5-19-2017

Preferential Binding Effects On Protein Structure and Dynamics Revealed by Coarse-Grained Monte Carlo Simulation

Ras B. Pandey
University of Southern Mississippi, ras.pandey@usm.edu

D.L. Jacobs
University of North Carolina

Barry L. Farmer
Air Force Research Laboratory

Preferential binding effects on protein structure and dynamics revealed by coarse-grained Monte Carlo simulation

R. B. Pandey,[1] D. J. Jacobs,[2] and B. L. Farmer[3]
[1]*Department of Physics and Astronomy, University of Southern Mississippi, Hattiesburg, Mississippi 39406, USA*
[2]*Department of Physics and Optical Science, University of North Carolina, Charlotte, North Carolina 28223, USA*
[3]*Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright Patterson Air Force Base, Ohio 45433, USA and Materials Science and Engineering, North Carolina State University, Raleigh, North Carolina 27606, USA*

(Received 29 January 2017; accepted 26 April 2017; published online 19 May 2017)

The effect of preferential binding of solute molecules within an aqueous solution on the structure and dynamics of the histone H3.1 protein is examined by a coarse-grained Monte Carlo simulation. The knowledge-based residue-residue and hydropathy-index-based residue-solvent interactions are used as input to analyze a number of local and global physical quantities as a function of the residue-solvent interaction strength (f). Results from simulations that treat the aqueous solution as a homogeneous effective solvent medium are compared to when positional fluctuations of the solute molecules are explicitly considered. While the radius of gyration (R_g) of the protein exhibits a non-monotonic dependence on solvent interaction over a wide range of f within an effective medium, an abrupt collapse in R_g occurs in a narrow range of f when solute molecules rapidly bind to a preferential set of sites on the protein. The structure factor $S(q)$ of the protein with wave vector (q) becomes oscillatory in the collapsed state, which reflects segmental correlations caused by spatial fluctuations in solute-protein binding. Spatial fluctuations in solute binding also modify the effective dimension (D) of the protein in fibrous ($D \sim 1.3$), random-coil ($D \sim 1.75$), and globular ($D \sim 3$) conformational ensembles as the interaction strength increases, which differ from an effective medium with respect to the magnitude of D and the length scale. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4983222]

I. INTRODUCTION

Within aqueous cellular environments, living organisms exquisitely maintain suitable levels of composition of solute molecules for proteins to function properly.[1] The composition of ions, osmolytes, and/or other crowding agents within a cell is important to the conformational dynamics of a protein and how it functions.[2] In simpler *in vitro* scenarios, the dynamics and stability of proteins are sensitive to the type and concentration of solute in single component aqueous solutions.[3–5] Despite heterogeneity in protein polymers and the large chemical space of possible solute molecules, thermodynamic descriptions of preferential binding[6] provide a simple explanation for why solute molecules favorably interact with specific protein sites, while having a tendency to avoid certain segmental regions. Unfortunately, structural details of preferential binding are not adequately addressed within thermodynamic models. Although more detailed statistical mechanics treatments are available, such as Kirkward-Buff theory,[7,8] these approaches are also limited by the need to have an accurate ensemble of protein conformations. Therefore, molecular simulation is necessary to generate thermodynamic ensembles, and to investigate how the characteristics of protein structure and dynamics depend on the nature of the interaction of the solute molecules with the protein.

Experimentally, concentration is a natural control variable for a one component aqueous solution. Furthermore, there are many types of solute molecules that can be considered, classified broadly as denaturants or as stabilizers.[6] Computationally, the nature of the solute molecule at a fixed concentration of solute can be continuously adjusted. Many molecular dynamic models are available to investigate how residue-solvent interaction affects the structure and dynamics of a protein. In particular, solvent can be explicitly[9–12] or implicitly[13–19] modeled as the dynamics of a protein are simulated as well as various levels of coarse graining can be employed. The GROMACS package[20] is an example of a software suite that provides many different options for users to run molecular dynamics at various accuracy levels of description. Explicit solvent models track all solvent degrees of freedom, whereas implicit solvent models embed a protein into a continuum medium where the strength of protein-solvent interaction is adjusted through an effective force field (or interaction energy). The choice of model and method to employ depends largely on the questions being asked. Here, we are interested in the salient features of how structure and the long-time dynamics of proteins depend on the role of preferential binding.

While conformational ensembles of proteins derived from simulations employing explicit or implicit solvent models agree on general trends,[20] the resulting free energy landscapes will differ.[21] However, differences also result due to different force fields within explicit solvent models.[22] Improvement of force fields with explicit and implicit solvent models continues to be an active area of research. Moreover, extensive

conformational sampling is required to extract thermodynamic properties and functional mechanisms, regardless of the underlying molecular mechanics model. To reduce computational costs, multiscale modeling is often employed to improve conformational sampling,[23,24] including replica exchange methods that merge implicit and explicit models together.[21] Other methods account for cosolvents in aqueous solutions by reweighting conformations that are generated in explicit water simulations.[25]

Implicit solvent models have been successful in uncovering conformational pathways critical to the function of a protein[26] and the formation of protein complexes.[27] Unfortunately, with an all-atom description of a protein, implicit solvent models often do not substantially improve performance over explicit solvent models.[28] As such, coarse-grained bead models remain popular.[12,29] Similarly, a bond fluctuation Monte Carlo method that was originally developed to simulate polymer melts[30] provides an alternative approach for proteins. To that end, a novel coarse-grained implicit solvent model was recently developed for globular proteins that efficiently explores protein conformations[31] and extracts long-time dynamics and thermodynamic properties. Clearly it would be computationally desirable to model aqueous solutions implicitly due to the substantial reduction in degrees of freedom that occurs. This approach is formally working with a potential of mean force where the configuration of solvent molecules and cosolvent molecules is integrated out for a specified set of coordinates for the protein, usually called solute. All implicit models rely on this concept of potential of mean force, and they assume that the effective interaction between pairs of particles is transferable. When the solvent has a large number of degrees of freedom, such as water being a large component in the system, this approach works well and forms the basis of all implicit models, which have been proven to work well, albeit an approximate theory.[13–19]

Although modeling solvent implicitly provides a large reduction of degrees of freedom, there is a key difference between water and solute molecules in terms of molar concentrations. The mean-field approach of using an effective-medium for pure water is justified because water is the dominant component (i.e., the solvent) at about 55.5 molar (M). Working at lower concentrations of solute is likely to produce correlated spatial fluctuations that may have direct impact on how the protein structure and dynamics are modified by the solute molecules. That is, different protein segments will have a varying number of bound solute molecules (if any). In addition, the excluded volume effects from these solute molecules are likely to be significant since they will generally be of a larger size than water molecules.

