Clusters of lysozyme in aqueous solutions

A. Baumketner and W. Cai

Phys. Rev. E **98**, 032419 — Published 28 September 2018

DOI: 10.1103/PhysRevE.98.032419

Clusters of lysozyme in aqueous solutions

A. Baumketner[1,*] and W. Cai[2]

[1] Institute for Condensed Matter Physics, NAS of Ukraine, 1 Svientsistsky Str, Lviv, 79011, Ukraine.

[2] Department of Mathematics, Southern Methodist University, 3200 Dyer Street, Dallas, TX 75275, USA

*Corresponding Author. Email address: andrij@icmp.lviv.ua

Abstract

Equilibrium clusters of protein lysozyme are at the center of an ongoing scientific debate. Previous attempts to provide a microscopic description of the clusters that is consistent with all experimental evidence have not been fully successful. The primary reason is the use of model potentials that have a pre-defined shape. In this paper we derive a model-free inter-protein potential directly from experimental structure factor.The derived potential is globally repulsive but has a local minimum at short distances. The minimum is essential for the correct behavior of the structure factor with protein concentration, in particular the shifting pattern of the signature maximum at short wave vectors. Equilibrium clusters are observed throughout the entire range of concentrations but their nature differs in the low and high concentration limits. At low concentrations, the clusters are extended in shape. As the concentration is increased, small clusters collapse while large clusters are assembled from the small ones. Hydrodynamic interactions drive a kinetic slow down at high concentrations, where a transition into a fluid of permanent clusters of specific size is observed. In good agreement with the available experimental data, our simulations shed new light on the microscopic nature of protein clusters.

Introduction

Whether protein lysozyme can make equilibrium clusters in aqueous solutions is the subject of a vigorous scientific debate. It started in 2004 with the paper of Stradner et al[1], presenting the results of a small-angle neutron scattering (SANS) study of aqueous solutions of protein lysozyme at varying concentration *c*. The reported static structure factor exhibited a second maximum located at a short wave vector k_{max}, in addition to the main maximum at a longer wave vector corresponding to the nearest-neighbor distance between the proteins. Since the second maximum indicates longer-range correlations it was concluded that it reports small protein assemblies, or clusters, and corresponds to the inter-cluster distances. Furthermore, the hallmark of the experimental observations - that k_{max} does not shift with *c*, implies that the cluster-cluster distance remains the same for all concentrations. The only way this can be achieved is when the clusters grow in size at increasing concentration. Specifically, a proportionality relationship $N_c \sim c$ was assumed, where N_c is the cluster size.

The cluster model was questioned by Shukla et al[2, 3]who repeated the original scattering experiments and found that k_{max} actually shifts to higher values as the concentration is increased. Based on this observation the authors suggested that k_{max} is the main maximum in the structure factor and that this maximum corresponds to the nearest-neighbor distance between the proteins. Since the proteins are believed to experience strong mutual repulsion under the conditions of the experiment they choose to remain at a maximal possible distance one from another. This leads to the uniform compression of the solution at increasing concentration, causing all inter-protein distances to shrink. As a consequence, the position of the maximum in the structure factor shifts upwards, in good agreement with the measurements. This model does not envision the formation of any clusters.

An alternative explanation seeking to reconcile the above two scenarios was proposed by Liu et al[4]. These authors varied systematically the shape of a model potential and used integral-equation theory to identify systems that a) display a maximum at k_{max} but lack clusters at low protein concentrations and b) display clusters but lack the maximum at high protein concentrations. Either case provides evidence against a strong correlation between a maximum at k_{max} and the appearance of clusters, as was suggested earlier[1, 5]. Instead, it is argued that the maximum arises because of the specific shape of the inter-protein potential which causes strong protein-protein correlations at intermediate distances (compared to the nearest-neighbors contacts) that may be realized through clusters but also through other structures, for instance protein gels. Correlation between the appearance of clusters and the maximum at $k_{max}$is seen when $S(k_{\max})$ is larger than 2.7[6]. An important conclusion following from this observation is that the presence of clusters cannot be determined based on structural information alone. Instead, it is argued that additional studies are needed, in which the size of the various species in the system can be evaluated directly.

Such studies were carried out recently by the neutron spin-echo (NSE) method[7-9], measuring hydrodynamic radius, and by NMR[8, 10], reporting long-time diffusion constant. Both approaches find evidence that, at increasing protein concentration, a new species emerges, which has a radius larger than that of the monomer and, therefore, can be associated with clusters. A consensus opinion emerging from these studies[8, 11] is that lysozyme clusters do exist but only at high protein concentrations, where k_{max} exhibits no shifting. This is to be contrasted with the low protein concentrations, where k_{max} shifts and no clusters are observed.

While the existence of clusters does not seem to be contested anymore, their nature remains elusive. A major obstacle for the structural characterization of clusters at the microscopic level is the lack of reliable inter-protein potentials. Lysozyme has been studied extensively over the last decade under a

wide range of conditions[2, 12-19]. However, in all these studies either a model potential[2, 12-17] or an integral-equation theory was used[18, 19], both of which are approximate. As a consequence, there is no potential for lysozyme to date that produces good agreement with experiment. In the most accurate approaches, potentials are obtained directly from SANS data by fitting [2, 4, 9, 20, 21]. The quality of such potentials remains untested, however. In one case[20], simulations experience very slow dynamics, presumably because of a deep minimum in the employed potential, reaching $\sim 8k_BT$, where k_B is the Boltzmann constant and T is the temperature. As a consequence, comparison with experiment is made difficult by poor equilibration. Studies conducted by another group[9, 21] do not suffer from the convergence problem but produce results that do not completely agree with experiment. Indeed, it is found[21] that the maximum in the obtained structure factor does not shift with concentration. In addition, the system does not experience a kinetic slow down near a certain critical concentration[9] seen experimentally. These shortcomings leave the question about the nature of lysozyme clusters unanswered.

Equilibrium clusters are a relatively new and little studied phenomenon[1, 22, 23]. In physics and chemistry clusters represent a particular example of particle self-assembly. In biology, protein clusters may have a well-defined functional role, for instance, serving as a first step in the polymerization reaction responsible for the sickle cell disease[24, 25]. Since recently clusters have also been researched for use in nanotechnological applications, in particular, drug delivery[26]. In the wider context, a comprehensive understanding of why and how clusters are made is needed for both advancing the frontiers of basic science and developing new technologies.

In this paper, we present the first microscopic description of lysozyme clusters that agrees well with all available experimental data. As in our prior work[27], we use SANS data to derive inter-protein potential by Boltzmann inversion. We focus on the structure factor[28] that has the characteristic second maximum at k_{max}. The potential obtained by our procedure is repulsive everywhere except at short distances,

where it has a small local minimum. The minimum, in contrast to a global minimum considered earlier[2, 4, 9, 21], ensures that the experimental concentration behavior of k_{max} is correctly reproduced: shifting is seen at low concentrations and there is no shifting at high c. The model predicts small and extended clusters at low concentrations while at high concentrations, the clusters are large and collapsed. At increasing concentration, a transition into a non-ergodic state is seen. Beyond a critical concentration $c_{cr} = 300mg/ml$ the clusters can be considered frozen, or permanent. Hydrodynamic interactions, included in our model through the dissipative particle dynamics (DPD), cause a kinetic slow down on approach to c_{cr}, in good agreement with experiment.

Model and methods

The inter-protein potential $v(r)$ was obtained as described in detail previously[27]. Briefly, our method relies on the one-to-one correspondence between a potential $v(r)$ and the associated pair distribution function (PDF)[29]. The potential is determined in successive iterations[30-32] as $v_{l+1}(r) = v_l(r) - \lambda k_B T log(g_R(r)/g_l(r)))$, where $v_l(r)$is the approximation at iteration l, $g_R(r)$ is the experimental PDF, $g_l(r)$ is the PDF obtained at iteration l and λ is a certain adjustable parameter whose purpose is to control the rate of convergence. The iterations were started from a purely repulsive initial guess $v_1(r)$. Pair distribution function was computed at each iteration by molecular simulations using stochastic dynamics algorithm. The temperature was set at $T = 298K$ while the number of particles was set to match the experimental density. It took 48 iterations to obtain the potential shown in Figure 1(b). The iterations used a cut-off distance of R_c =250Å. The simulation box contained 800 particles.

