Equilibration of experimentally determined protein structures for molecular dynamics simulation

Emily B. Walton and Krystyn J. VanVliet*
Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
(Received 2 June 2006; published 5 December 2006)

Preceding molecular dynamics simulations of biomolecular interactions, the molecule of interest is often equilibrated with respect to an initial configuration. This so-called equilibration stage is required because the input structure is typically not within the equilibrium phase space of the simulation conditions, particularly in systems as complex as proteins, which can lead to artifactual trajectories of protein dynamics. The time at which nonequilibrium effects from the initial configuration are minimized—what we will call the *equilibration time*—marks the beginning of equilibrium phase-space exploration. Note that the identification of this time does not imply exploration of the entire equilibrium phase space. We have found that current equilibration methodologies contain ambiguities that lead to uncertainty in determining the end of the equilibration stage of the trajectory. This results in equilibration times that are either too long, resulting in wasted computational resources, or too short, resulting in the simulation of molecular trajectories that do not accurately represent the physical system. We outline and demonstrate a protocol for identifying the equilibration time that is based on the physical model of Normal Mode Analysis. We attain the computational efficiency required of large-protein simulations via a stretched exponential approximation that enables an analytically tractable and physically meaningful form of the root-mean-square deviation of atoms comprising the protein. We find that the fitting parameters (which correspond to physical properties of the protein) fluctuate initially but then stabilize for increased simulation time, independently of the simulation duration or sampling frequency. We define the end of the equilibration stage—and thus the equilibration time—as the point in the simulation when these parameters attain constant values. Compared to existing methods, our approach provides the objective identification of the time at which the simulated biomolecule has entered an energetic basin. For the representative protein considered, bovine pancreatic trypsin inhibitor, existing methods indicate a range of 0.2–10 ns of simulation until a local minimum is attained. Our approach identifies a substantially narrower range of 4.5–5.5 ns, which will lead to a much more objective choice of equilibration time.

DOI: 10.1103/PhysRevE.74.061901 PACS number(s): 87.15.Aa, 02.70.Ns

I. INTRODUCTION

Classical molecular dynamics techniques are commonly applied to computational simulations of molecular interactions that do not include explicit chemical reactions [1]. As computational resources have increased, simulations of many aspects of protein behavior have become increasingly well studied, ranging from refinements of static x-ray crystallography structures to dynamic conformational changes in ion channels, as recently reviewed by Karplus and McCammon [2].

The statistical mechanical underpinnings of many algorithms used to analyze classical molecular-dynamics simulations assume thermodynamic equilibrium [3]. When simulating biomolecules, the initial structure is often an experimentally determined structure acquired from the Protein Data Bank (PDB) [4]. The data gathered from an x-ray-diffraction (XRD) experiment is not the equilibrium structure of the protein; rather, it is a map of electron probability density, averaged over many unit cells, from which an average structure can be determined. This provides an initial guess for a set of atomic positions but, due to the ensemble averaging inherent in XRD, it is not an equilibrium structure; nuclear magnetic resonance has similar limitations [5]. Therefore, simulations utilizing these experimental structures as input will always require a so-called equilibration stage before usable trajectories can be collected. It is understood that the purpose of equilibration is limited to the minimization of the structural and dynamic artifacts of the nonequilibrium initial conditions. Thus, although this stage of the simulation does not make any claims as to exploration of the entire equilibrium phase space, it is necessary to identify the *starting point* of that exploration. Others have described methods that help determine the extent of phase-space sampling; see, for example, Refs. [6–9]. It should be noted that any method for equilibration that begins with an experimentally determined structure has one major flaw: the system may equilibrate into an energy basin that is not the native basin of the protein, depending on the physical properties of the protein and the quality of the experimental structure. Our approach does not mitigate this possibility; rather, it provides an objective, physically grounded method for determining equilibration of an experimentally derived initial structure.

Each application of molecular-dynamics (MD) simulation will require its own method of equilibration and its own accuracy threshold. These methods are fairly straightforward for enthalpic systems such as infinite crystals, and are more challenging for entropic systems such as proteins. We note that if the time scale of the process of interest exceeds the time scale accessible with molecular-dynamics simulation (currently less than one microsecond), full equilibration is not possible and other simulation methods should be consid-

*Present address: Massachusetts Institute of Technology, 8-237, 77 Massachusetts Ave., Cambridge, MA 02139. Electronic address: krystyn@mit.edu

