. In corresponding simulations of chymotrypsinogen at pH 6.8, 5mM salt, similar behavior is obtained: here, the short-range interaction is net favorable, and a very large value of $g(r)$ results, indicative of a substantial amount of dimerization; however, the $g(r)$'s for 1.25 and 10 g/L are again very similar to one another (Figure S3A).

A second set of conditions in which long-range electrostatic interactions are expected to be weakened is at higher salt concentrations. Figure 3D shows the $g(r)$'s obtained from pH 3, 100 mM salt simulations of HEWL performed at 1.25 and 10 g/L concentrations. As expected, the two $g(r)$'s obtained from the BD simulations are very similar to one another, suggesting that the effective interaction between pairs of HEWL molecules in pH 3, 100 mM salt conditions is likely to be independent of the protein concentration in the range explored experimentally by Velev et al. The same finding is obtained in all other simulations performed at 100 mM salt (e.g., Figure S3B): no significant differences are observed between the effective interactions of protein molecules at protein concentrations of 1.25 and 10 g/L.

The differences in the $g(r)$ values obtained at different protein concentrations in 5 mM salt conditions have profound consequences for the estimated B_{22} values. The computed B_{22} values obtained for HEWL in 5 mM salt are plotted as a function of pH in Figure 4A for the protein concentrations of 1.25 and 10 g/L. Also plotted in this figure are the experimental B_{22} values obtained by Velev et al. from a linear regression of SLS data in the range 2 to 10 g/L. The B_{22} values obtained from simulations performed at 1.25 g/L (△) clearly far exceed the experimental estimates and exhibit an exaggerated dependence on pH. The B_{22} values obtained from simulations performed at 10 g/L (○) also exceed the experimental estimates though less so and, intriguingly, have a pH dependence that closely matches that observed experimentally. Consistent with the $g(r)$'s plotted in Figure 3A and 3C, the difference between the computed B_{22} values obtained at 1.25 and 10 g/L is greatest at pH 3 and smallest at pH 9. A similar picture emerges when the same kind of comparison is performed for chymotrypsinogen (Figure 4B): the absolute B_{22} values and their pH dependence are both drastically overestimated in the BD simulations performed at 1.25 g/L, and while the computed B_{22} values at 10 g/L are again too high, their pH dependence is again in much closer agreement with experiment.

As is considered in detail in the Discussion, the above results provide a potentially straightforward explanation for the overestimated B_{22} values obtained by others in 5 mM salt and also

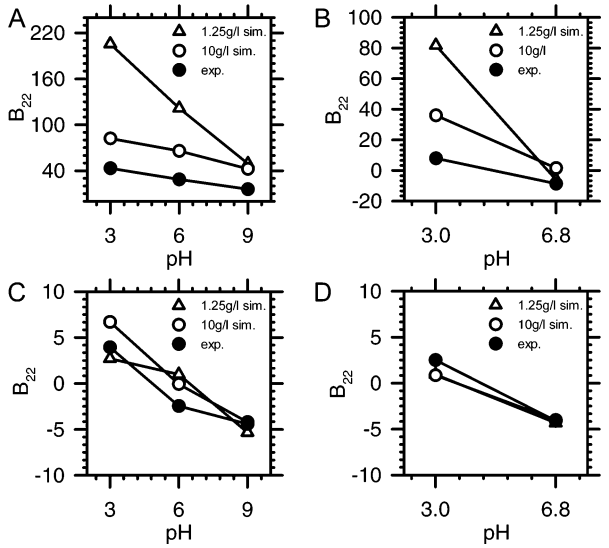

Figure 4. Comparison of computed and experimental B_{22} values as a function of pH showing the dependence on protein concentration. Experimental data are taken from Velev et al.[4] (A) HEWL in 5 mM salt (units of B_{22} are $10^4 \times$ mol mL/g^2). (B) Chymotrypsinogen in 5 mM salt. (C) HEWL in 100 mM salt. (D) Chymotrypsinogen in 100 mM salt.

suggest that estimating B_{22} by linearly regressing low-salt experimental SLS data at protein concentrations in the range 1−10 g/L is likely to be problematic. Fortunately, however, the simulations performed at moderate salt concentrations (100 mM) indicate that such problems are likely to be restricted to the very low salt regime: consistent with $g(r)$'s plotted earlier, the computed B_{22} values in 100 mM salt are very similar for both 1.25 and 10 g/L protein concentrations for HEWL and chymotrypsinogen (Figure 4C and 4D).