To obtain structural functions simulations were performed at three temperatures as indicated in Figure 2(a). The size of the system was reduced to 512 particles, which had no adverse effects on the results. Additionally, the cut-off distance was reduced to 75Å. The potential is seen to decay to zero at that distance. The truncated part at $r > R_c$ contains small undulations, which we ignored as they were

judged an artifact of the truncation of the experimental structure factor at a finite wave vector during the numerical Fourier transform.

All simulations were performed for the potential in which the distance was multiplied by 0.1. The size of the simulation box was scaled appropriately. This allowed us to bring the system to the atomic scale, making it easier to set simulation parameters and manage and store the data. We used the stochastic algorithm implemented in Gromacs[33, 34] to maintain constant temperature. The inverse friction coefficient was set to $\tau_t = 2ps$. All trajectories contained $5 \cdot 10^5$ time steps. The time step was set at $\delta t = 2fs$.

Dissipative particle dynamics (DPD) method was used to conduct dynamics simulations[35]. A Fortran code was written specifically for this purpose. The method takes into account viscous forces between particles that are mediated by the solvent flows, or the hydrodynamic (HD) interactions. Stochastic forces are added to maintain constant temperature. The equations of motions to be solved are as follows:

$$\frac{d\vec{r}_i}{dt} = \vec{v}_i(t) \tag{1}$$

$$m\frac{d\vec{v}_i}{dt} = \vec{f}_i(t) = \sum_{i\neq j, r_{ij}<Rc} \vec{F}^C(\vec{r}_{ij}) + \sum_{i\neq j, r_{ij}<Rcv} \{\vec{F}^D(\vec{r}_{ij}, \vec{v}_{ij}) + \vec{F}^R(\vec{r}_{ij})\} \tag{2}$$

where m is the mass of the particles, $\vec{r}_{ij} = \vec{r}_i - \vec{r}_j$,$\vec{v}_{ij} = \vec{v}_i - \vec{v}_j$, $\vec{r}_i$ is the radius vector of particle i, $\vec{v}_i$ is its velocity and $\vec{f}_i(t)$ is the total force acting on that particle. The force is pair-wise additive and composed of three contributions. The first is the conservative force due to inter-particle potential $\vec{F}^C(\vec{r}_{ij})$. The second is the viscous drag force $\vec{F}^D(\vec{r}_{ij}, \vec{v}_{ij})$, which describes how movements of one particle are transmitted to the other particle through the flow of solvent. This force depends on relative velocity of the affected particles. The third is a random force $\vec{F}^R(\vec{r}_{ij})$, designed to maintain constant

temperature in the system. Stochastic and viscous forces are related through the fluctuation-dissipation theorem. In our implementation the relevant expressions are: $\frac{\vec{F}^R(r_{ij})}{m}\delta t = f(r_{ij})\, g \frac{\vec{r}_{ij}}{r_{ij}}\xi$ and $\frac{\vec{F}^D(r_{ij})}{m}\delta t = -f(r_{ij})^2 \frac{(\vec{v}_{ij}\vec{r}_{ij})}{r_{ij}}\frac{\vec{r}_{ij}}{r_{ij}}$, where ξ is a random variable uniformly distributed between 0 and 1, δt is the time step to be used in numerical integration of the equations of motion, $g = \sqrt{2k_BT/m}$ and $f(r_{ij})$ is a certain dimensionless function that controls the strength of the viscous forces. We used the self-consistent leap-frog algorithm of Pagonabarraga et al [36] to integrate the equations of motion(**1**)-(2). The algorithm consists of two consecutive steps as shown below:

Step 1

$$\vec{v}_i\left(t+\frac{\delta t}{2}\right) = \vec{v}_i\left(t-\frac{\delta t}{2}\right) + \delta t\frac{\vec{f}_i(t)}{m} =$$

$$= \vec{v}_i\left(t-\frac{\delta t}{2}\right) + \delta t\frac{\vec{F}_i^{\,C}}{m} + \sum_{j\neq i, r_{ij}<R_{cv}} \left\{ f(r_{ij}) g\, \frac{\vec{r}_{ij}}{r_{ij}}\, \xi - f(r_{ij})^2 \frac{(\vec{v}_{ij}\vec{r}_{ij})}{r_{ij}}\frac{\vec{r}_{ij}}{r_{ij}} \right\}$$

Step 2

$$\vec{r}_i(t+\delta t) = \vec{r}_i(t) + \vec{v}_i(t)\delta t + \frac{\vec{f}_i(t)}{m}\frac{\delta t^2}{2}$$

In step 1 the velocities are propagated by t . Note however, that the right-hand side contains a contribution that depends on velocities at moment of time t. The latter can be estimated as $\vec{v}_i(t) = \frac{\vec{v}_i\left(t+\frac{\delta t}{2}\right)+\vec{v}_i\left(t-\frac{\delta t}{2}\right)}{2}$, which turns the single equation for particle i in Step 1 into a system of coupled linear equations for all particles[36]. The equations can be efficiently solved by matrix inversion. However, this method becomes time consuming for large systems so instead we chose the iterative solution. It took no more than 5 iterations for all densities to obtain converged velocities. The second step is for the propagation of coordinates. It is straightforward to perform once the velocities are known.

The summation in eq. (2) is carried out over pairs of particles with mutual separations up to certain cut-off distance. For the conservative forces, the cut-off is $R_c = 75\text{Å}$ as was discussed earlier. For the hydrodynamic interactions we used a different cut-off R_{cv}, which was set according to the following considerations. Function $f(r_{ij})$ controls the strength of the viscous force acting between particles at a distance r_{ij}. It is defined by the properties of the solvent contained in the space between the particles. If something other than solvent, for instance another particle, is allowed to enter that space, $f(r_{ij})$ is expected to change drastically. In this case the description with a single $f(r)$ function may no longer be a good approximation for HD interactions and should be avoided. We set Rcv according to this criterion. It is seen in Figure 1(a) that the nearest-neighbor distance in a pair of particles is 26Å, which means that the distance between furthermost particles in a 3-particle linear cluster is 52Å or more. To eliminate the error in HD interactions caused by such clusters we set the cut-off at a shorter distance, $Rcv = 50\text{Å}$.

For simplicity we ignored the distance dependence in $f(r)$[36] and replaced that function with a constant f[37]. We varied f, which is the only free parameter in the algorithm, systematically in order to determine its effect on the dynamics of the system. Note that the static structure is not affected by f. The value of 0.4 was seen to produce best agreement between simulation and experiment. Lower f values underestimate $D_0/D_s(k)$ (faster dynamics) while higher values – overestimate it (slower dynamics). To the extent that the employed model of HD interactions is correct, the determined value of f reflects the true dynamics of the studied system.

The dissipative force in eq. (2) is along the vector connecting the particles[35]. We also tested models in which the force lies along a perpendicular direction[37] or along the vector of relative velocities[38]. Quantitatively the results differed among all three approaches but qualitatively – remained the same.

Results and Discussion

Inter-protein potentials

Structural functions of protein solutions intricately depend on a number of parameters, including pH, the type and concentration of buffer used, counter ions etc. In this study we use the SANS data of Abramo et al[28]obtained for lysozyme solutions at pH 2, temperature T=298K and protein number density $\rho = 4.2 \cdot 10^{-6}$ Å^{-3}. In common with the results of other groups[1, 2, 39], the studied static structure factor $S(k)$ has a second maximum at $k_{max} \sim 0.1 Å^{-1}$, indicating the presence of long-range correlations. To derive the inter-protein potential $v(r)$ we follow the Boltzmann inversion procedure as described in detail in our previous work[27]. Briefly, the structure factor is first converted into the pair-distribution function $g(r)$, which is then used in an iterative fitting procedure to find the corresponding $v(r)$. The experimental structural functions are shown in Figure 1(a) in comparison with their theoretical counterparts. Overall, there is a very good agreement between theory and experiment. The pair distribution function displays some ripples for $r > 8nm$ which, most likely, are an error resulting from the truncation of $S(k)$ at a finite wave vector.