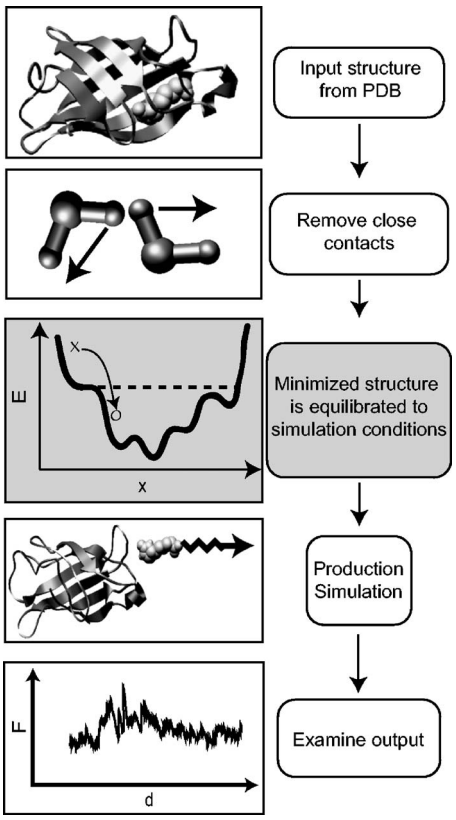

FIG. 1. Outline of the steps in a typical MD protein simulation, illustrated for a steered molecular dynamics simulation of biotin unbinding from streptavidin [10]. The first step is to import an initial structure from the PDB. Next, solvent is added and the structure is energy-minimized via a non-MD algorithm such as steepest-descents minimization to remove close contacts. Then the system is *equilibrated*: the minimized experimental structure undergoes MD simulation to bring it into equilibrium with the simulation conditions. Only after this equilibration stage can production simulations be run that will yield realistic results. We are proposing a new procedure for the equilibration stage, which is indicated by a shaded background.

ered. However, many important biomolecular interactions are computationally accessible and depend critically on simulated protein structure. Computational predictions of such interactions should adopt robust and reproducible approaches to the equilibration step of the simulation.

Figure 1 outlines the steps for a typical MD simulation of a protein, starting from a structure file obtained from the PDB before proceeding through minimization, equilibration, and production to obtain meaningful results. Here, we are concerned with only the equilibration phase, and not the production phase, in which enough of the equilibrium phase space would be sampled to achieve convergence of simulation results. The goal of the equilibration stage of an MD simulation is to separate the trajectory into two portions: one containing nonequilibrium fluctuations due to the initial nonequilibrium structure, and one which is free of those fluctuations. Logically, a robust equilibration method will have three important characteristics. First, the method should be based on a model which accurately reflects the physics of the system of interest. Second, it should unambiguously identify the point in the trajectory where the initial nonequilibrium motions due to the initial structure have dissipated. Third, the method should provide an equilibration time that is independent of the duration of the simulation—the time required for nonequilibrium motions to cease will of course not depend on the amount of computational time used, but solely on the physics of the system.

Interestingly, equilibration procedures for structural studies of proteins have not changed significantly as computational resources have increased. The root-mean-square deviation (RMSD) from the original structure was introduced as an equilibration metric by Daggett *et al.* in 1993 [11], and is still very common in protein simulation. When the RMSD reaches a plateau value, the system has reached a basin in the potential energy surface. This energetic basin is assumed to correspond with the equilibrium phase space, and the system is deemed suitable for the production stage of molecular-dynamics simulation.

A variation of this simple RMSD-based technique was proposed by Stella *et al.* [12]. This variant considers not only the RMSD from the initial structure, but also the RMSD referenced from several different intermediate structures of the trajectory, which are taken at equal intervals throughout the simulation trajectory.

The authors suggested that the RMSD plateau value would decrease sharply from an initial value to a final value when the artifacts from the initial structure had dissipated, such that the set of plateau values would consist of two distinct clusters. The end of the equilibration portion of the trajectory could then be identified as the time step corresponding to the first intermediate structure with the lower RMSD plateau value. Stella *et al.* predicted that equilibration stage durations as determined by this method would be shorter than those determined by the simple RMSD approach, and demonstrated this reduction for the protein, human glutathione *S*-transferase P1-1 [12]. However, as discussed below, this approach cannot be applied to all biomolecules and includes ambiguities such as the choice of intermediate structures.

Both the simple RMSD and Stella methods yield equilibration times that depend not only on the duration of the simulation, but also on the judgment of the researcher applying the algorithm. The model of the protein settling into an energy basin is phenomenologically accurate, but it does not enable the quantitative analysis required for objective comparison among simulation conditions or among researchers. Here, we propose a method for identifying the beginning of a protein's exploration of its equilibrium phase space. While our method has the same goal as the simple RMSD and Stella methods, it is based on a specific physical model that supports an objective equilibration algorithm. We base our approach on the physical model used in Normal Mode Analysis (NMA), which portrays the protein as a linear combination of independent harmonic oscillators. This allows us to fit a physically meaningful functional form to the RMSD of the protein, and monitor the changing properties of that fit as a function of reference time. We find that the fitting parameters (which correspond to physical properties of the protein) fluctuate initially but become roughly constant later in the simulation. We define the end of the equilibration stage—

and thus the equilibration time—as the point in the simulation when these parameters attain constant values that are independent of simulation duration and sampling frequency.