Translational and Rotational Diffusion. A key advantage of the present computational model is that in addition to providing structural data in the form of $g(r)$'s, and through them thermodynamic data in the form of B_{22} values, the BD simulations also naturally yield a large amount of information on the dynamic behavior of individual protein molecules. In this regard, it is important to note that although the Ermak−McCammon algorithm requires that the infinite-dilution values of the proteins' translational and rotational diffusion coefficients are specified prior to simulations being performed, the *effective* diffusion coefficients actually exhibited by the proteins during the simulations can differ significantly depending on the nature of their interactions with other molecules. For HEWL at a concentration of 10 g/L, a number of interesting trends are obtained when these effective diffusion coefficients are plotted against the computed B_{22} values (Figure 5A). The effective translational diffusion coefficient shows an approximately parabolic dependence on the computed B_{22} and is noticeably decreased from its infinite-dilution value both at negative and very positive values of B_{22}. Negative B_{22} values (which for HEWL are obtained at 100−500 mM salt and high pH) reflect the presence of significant favorable intermolecular interactions, and the accompanying decrease in translational diffusion coefficient therefore results from the formation of more slowly diffusing dimers and higher-order oligomers (see below). Very positive B_{22} values on the other hand (obtained at 5 mM salt and low pH) result from the presence of long-range repulsive interactions; the decreased translational diffusion coefficient therefore suggests that, in addition to having consequences for

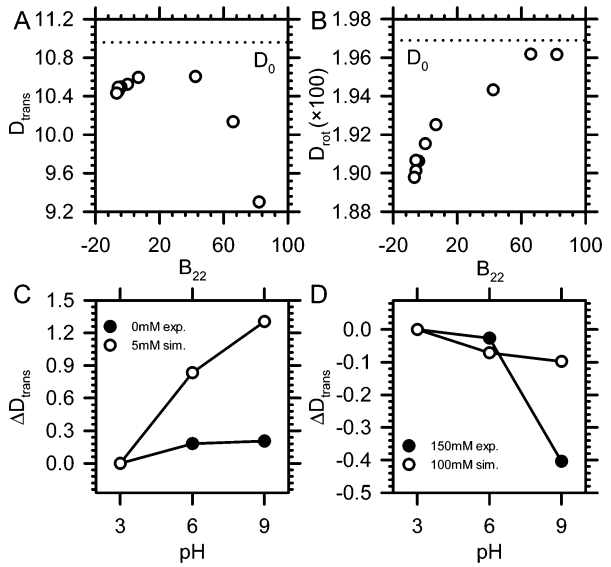

Figure 5. Dependence of translational and rotational diffusion coefficients on solution conditions. (A) Effective translational diffusion coefficient (Å^2/ns) of HEWL versus computed B_{22} value plotted for all simulated conditions; the dotted line indicates the infinite-dilution value (D_0) assigned to the proteins during simulations. (B) Same, but plotting effective rotational diffusion coefficient (/ns). (C) Comparison of effective translational diffusion coefficient from 5 mM salt HEWL simulations with experimental data (at 21 g/L) taken from Figure 2A of Price et al.[3] (D) Same, but comparing 100 mM salt HEWL simulations; experimental data taken from Figure 2B of Price et al.[3]

B_{22}, long-range electrostatic interactions may also significantly restrict diffusive movement.

The dependence of the effective rotational diffusion coefficient on the computed B_{22} values presents an interesting counterpoint to that of the translational diffusion coefficients (Figure 5B). For negative B_{22} values, the rotational diffusion coefficient is decreased from its infinite-dilution value, again reflecting the slowed diffusion that occurs when monomers become part of transient oligomeric clusters. For very positive B_{22} values, however, no decrease in rotational diffusion coefficient is observed. This result stands in contrast to the significant decrease in the translational diffusion coefficient that occurs under the same conditions but can be understood by considering the fact that at long distances the electrostatic potential generated by protein molecules becomes increasingly centrosymmetric (see Figure S4). In the 10 g/L protein concentrations studied here, each protein molecule is effectively surrounded on all sides by neighbors; all angular orientations of the molecules are therefore approximately isoenergetic, with the result that their rotational motion, in contrast to their translational motion, is largely unrestricted.

Based on the above results, it would be predicted that the translational diffusion coefficient of HEWL in very low salt conditions would decrease as the pH decreases due to the increased long-range repulsion progressively limiting the molecules' opportunities for translational movement. On the other hand, it would also be predicted that in higher salt conditions (e.g., 100mM) where long-range repulsions are suppressed, the translational diffusion coefficient should *increase* as the pH decreases because the increasing net charge would be expected to disfavor the formation of dimers and higher-order oligomers (see below). These predictions are in fact qualitatively (but only qualitatively) borne out in the experimental diffusion coefficient data reported by Price et al.[3] for HEWL in slightly more concentrated 21 g/L solutions (Figure 5C and 5D).