Globally, the generated potential, shown in Figure 1(b), is repulsive but has a small local minimum at $r = 2.6nm$. For further analysis we split the potential into a short-range, $v_{sr}(r)$, and a long-range, $v_{lr}(r)$, components, $v(r) = v_{sr}(r) + v_{lr}(r)$. The long-range part is approximated by the electrostatic solvation energy of a charged colloid computed within the framework of the one-component model[40] $v_{lr}(r) = \gamma v_{ocm}(r)$, where $v_{ocm}(r) = k_B T L_B Z_0^2 \chi^2 \frac{e^{-\kappa r}}{r}$ and γ is a coefficient introduced by us to correct for the errors in the model. The protein's density ρ, its charge Z_0, and the counter-ion density ρ_s can be used to compute other quantities involved in this expression, including : a) the Bjerrum's length $L_B = \frac{e^2}{4\pi\epsilon_0\epsilon(T)} \frac{1}{k_B T}$, where e is the electron charge, ϵ_0 is the dielectric permittivity of vacuum, $\epsilon(T)$ is the

dielectric constant of water and T is the temperature, b) the screening constant $\kappa = \sqrt{4\pi L_B(\rho|Z_0| + 2\rho_s)}$, and c) a scaling constant χ that can be computed numerically[40] from a, the presumed radius of the protein and $\phi = \frac{4}{3}\pi a^3\rho$, the protein volume fraction. With the following values adopted for the model's parameters: $Z_0 = 17$, $\rho_s = 6$ mM[28], $a = 1.7nm$[9], $T = 298K$ and ϵ=78.5[41] we get $L_B = 0.713nm$, $\kappa = 0.837nm^{-1}$and $\chi^2 = 2.98$. The correction coefficient γ is determined from the condition that $v_{lr}(r)$ reproduces the tail of the full potential $v(r)$ in the long- r range. This means that the short-range part in the concerned range is zero, which is a reasonable assumption. For γ=0.35 the model and the actual potential match at $r_m = 0.42nm$, as shown in Figure 1(b). For larger distances the two potentials are in good agreement. The fitted potential displays some ripples which most likely are unphysical. Its long-range region is highlighted in the inset of Figure 1(b), which plots $rv(r)$ and thus should exhibit an exponential decay. Indeed this is what happens over the range of distances from 4 to 7nm. The decay constant $0.78nm^{-1}$ agrees remarkably well with the value of $0.837nm^{-1}$ predicted by the theory. Further evidence of the high accuracy of the OCM theory is that it overestimates (since γ<1) the strength of the repulsive potential only by about a factor of 2.

Temperature-driven structural transformation

The short-range part of the potential acting between charged colloids solvated in water is the sum of two terms: dispersion and hydrophobic interactions. The first term is temperature independent. The second term does depend on temperature. However, if solvation of small hydrocarbons is used as a phenomenological model[42], the dependence is strong around $T = 373K$ but moderates significantly below $T = 298K$. Theoretical calculations seem to support this assessment. Numerical estimate of the surface tension, for instance, obtained by Huang and Chandler[43] for hard-sphere solutes in water varies only by about 2% when the temperature changes between 277K and 298K. As a consequence, the associated change of the hydrophobic solvation energy can be considered small.

The electrostatic long-range potential, on the other hand, appears to be more sensitive to temperature variations when *T* is in the room temperature range. If at $T = 298K$ the dielectric constant, which describes how strongly charge-charge interactions are screened, is 78 then at $T = 277K$ it increases to 87. As a result, the electrostatic repulsion is expected to weaken by more than 10%.

Taking these arguments into consideration we retain temperature dependence only in the long-range part of the potential $v_{lr}(r;T)$ while the short-range part, $v_{sr}(r)$, is treated as temperature independent. The short-range part then can be evaluated as $v_{sr}(r) = v(r;T_{MD}) - v_{lr}(r;T_{MD})$, where $v(r;T_{MD})$ is the potential extracted from SANS data while $v_{lr}(r;T_{MD})$ is the model long-range potential. Both potentials are computed for some reference temperature T_{MD}. With these notations, the full potential can be computed for any temperature T as $v(r;T) = v(r;T_{MD}) - v_{lr}(r;T_{MD}) + v_{lr}(r;T)$. We used this formula together with the simulation data for $T_{MD} = 298K$ to compute potentials for two other temperatures: $T = 278$ and $273K$. The results are shown in Figure 2(a). Like for $T = 298K$, these potentials have a local minimum at short distances. As the temperature is decreased the position of the minimum shifts to the left while its depth increases. As anticipated, this is the consequence of the weakening electrostatic repulsion at lower temperatures[41]. Note also that this trend would be further enhanced by the temperature dependence in the hydrophobic forces[44], which was neglected in the present model. How justifiable are the assumptions made in the derivation of the model can be tested directly in SANS experiments at varying *T*.

The position of the cluster maximum k_{max} in the structure factor evaluated for all three temperatures is shown in Figure 2(b) as a function of the protein concentration c, calculated from the numerical density as $c[mg/ml] = 2.37 \cdot 10^7 \rho[\text{Å}^{-3}]$. As the concentration is increased, k_{max} rises rapidly until certain transition point c_T, after which a flat plateau follows. While this shape is consistent for all studied temperatures, its details are specific for each *T*. In particular, the transition concentration is

close to 150 mg/ml for $T = 298K$ but declines rapidly for lower temperatures. For $T = 273K$, c_T drops to 50 mg/ml, making the maximum-position curve flat in a very wide range of concentrations.

While the concentration data for k_{max} measured under the same experimental conditions as those for which the inter-protein potential was derived[28] are not available, comparison can be made with the experimental data of Cardinaux et al[9], obtained under different pH and salt concentration. Remarkably, these experiments report exactly the same behavior for k_{max}as our simulations, as can be seen from Figure 2(b). There is an upward shift in both theoretical and experimental curves as the temperature is increased, while the point at which plateau sets in is moving to a higher concentration. Numerically, k_{max} observed in the two curves differ by 0.02Å^{-1}, which is the consequence of different experimental conditions under which the two data sets were obtained [1, 2, 9]. Importantly, the experimentally observed concentration dependence of k_{max}, and its behavior with temperature, are correctly reproduced by the theory.

Evidence for clusters

Clustering analysis was conducted using a criterion according to which the given particle belongs to a cluster if it is separated from it by a distance $r < r_b$, where $r_b = 3.6\ nm$ and corresponds to the position of the barrier in the inter-protein potential. The results, shown in Figure 3 for $T = 273K$, demonstrate that the studied system makes clusters at increasing protein concentration. Panel a) displays fraction of particles $P(s)$ belonging to a cluster of size s computed at varying protein concentration. While at low c mostly monomers are observed, their population is seen to decline as the concentration is increased. Concomitantly, clusters of larger sizes, dimers, trimers and so on, begin to appear. Panel b) displays population of the four smallest clusters over a range of concentrations. It is seen that at $c = 100\ mg/ml$, the percentage of monomers drops below 50%. This concentration can be regarded as the transition point into the cluster fluid state. At $> 125mg/ml$, there are more particles engaged in

dimers than in monomers. Further concentration increases result in further reduction of monomers in favor of clusters. Note that the appearance of clusters can not explain the observed concentration trends in k_{max}. Indeed, at $T = 273K$ the plateau in $k_{max}(c)$ begins at $c = 50mg/ml$. At that point, the system is comprised mostly of monomers with a small percentage of dimers while a significant population of clusters is not observed until $c = 100\ mg/ml$. Therefore, there must be another mechanism responsible for the shifting patterns of k_{max}.

To learn more about this mechanism, it is instructive to consider a potential without a minimum, as shown in Figure 1(b).We computed structural functions for this potential using the same protocols as for the original system. Figure 2(b) shows the obtained k_{max} at $T = 273K$. At high concentrations, there is no sign of leveling off. Instead, after starting at low c from values very close to those of the original potential, k_{max} continues to grow uncontrollably as the concentration is increased. In real space, this behavior indicates a shrinking correlation length, in agreement with the predictions made earlier for systems interacting via purely repulsive potentials[2, 9]. This structural transformation can be visualized by projecting the inter-protein distances onto a plot of potential energy. Figure 4(a) shows one such visualization along with a cartoon illustrating anticipated structural changes. Since the interaction among particles is repulsive, the system tends to maximize its inter-particle distances. When binned into a distribution these distances are centered around a certain average value. Each pair of particles makes a contribution to the total potential energy of the system. As the concentration is increased, the system undergoes a uniform compression, driven by the need to minimize the potential energy. As a result, the average inter-protein distance shifts to the left while k_{max} shifts to the right; the average energy of a pair of particles goes up together with the total potential energy. Further increase of density does not bring about any new behavior: the distances still continue to shrink while pushing the particles further uphill on the potential energy surface.