II. RELAXATION MODELS

The RMSD of the atoms of a protein is defined as a function of both reference time and simulation time,

$$\xi(t_{\mathrm{ref}},t)=\left[\frac{1}{M}\sum_{i=0}^{N} m_i\|r_i(t_{\mathrm{ref}})-r_i(t)\|^2\right]^{1/2}, \tag{1}$$

where M is the total mass of the system, m_i is the mass of atom i, N is the number of atoms in the system, $r_i(t)$ is the position of atom i at simulation time t, and t_{ref} is the time step in the simulation corresponding to the reference structure; rigid body translation and rotation are excluded [13]. A very commonly used equilibration protocol for molecular dynamics simulations of proteins is to monitor the RMSD from the initial structure (often a PDB structure that is first solvated and then minimized to an arbitrary energy level). In such an approach, the reference structure is always the structure at the initial time, $t_{\mathrm{ref}}=0$. In the Stella *et al.* method [12], the RMSD from many different intermediate reference structures are considered. Both of these methods rely upon the RMSD attaining a plateau value, typically determined by visual inspection. This simple RMSD-based method relies on the physical picture of the protein settling into a basin in the potential-energy surface of unspecified shape and size. This has both the advantage and the disadvantage of generality—the idea that the protein settles into some type of energy minima is irrefutable, but without specifying a functional form for the potential, this model is far too general to support quantitative analysis of data from the simulation. As such, it cannot be used to determine when nonequilibrium fluctuations have dissipated sufficiently.

In proteins well above the corresponding glass transition temperatures, many different processes contribute to relaxation, including interactions with solvent molecules, transitions between conformations of side chains, and the vibrations of individual covalent bonds [14]. Each of these processes occurs on a different time and length scale. Although atomic motion is necessarily coupled in a close-packed biomolecule like a protein, different vibrational and rotational modes are not necessarily coupled to each other. Assuming that the processes contributing to the RMSD can be represented by decoupled modes, the internal movements of the protein could be modeled as a collection of decoupled harmonic oscillators. This is equivalent to the well-established method of analyzing protein dynamics known as NMA [15–18]. Here, we aim only to identify the time at which the protein is represented by a stable set of modes, and thus the harmonic approximation. While this assumption is imperfect (there is ample evidence that some modes of protein motion are actually anharmonic [19]), the decoupled multiple harmonic oscillator model is adequate because we do not need to identify the character of these modes. This model provides a substantial advantage over alternatives such as principle component analysis because it results in a mathematically tractable expression for the RMSD of the protein.

In this multiple simple harmonic oscillator (MSHO) model, the RMSD of the protein (ξ) can be represented as

$$\xi(t_{\mathrm{ref}},t)=\sum_{i=0}^{N} A_i(1-e^{-(t-t_{\mathrm{ref}})/\tau_i}), \tag{2}$$

where N is the number of independent harmonic oscillators, and A_i and τ_i are the preexponential and the time constant for each harmonic oscillator, respectively. Fitting Eq. (2) to the RMSD of the protein calculated from Eq. (1) would require knowledge of the number of processes contributing to the motion of the protein as well as a preexponential factor and a time constant for each process, leading to an excessive number of parameters for even small proteins.

However, Eq. (2) can be well approximated under certain conditions using the Kohlrausch-Williams-Watts (KWW) function (also known as a stretched exponential) as

$$\xi(t_{\mathrm{ref}},t)\approx A_e(1-e^{-[(t-t_{\mathrm{ref}})/\tau_e]^{\beta}}), \tag{3}$$

where N is the number of independent harmonic oscillators, A_e is the effective preexponential factor, τ_e is the effective time constant, and β is a scalar value between 0 and 1 that represents the complexity of the system [20,21]; $\beta=1$ would be a single simple harmonic oscillator. This is a significant simplification of Eq. (2), which is only valid when the distribution of time constants τ_i consists of wide, overlapping peaks, as described by Apitz *et al.* [22]. If the distribution of time constants consisted of discrete narrow peaks, that would represent a very small number of harmonic oscillators, which might be an adequate model for a small molecule with a limited number of vibrational and torsional modes. A typical biomolecule, however, will have an enormous number of vibrational and torsional modes operating on different time and length scales. This will generally lead to wide, overlapping peaks in the time-constant distribution, as is required for the use of Eq. (3). Results from normal-mode analysis support this: simulations of various proteins have shown that the modes contributing most to the RMSD exist over a large spectrum of frequencies [19,23,24], corresponding to a large spectrum of the associated time constants.