The relationship between the diffusive behavior of a protein molecule and its immediate environment can be investigated further by computing an "instantaneous" diffusion coefficient, by which we mean the diffusion coefficient of a molecule computed during a relatively short period of the simulation (e.g., 1 μs), and correlating this with properties describing the molecule's state of association with other molecules. Figure 6A shows how the instantaneous translational diffusion coefficient of a "typical" HEWL molecule changes during the course of a 10 μs simulation and compares this with (a) the molecule's average interaction energy with all other molecules and (b) its average oligomerization state at the same point in the simulation. As might be expected, there is a clear connection, and the time evolution of the diffusion coefficient of the molecule tracks closely with its complexation state.

This appears to be a surprisingly general result: when a similar analysis is conducted on all 1000 molecules in *all* of the HEWL systems simulated (at 100 mM salt and higher) and the results are averaged, a simple linear relationship emerges between the average oligomerization state of a molecule and its "instantaneous" translational diffusion coefficient (Figure 6B). Moreover, when this diffusion coefficient is expressed in ratio form relative to the infinite-dilution value of the diffusion coefficient (D_o), an identical *quantitative* dependence on the average oligomerization state is also obtained with chymotrypsinogen (Figure 6B).

Thermodynamics and Kinetics of Oligomerization. With the exception of the most repulsive solution conditions simulated (pH 3 and 5 mM salt), transient oligomeric clusters are formed in all BD simulations, and sampling is sufficient that at 10 g/L every one of the 1000 molecules becomes involved in a cluster at least once during the 10 μs of simulation (Figure S5). For HEWL in 100 mM salt, the association constants for oligomers obtained from the simulations (Figure 7A) are consistent with experimental estimates which range from 10 M^{-1} to ~300 M^{-1} (see discussions in refs 3 and 48), and there is a small, but statistically significant, increase in the association constant with increasing size of oligomer; similar behavior was seen in Monte Carlo simulations of spherical models of HEWL.[48] At pH 6, we also observe a modest (~25%) increase in the K_a values of all oligomers when going from 100 mM salt to 500 mM salt (data not shown); this result is also very similar to that obtained with the Linse group's spherical models.[48] For chymotrypsinogen, the simulated behavior is similar, albeit with an apparently smaller dependence of the association constant on the oligomer size (data not shown). The lifetimes of HEWL oligomers in 100 mM salt are shown in Figure 7B, from which it is apparent that the dissociation kinetics are quite rapid. Interestingly, the lifetimes of HEWL oligomers are essentially identical at pH 6 and pH 9 (Figure 7B), despite the fact that their thermodynamic association constants are significantly greater at pH 9 (Figure 7A). The pH independence of the dissociation kinetics is consistent with oligomerization being driven primarily by pH-independent van der Waals/hydrophobic interactions (see Methods). Since the dissociation kinetics are pH-independent, the pH dependence of the thermodynamic association constant must result from changes in the association kinetics. Interestingly, this pH dependence of association kinetics and pH independence

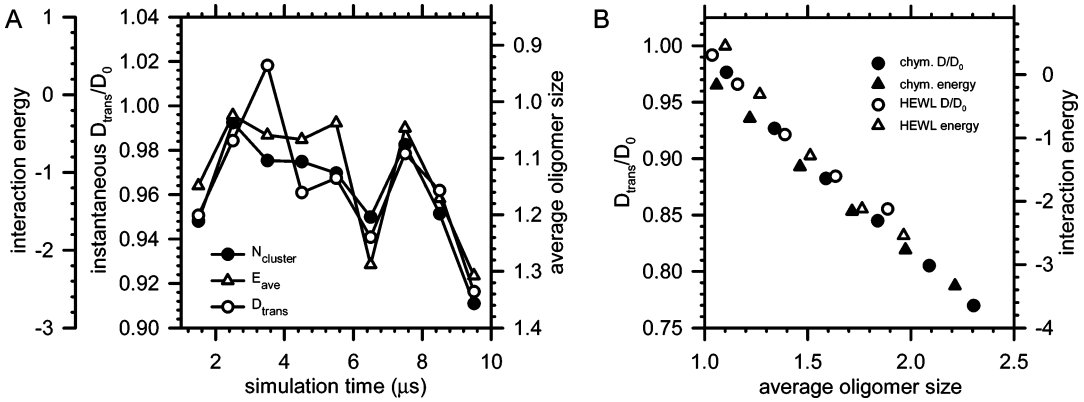

Figure 6. Dependence of diffusional behavior on intermolecular interactions. (A) (○) "Instantaneous" translational diffusion coefficient of a typical HEWL molecule during the course of a simulation in 100 mM salt at pH 9; (●) average oligomerization state of the same molecule; (△) average interaction energy (kcal/mol) of the same molecule with all other molecules. (B) (○) Average "instantaneous" translation diffusion coefficient of molecules in all HEWL simulations performed in 100 mM and 300 mM salt plotted versus their average oligomerization state. (●) Same, but for chymotrypsinogen. (△) Average interaction energy of molecules in all HEWL simulations plotted versus their average oligomerization state. (▲) Same, but for chymotrypsinogen.