In recent years, we have examined the structure and dynamics of a protein (H3.1: ^{1}M ^{2}A ^{3}R . . . ^{136}A) in an effective solvent medium.[32] Herein, we re-examine the same protein, but model solute molecules explicitly to address the role of spatial fluctuations due to preferential binding. The solute concentration is fixed as the protein-solute interaction strength is varied. This setup provides a means to compare results from an effective solvent medium to a semi-implicit water solvent model where all solute molecules are explicitly modeled. A similar type of setup has been employed previously for the study of RNA and DNA.[33] Here, we will compare the diffusive and sub-diffusive regimes of protein dynamics[34,35] while employing an effective medium following our previous work to the new case considered here where solute molecules are modeled explicitly. Differences in the results will inform on the effect of using an effective medium to account for the quality of an aqueous solvent. Insights from this comparison will help better understand the spatial fluctuations in preferential binding of solute molecules as a protein unfolds or partially unfolds in a cellular environment or within a liquid formulation.[36,37]

The histone H3.1 is an intrinsically disordered protein (IDP)[38–42] that does not have a well-defined tertiary structure in pure water, meaning that it is completely intrinsically disordered. In general, IDPs exhibit a diverse-range of characteristics such as partial unfolding in intrinsically disorder regions, and the degree of disorder can be sensitive to solvent conditions. For example, adding a high salt concentration to water drives histone H3.1 to favor a native fold,[38] while in a cellular environment its tails are flexible and intrinsically disordered, which is believed to be a key element for it to function. Experimentally exhibiting the full range of disorder characteristics due to changes in solvent conditions, histone H3.1 serves as a model protein because partially unfolded regions are more susceptible to solvent-protein interactions and a greater heterogeneity in the conformational ensemble is sampled. In contrast, a protein with a well-defined native fold and solvent accessible surface will preferentially bias where cosolutes bind, in part, due to steric constraints. Said another way, if an implicit solvent model holds up well for an IDP, the concerns about the role of fluctuations in solute concentration will be lessened, suggesting a model with an effective medium is likely sufficient. If the fluctuations in solute concentrations are shown to be important, the consequences are applicable to proteins with their native structures disrupted due to the onset of solvent penetration, which is certain to happen if the cosolvent is a denaturant.

Complex biological systems often exhibit self-assembly and aggregation. To model such effects is challenging because of the required computer resources to simulate long-time scales of large systems. The bond fluctuation model is well suited to meet these challenges because it requires far less computer resources than methods that solve dynamical equations of motion. In addition, solvent-solute interactions are parameterized differently from most implicit solvent models (see below) making it straightforward to include solute constituents explicitly while accounting for excluded volume effects. This approach enables conformational ensembles in thermodynamic equilibrium to be explored for different scenarios where solute-water and solute-residue interactions are altered to control self-assembly properties as well as changing the solute composition, concentration, and temperature. It is worth pointing out that explicitly modeling solute molecules expand phase space considerably, which is a major motivating factor for our on-going efforts to generalize the bond fluctuation method. Hence, the significance of explicitly modeling solute in aqueous solution on the structure and dynamics of a single protein will be important for all models/methods involving implicit solvent, including the Langevin dynamics approach.[43] After describing the model in Sec. II, results and

discussions are given on a series of simulations, followed by conclusions.

II. METHOD

The bond fluctuation method simulates conformations of a protein chain of nodes on a periodic cubic lattice. Each node occupies a cubic cell of 8 lattice sites. A node represents a residue and captures its specificity via an interaction matrix[32,44,45] involving pairs of nodes that fall within a cutoff distance. Consecutive nodes along the protein chain are connected by allowed distances that vary between 2 and $\sqrt{(10)}$ in units of the lattice constant. The allowed discrete distance sets define backbone flexibility. Despite being confined to a cubic lattice, it has been shown that there is ample phase space coverage so that results from the bond fluctuation model recover continuous space simulations markedly well.[30,31] Since the histone H3.1 protein has 136 residues, it takes 136 nodes to define its conformation on the cubic lattice. When modeling solute explicitly, a node can also represent a solute molecule. There is no difference in how solute molecules and protein residues are treated as far as the bond fluctuation model is concerned, except no chain of solute molecules is contiguously constrained as peptides. It is worth noting that a solute molecule in the form of a peptide would be straightforward to model using this approach. As an initial condition, the protein chain is first placed on the cubic lattice in a random conformation. Next, the explicit solute molecules are placed on the lattice one at a time by uniformly randomly selecting among all remaining unoccupied lattice sites.

Each residue and solute molecule interacts with nearby residues and solute molecules within a range (r_c) with a generalized Lennard-Jones potential,

$$U_{ij} = \left[\left|\varepsilon_{ij}\right| \left(\frac{\sigma}{r_{ij}}\right)^{12} + \varepsilon_{ij} \left(\frac{\sigma}{r_{ij}}\right)^{6} \right], \quad r_{ij} < r_c, \tag{1}$$

where r_{ij} is the distance between the residues at site i and j or between the residue at site i and solute molecule at site j; $r_c = \sqrt{8}$ and $\sigma = 1$ in units of lattice constant. Note that the range of interaction includes ample number of lattice sites that can be occupied by solute molecules or residues of the protein. The degrees of freedom can be enhanced dramatically if needed with a fine-grain representation of each residue.[32] We use a knowledge-based interaction matrix[31,44,45] for the residue-residue pair interaction (ε_{ij}), which is derived from an ensemble of a large number of protein structures from the protein data bank (PDB). The strength ε_{ij} of the potential is unique for each interaction pair with appropriate positive (repulsive) and negative (attractive) values.[31,32,44,45] We have used the classic Miyazawa-Jernigan (MJ)[44] interaction matrix for most of the data presented here. The Betancourt and Thirumalai (BT)[45] interaction matrix is also used to verify the trend in data. Both the MJ and BT interactions take water as the solvent, and all possible pairs of residues are parameterized. For an implicit solvent model for pure water, only effective interactions between residue pairs need to be considered in the simulation. However, generalizing to aqueous solutions, all pair interactions between a solute molecule and each residue type, and between pairs of solute molecules must be parameterized.