In the case of the potential with a local minimum, there are no clusters (all distances are greater than r_b, the position of the barrier) in the infinite dilution limit. At finite densities clusters begin to form and, as Figure 3(a) demonstrates, this happens for a very low c. Note that the potential energy minimum is still attained when the inter-particle distances are at their maximum. Consequently, the formation of clusters necessarily costs potential energy. In the analysis of a potential with a local minimum similar to that studied here[45], we explained that in the limit of low *c* clusters in such systems are stabilized by entropy. This is the consequence of the volume of the configuration space corresponding to cluster states being greater than the volume of the space available to the monomeric states. An illustration of what happens at finite but low *c* is shown in Figure 4(b), top row. Since the cluster population is low, $r > r_b$ for the majority of particles. The system responds to an increase in density in two ways. First, the number of particles belonging to a cluster increases. This is evident from Figure 3(a) and can be explained in terms of the thermodynamic balance shifting away from monomers and toward clusters due to the loss of entropy by the former. The percentage of particles affected by this process is still low, however. Second, the particles with $r > r_b$ undergo a uniform compression as they try to minimize the potential energy. This results in the shift of the average distance to the left, see Figure 4(b), middle row, just as for the model without the local minimum. Also in common with that model, there will be a shift in k_{max} to the right. As the average distance continues to shrink, the average energy of a pair of particles continues to rise. At some density, it becomes beneficial for the particles to cross over the barrier into the local minimum instead of continuing moving uphill on the potential energy surface, see Figure 4(b), bottom row. At that point, the assembly of clusters becomes driven by the potential energy, as observed earlier[45], while the shrinking of the average distance stops, since it leads to a higher potential energy compared to that offered by the clustering route. It follows from this analysis that the concentration c_m at which assembly of clusters changes its mechanism should be equal to c_T, the density at which the shifting in $k_{max}(c)$ stops. A key prediction of this model is that for $c > c_m$, the potential energy

produced by the potential with the local minimum should be lower than that of the potential without the minimum. The results of a direct test of this prediction in simulations are shown in Figure 4(c) for $T = 273$ and $298K$. The potential with the minimum indeed begins to yield lower energy starting at $c = 150\ mg/ml$ for $T = 298K$ and at $c = 50\ mg/ml$ for $T = 273K$, in full agreement with Figure 2(b).

Cluster statistics

In the proposed model k_{max} corresponds to the distance between cluster edges, not their centers as suggested earlier[1]. Since that distance does not change for $c > c_T$, a pertinent question to ask is this: "How can clusters accommodate ever more particles at the growing density, while keeping a constant distance to their neighbors? " The only reasonable answer is that they must shrink. A detailed analysis of the clusters' dimensions, measured by the radius of gyration Rg, is presented in Figure 5(a) for $T = 273K$ and varying cluster size s. Scaling properties are extracted by fitting the simulation data to the function $f(s) = a(s-1)^b$, where s is the cluster size, b is the scaling exponent and a is the gyration radius of the dimer. It is instructive to draw comparisons with the statistics of polymers[46] in order to better understand the scaling properties of clusters. Polymers with constitutive units that are able to penetrate each other experience an entropic collapse, which leads to the so-called ideal chain statistics. The size of ideal-chain globules scales with the exponent *b*=0.5. Polymer units with finite size experience repulsion, which causes the chain to increase its size as it performs a self-avoiding walk. The corresponding scaling constant, known as the Flory exponent, has been evaluated at *b*=0.588[46]. Adding sufficiently strong attraction to the interaction energy triggers a collapse of the polymer chain into the minimum-size conformations, permitted by the excluded volume of the monomers. The size of the resulting maximally compact globule scales with the exponent *b*=0.33.

We find that at low concentrations the scaling constant is larger than the Flory exponent. At $c = 49.5\ mg/ml$, for instance, $b = 0.68$ and the typical cluster observed at this concentration has an

expanded shape. As an illustration, Figure 5(b) shows a pentamer which has the radius of gyration 3.0nm. In agreement with our prediction, the clusters undergo collapse, while the scaling constant goes down, as the density of the solution is increased. At $c = 346\ mg/ml$ the size of small clusters drops by a factor of 2, as data in Figure 5(b) for the pentamer illustrate. Simultaneously, the scaling constant b declines to 0.36, which is close to the exponent of the maximally compact object 0.33. Interestingly, large clusters at high concentrations obey a different statistics from the small ones. Figure 5(a) shows that for $s > 20$ the simulation data can be well reproduced by the following function $Rg(s) = 1.5(s-1)^{0.5}$. The two scaling regimes in $Rg(s)$ are separated by an inflection point at some intermediate s_i, which, as Figure 5(a) shows, depends on protein concentration. The higher the concentration, the larger the size s_i.The emergence of two different scaling laws, defined by one small and one large exponent b, suggests a hierarchical organization of the clusters, in which small clusters are used as building blocks for the assembly of the larger clusters. This model is confirmed by the visual inspection of the trajectories. Figure 5(b) shows a decamer observed in our simulations at $c = 346\ mg/ml$. Its radius of gyration $Rg = 3.5nm$ is much greater than the value of 2.5nm expected according to the small-s scaling function, see extrapolation data in Figure 5(a) for $s = 10$ and $= 346\ mg/ml$, indicating an expanded shape. It is easy to see that this cluster is built from two small pentamers joined together. This model of cluster organization was reported by us earlier [45]. Its key feature - the large-cluster exponent of 0.5, that corresponds to the ideal-chain statistics and implies no excluded volume interactions, can be explained by the ability of clusters to "pass through" each other via the exchange of particles mechanism.

Nature of the cluster fluid

The initial solution with individually dispersed proteins (monomers) effectively turns into a cluster fluid at large c, where the population of monomers is low. The nature of this fluid can be gleaned from the

cluster distribution function $P(s)$, shown in Figure 6. At $c = 178\ mg/ml$, $P(s)$ has a broad shape indicating the presence of clusters in a wide range of sizes, $2 < s < 20$, with a maximum observed for dimers, $s = 2$. An increase in the concentration has two effects. First, the distribution becomes more sharply peaked. The distribution half-width Δs, defined as $\Delta s = s_o - s_{max}$, where s_{max} is the position of the maximum in $P(s)$ and s_o is the size, where the distribution falls 50% off of its maximum, $P(s_o) = 0.5P(s_{max})$, is seen to decline from 2, observed for $c = 178\ mg/ml$, to 1.7, observed for $c = 456\ mg/ml$. Second, the distribution maximum s_{max} shifts to higher values: while s_{max}=2 for $c = 178\ mg/ml$ it increases to 6 for $c = 456\ mg/ml$. An additional feature emerges at c_s~$300\ mg/ml$, see Figure 6, where a second maximum (first as a shoulder) appears in the distribution function, located at the position double that of the main maximum. For $c = 291.6\ mg/ml$ the maximum is at $s = 8$, while for $c = 456.1\ mg/ml$ - at $s = 12$. Note that, according to the analysis of the radius of gyration, small clusters begin to self-associate into larger clusters starting at a much lower concentration. Indeed Figure 5(a) shows that two scaling regimes are present in $Rg(s)$ starting from $c = 178\ mg/ml$ onwards. So c_s must signal a different transformation. What happens specifically at this concentration is that, out of the sea of clusters of different sizes, a cluster of preferred size emerges. The specific cluster, and its assemblies, begin to dominate the ensemble, in a sign of the structural phase transition.

Structural relaxation

Structural dynamics can be probed directly by neutron spin-echo (NSE) scattering, a technique that measures intermediate scattering function $F(k,t)$, where k is the wave vector and t is time. For the purpose of the cluster discussion, we are interested in the short-time dynamics, where the following approximation applies $F(k,t) = S(k)e^{-k^2 D_s(k)t}$. Here $D_s(k)$ is a transport quantity that has the meaning of generalized diffusion coefficient for density fluctuations at wave vector k and can be

extracted from experimental data by fitting. Dimensionless quantity $D_0/D_s(k)$, the inverse of this coefficient multiplied by the diffusion constant in the free diffusion limit, D_0, can be used as a generalized relaxation time; for lysozyme it has a minimum[9] at the location where a maximum in $S(k)$ is seen, the so-called de Gennes narrowing[47]. At $k = k_{max}$, $D_0/D_s(k_{max})$ undergoes a rapid growth with protein concentration that can be attributed to a kinetic phase transition[9]. Interestingly, earlier theoretical models fail to predict this transition[9], finding only a moderate increase in the relaxation time. We used the same stochastic dynamics (SD) simulations as for the structural studies, see "Methods" section, to compute $F(k_{max}, t)$ for $k_{max} = 0.118\text{Å}^{-1}$ and $T = 278K$. Fitting over the time domain where clear exponential decay is observed permitted the calculation of $D_s(k_{max})$. We applied the same analysis to the experimental data[9], to make that sure that comparison between theory and experiment is consistent. Figure 7(a) shows our results, scaled to match the experiment at $c = 50mg/ml$, in comparison with the experimental data. The simulations significantly overestimate the rate of structural relaxation. For $c > 200mg/ml$, the theoretical $D_0/D_s(k_{max})$ is more than twice smaller than its experimental counterpart. In agreement with prior work[9], we see here that direct protein-protein interactions are not sufficient to provide a proper description for the slowdown of structural relaxation taking place in the simulated system at increasing concentration.