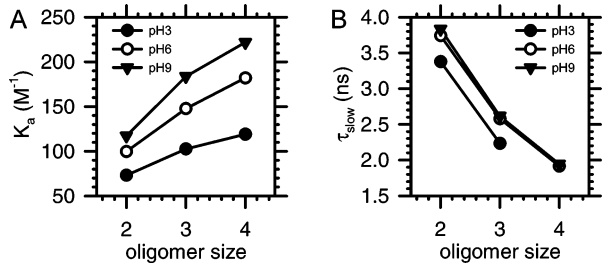

Figure 7. Thermodynamics and dissociation kinetics of oligomers in HEWL systems in 100 mM salt. (A) Association constants plotted versus oligomer size. (B) Lifetimes plotted versus oligomer size.

of dissociation kinetics are mirrored experimentally in the differing salt dependences of protein−protein association and dissociation kinetics.[55]

Intermolecular Contacts. The atomic detail of the simulation model allows a direct view of the relative propensities of different surface atoms to participate in interactions with other molecules. Following a straightforward conversion of atomic contact frequencies into free energies (see Methods), a simple coloring scheme can be used to illustrate preferred sites of interaction, and these can in principle be used to obtain insights into the types of interactions (electrostatic or hydrophobic) that drive the associations. Of course, an additional factor that determines an atom's propensity to be involved in interactions with other proteins is simply its accessibility, but it is possible to control for this effect by comparing with BD simulations in which all surface atoms are treated as hard spheres incapable of engaging in favorable interatomic interactions. Interestingly, even when accessibility is controlled for, it is not always clear that there is a simple connection between an atom's propensity to be involved in intermolecular contacts and the local hydrophobicity or electrostatic potential. There are however cases where straightforward relationships can be discerned, with the most blatant example being found in a comparison of wild-type T4 lysozyme with its S44F mutant: in the case of the wild-type protein (Figure 8A, left), atoms with high contact propensities are relatively evenly spaced over the entire surface, whereas, in the mutant (Figure 8A, right), they are highly concentrated in the region of the Phe 44 side chain (see Figures S6 and S7 for corresponding views of other systems).

(55) Zhou, H. X. *Biopolymers* **2001**, *59*, 427.

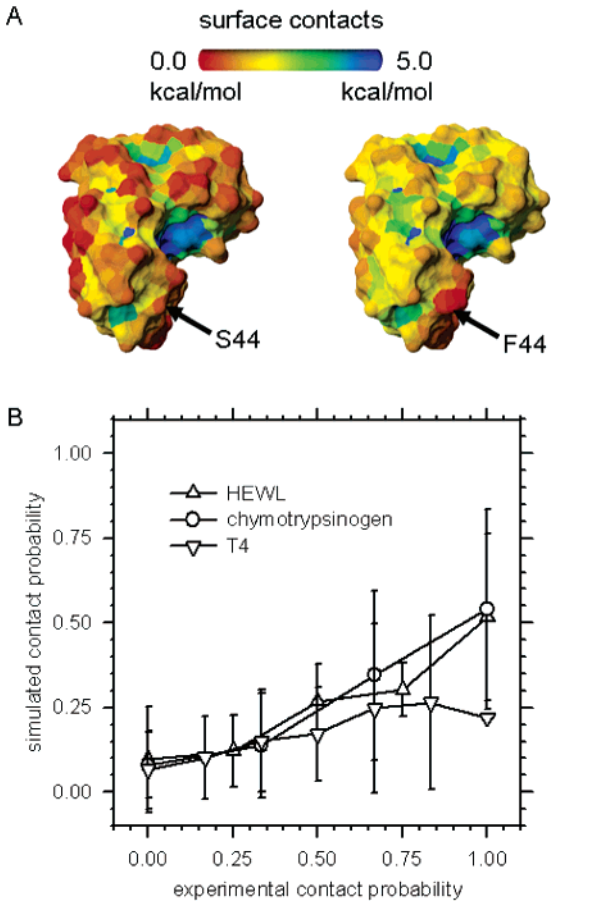

Figure 8. (A) Relative contact probabilities plotted in free energy form ($\Delta G°_{contact}$) for surface atoms in wild-type T4 lysozyme (left) and the S44F mutant (right). (B) Relative probability of an atom being involved in a contact with another protein in BD simulations plotted against relative probability of an atom being involved in a crystal contact (see text).