In general, 21 parameters are needed to model a solute molecule in a single component aqueous solution, i.e., solute-solute interaction and solute-residue interactions to capture its specificity. For purpose of demonstration and to simplify the problem of parameterization, only polar solute molecules are considered here. This means that the solute molecule will interact with different residues in a similar way water interacts, except the magnitude of the strength will be modulated. Parameterization for the interaction between a solute molecule (at site i) and a residue (at a site j) is based on the hydropathy index[32] of each residue, $\varepsilon_{ij} = f\ \varepsilon_r A_{h/p/e}$. Furthermore, interaction between pairs of solute molecules will be ignored ($\varepsilon_{ij} = 0$) apart from their excluded volume effect. The hydropathy index defines whether the interaction between the solute molecule and a residue (ε_r) will be attractive or repulsive. The residue-solute interaction is repulsive ($\varepsilon_r = 0.1$) for all hydrophobic (H) residues, attractive ($\varepsilon_r = -0.2$) for all polar (P) residues, and is even more attractive ($\varepsilon_r = -0.3$) to all electrostatic (E) residues. Notice that the hydropathy index is binned into three groups (H, P, E) for simplicity. The weight $A_{h/p/e}$ of a residue varies within each group (H, P, E) according to its relative hydropathy index[22] (see Table I). For example, the interaction ε_{ij} of a solute molecule with a hydrophobic residue, say Cysteine $A_{h/p/e} = H_5$, $\varepsilon_{ij} = f\ (-0.1)H_5$. Similarly for a polar residue such as tryptophan, $A_{h/p/e} = P_3$, $\varepsilon_{ij} = f\ (-0.2)P_3$ and for Arginine, $A_{h/p/e} = P_3$, $\varepsilon_{ij} = f\ (-0.3)E_4$ (see Table I). Thus, the interactions among different types of residues and solvent constituents are unique. The empirical parameter f introduced above modulates solvent quality. It was suggested previously[32] that changes in f could be attributed to changes in pH. However, the interpretation of changing f is considered here to test different types of potential polar solutes with a continuously varying strength. In essence, we are exploring a 21 dimensional

TABLE I. Hydropathy H-index and corresponding weights.

Residue	H-index	Weight ($A_{h/p/e}$)
Ile	4.5	H1 = 1.000
Val	4.2	H2 = 0.933
Leu	3.8	H3 = 0.844
Phe	2.8	H4 = 0.622
Cys	2.5	H5 = 0.556
Met	1.9	H6 = 0.422
Ala	1.8	H7 = 0.400
Gly	−0.4	H8 = 0.089
Thr	−0.7	P1 = 0.200
Ser	−0.8	P2 = 0.229
Trp	−0.9	P3 = 0.257
Tyr	−1.3	P4 = 0.371
Pro	−1.6	P5 = 0.457
His	−3.2	P6 = 0.914
Gln	−3.5	P7 = 1.000
Asn	−3.5	P8 = 1.000
Asp	−3.5	E1 = 0.778
Glu	−3.5	E2 = 0.778
Lys	−3.9	E3 = 0.867
Arg	−4.5	E4 = 1.000

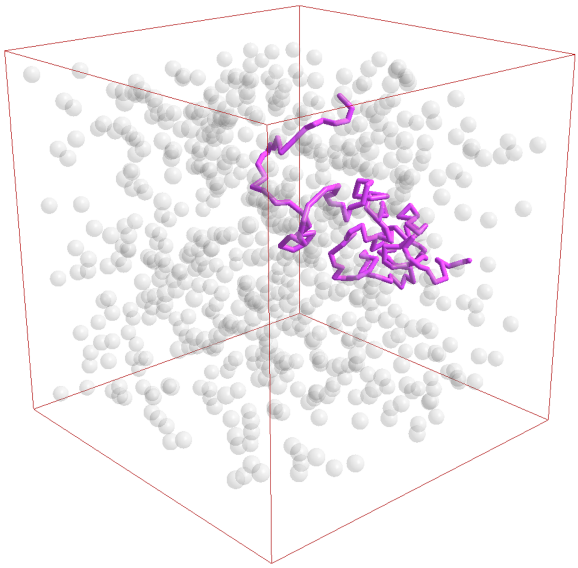

FIG. 1. An initial setup of the protein (H3.1) chain in the simulation box with solvent particles (spheres). The box size is 64^3 and the volume fraction of solute molecules is 0.02.

parameter space using a parametric line with $f \geq 0$, restricting the aqueous solution to polar solutes only. Note that $f = 0$ corresponds to a solute molecule with identical interaction strength as that of water.

The number of solute molecules is fixed throughout the simulation (see Figure 1). Each residue and solute molecule executes stochastic motion using the Metropolis algorithm within the constraints of excluded volume and the allowed discrete set of states for backbone flexibility in the protein. Simulations are carried out for sufficiently long time to generate conformational ensembles in thermodynamic equilibrium, which typically reach five million time steps. Such simulations are run for each solute interaction strength f for a few different temperatures at two different solute concentrations. At each condition, 10-50 independent simulations are performed for statistical averaging for a variety of local and global properties. Different lattice sizes are used to verify no finite size effect on any qualitative trends. The presented results are for a lattice of 64^3 sites, which gives ample sampling at long-time scales without using excessive computer resources. Temperature, time step, and spatial length scales are reported in natural units for the simulation since our focus is on changes in physical quantities in response to changing the solute interaction strength (f) affecting preferential binding. To match with experimental conditions, the simulated temperature range in reduced units from 0.010 to 0.030 corresponds to 150 to 450 K. A volume fraction of 0.01 or 0.02 (1 solute molecule for every 100 or 50 sites) corresponds to roughly 2.2 mM or 4.4 mM of solute, respectively, and 55.5M of water solvent.

III. RESULTS AND DISCUSSION

A set of typical snapshots of the protein conformation and solute at positions that are within the range of interaction of any residue are presented in Figure 2 for different interaction strengths (f) at the temperature $T = 0.030$. At this temperature, the conformation of the H3.1 protein assumes a random coil.[31] Previously, we observed a non-monotonic response of the radius of gyration (R_g) with the solvent interaction in an effective solvent medium[32] at this temperature. The open protein conformation provides the greatest cross section for solute molecules to interact with any residue. Therefore, changes in structure due to the presence of solute will be quantified as a deviation from a random coil conformation. These snapshots provide visual indication for preferential affinity of solute molecules to specific residues together with an overall spread around the protein structure. Figure 2 shows that an increase in solute interaction (e.g., at $f = 3.3, 3.4$) leads to pinning down some of the residues selectively in configurations that form clusters of local segmental structures.