The stretched exponential function has been used to model many relaxation processes, from the dielectric relaxation of polymers [21] to the compaction of granular systems [25] and the long time-scale density fluctuations of supercooled liquids [26]. These same types of systems have also been analyzed with NMA (relaxation of polymers [27] and supercooled liquids [28]). However, although both the KWW approximation and NMA are appropriate to the physics of the system [22], to the best of our knowledge they have not been utilized in a previously published study to gauge protein equilibration in MD simulations [29].

As discussed in Sec. III, we found that the KWW model is applicable for the RMSD of our sample protein, and we have found it also to be applicable for the larger biotin-streptavidin system (see Appendix A). In evaluating this multiple harmonic oscillators model, we fit RMSD data to Eq. (3) to obtain quantitative measurements of A_e, τ_e, and β.

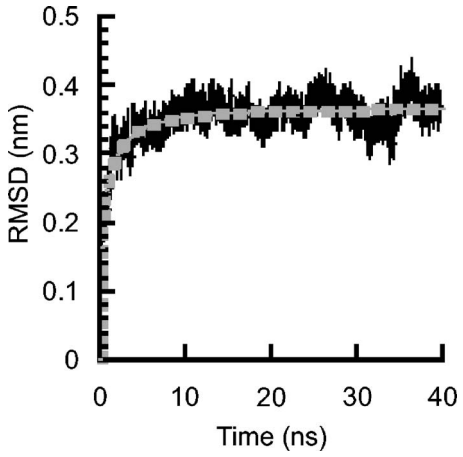

FIG. 2. Raw trajectory of solvated BPTI at 300 K as a function of time (black), showing the corresponding KWW model fit (gray). The correlation coefficient for this fit is 0.843, indicating good agreement between the data and the model.

We will refer to this model as the KWW model. We chose to use the correlation coefficient (r) as our parametrization of the quality of this fit. The correlation coefficient is a measure of the error of a least-squares fit that ranges between -1 and 1, with $r=\pm 1$ indicating a perfect positive (negative) linear relationship between the data and the model and 0 indicating no linear relationship between the data and the model. A good model for the RMSD (t_{ref},t) data of our sample protein will therefore have a correlation coefficient approaching one. For our complete set of reference structures, correlation coefficients averaged 0.886 ± 0.027. The fit of the KWW model to a representative trajectory is shown in Fig. 2.

III. METHODS

A sample protein was chosen to generate representative molecular dynamics trajectories. Bovine pancreatic trypsin inhibitor (BPTI) was selected for its small size (58 amino acids), which permits extended simulation times in reasonable real-time durations. A schematic view of BPTI can be seen in Fig. 3. A mutant of BPTI with an altered binding loop, for which the XRD structure can be found in the Protein Data Bank with identifier 1QLQ [4,30], was simulated using the GROMACS molecular dynamics package, version

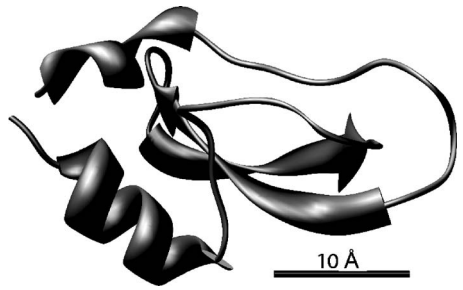

FIG. 3. Bovine pancreatic trypsin inhibitor was chosen because it is a small (58 amino acids) protein, which has been widely studied computationally. Thus, large simulated times could be considered for reasonable real-time durations. This image was produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081) [36].

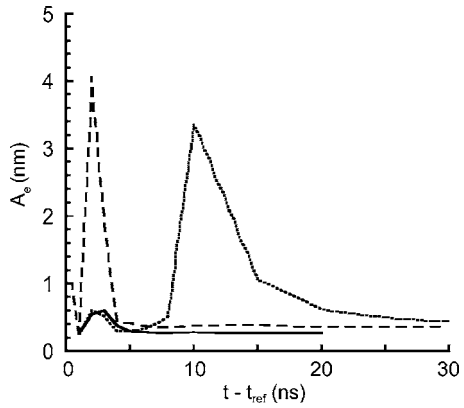

FIG. 4. For simulated BPTI at $T=300$ K, parameters of the KWW model of the RMSD referenced to $t_{\mathrm{ref}}=0$ ns (dashed line) and $t_{\mathrm{ref}}=11$ ns (solid line) became independent of simulation duration when fitting Eq. (3) to only 5 ns of the trajectory ($t=0-5$ ns or $t=11-16$ ns, respectively). However, the RMSD referenced to $t=2$ ns (dotted line) required 20 ns of the trajectory ($t=2-22$ ns) before the fitting parameters A_e, τ_e, and β stabilized to unique values.