While they can be visually informative, it is difficult to directly compare the computed interaction propensities with experiment, although this might be done in future by experimentally mutating those residues predicted to be most responsible for intermolecular contacts. However, one indirect way of evaluating the predictions is by comparison with the atoms involved in interprotein contacts in high-resolution crystal structures of the proteins. The structures of all three proteins

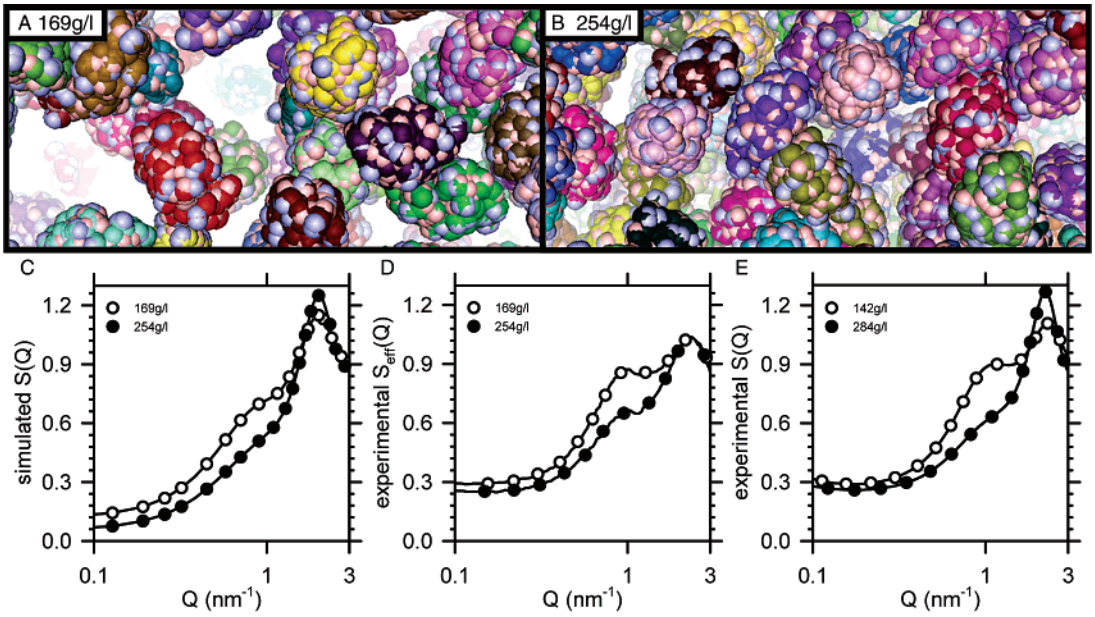

Figure 9. (A) Image of HEWL system simulated at 169 g/L protein concentration. (B) Same, but for 254 g/L. These figures were prepared with PyMol.[56] (C) Structure factor, $S(Q)$ computed from BD simulation data plotted versus the wavevector Q. (D) Same, but showing experimental "effective" structure factors taken from Stradner et al.[20] (E) Same, but showing experimental structure factors taken from Liu et al.[21]

studied here have been solved in multiple space groups, and a relatively straightforward comparison can therefore be made by plotting the probability of an atom being involved in a contact in the simulations against its probability of being involved in a crystal contact in one of the experimentally crystallized space groups. These comparisons are shown in Figure 8B, from which it can be seen that for all three proteins there is a significant correlation between the simulated and experimental probabilities. Interestingly, a similar degree of correlation is also obtained using contact probabilities obtained from hard-sphere-only simulations (Figure S8); this suggests that the probability of an atom being involved in a crystal contact is primarily determined by its accessibility to contact from other protein molecules. It should be realized however that this measure of accessibility, in which the probe is another protein molecule, is not the same as the conventional solvent accessibility measured with a probe the size of a water molecule.

Behavior at Very High Protein Concentrations. One final aspect that was investigated was whether the simulation model parametrized to fit B_{22} data for HEWL at a 10 g/L protein concentration could also capture solution behavior observed at higher protein concentrations; in fact, simulations were performed at concentrations up to 254 g/L, which, relative to the parametrization conditions, represents an ambitious 25-fold increase in the simulated protein density. Although much more computationally expensive than simulations performed at lower protein concentrations, simulations could be run for sufficiently long time periods that reasonably converged estimates of $g(r)$, and hence, the structure factor, $S(Q)$, could be obtained for Q down to $\sim$0.1 nm^{-1}. Figure 9C compares the computed $S(Q)$ obtained from simulations performed at 169 and 254 g/L with the corresponding experimental structure factors recently reported by two groups[20,21] (Figure 9D and 9E); close-up views of the simulated systems are shown in Figure 9A and 9B. The agreement is good, though not perfect. In the experimental $S(Q)$ data reported by Stradner et al.,[20] a small but discernible peak is obtained in the 169 g/L plot at $Q \approx 1$ nm^{-1} that disappears in the 254 g/L plot; the same behavior is also seen in the data of Liu et al.[21] obtained at slightly different protein concentrations. In the simulated $S(Q)$ data, the peak manifests itself instead as a shoulder (again at $Q \approx 1$ nm^{-1}), but its disappearance at the higher protein concentration is correctly captured (for $S(Q)$ plots obtained at somewhat lower protein concentrations, see Figure S9). Less easy to interpret is the behavior of the major peak in $S(Q)$ at $\sim$2 nm^{-1}. In the data of Stradner et al.[20] the amplitude of this peak is concentration-independent; in the data of Liu et al.[21] however, the peak increases significantly in magnitude as the concentration increases. It is not clear why there is this discrepancy between the two experimental curves, but the latter behavior is qualitatively reproduced in the $S(Q)$ computed from the BD simulations. It is further worth noting that the study of Liu et al.[21] also reports an increase in $S(Q)$ at very low Q (0.004 nm^{-1}); however in these preliminary simulations it has not been possible for us to obtain accurate estimates of $g(r)$ at the long distances necessary for computing $S(Q)$ with confidence at very low Q.