To quantify the solvation profile of the protein, the average number of solute molecules interacting with each residue is calculated. Figure 3 shows the solvent profile of the protein for $f = 3.1$ and 3.4 at the temperature $T = 0.030$ with the volume fraction of solute at 0.01 and 0.02. We see that solute constituents are drawn toward specific residues and as the volume fraction is increased the clustering effect of localized structures also increases. Particularly, most of the residues in the middle segment of the contour (range ^{35}G–^{85}F) attract an increasingly large fraction of solute along with the onset of accumulation towards the end segment (^{129}R-^{136}A) at a higher interaction strength. For example, some of the electrostatic residues (R, K) (^{41}R, ^{43}R, ^{50}R, ^{51}E, ^{53}R, ^{54}R, ^{56}Q, ^{57}K, ^{64}R, ^{70}R, ^{73}R, ^{80}K, ^{82}D, ^{84}R, ^{116}K, ^{123}K, ^{129}R, ^{130}R, ^{132}R, ^{134}E, ^{135}R) become pinned down in the formation of a clustered structure because these electrostatic residues are prone to interact with the highly polar solute. Most solvated segments include ^{35}G – ^{45}G, ^{50}R – ^{60}E, ^{70}R – ^{75}I, ^{77}Q – ^{86}Q, ^{129}R-^{135}R. With the locations of these electrostatic residues constrained by protein sequence,

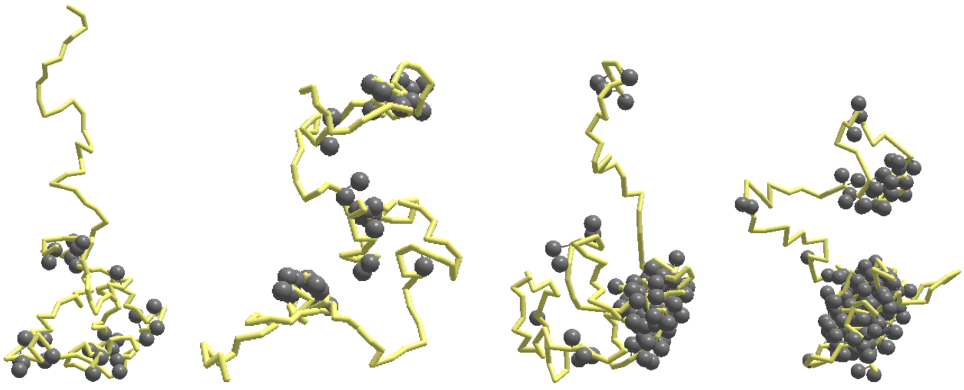

FIG. 2. Snapshots of protein in presence of explicit solvent (spheres) after 5×10^6 time steps on a 64^3 lattice at the volume fraction of 0.01 for solute molecules. The interaction parameter is varied as $f = 3.1, 3.2, 3.3, 3.4$ shown from left to right. Only solute molecules that are within the range of interaction of any residue are shown for clarity.

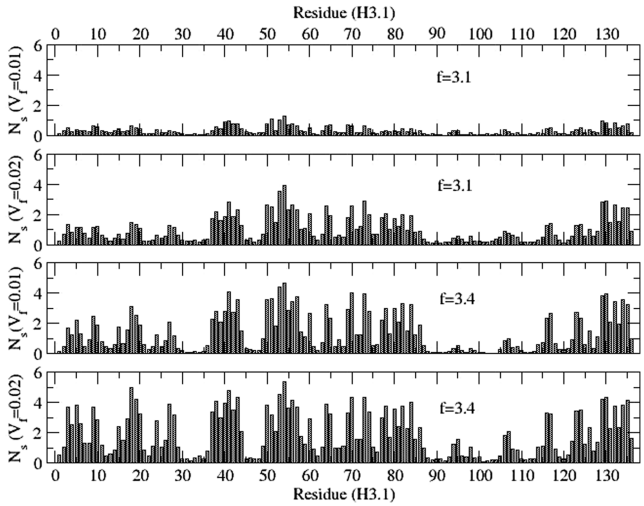

FIG. 3. Average number of solute molecules around each residue at the different solute concentrations with the volume fraction (V_f) set at 0.01 and 0.02. Data are generated on a 64^3 lattice with 50 independent samples for averaging, each for 5×10^6 time steps with MJ interaction.

together with the solute propensity to form proximal bridge interactions that effectively cross link the protein conformation, it is clear that solute molecules orchestrate both local and global structures of the H3.1 protein. This result suggests that the folding pathway to either a collapsed molten globular state or a well-defined native state will be dramatically modified as the starting conformation of the protein is preset by a structured clustering effect. Moreover, it is likely that a folding pathway cannot be achieved in the presence of a large concentration of solute molecules that interact favorably with polar residues, and hence protein denaturation can be expected at lower temperatures. Essentially the preferential binding between a polar solute and the electrostatic residues in a protein in particular can completely reshape the free energy landscape of the protein.

The mobility profile of the H3.1 protein is shown in Figure 4. The mobility of a residue is quantified as the probability of its successful attempts to hop per Monte Carlo step

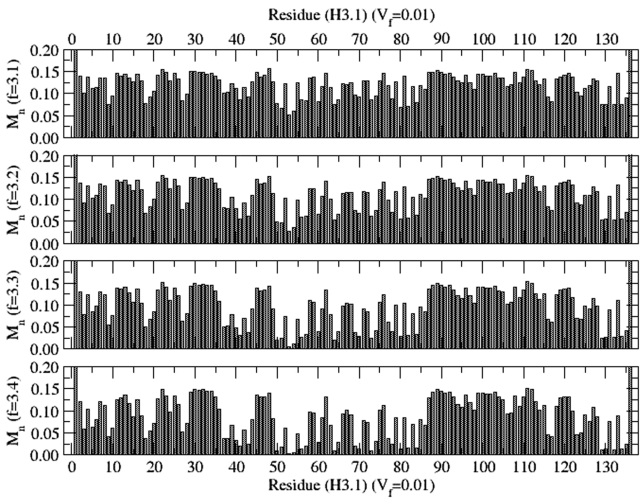

FIG. 4. Average mobility of each residue at a 0.01 volume fraction of solute. Data are generated on a 64^3 lattice with 50 independent samples for averaging, each for 5×10^6 time steps with MJ interaction.

time. We see that all residues increase in mobility as the solvent interaction is weakened ($f = 3.1$) and also at a lower solute concentration. Following intuition, Figure 4 is complementary to Figure 3 because the mobility in a residue comparatively decreases as the solute interaction increases ($f = 3.4$) and as its local surrounding becomes dominated by solute. The mobility profile of the protein with explicit solute (Figure 4) is different from that obtained using an effective solvent.[32] Modeling the dynamics of solute explicitly with a physical presence or not (versus employing an effective omnipresence solvent environment) produces spatial fluctuations that are observed to be critical to the formation of local structured clusters that affect the radius of gyration.