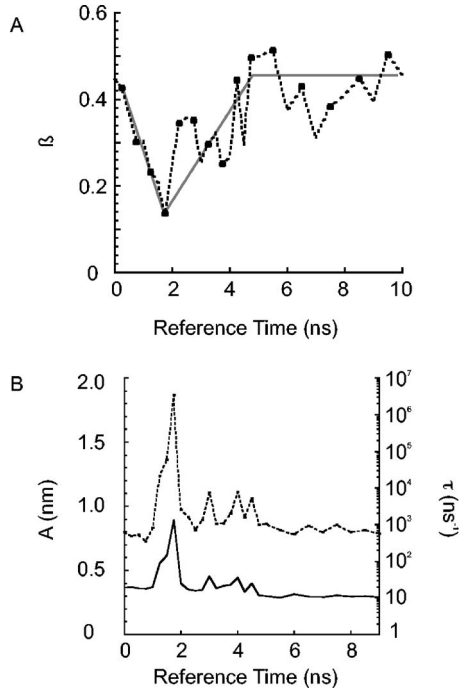

FIG. 5. Consideration of BPTI fluctuations via the KWW model. (A) The complexity parameter β decreases before increasing to a steady value of about 0.43 at 5 ns. (B) The exponential prefactor A_e (solid line) and the effective time constant τ_e (dashed line) fluctuated before attaining steady values at 5 ns. These parameters were obtained at each reference time t_{ref} by fitting Eq. (3) to the 20 ns of the trajectory following each t_{ref}, as this was the maximum sampling duration required for the KWW model of the RMSD to attain a stable, unique functional form.

3.2.1 [31,32]. The protein was solvated in a cubic box of edge length 5.36 nm with 4640 simple point charge (SPC216) water molecules. Steepest descents minimization was used on the XRD structure to reduce the maximum force in the system to 2000 kJ mol^{-1} nm^{-1}. The solvated, steepest descents minimized structure represents the reference structure at $t_{ref}=0$. After minimization, 40 ns of unconstrained molecular-dynamics simulation was carried out in two sequential 20 ns simulations. The time step was 2 fs, using Berendsen temperature coupling with $\tau_T=0.1$ ps to maintain the temperature at 300 K, and isotropic pressure coupling with $\tau_P=0.5$ ps to maintain the pressure at 1 atm [33]. LINCS constraints were used on all protein covalent bonds to maintain constant bond length [34], and the SETTLE algorithm was used to constrain the intramolecular water bonds to their equilibrium length [35]. By eliminating very high frequency vibrations, these constraints allowed for a longer time step than is otherwise possible. Structures were recorded every 1 ps. Each recorded structure was subsequently fit to the initial structure using a least-squares fit, allowing only overall rotation and translation to remove the effects of global system movement. Twenty days of CPU time on a single cluster node with a Intel Xeon 3.20 GHz processor were necessary to carry out these simulations.

A second system, streptavidin-biotin, was chosen to test the method on a larger and more rigid protein. A tetramer of the structure 1STP [37] from the PDB was simulated according to the procedure above, with the following changes: the cubic box had edge length 8.590 nm, 18 533 water molecules were added along with 50 sodium ions and 42 chlorine atoms (for charge neutrality and an approximation of physiological conditions), and the total length of the simulation was 95 ns, for which 6 weeks of CPU time on 14 of the same cluster nodes was used.

Using analysis tools included with GROMACS, we calculated the RMSD of each structure in the trajectory from reference structures spaced in simulation time 0.5 ns apart from 0 to 6 ns, 1 ns apart from 6 to 15 ns, and then 5 ns apart until 40 ns, starting with $t_{ref}=0$. This resulted in a total of 26 RMSD time series, each calculated from a different reference structure. The intervals between reference structures were not kept constant because greater precision was desired in time regions where the fitting parameters were changing rapidly. The choice of the interval between reference structures will determine the accuracy with which the equilibration time can be determined: the shorter the interval between reference structures, the more precise the equilibration time will be. All fits of Eq. (3) to the RMSD time series of the protein were performed in GRACE version 5.1.14, using the nonlinear curve-fitting feature, which is an implementation of the Levenberg-Marquardt algorithm. All parameters (A_e,τ_e,β) were initially set to one, and then the fit algorithm was repeated until there was no change in the first five significant digits of each parameter.