If not direct then, perhaps, it is interactions mediated by the solvent that are missing? Note that solvent is present implicitly in stochastic dynamics through friction force designed to keep the temperature constant. It is assumed to be the same for all particles and independent of their positions or velocities. This approximation ignores viscous forces created by solvent flows due to the movements of particles with respect to one another, or the so-called hydrodynamic interactions[48]. Earlier studies of colloidal suspensions indicate that hydrodynamic interactions may cause slow dynamics[49]. To investigate their effect in the context of lysozyme solutions we employed the dissipative particle dynamics (DPD) method, as discussed in detail in the "Methods" section. The method contains one free parameter f,

which controls the strength of the friction force. We varied that parameter systematically to determine its influence on structural relaxation. Two main effects were revealed in these studies: a) $D_0/D_s(k_{max})$ curve grows steeper with f, and b) it becomes strongly non-linear. By treating f as an adjustable parameter, it is possible to achieve a very good agreement between theoretical and experimental $D_0/D_s(k_{max})$. Figure 7(a) shows our DPD data for $= 0.4$. Deviations on the order of 0.1 from this value do not lead to noticeable differences in the dynamics. It is seen that the experimental function can be reproduced very well for $c < 300mg/ml$. At $c{\sim}300mg/ml$ our model slightly underestimates the relaxation time. For higher concentrations, experimental data are missing, where it is concluded[9] that the system undergoes a kinetic arrest. In the same limit our simulations show that $D_0/D_s(k_{max})$ has large but finite values. We note that in an unrelated study[50], viscosity - another property that can detect kinetic transitions, grows rapidly with concentration, but nevertheless lacks a mathematical singularity for $c > 300mg/ml$. So the precise nature of the observed kinetic transition seems to be unclear at the moment.

To learn more about the dynamics in the high- c regime we carried out tests designed to determine whether the system remains ergodic in that limit, on the simulation time scale. Specifically, we computed the mean-square particle displacements $< \Delta\vec{R}^2(t) >$as a function of time and fitted them to the following template $f(t) = \alpha + \delta t^{\beta}$, where t is time and α, δ and β are certain adjustable parameters. When the system is ergodic, the displacement is described by the diffusion law: $\beta = 1$ and $\delta = 6D$, where D is the diffusion coefficient. Fitting has to be done for sufficiently long times as the linear regime is preceded by a short ballistic phase where $< \Delta\vec{R}^2(t) > {\sim}t^2$. The point separating the two types of dynamics, t_d, was determined for $c = 49.5\ mg/ml$. For consistency, fitting was performed over the same time interval for all concentrations. The computed exponent β is shown in Figure 7(b). Five independent simulations were carried out to test whether the results are reproducible. The error bars in the figure were computed from the analysis of these trajectories. It is seen that $\beta = 1$ for small

concentrations, where ergodic (on the simulation time scale) behavior is observed. Statistically meaningful deviations begin at $c_{cr}{\sim}300mg/ml$, which is close to c_s. Since the exponent β is less than unity, it indicates a confinement of particles, similarly to supercooled and glassy liquids, where the cage effect is observed[51].In our case, the steric effects due to the protein size can be ruled out as the source of the confinement. It is known that lysozyme is able to transition into high-density states[52, 53] with $c = 400mg/ml$ and higher and still remain an ergodic fluid. The confinement, therefore, must be due to a different mechanism. In one likely scenario, one can argue that particles are expected to diffuse much more slowly as part of a cluster than when they are in the monomeric state. It is reasonable to expect then that low β values are a direct consequence of cluster formation. There is one problem with this explanation, however. Clusters start forming at $c < 50mg/ml$ and at $c > 100mg/ml$ (see Figure 3(b)) they represent the most populated species in the system. It is not clear then why there is no confinement at $100mg/ml < c < 300mg/ml$, despite preponderance of clusters. Perhaps, it is not the clusters but their dynamics that leads to non-ergodicity? If clusters are allowed to make and break multiple times on the simulation time scale, their constituent particles should exhibit the same dynamics as that of the monomers. Clusters that do not change in the course of the simulations, on the other hand, should experience the slow down. To test this hypothesis we examined the autocorrelation function $\varphi(t) =< N_c(0)N_c(t) >/< N_c^2 >$, where $N_c(t)$ is the number of clusters at time t. The function reports the time scale on which clusters assemble and fall apart. The characteristic time constant, or relaxation time, for this process, τ, was extracted from $\varphi(t)$ by fitting it to an exponential template. The results, plotted in the inset of Figure 7(b), show that τ remains flat for low concentrations but begins to rise sharply at $c{\sim}300mg/ml$, in a clear sign of the kinetic transition. The transition leads to the creation of clusters that are unable to exchange particles among themselves and, thus, can be considered permanent. As specific cluster distributions are unable to equilibrate over the available simulation time, they can be considered frozen, or non-ergodic. As any collective or phase-change

phenomena, the transition is expected to be strongly influenced by the size of the simulation cell. To assess that influence, we repeated our simulations for boxes in which the number of particles was increased more than three-fold, from 512 to 1728. The results, shown in Figure 7(b), demonstrate that the transition becomes much sharper. This trend is likely to continue for larger systems. A careful finite-size analysis is needed to determine the exact transition density. Our estimate based on the present simulations is $c_{cr}{\sim}300mg/ml$. Since at $c > c_{cr}$ the system is non-ergodic, relaxation times in that limit can not be determined reliably. For that reason the relevant non-converged data in Figure 7(a) are shown by different symbols from those used for other data, specifically, full circles with white squares inside.

Whether the hydrodynamic interactions provide the only feasible mechanism for the kinetic slowdown remains to be seen. At least two other models have been discussed in the literature[9]. The first is the Wigner glass, which relies on the repulsive interactions. We mentioned earlier that steric interactions are unlikely to cause kinetic arrest at the studied densities. The same is true for the electrostatic repulsion, as in that case stochastic dynamics simulations, lacking the hydrodynamic forces, would also uncover non-ergodic behavior, which they did not. The other model is attractive glass, in which protein-protein attraction causes slow dynamics. Here again sufficiently strong attraction would reveal itself in stochastic dynamics simulations and it did not. Nevertheless, either model may still apply if the inter-protein potential depends strongly on protein density. That dependence is ignored in the present paper. In the case of repulsive glass, electrostatic interactions may gain strength at high densities, although it is not clear to us at the moment what mechanism could cause this effect. For attractive glass, protein-protein attraction may become stronger at high c. This could happen because of specific density dependence of the hydrophobic interactions. Or, alternatively, strong counter-ion mediated attraction between proteins may arise. Clearly, more research is needed to assess the likelihood of these different scenarios.

Conclusions

In this paper we report the formation of equilibrium clusters in aqueous solutions of protein lysozyme. Proteins are modeled as soft spherical particles for which interaction potential is derived directly from experimental structure factor. The potential is overall repulsive but contains a local minimum at short distances. Computer simulations reveal that this potential leads to the formation of clusters at varying protein concentration c. Clusters are small, mostly dimers and trimers, and only weakly populated at small concentrations. They grow in size as the concentration goes up, while their population is steadily increasing. At some point clusters become more populous than monomers, signaling a structural transition into the cluster-fluid phase.

Structure factor of lysozyme solutions studied in this paper has a secondary maximum at short wave vectors. The specific dependence of the position of this maximum on concentration $k_{max}(c)$ has been used previously[1, 9] to characterize clusters at the quantitative level. This dependence can be correctly reproduced by our simulations, which lead to the following clustering scenario. At small concentrations most of the proteins are found at maximal distances from another one, since such configurations are the lowest-energy states in systems interacting via repulsive potentials. There are only a few clusters in the solution and their assembly is favored by entropy. The wave vector k_{max} corresponds to the average distance between proteins. When the protein concentration is increased the solution undergoes uniform compression. This causes the average distance to shrink, leading to an upward shift in $k_{max}(c)$, in excellent agreement with experiment[1]. As the concentration is increased beyond a certain threshold value $c \geq c_T$ the assembly process becomes driven by enthalpy. It then becomes energetically beneficial for the monomers to join an existing cluster (or make a new one) instead of continuing to approach other proteins at ever shorter distances, as happens in the compression scenario. As a consequence, the

average inter-protein distance stops shrinking while $k_{max}(c)$ stops shifting at high concentrations, again in excellent agreement with experiment[1]. Our simulations produce quantitatively accurate estimates for c_T: $50mg/ml$ at $T = 273K$ and $150mg/ml$ for $T = 298K$.