The global dynamics of the protein and solute molecules is analyzed by the variations of the root mean square displacements R_p and R_s, respectively. Examples are presented in Figure 5 with a 0.01 volume fraction of solute. By examining the power-law dependence (i.e., $R \sim t^{\upsilon}$), we characterize the global dynamics as diffusive with $\upsilon = 1/2$ or as sub-diffusive with $\upsilon < 1/2$. We see that the nature of the protein dynamics is diffusive before saturation takes place for weak polar solute-residue interaction (i.e., $f = 3.1$ and 3.2). As the polar solute-residue interaction increases, sub-diffusive dynamics appears before saturation takes place ($f = 3.4$). Interestingly, a sharp transition occurs involving a critical slowing down as the polar solute-residue interaction further increases ($f = 3.4$). Increasing the polar solute-residue interaction beyond this value simply allows the protein to reach saturation more rapidly.

The dynamics of the solute molecules exhibit the same general trend regardless of the polar solute-residue interaction strength, at least over the entire range that is otherwise extremely sensitive for protein dynamics. In particular, there is a sub-diffusive dynamics for the solute to reach target residues for about a million time steps (see the inset in Figure 5 for mean squared displacement). The sub-diffusive dynamics is not

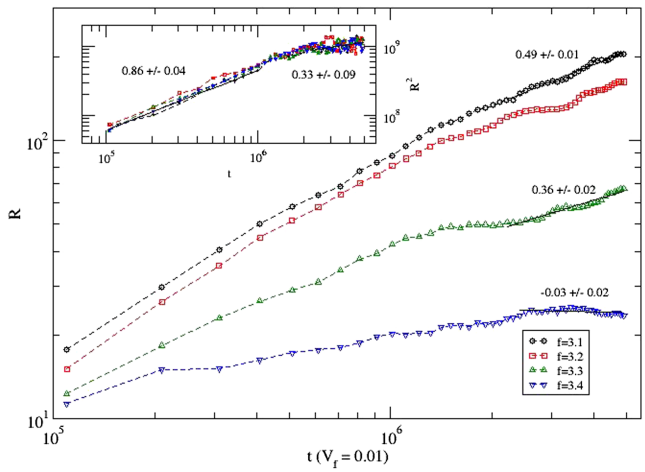

FIG. 5. Variation of the root mean square displacement (R) of the protein chain with a 0.01 volume fraction of solute molecules, $T = 0.030$, and for a narrow range of polar solute interaction strength ($f = 3.1$–3.4) with the MJ interaction matrix. Fit of the data points in the asymptotic time regime is included. The inset shows the variation of the mean square displacement (R^2) of the center of mass of the solute constituents with the time step. Slopes of the data for the pre-asymptotic and asymptotic regimes with $f = 3.1$ are also included.

surprising because as solute molecules diffuse toward target binding sites, the protein conformation itself together with solute molecules that have already preferentially landed on high affinity protein sites makes it more difficult for solute molecules to diffuse through the obstructions. Eventually all polar solute molecules are absorbed by the protein and become localized ($\upsilon \sim 0$), for the regime of high interaction strength, despite the high temperature. This effect is also intuitive because the electrostatic residues are essentially absorbing sites, which are maximally exposed at high temperature. The cross linking that occurs reduces the openness of the protein conformation, but as more solute molecules diffuse into the protein to interact at a preferential site, the protein conformation can adjust to further maximize these favorable interactions. Not caused by electrostatic steering, this clustering effect will wash out at this high temperature when the polar solute-residue interaction is too low.

Variation of the radius of gyration (R_g) with the solvent interaction strength is presented in Figure 6 at the temperature $T = 0.030$. The radius of gyration at saturation undergoes a sharp transition as a function of solvent interaction strength. At weak polar solute-residue interaction strength ($f \sim 1.0 - 2.5$), it is seen that $R_g \sim 20$. This result indicates that the polar solute molecules are behaving energetically too similar to the way water interacts with the protein, such that the structural properties of the protein are not disturbed. Conversely, the strong attractive interaction ($f \geq 3.5$) between the polar solute molecules with electrostatic residues (especially) causes the extended conformations to become compact with $R_g \sim 10$–12. This dramatic change from an extended to compact globular form occurs within a narrow range of interaction strength ($f \sim 3.0$–3.5). We can understand this transition thermodynamically. The weak attraction of polar solute molecules cannot overcome the entropy of mixing of solute molecules diffusing throughout the solution. At some point, increasing favorable polar interactions can overcome mixing entropy.

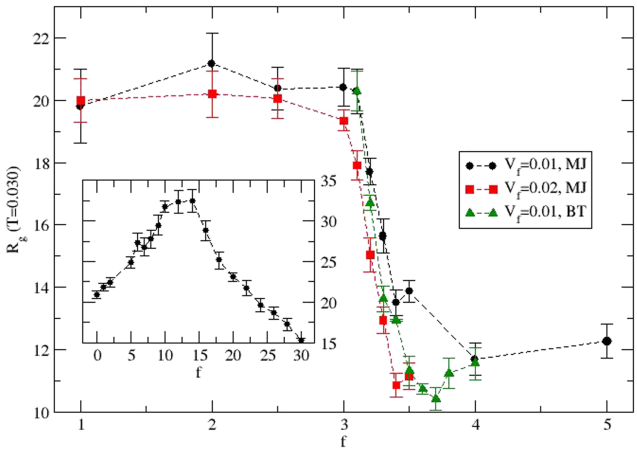

FIG. 6. Variation of the radius of gyration (R_g) as a function of polar solute interaction strength (f) at the temperature $T = 0.030$. Simulations are performed using the knowledge-based residue-residue interactions with MJ and BT interaction matrices at 0.01 volume fraction of solute molecules. A 0.02 volume fraction of solute molecules is also used with the MJ interactions. The inset figure shows the corresponding variation of R_g for the protein. In this case, every empty site represents solvent. Over the dataset, 10-50 independent samples on a 64^3 lattice are considered.

This sharp transition can also be understood dynamically. The solute molecules execute their stochastic motion rather fast at $T = 0.030$. As a result, they quickly reach their targets, which are mainly the electrostatic residues. The sticking probability of the solvent particles around the target residues depends on their interaction strength, which competes with thermal noise. The solute molecules are unable to stick around the target sites with weak interactions ($f \sim 1.0$–2.5) for a long period of time. With increased interaction strength, solute molecules stay around the target sites for most of the time and pin down sites that subsequently act as seeds for protein segments to collapse in a globular form. It is worth noting that the compact structure that forms is not a normal folded state, because the structure comprises penetrating solute molecules that are well positioned to maintain the stability of the structure. This mechanism of increasing stability by binding partners occurs in the tails in the histone protein. They are intrinsically disordered until stabilized by interactions with DNA, which allows the compaction of DNA to occur. Both binding partners play a critical role in stabilizing the complex as conformational flexibility is reduced. In this system, the stabilization is derived from binding to cosolvent, which suppresses conformational flexibility at cross-linking locations. The reduction in mobility is evident in Figure 4 as well as dynamical slowing down in the compact state and the onset of the compact state quantified by sub-diffusion in Figure 5.