Our approach for determining when the nonequilibrium displacements from the initial structure have dissipated is comprised of two parts. First, the RMSD of the protein referenced to each chosen intermediate structure must be accurately fit to the KWW model of Eq. (3). Then, the fitting parameters A_e, τ_e, and β from each reference structure are considered as a function of t_{ref}, the time step in the trajectory corresponding to the reference structure. The end of the equilibration portion of the trajectory can be unambiguously determined as the time at which A_e, τ_e, and β attain steady-state values. Physically, this corresponds to the protein settling into a steady set of modes—the nonequilibrium modes that arise due to the initial structure have dissipated, and have been replaced by modes corresponding to an exploration of an energetic minimum assumed to be included in equilibrium phase space.

The KWW fitting parameters are physical properties of the system, and as such are always independent of the method used to measure them. It is important to note that, like any value derived from a simulation, the KWW fitting parameters must be based on sufficient sampling from the trajectory such that the fit is no longer a function of sampling duration. When examining the RMSD of the protein referenced to different intermediate structures (i.e., $t_{ref}\neq 0$), we found that while for some reference structures only 5 ns of the trajectory following t_{ref} was sufficient for the fit parameters to become independent of simulation duration, other reference structures required up to 20 ns of the trajectory to be included before the fit parameters stabilized. Figure 4 shows three cases: two where the fitting parameters stabilized after sampling 5 ns of the trajectory, and one where the fitting parameters did not stabilize until 20 ns of the trajectory was sampled. We thus used 20 ns of the trajectory following each reference structure when determining A_e, τ_e, and β by fitting Eq. (3) to each RMSD time series. It is paramount to our method that the necessary sampling duration for each reference structure be determined and the maximum duration among these reference structures be used to establish unique values of A_e, τ_e, and β for all RMSD time series; otherwise, the fitting parameters are meaningless.

IV. RESULTS AND DISCUSSION

The goal of the equilibration stage of an MD simulation is to separate artifactual motions due to the initial nonequilibrium structure from the rest of the trajectory. Evaluating the RMSD of the protein's internal fluctuations in terms of the KWW model can accomplish this. Initially, protein motion is dominated by nonequilibrium displacements away from the high-energy portions of the initial structure. As the simulation continues, the driving force for these displacements decreases, and the number of modes (which would each be represented by a separate harmonic oscillator in the KWW model) contributing to the RMSD that can be attributed to initial conditions also decreases exponentially with time. At the same time, the equilibrium fluctuations will not be present in the initial time steps of the simulation, because velocities are assigned randomly from a Gaussian distribution. As the simulation progresses, velocities of atoms can become coupled, creating modes that correspond to equilibrium fluctuations. The number of modes that can be attributed to these equilibrium fluctuations will reach a constant value when the system has equilibrated.

These two processes—the decrease in displacements due to initial conditions and the increase in displacements due to

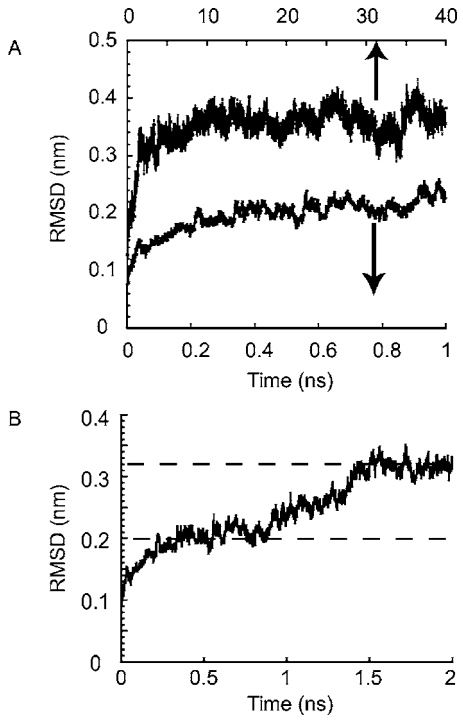

FIG. 6. Identification of an equilibrated BPTI structure via the RMSD plateau method. (A) Examining the trajectory between 0 and 1 ns (bottom) gives an equilibration time of ~200 ps, but examining the trajectory between 0 and 40 ns (top) gives an equilibration time of ~10 ns. (B) Ambiguity arises from examining the trajectory between 0 and 2 ns due to a double plateau, one at RMSD ~0.2 nm and the other at RMSD ~0.33 nm (dashed lines). It is not obvious which, if either, corresponds to the end of nonequilibrium fluctuations.

equilibrium fluctuations—will cause the total number of modes to increase to a peak and then fall to a constant value. This rise and fall should be reflected by changes in the fitting parameters of the KWW model. However, once the protein is at least locally equilibrated, we expect the internal displacements to occupy a single set of modes within the equilibrium phase space, and this will be reflected by the KWW fitting parameters attaining steady-state values. Because β scales inversely with the complexity of the system, when the number of modes increases, β will decrease, and vice versa. We expect, therefore, to observe a fall followed by a rise in β, accompanied by fluctuations in A_e and τ_e, before all three parameters attain steady-state values.