Discussion

The simulation method discussed here is intended to model the diffusion and association of macromolecules on a length scale of thousands of angstroms and on a time scale of microseconds to milliseconds. As such, its ultimate purpose is to provide a realistic description of macromolecular behavior in the types of complex mixtures that are encountered physiologically, while retaining a high level of structural detail in the modeled molecules.[57] In this application of the methodology to single-component protein solutions, one of the central goals has been to reproduce experimental B_{22} data for three model proteins, with the idea that this should provide an important indication of the method's ability to describe weak, nonspecific macromolecular interactions. Although there are a number of

(56) DeLano, W. L. *The PyMOL; User's Manual*; DeLano Scientific; San Carlos, CA, 2002.
(57) Takahashi, K.; Arjunan, S. N. V.; Tomita, M. *FEBS Lett.* **2005**, *579*, 1783.

possible applications of the methodology, one for the immediate future is modeling the concentration dependence of protein rotational diffusion coefficients studied experimentally by Krushelnitsky, Fedotov, and colleagues;[58] a somewhat simplified BD model developed by these authors has already proven useful for qualitatively describing aspects of the experimental behavior.[59] A second attractive application of the method would be modeling the early stages of protein crystallization, a process that has already attracted computer modeling work,[60−62] and for which an interesting connection between crystallization conditions and B_{22} values has been reported.[63]

In assessing the success of the present application, it should be remembered that B_{22} computations are extremely sensitive to the parameters of energy models.[16,17] This can be nicely illustrated by comparing the parameter sensitivity of B_{22} with the parameter sensitivity of an alternative measure of protein−protein interaction thermodynamics, the free energy of association of two monomers to form a dimer, ΔG°_{assoc}. As an example, in our attempts to parametrize the energy well-depth ϵ_{LJ} for HEWL, three different values were investigated: $\epsilon_{LJ} = 0.26$, 0.28, and 0.30 kcal/mol. The computed B_{22} values obtained with these parameters were -1×10^{-4}, -4×10^{-4}, and -11×10^{-4} mol mL/g^2, respectively, all of which values are sufficiently well spaced that they should be experimentally distinguishable.[4] The computed ΔG°_{assoc} values obtained with these same parameters (obtained simply from the relative populations of dimers and monomers in the simulations) are -2.70, -2.88, and -3.11 kcal/mol, respectively. Even if such weak binding constants could be measured experimentally, the small differences would likely remain unresolvable. In other words, apparently drastic errors in computed B_{22} values may actually correspond to rather small errors in ΔG°_{assoc} values.

This point should in particular be remembered when considering the apparently disappointing result that different ϵ_{LJ} parameters were derived for the three proteins studied here. Since an attempt to use a single "best-fit" ϵ_{LJ} parameter would lead to poor predictions of B_{22} values for one or more of the proteins, it is clear that further refinement of the current model will be required if B_{22} values are to be quantitatively reproduced (see below). However, for investigating less sensitive properties of a system, it may be that the existing level of correspondence between the ϵ_{LJ} parameters is already sufficient to arrive at a single compromise value that might be used in studies of other protein systems. For example, if we extrapolate the computed ΔG_{assoc} values for HEWL to estimate what might be obtained using $\epsilon_{LJ} = 0.22$ kcal/mol, a ϵ_{LJ} value that produces very good estimates of B_{22} for T4 lysozyme and chymotrypsinogen, we predict a value of $\Delta G^\circ_{assoc} = -2.49$ kcal/mol, which differs by only 0.39 kcal/mol from the value obtained with our "best" value for HEWL of ϵ_{LJ} (0.28 kcal/mol); it may be therefore that the former value could be used to compute properties of HEWL systems (other than B_{22}) without significantly sacrificing accuracy.