The radius of gyration as a function of solute interaction is very different between explicit and implicit solvent models. From the effective solvent model,[32] the inset in Figure 6 shows that R_g has a non-monotonic dependence on the interaction strength. Only at $f = 0$ for the case of pure water, both solvent models are in agreement, showing that $R_g \sim 20$. For the implicit solvent model, R_g initially increases as polar interaction strength increases, until R_g peaks at ($f \sim 15$), followed by a slow continuous decay that dips below $R_g \sim 20$, where $R_g \sim 15$ at $f = 30$. It is worth mentioning that the range of interaction strength over which the changes in radius of gyration occur is much larger than that in presence of explicit solute molecules at the same temperature. In our previous work,[32] no attempt was made to identify the interaction strength parameter f to solute concentration. Therefore, our prior result over-emphasizes the changes that occur in the protein structure and dynamics due to the effective media. However, it is clear that the explicit modeling of solute molecules causes much more dramatic differences because of the preferential sites that the solute molecules tend to associate with, and causing partial unfolding.

Increasing the polar nature of the solvent everywhere in a uniform manner as mean-field assumes would encourage the extended random coil state to extend further because of the effect of solute molecules pulling on the protein conformation outwards. The entropy of mixing is not part of the considerations within the implicit model. Conversely, the explicit model of solute molecules shows that this pulling cannot be uniform simultaneously across the protein surface. Rather, a protein conformation would experience intermittent tugs that would tend to open up the protein structure like the implicit solvent model does, but the conformation is rapidly repaired because of intramolecular interactions with the protein. At some point when the interaction is sufficiently strong (attractive to

electrostatic residues), the protein conformations open up enough for the solute particles to reach favorable target sites on the protein, which as indicated above, crosslink the protein, and cause it to become compact. This compaction is seen in the implicit model at high interaction strength as well.

In order to quantify the structural response of the protein to solvent interaction, we analyze the structure factor,

$$S(q) = \langle \frac{1}{N} | \sum_{j=1}^{N} e^{-i\vec{q}\cdot\vec{r_j}} |^2 \rangle_{|\vec{q}|},$$

where r_j is the position of each residue and $|q| = 2\pi/\lambda$ is the wave vector of wavelength, λ, which informs on the spatial spread of residues in the protein's conformation. Assuming the structure factor exhibits a common power-law scaling as a function q, i.e., $S(q) \propto q^{-1/\gamma}$, then the spatial distribution of residues can be estimated. For example, the scaling of the radius of gyration (R_g) of the protein with the number (N) of residues is described by $R_g \propto N^{\gamma}$, i.e., $N \propto R_g{}^{1/\gamma}$ which implies that the effective dimension of the protein $D \approx 1/\gamma$. Figure 7 shows the variation of the structure factor with the wave vector q for different interaction strengths (f = 3.1–3.4) over the narrow range where the radius of gyration exhibits the maximum response. Moreover, the variation of $S(q)$ as a function of q changes systematically as the interaction strength is increased. Note that $q \sim 0.3$-0.4 is probing spatial scales that of the radius of gyration of the protein having $R_g \sim 20$. The effective dimension of $D \sim 1.3$ found when the polar solute-residue interaction is weak (f = 3.1) implies that the protein and surrounding solute structure is linear (i.e., fibrous with less loops) than the random-coil configurations ($D \sim 1.75$) with f = 3.2 (Figure 7). At a higher interaction strength (f = 3.4), the scaling of the structure factor leads to $D \sim 3$, a measure of a well-packed compact globular structure.

Also of interest is the nature of the protein structure, or its ensemble of conformations. Below the transition, there is no well-defined structure and the protein is essentially in an unfolded state. However, at the onset of the transition, oscillation in the variation of $S(q)$ with q indicates the presence of short range segmental correlations. Strong oscillations indicate that the relative positions of the solute molecules are quenched, an opposite effect from random diffusion as a stable structure forms. As solute molecules come into position, conformation fluctuations are suppressed and the structure no