At low concentrations the radius of gyration of the observed clusters $Rg(s)$ can be well described by a single scaling function. The scaling exponent obtained from fitting indicates that the clusters occupy mostly expanded, or stretched, configurations. As the concentration grows, clusters begin to collapse. For sufficiently large c their scaling exponent approaches 0.33, which is characteristic for maximally compact objects. At the same time, large clusters develop a distinct statistics. Their scaling exponent increases in comparison with the small clusters and approaches 0.5, the ideal-chain value. Two separate regimes in $Rg(s)$ suggest that clusters obey a hierarchical structural model, according to which large clusters are assembled from the small ones as building blocks. At high concentrations cluster fluid is composed of a large variety of small and large clusters. At $c = c_s$ a structural transition is seen into a state in which a cluster of specific size, and its assemblies, begin to dominate the entire ensemble. The transition is manifested in the appearance of multiple peaks in the cluster-size distribution. At $T = 273K$ we estimate that c_s is approximately $300mg/ml$.

In agreement with experiment, the studied system experiences a kinetic phase transition at sufficiently high concentrations. Our simulations find that for $c \geq c_r$, where c_r is a certain critical concentration, clusters are unable to exchange particles among themselves, which leads to the breakdown of ergodicity. The transition is accompanied by a strong growth in the relaxation time of various processes, including density fluctuations. Our simulations find that hydrodynamic interactions are critical for the kinetic slow down. The parameter controlling the strength of these interactions, f, is seen to strongly influence the value of c_r. Greater f's lead to lower c_r's and vice versa. The kinetic transition is not observed when the hydrodynamic interactions are switched off, $f = 0$. When this parameter is

calibrated against experimental data for low protein concentrations, we find that $c_r \approx 300mg/ml$ at $T = 278K$.

Acknowledgement

The authors warmly thank Prof. U. Wanderlingh for his permission to use the neutron diffraction data shown in Figure 1(a). We are also grateful to A. Stelmakh for his comments on the early version of the manuscript. Computational resources of the research cluster managed by the Institute for Condensed Matter Physics, National Academy of Science of Ukraine and the Ukrainian National Grid were used in the completion of this work. W. Cai is supported by US National Science Foundation (grant DMS-1764187).

References

1. A. Stradner, H. Sedgwick, F. Cardinaux, W. C. K. Poon, S. U. Egelhaaf and P. Schurtenberger, Nature **432** (7016), 492-495 (2004).
2. A. Shukla, E. Mylonas, E. Di Cola, S. Finet, P. Timmins, T. Narayanan and D. I. Svergun, Proc. Natl. Acad. Sci. U. S. A. **105** (13), 5075-5080 (2008).
3. A. Shukla, E. Mylonas, E. Di Cola, S. Finet, P. Timmins, T. Narayanan and D. I. Svergun, Proc. Natl. Acad. Sci. U. S. A. **105** (44), e76 (2008).
4. Y. Liu, L. Porcar, J. Chen, W.-R. Chen, P. Falus, A. Faraone, E. Fratini, K. Hong and P. Baglioni, The Journal of Physical Chemistry B **115** (22), 7238-7247 (2011).
5. B. Lonetti, E. Fratini, S. Chen and P. Baglioni, Phys. Chem. Chem. Phys. **6** (7), 1388-1395 (2004).
6. P. D. Godfrin, N. E. Valadez-Pérez, R. Castaneda-Priego, N. J. Wagner and Y. Liu, Soft Matter **10** (28), 5061-5071 (2014).
7. L. Porcar, P. Falus, W.-R. Chen, A. Faraone, E. Fratini, K. Hong, P. Baglioni and Y. Liu, The Journal of Physical Chemistry Letters **1** (1), 126-129 (2010).
8. P. Falus, L. Porcar, E. Fratini, W.-R. Chen, A. Faraone, K. Hong, P. Baglioni and Y. Liu, J. Phys.: Condens. Matter **24** (6), 064114 (2012).
9. F. Cardinaux, E. Zaccarelli, A. Stradner, S. Bucciarelli, B. Farago, S. U. Egelhaaf, F. Sciortino and P. Schurtenberger, The Journal of Physical Chemistry B **115** (22), 7227-7237 (2011).
10. S. Barhoum and A. Yethiraj, The Journal of Physical Chemistry B **114** (51), 17062-17067 (2010).
11. A. Stradner, F. Cardinaux, S. U. Egelhaaf and P. Schurtenberger, Proc. Natl. Acad. Sci. U. S. A. **105** (44), E75; author reply E76 (2008).
12. M. Malfois, F. Bonneté, L. Belloni and A. Tardieu, The Journal of chemical physics **105** (8), 3290-3300 (1996).

13. A. Ducruix, J. P. Guilloteau, M. Riès-Kautt and A. Tardieu, J. Cryst. Growth **168** (1-4), 28-39 (1996).
14. M. Niebuhr and M. H. J. Koch, Biophys. J. **89** (3), 1978-1983 (2005).
15. M. A. Schroer, J. Markgraf, D. F. Wieland, C. J. Sahle, J. Möller, M. Paulus, M. Tolan and R. Winter, Phys. Rev. Lett. **106** (17), 178102 (2011).
16. J. Möller, M. A. Schroer, M. Erlkamp, S. Grobelny, M. Paulus, S. Tiemeyer, F. J. Wirkert, M. Tolan and R. Winter, Biophys. J. **102** (11), 2641-2648 (2012).
17. A. Chinchalikar, V. Aswal, J. Kohlbrecher and A. Wagh, Phys Rev E **87** (6), 062708 (2013).
18. M. C. Abramo, C. Caccamo, M. Cavero, D. Costa, G. Pellicane, R. Ruberto and U. Wanderlingh, J. Chem. Phys. **139** (5), 054904 (2013).
19. T. Sumi, H. Imamura, T. Morita, Y. Isogai and K. Nishikawa, Phys. Chem. Chem. Phys. **16** (46), 25492-25497 (2014).
20. P. Kowalczyk, A. Ciach, P. Gauden and A. Terzyk, J. Colloid Interface Sci. **363** (2), 579-584 (2011).
21. F. Cardinaux, A. Stradner, P. Schurtenberger, F. Sciortino and E. Zaccarelli, EPL (Europhysics Letters) **77** (4), 48004 (2007).
22. D. Soraruf, F. Roosen-Runge, M. Grimaldo, F. Zanini, R. Schweins, T. Seydel, F. Zhang, R. Roth, M. Oettel and F. Schreiber, Soft Matter **10** (6), 894-902 (2014).
23. Y. Yan, D. Seeman, B. Zheng, E. Kizilay, Y. Xu and P. L. Dubin, Langmuir **29** (14), 4584-4593 (2013).
24. O. Galkin, W. Pan, L. Filobelo, R. E. Hirsch, R. L. Nagel and P. G. Vekilov, Biophys. J. **93** (3), 902-913 (2007).
25. W. Pan, O. Galkin, L. Filobelo, R. L. Nagel and P. G. Vekilov, Biophys. J. **92** (1), 267-277 (2007).
26. K. P. Johnston, J. A. Maynard, T. M. Truskett, A. U. Borwankar, M. A. Miller, B. K. Wilson, A. K. Dinin, T. A. Khan and K. J. Kaczorowski, ACS Nano **6** (2), 1357-1369 (2012).
27. A. Baumketner, R. Melnyk, M. F. Holovko, W. Cai, D. Costa and C. Caccamo, J. Chem. Phys. **144** (1), 015103 (2016).
28. M. C. Abramo, C. Caccamo, D. Costa, G. Pellicane, R. Ruberto and U. Wanderlingh, J. Chem. Phys. **136** (3), 035103 (2012).
29. R. L. Henderson, Phys. Lett. A **49** (3), 197-198 (1974).
30. W. Schommers, Phys. Lett. A **43** (2), 157-158 (1973).
31. W. Schommers, Phys. Rev. A **28** (6), 3599-3605 (1983).
32. A. K. Soper, Chem. Phys. **202** (2-3), 295-306 (1996).
33. D. Van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen, J. Comput. Chem. **26** (16), 1701-1718 (2005).
34. B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl, J. Chem. Theory Comput. **4** (3), 435-447 (2008).
35. D. Frenkel and B. Smit, *Understanding molecular simulation*. (Academic Press, San Diego, 2002).
36. I. Pagonabarraga, M. Hagen and D. Frenkel, EPL (Europhysics Letters) **42** (4), 377 (1998).
37. C. Junghans, M. Praprotnik and K. Kremer, Soft Matter **4** (1), 156-161 (2008).
38. N. Goga, A. Rzepiela, A. De Vries, S. Marrink and H. Berendsen, J. Chem. Theory Comput. **8** (10), 3637-3649 (2012).
39. A. Stradner, F. Cardinaux and P. Schurtenberger, The Journal of Physical Chemistry B **110** (42), 21222-21231 (2006).
40. L. Belloni, The Journal of Chemical Physics **85** (1), 519-526 (1986).
41. D. P. Fernández, Y. Mulev, A. Goodwin and J. L. Sengers, J. Phys. Chem. Ref. Data **24** (1), 33-70 (1995).
42. R. L. Baldwin, Proceedings of the National Academy of Sciences **83** (21), 8069-8072 (1986).