The major advantage of the present method is the fact that it allows simulations of large numbers of macromolecules to be performed. This feature has turned out to be critical for uncovering an important result of the present work, which is that long-range interactions between many molecules in low salt conditions can significantly affect their apparent pairwise interaction. Before considering what this means for previous attempts to computationally model B_{22} values for low-salt conditions, it is obviously crucial to establish whether this result has any experimental support. The answer is "yes". To see this, it is important to appreciate that the experimental estimates of B_{22} for protein solutions are usually obtained as the *gradient* of static light scattering (SLS) data plotted as a function of protein concentrations in the range 1−10 g/L. If the pairwise interactions of protein molecules are truly independent of protein concentration in this concentration range, the gradient will also be constant, and the resulting plot should therefore be linear. The raw data shown in Figure 1 of Velev et al.[4] for HEWL in moderate salt concentrations (100 and 300 mM) do indeed fit this scenario, and a linear regression is supported by the fact that an extrapolation to zero protein concentration leads to an accurate estimate of HEWL's molecular weight. The raw data reported in the same figure for low salt conditions (5 mM) were also assumed to be linear by Velev et al., but an indication that this may not have been appropriate for at least the pH 3 data is that, as noted by the authors, its extrapolation to zero protein concentration leads to an inaccurate molecular weight estimate. Perhaps more tellingly, in more recent works reported by the same group, SLS data obtained in low-salt conditions have been fitted to quadratic rather than linear functions,[19,52] and B_{22} values have been obtained as the gradients of these functions evaluated at zero protein concentration; certainly significant curvature is now apparent in newer data reported for HEWL at low salt by the same group (see Figure 2 of Paliwal et al.[52]). Importantly, both the presence and the sign of curvature in these plots are consistent with the behavior obtained in the present HEWL simulations: in 5 mM salt and low pH (Figure 3A), our computed B_{22} values are much smaller in magnitude (consistent with a smaller gradient in SLS data) at high protein concentration (10 g/L) than at a lower concentration (1.25 g/L).

Velev et al.'s use of a linear regression of SLS data in low salt conditions means that their reported B_{22} values for both HEWL and chymotrypsinogen are likely to be significantly underestimated at the lower pH values (since the regression included high concentration data points for which the apparent B_{22} is lower). If so, this will have had unfortunate consequences for the previous computational studies that have aimed to reproduce their data, all of which have obtained values that are significantly more positive than the reported experimental values. Previously published computations of low-salt behavior include the simple but effective DLVO model calculations reported by Velev et al. themselves,[4] the calculations of the Linse group[48] which employed a spherical protein model for HEWL with a charge distribution closely approximating the distribution found in the crystal structure, and the calculations of Lund and Jönsson,[18] which used a protein model in which individual residues were modeled as spheres, and which explicitly modeled the dissolved salt ions. At least some of the overestimation of the low-salt B_{22} obtained in these previous studies might now be explained by the fact that the calculations considered only a *pair* of interacting molecules and were

(58) Krushelnitsky, A. *Phys. Chem. Chem. Phys.* **2006**, *8*, 2117.
(59) Ermakova, E.; Krushelnitsky, A. G.; Fedotov, V. D. *Mol. Phys.* **2002**, *100*, 2849.
(60) Pellegrini, M.; Wukovitz, S. W.; Yeates, T. O. *Proteins: Struct., Funct., Genet.* **1997**, *28*, 515.
(61) Kierzek, A. M.; Zielenkiewicz, P. *Biophys. Chem.* **2001**, *91*, 1.
(62) Auer, S.; Frenkel, D. *J. Phys.: Condens. Matter* **2002**, *14*, 7667.
(63) George, A.; Wilson, W. W. *Acta Crystallogr., Sect. D* **1994**, *50*, 361.

therefore incapable of capturing the many-body effects that were probably present in the experimental data (for an interesting discussion of an additional issue that may have been overlooked in some of these studies, see Asthagiri et al.[19] and Paliwal et al.[52]). If this is so, then it may well be that a comparison with experimental data obtained at (or extrapolated to) lower protein concentrations would show that the computational models developed in these previous works are actually more accurate at low salt than previously thought.

An alternative way to explore this issue would be to attempt to incorporate many-body effects into two-molecule calculations using integral equation approaches:[53,54,64,65] this would enable the range of validity of these previous models to be extended to higher protein concentrations. In this context it is interesting to note that it has recently been shown that integral equation calculations using the hypernetted chain closure can provide $g(r)$ estimates that are in good agreement with the results of Monte Carlo simulations of charged spheres in low-salt conditions similar to those studied here.[54] It is also worth noting that some of these same authors also anticipated,[66] on purely theoretical grounds, the idea that long-range repulsive electrostatic interactions might contribute to an effective favorable interaction between protein molecules in the experiments of Velev et al.[4]