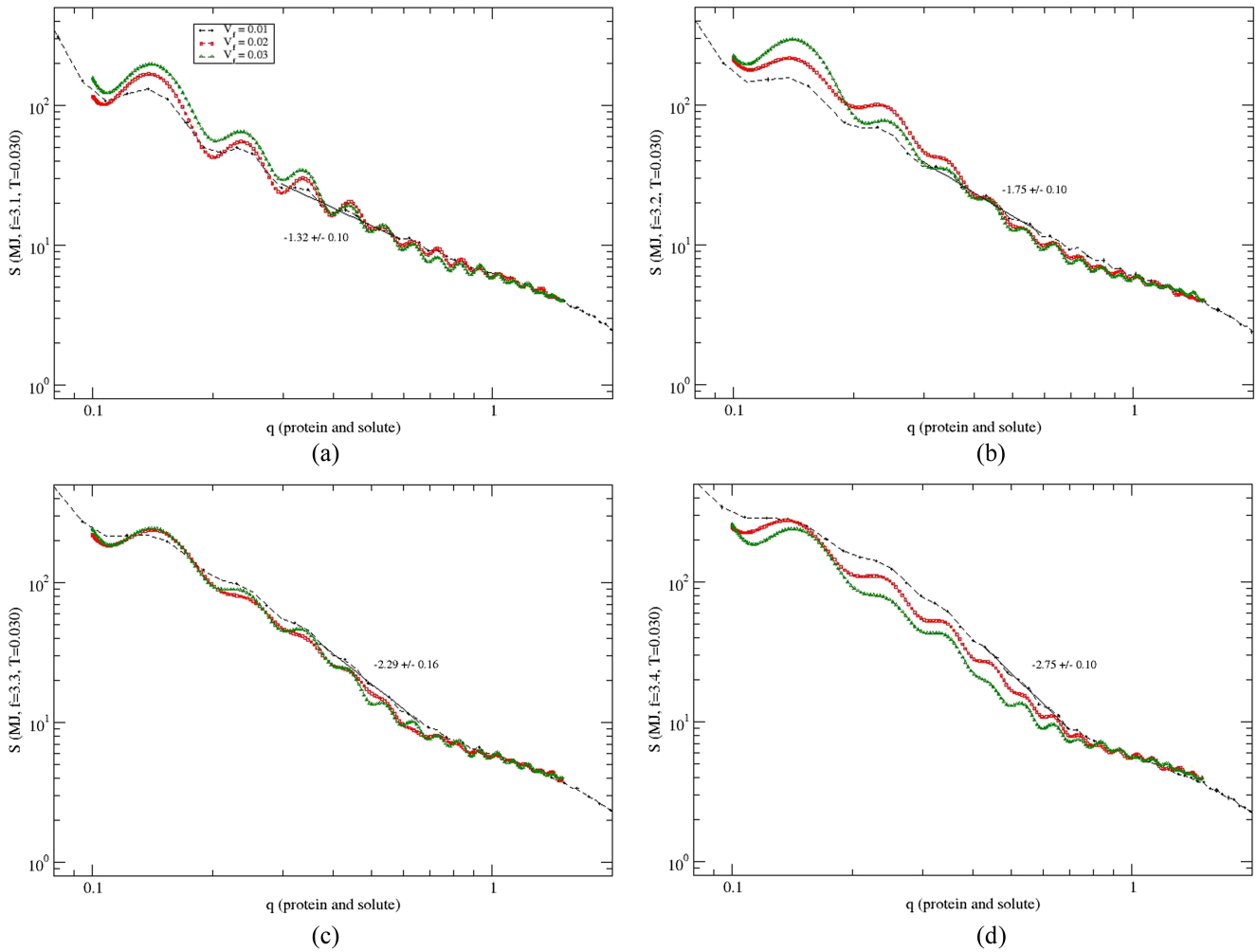

FIG. 7. The structure factor $S(q)$ is shown for the H3.1 protein with wave vector q in aqueous solvent with a 0.01-0.03 volume fraction for the solute molecules and with varying polar solute interaction strengths, i.e., f = 3.1–3.4 ((*a*)-(*d*)) at temperatures T = 0.030. Slopes of the fitted data (covering the spread of the radius of gyration, see Figure 7) are included to guide the eye. All simulations are performed on a 64^3 lattice for 5×10^6 time steps with MJ interaction using 50 independent samples.

longer undergoes large conformational changes on long time scales (its folded). This can be seen by the fine scale oscillations that persist even at large q, corresponding to short distances. Furthermore, these correlations become pronounced as solute concentration increases, causing mobility to be further suppressed.

IV. CONCLUSIONS

The structure and dynamics of the histone H3.1 protein are examined using an implicit water solvent model that models solute molecules explicitly. The results of this model are compared with previous results using an effective medium.[32] In particular, the structural response to the polar solute-residue interaction strength (i.e., solvent quality) for various solute concentrations is investigated using a coarse-grained representation to describe the protein chain and also for the solute molecules when modeled explicitly. While varying molar concentration of solute through volume fraction, and varying the polar solute interaction strength (f), both local (solvation and mobility profiles) and global (radius of gyration and structure factors) quantities are analyzed.

We find that the structural response of the H3.1 protein to solvent quality in the presence of explicit solvent substantially differs from that found in simulations using an effective medium. This demonstration clearly points to the important role that spatial fluctuation of solute-protein interactions plays in regards to preferential binding. Visual inspections together with the solvation profile indicate that preferential binding of polar solute molecules occurs at electrostatic residues. Specifically, the segments of the protein with the highest propensity to bind with the polar solute include ^{35}G–^{45}G, ^{50}R–^{60}E, ^{70}R–^{75}I, ^{77}Q–^{86}Q, ^{129}R–^{135}R. The solvation profile with a higher solute concentration remains similar to that with low solute concentration (Figure 3). However, it is easier to identify segments of preferential binding due to accentuated profiles at a higher polar solute concentration even in cases of lower interaction strength.

Mobility profiles of the H3.1 protein calculated from an effective medium[32] versus explicitly modeling the polar solute molecules also differ. In the former case, the saturated value for the radius of gyration, R_g, has a non-monotonic behavior as the attractive polar solute interaction strengthens. Using an effective medium, R_g initially increases when the attractive polar solute interaction is weak, although still stronger than water. Conversely, R_g is unmodified when modeling solute explicitly. The differences in response from the two models suggest that spatial fluctuations of explicitly modeled solute molecules remove the artificial coherency in the collective affect from solute molecules that is created by an effective medium. As the attractive polar solute interaction further strengthens, R_g decreases in both solvent models. However, the spatial fluctuations from explicitly positioned solute molecules open up specific pathways for solute to penetrate the random coil conformation. The solute molecules reach residue binding-partners rather quickly during this process, exhibiting preferential binding to specific residues.

Variations in the structure factor $S(q)$ of the protein with wave vector (q) are different from an effective solvent medium[32] compared to modeling solute molecules explicitly. Unlike in effective medium, the structure factor becomes oscillatory in the presence of explicit solute molecules. We observe fast passage of solute molecules to targeted residues followed by a critical slowing down due to pinning and cross-linking as the main reasons for the oscillatory patterns in $S(q)$. Oscillation in $S(q)$ as a function of q is a sign of segmental correlations which diminishes as solute concentration decreases. The relevant length scale is comparable to R_g of the protein. The scaling of the structure factor with wave vector, q, reveals fibrous ($D \sim 1.3$), random-coil ($D \sim 1.75$), and globular ($D \sim 3$) conformational ensembles as the polar solute interaction strength is increased. Thus, solute properties are critical in orchestrating the structural response of a protein, which is sensitive to the type and manner in which the solute molecule interacts with the protein.

Interestingly, attractive polar solute interactions promote structurally stable clusters of residues due to cross-linking by solute molecules. This cross-linking zips up the random coil, creating a sharp transition from an extended random coil to a compact unfolded globular structure as a function of interaction strength. The effect of an external agent stabilizes an IDP which is common[46] and indeed essential for the biological function of histone in its compaction of DNA within the cellular environment. While our simulation results are consistent with the known properties of histone proteins, much more detailed simulation results from all-atom molecular dynamics on the intrinsically disordered tails in the histone family have been previously reported to quantify conformational ensemble characteristics regarding frequency of occurring secondary structure elements and the degree of disorder,[40] and, therein, differences between explicit and implicit models were also noted. We do not attempt to use the hybrid model introduced here to predict free energy landscapes that have already been calculated more accurately with extensive all atom MD simulation. We also defer making any claim about the kinetic mechanism that leads to the stable compact structures, because the model oversimplified in its parameterization, and it is not the focus of this work. Rather, the parameterization used here provides a minimalist model to address the question of suitability of an effective medium approximation when cosolute molecules penetrate partially the unfolded structure within a protein. When studying preferential binding or the effect of formulation on protein stability, the cosolvent should be modeled explicitly to account for spatial fluctuations. A caveat is when a protein undergoes small conformational fluctuations around a native state. This case was not considered here because histone in water is intrinsically disordered and we performed simulations at high temperature. The remarkable finding is at sufficiently high concentration of cosolvent, protein-cosolvent binding forms ordered conformations.