43. D. M. Huang and D. Chandler, Proceedings of the National Academy of Sciences **97** (15), 8324-8327 (2000).
44. D. M. Huang and D. Chandler, Proc. Natl. Acad. Sci. U. S. A. **97** (15), 8324-8327 (2000).
45. A. Baumketner and W. Cai, Condensed Matter Physics **19** (1), 13605 (2016).
46. P. G. De Gennes, *Scaling Concepts in Polymer Physics*. (Cornell University Press, New York, 1979).
47. J.-P. Hansen and I. R. McDonald, in *Theory of Simple Liquids (Third Edition)* (Academic Press, Burlington, 2006), pp. 255-290.
48. M. P. Allen and D. J. Tildesley, *Computer simulations of liquids*. (Oxford University Press, Oxford, 1987).
49. J. Riest and G. Nägele, Soft Matter **11** (48), 9273-9280 (2015).
50. P. D. Godfrin, S. D. Hudson, K. Hong, L. Porcar, P. Falus, N. J. Wagner and Y. Liu, Phys. Rev. Lett. **115** (22), 228302 (2015).
51. W. Gotze and L. Sjogren, Reports on Progress in Physics **55** (3), 241 (1992).
52. C. Ishimoto and T. Tanaka, Phys. Rev. Lett. **39** (8), 474 (1977).
53. M. Kastelic, Y. V. Kalyuzhnyi, B. Hribar-Lee, K. A. Dill and V. Vlachy, Proceedings of the National Academy of Sciences **112** (21), 6766-6770 (2015).

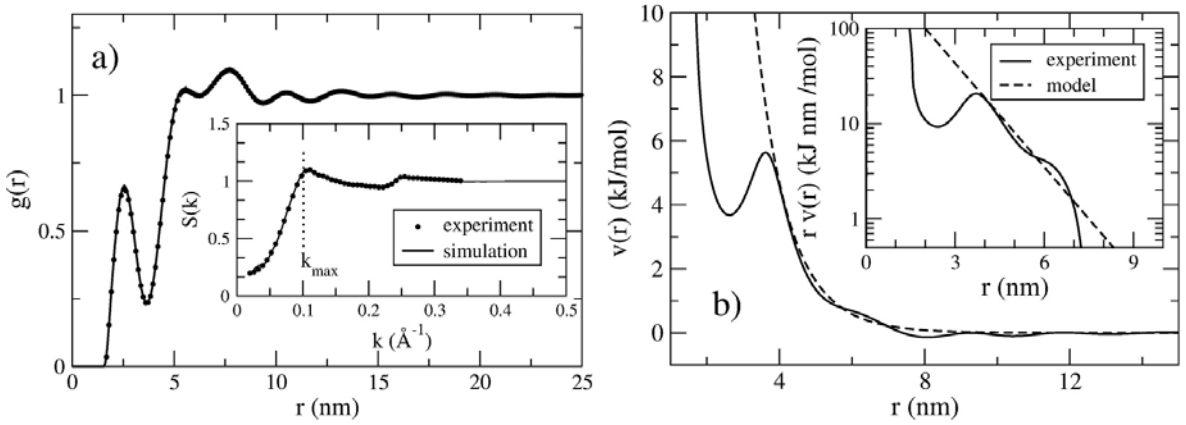

Figure 1 Panel a) shows structural data for lysozyme solutions: pair distribution function, main figure, and the static structure factor, the inset. Both experimental[28] and theoretical data are shown. Panel b): inter-protein potential obtained for T=298K. For comparison, a model potential is also shown, see main text for details. The inset shows $rv(r)$ which highlights the exponential decay of the potential

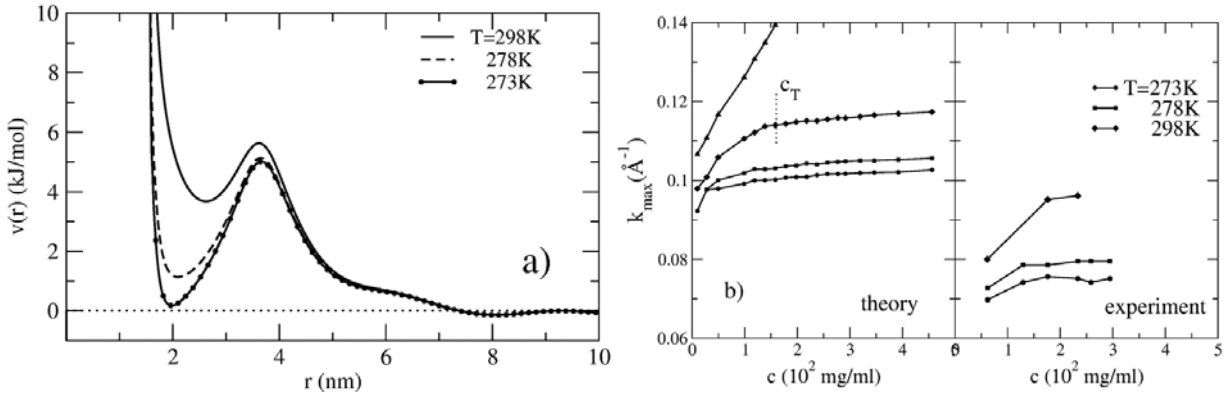

Figure 2 Panel a): inter-protein potentials obtained for lysozyme at three temperatures, T=298, 278 and 273K. Panel b): position of the cluster maximum k_{max} computed for the chosen temperatures as a function of protein concentration c. Data for the purely repulsive model potential from Figure 1(b) are shown by filled triangles. Experimental data of Cardinaux at al[9] are shown for comparison.

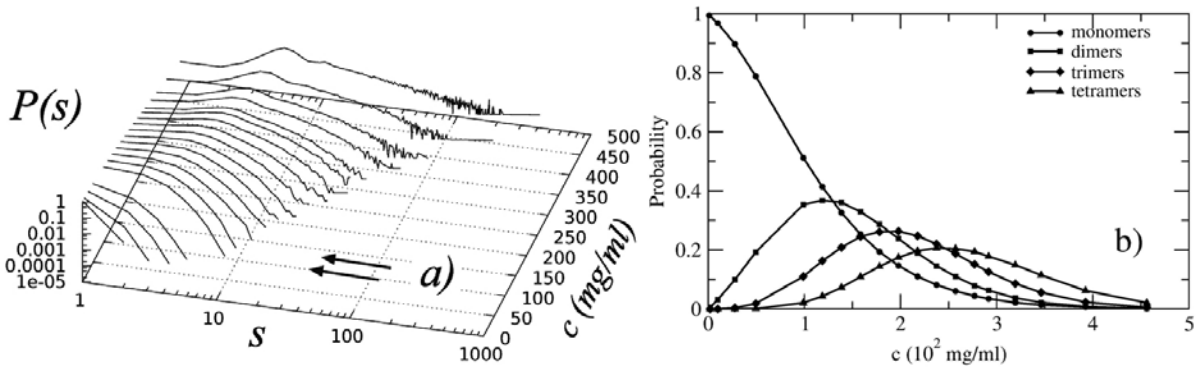

Figure 3 Results of the clustering analysis for T=273K. Panel a) shows the fraction of particles *P(s)* involved in clusters of size *s* for varying concentration *c*. Arrows indicate concentrations at which, first, the population of the monomers drops below 50% and then, second, the distribution develops a non-monomeric peak. Population of four smallest clusters over a concentration range is shown in panel b).