This is observed in our simulation of BPTI, as shown in Fig. 5: β falls from a value of 0.4 to a value of 0.15 at $0 \leqslant t_{\mathrm{ref}} \leqslant 2$ ns, and rises to a steady-state value of about 0.43 at $t_{\mathrm{ref}} \approx 5$ ns; A_e and τ_e undergo large fluctuations before settling to steady-state values at $t_{\mathrm{ref}} \approx 5$ ns. We have therefore identified the equilibration stage of the trajectory as occurring over the first 5 ns of simulated time. At reference times less than 5 ns, the dissipation of nonequilibrium motions due to the initial structure and the growth of equilibrium fluctuation modes caused the KWW parameters to fluctuate. Once the protein has begun to sample its equilibrium phase space (after $t_{\mathrm{ref}}=5$ ns), the protein is represented by a constant set of harmonic oscillators.

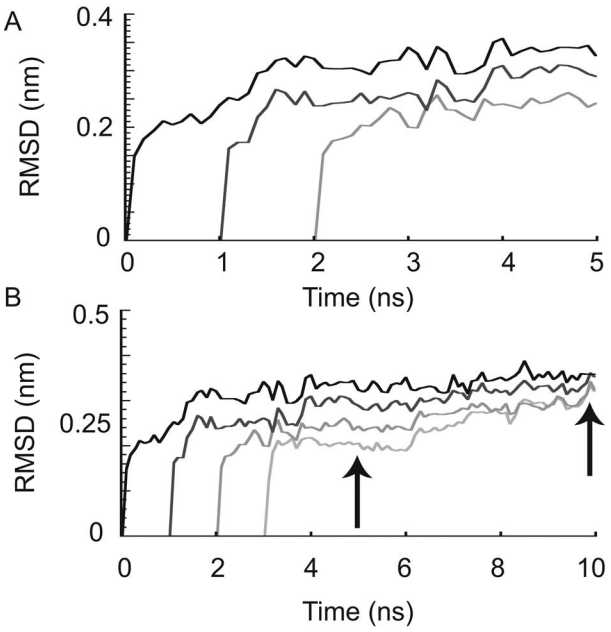

FIG. 7. Identification of an equilibrated BPTI structure via the Stella method of sharp decreases in RMSD plateau values. (A) Here, two different reference structures ($t_{\mathrm{ref}}=1$ and 2 ns) both give distinct drops in the plateau value, leading to ambiguity in determining the proper equilibration time. Black: RMSD ($t_{\mathrm{ref}}=0$ ns, t); dark gray: RMSD ($t_{\mathrm{ref}}=1$ ns, t); medium gray: RMSD ($t_{\mathrm{ref}}=2$ ns, t). (B) While the RMSD plateau values of all the reference structures are distinct at $t=5$ ns (left arrow), differences decrease when considering fluctuations to $t=10$ ns (right arrow). Examining only part of the data shown, e.g., from 0–5 ns, would lead to choosing an equilibration time from this trajectory, but when longer simulations are considered it becomes unclear as to whether equilibration has occurred according to this method. Black: RMSD ($t_{\mathrm{ref}}=0$ ns, t); dark gray: RMSD ($t_{\mathrm{ref}}=1$ ns, t); medium gray: RMSD ($t_{\mathrm{ref}}=2$ ns, t); light gray: RMSD ($t_{\mathrm{ref}}=3$ ns, t).

A. Evaluation of and comparison to current equilibration protocols

As noted, the end of artifactual displacements due to the initial structure should be unambiguously identifiable, with no dependence on simulation duration. We found determination of the equilibration time via the simple RMSD and Stella methods to be ambiguous for BPTI (and for

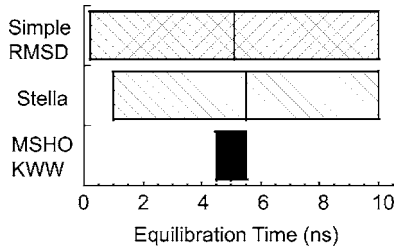

FIG. 8. The simple RMSD-based method and the Stella method both result in a considerable range in potential equilibration times and thus in potential structures for solvated BPTI, reflecting the ambiguity present in those methods. Our KWW-based method gives a comparatively unambiguous equilibration time of 4.5–5.5 ns.