Of course, the present simulation model, by explicitly modeling the interactions of multiple molecules, provides a more natural way of exploring the concentration dependence of protein−protein interactions. That said, it should not be thought that it overcomes all of the problems encountered in other studies: our own computed B_{22} values are clearly far from perfect in 5 mM salt, and although it is intriguing that the pH dependence of B_{22} obtained at 10 g/L concentration is in rather good agreement with that obtained by Velev et al., we should be careful not to overinterpret this result. In fact, in previous work, one of us has argued that correct reproduction of pH dependent effects would almost certainly require that modeled proteins be allowed to assume variable protonation states during simulations,[17] and subsequent work by others has reached a similar conclusion.[67] This may be one reason why the pH dependence of the effective translational diffusion coefficients obtained in our simulations is significantly greater than that seen in the experiments of Price et al.[3] at low salt (Figure 5C). An efficient way of incorporating protonation state changes during simulations remains to be developed. It is also worth noting that the mere ability to model multiple molecules does not guarantee that many-body effects will be properly captured; instead, as always with simulations, there can be technical issues that have unforeseen and undesirable consequences. An illustration of this particular aspect can be found in one of the previous B_{22} studies discussed above. In the same study performed by the Linse group[48] that described two-molecule B_{22} calculations, Monte Carlo (MC) simulations of 100 HEWL molecules were also reported (though not explicitly used to compute B_{22}). Since these simulations contained multiple molecules and were performed at the experimental concentrations studied by Velev et al., they should in principle have captured the same many-body effects observed in the present simulations. Crucially however, the Linse group's MC simulations truncated electrostatic interactions between proteins at 120 Å, which is precisely the region where the long-range attractive peak in $g(r)$ begins to appear in our simulations (Figure 3A); if electrostatic interactions had instead been truncated at a somewhat longer distance in that study, it is likely that an attractive peak in $g(r)$ would have been obtained.

In addition to the straightforward modeling of interactions between many molecules, a final, significant advantage of the present simulation method is that it yields a rather broad range of structural and dynamic data, much of which can also be accessed experimentally. Because of this, the method has the potential to provide a natural framework for interpreting experimental data such as translational and rotational diffusion coefficients for which the derivation of analytical theories is not straightforward, or for which analytical expressions have an uncertain range of validity. The comparisons that we have made between simulated and experimental translational diffusion coefficients and the structure factors of highly concentrated HEWL solutions, although not quantitatively accurate, clearly show a promising qualitative agreement. The ability to simulate a variety of properties is likely to be of considerable use in the future since a simultaneous comparison of several different simulated properties with corresponding experimental data should allow the parameters, and perhaps the form, of the energetic description used in the simulation model to be more tightly defined than is currently possible: it may be for example that a number of different energy models might be capable of reproducing B_{22} data in moderate salt conditions, whereas only one might be capable of simultaneously capturing additional data such as translational diffusion coefficients. Clearly, there is a number of different extant energy models that might be incorporated into the same basic framework used here;[68−75] even in its current form however the model presented here appears to have considerable potential for providing predictive rather than purely phenomenological descriptions of the behavior of concentrated macromolecular systems.

Acknowledgment. The authors would like to thank Professor Craig Kletzing for help with understanding the dynamics of rigid bodies, Professor Carlos J. Camacho for the idea of conducting simulations with only hard-sphere interactions, and Professor Harvey W. Blanch for discussions of his group's B_{22} measurements. This work was supported by a Research Grant from the Carver Trust.

Supporting Information Available: Additional figures; a discussion the simple approach to ensure conservation of forces and torques; a discussion of the methodology developed for dealing with steric overlaps of proteins. This material is available free of charge via the Internet at http://pubs.acs.org.

JA0614058

(64) Vlachy, V.; Prausnitz, J. M. *J. Phys. Chem.* **1992**, *96*, 6465.
(65) Lin, Y.-Z.; Li, Y.-G.; Lu, J.-F. *J. Chem. Phys.* **2002**, *117*, 407.
(66) Spinozzi, F.; Gazzillo, D.; Giacometti, A.; Mariani, P.; Carsughi, F. *Biophys. J.* **2002**, *82*, 2165.
(67) Lund, M.; Jönsson, B. *Biochemistry* **2005**, *44*, 5722.
(68) Cerutti, D. S.; Ten Eyck, L. F.; McCammon, J. A. *J. Chem. Theory Comput.* **2005**, *1*, 143.
(69) Wang, T.; Wade, R. C. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 158.
(70) Jiang, L.; Gao, Y.; Mao, F. L.; Liu, Z. J.; Lai, L. H. *Proteins: Struct., Funct., Genet.* **2002**, *46*, 190.
(71) Camacho, C. J.; Kimura, S. R.; DeLisi, C.; Vajda, S. *Biophys. J.* **2000**, *78*, 1094.
(72) Elcock, A. H.; Gabdoulline, R. R.; Wade, R. C.; McCammon, J. A. *J. Mol. Biol.* **1999**, *291*, 149.
(73) Dominy, B. N.; Brooks, C. L., III. *J. Phys. Chem. B* **1999**, *103*, 3765.
(74) Zhang, C.; Vasmatzis, G.; Cornette, J. L.; DeLisi, C. *J. Mol. Biol.* **1997**, *267*, 707.
(75) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623.