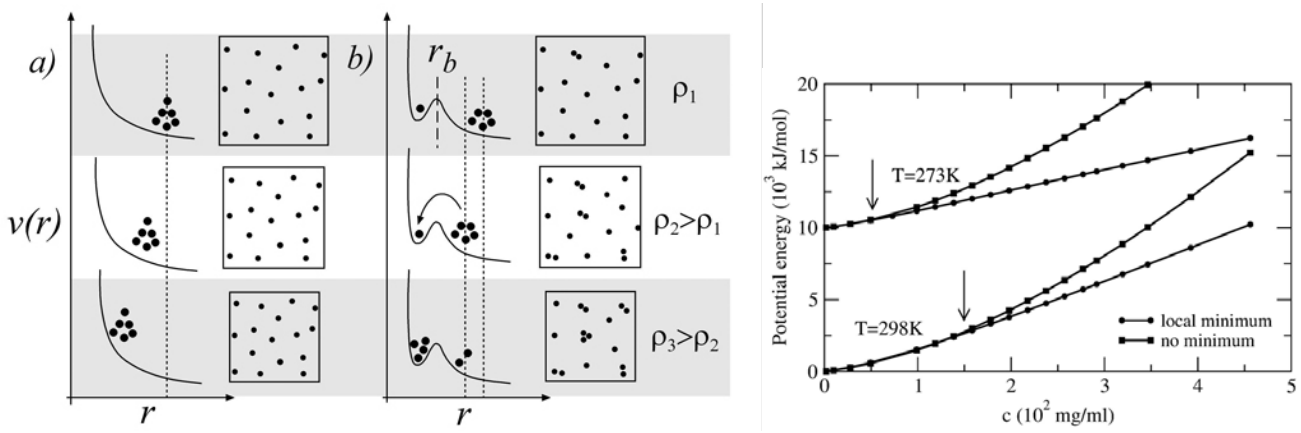

Figure 4 Graphical illustration of structural changes taking place in the system interacting via a purely repulsive potential, a), and a potential that contains a local minimum, b). Dotted lines represent average distances between particles. Panel c) shows potential energy computed for the two systems at two temperatures. The data for the lower temperature are shifted for better readability.

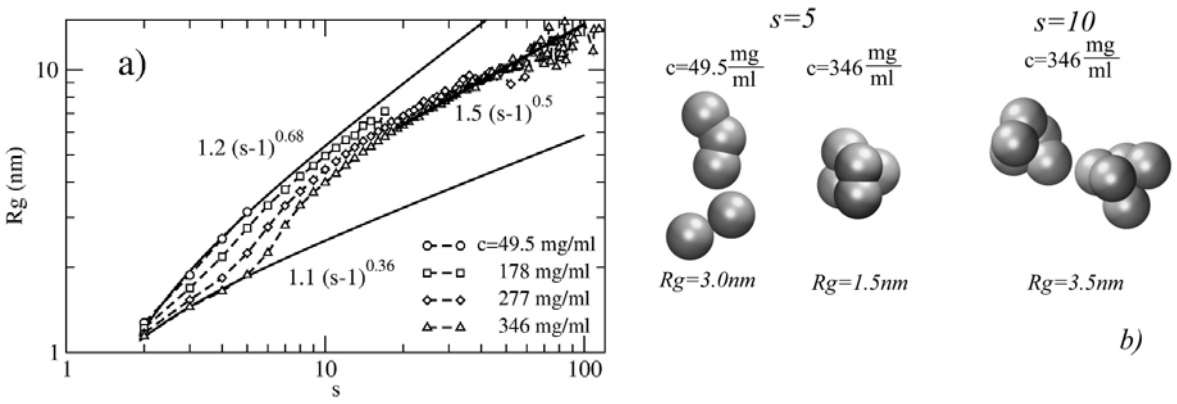

Figure 5 Panel a): Statistics of clusters at varying concentration and $T = 273K$. The radius of gyration is shown as a function of the cluster size. Broken lines with symbols denote simulation data for varying concentration *c*. Solid lines display scaling functions obtained for the simulation data by fitting. Two concentrations are considered: 49.5mg/ml and 346mg/ml. Panel b): illustration of clusters with different shapes and sizes observed in our simulations at varying protein concentration. A pentamer with *Rg*=3.0nm is shown for *c*=49.5mg/ml. A smaller pentamer with *Rg*=1.5nm is observed at a higher *c*=346mg/ml. A decamer at the same concentration is seen to consist of two small pentamers joined together.

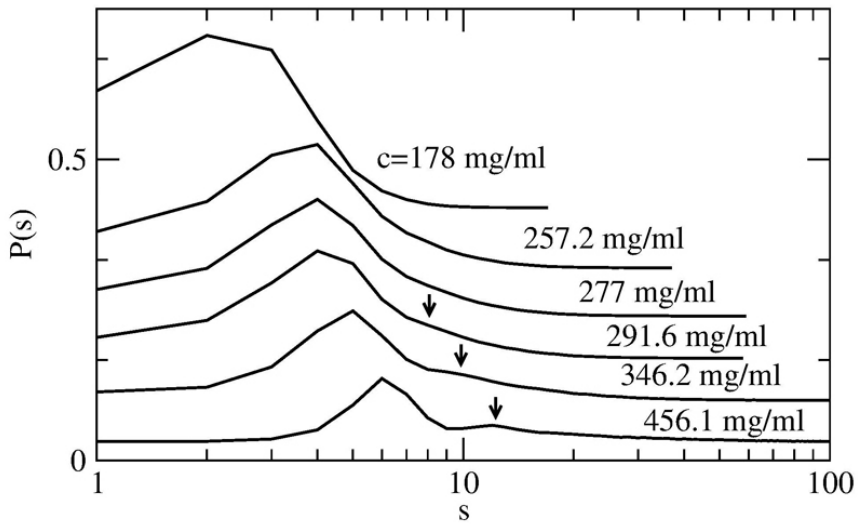

Figure 6 Cluster-size distribution $P(s)$, shifted appropriately for better readability, for varying concentration and $T = 273K$. Arrows indicate the position of the side peak/shoulder in the distribution, signaling cluster-cluster association.

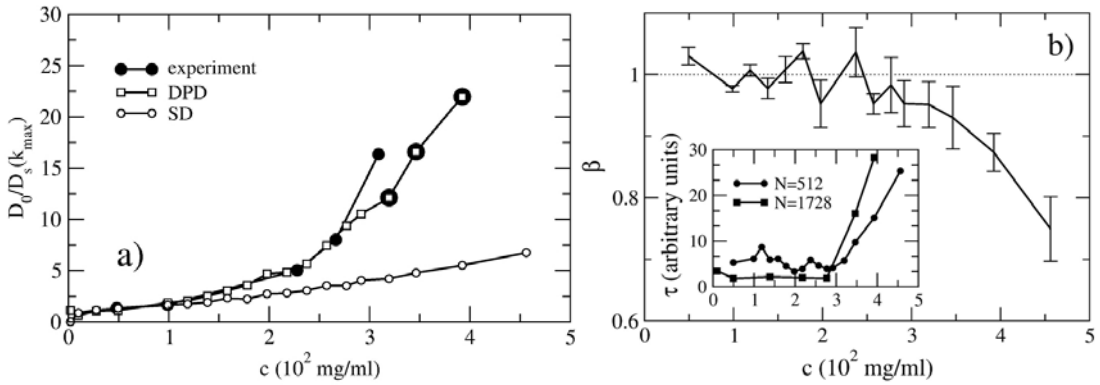

Figure 7 Panel a): Quantity $D_0/D_s(k_{max})$ characterizing the rate of structural relaxation at wave vector k_{max}, where D_0 is the diffusion coefficient in the free diffusion limit and $D_s(k_{max})$ is the generalized diffusion coefficient, see main text for definition. Experimental data[9] and the results of two simulation methods, DPD and SD, are shown. The parameter controlling the strength of the hydrodynamic interactions in the DPD simulations was calibrated against experimental data. Data points obtained from non-converged simulations are shown by full circles with white squares inside. Panel b): the exponent extracted from the mean squared displacement as a function of time. The inset plots the relaxation time extracted from the time auto-correlation function of the total number of clusters. Data for two system sizes are shown. All data are shown for $T = 278K$.