These results indicate that the application of implicit solvent models to biomolecular systems should account for spatial fluctuations in solute molecules within an aqueous solution. Furthermore, methods that suppress conformational flexibility in proteins, such as regarding a protein as a fixed rigid body, will suppress the effect of spatial fluctuations of solute molecules. While thermodynamic and statistical mechanics

models of preferential binding are useful for understanding protein stability, the bottleneck in applying these theories relies on methods that generate conformational ensembles. As such, the models/methods employed should retain the essential effects of spatial fluctuations of solute molecules that comprise aqueous solution and other crowding agents. The bond fluctuation model at a coarse grained description offers an alternative method to those that integrate equations of motion for the pragmatic simulation of large-scale biological systems on long-time scales. Future work along these lines is to consider a variety of solute molecule properties individually, as well as investigate multicomponent liquid formulations intended to stabilize proteins.

[1] P. H. Yancey, Am. Zool. **41**, 699–709 (2001).
[2] G. Wei, W. Xi, R. Nussinov, and B. Ma, Chem. Rev. **116**, 6516–6551 (2016).
[3] J. L. England and G. Haran, Annu. Rev. Phys. Chem. **62**, 257–277 (2011).
[4] G. Caliskan *et al.*, J. Chem. Phys. **121**, 1978–1983 (2004).
[5] D. R. Canchi and A. E. Garcia, Annu. Rev. Phys. Chem. **64**, 273–293 (2013).
[6] S. N. Timasheff, Proc. Natl. Acad. Sci. U. S. A. **91**, 9721–9726 (2002).
[7] S. Shimizu, J. Chem. Phys. **120**, 4989–4990 (2004).
[8] J. Rosgen, B. M. Pettitt, and D. W. Bolen, Biophys J. **89**, 2988–2997 (2005).
[9] J. Skinner *et al.*, Proc. Natl. Acad. Sci. U. S. A. **111**, 15975–15980 (2014).
[10] R. O. Dror *et al.*, Annu. Rev. Biophys. **41**, 429–452 (2012).
[11] S. Piana, J. L. Klepeis, and D. E. Shaw, Cur. Opin. Struct. Biol. **24**, 98–105 (2014).
[12] S. J. Marrink and D. P. Tieleman, Chem. Soc. Rev. **42**, 6801–6822 (2013).
[13] T. Lazaridis and M. Karplus, Proteins **35**, 133 (1999).
[14] R. B. Pandey and B. L. Farmer, J. Chem. Phys. **132**, 125101 (2010).
[15] J. Chocholousova and M. Feig, J. Phys. Chem. B **110**, 17240 (2006).
[16] J. Chen and C. L. Brooks III, Phys. Chem. Chem. Phys. **10**, 471 (2008).
[17] L. Li, K. Dill, and C. J. Fennell, J. Comput.-Aided Mol. Des. **28**, 259 (2014).
[18] J. Kleinjung and F. Fraternali, Cur. Opin. Struct. Biol. **25**, 126–134 (2014).
[19] F. Ding, D. Tsao, H. Nie, and N. V. Dokholyan, Structure **16**, 1010–1018 (2008).
[20] S. Pronk *et al.*, Bioinformatics **29**, 845–854 (2013).
[21] S. Chaudhury *et al.*, J. Chem. Theory Comput. **8**, 677–687 (2012).
[22] L. L. Duan, G. Q. Feng, and Q. G. Zhang, Sci. Rep. **6**, 31488 (2016).
[23] J. Zavadlav *et al.*, J. Chem. Theory Comput. **10**, 2591–2598 (2014).
[24] R. Harada, Y. Takano, T. Babad, and Y. Shigeta, Chem. Phys. Chem. Phys. **17**, 6155–6173 (2015).
[25] E. P. O'Brien *et al.*, Proc. Natl. Acad. Sci. U. S. A. **105**, 13403–13408 (2008).
[26] G. Cazzolli *et al.*, Proc. Natl. Acad. Sci. U. S. A. **111**, 15414–15419 (2014).
[27] S. Kimura *et al.*, Proteins **82**, 633–639 (2014).
[28] R. Anandkrishna, A. Drozdetski, R. C. Walker, and A. V. Onufriev, Biophys. J. **108**, 1153 (2015).
[29] C. Arnarez *et al.*, J. Chem. Theory Comput. **11**, 260–275 (2015).
[30] K. Binder, *Monte Carlo and Molecular Dynamics Simulations in Polymer Science* (Oxford University Press, New York, 1995).
[31] R. B. Pandey and B. L. Farmer, PLoS One **7**, e49352 (2012).
[32] R. B. Pandey and B. L. Farmer, PLoS One **8**, e76069 (2013).
[33] N. V. Prabhu, M. Panda, Q. Yang, and K. A. Sharp, J. Comput. Chem. **29**, 1113–1130 (2008).
[34] S. B. Kim, C. J. Dsilva, I. G. Kevrekidis, and P. G. Debenedetti, J. Chem. Phys. **142**, 085101 (2015).
[35] E. D. Nelson and N. V. Grishin, Phys. Rev. E **91**, 060701(R) (2015).
[36] R. J. Ellis and R. U. Hart, Cur. Opin. Struct. Biol. **9**, 102–110 (1999).
[37] K. E. S. Tang and V. A. Bloomfield, Biophys. J. **82**, 2876–2891 (2002).
[38] T. Frege and V. N. Uversky, Biochem. Biophys. Rep. **1**, 33 (2015).
[39] V. N. Uversky, J. Biol. Chem. **291**, 6681 (2016).
[40] D. A. Potoyan and G. A. Papoian, J. Am. Chem. Soc. **133**, 7405 (2011).
[41] Q. Qiao, G. R. Bowman, and X. Huang, J. Am. Chem. Soc. **135**, 16092 (2013).
[42] P. Mirau, B. L. Farmer, and R. B. Pandey, AIP Adv. **5**, 092504 (2015).
[43] E. Paquet and H. L. Viktor, BioMed. Res. Int. **2015**, 183918.
[44] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[45] M. R. Betancourt and D. Thirumalai, Protein Sci **2**, 361 (1999).
[46] L. Mollica *et al.*, Front. Mol. Biosci. **3**, 52 (2016).