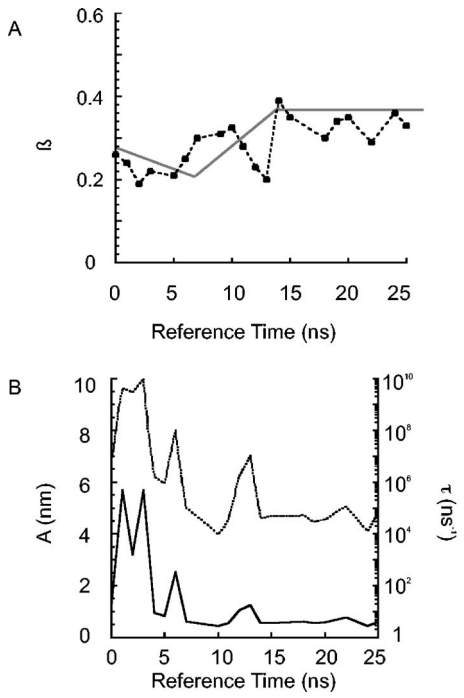

FIG. 9. Consideration of fluctuations in streptavidin alpha-carbon atoms via the KWW model. (A) The complexity parameter β fluctuates before attaining to a steady value of about 0.35 at 15 ns. (B) The exponential prefactor A_e (solid line) and the effective time constant τ_e (dashed line) fluctuated before attaining steady values at 15 ns. These parameters were obtained at each reference time t_{ref} by fitting Eq. (3) to the trajectory following each t_{ref}, checking at each t_{ref} that the parameters did not depend on the duration of trajectory used.

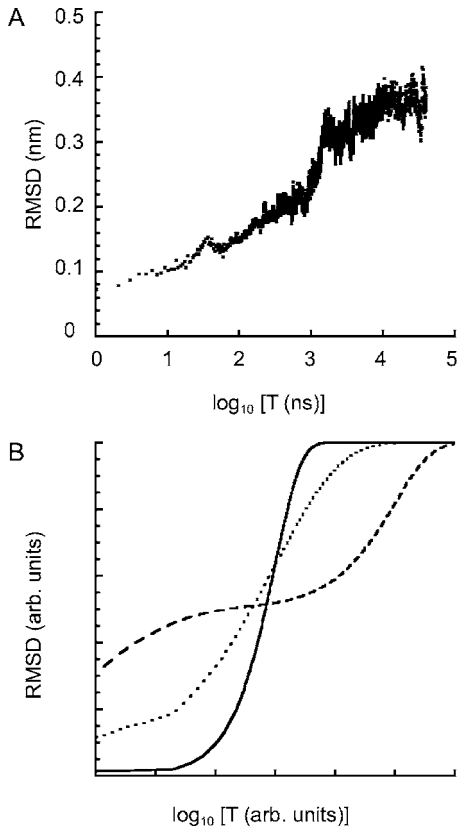

FIG. 10. Determining the applicability of the KWW approximation. (A) The RMSD of our computationally modeled protein (BPTI) on a logarithmic time scale (B) RMSD of a single simple harmonic oscillator (solid): KWW model is applicable, with the number of harmonic oscillators equal to one. RMSD of a system with two discrete narrow peaks in the time-constant distribution (dashed): KWW model is not applicable. RMSD of a system with broad overlapping peaks in the time-constant distribution (dotted): KWW model is applicable. The RMSD of our computationally modeled protein BPTI most resembles the dotted line, indicating that the time-constant distribution likely consists of broad, overlapping peaks. Therefore, we conclude that the KWW approximation is applicable to our system.

streptavidin-biotin, which is a larger, more rigid protein: see Fig. 9. Using the simple RMSD method, the starting time of the plateau was typically difficult to identify objectively and depended on the duration of simulation examined, as shown in Fig. 6.

Often, when examining longer durations of the trajectory, our estimate of the equilibration time moved to later times. Depending on our judgment and the duration of trajectory examined, equilibration times ranged from 200 ps to 10 ns. Using the Stella method, we noted multiple large drops in the RMSD plateau value, as well as large drops that actually decreased in magnitude at later times, as shown in Fig. 7.

Depending on how large of a drop constituted a "sharp decrease," on how frequently intermediate structures were sampled, and on the duration of the simulation examined, equilibration times ranged from 1 to 10 ns. In contrast, our method provided a comparatively unambiguous equilibration time of 4.5 to 5.5 ns, provided that the KWW parameters were determined accurately, as discussed in Sec. III. These ranges in equilibration times are illustrated in Fig. 8, which shows that our method results in a significantly narrower range and a more objective choice of the end of the equilibration stage of the trajectory.

B. Application to other protein systems

To further test our method, we applied it to a 95 ns simulation of the streptavidin-biotin system (PDB 1STP). In this case, we calculated the RMSD of only the alpha-carbon atoms for computational efficiency. We observed a similar trend in the KWW fitting parameters, as shown in Fig. 9: after large initial fluctuations, steady-state values were attained after 15 ns of simulation. As in the fluctuations of the BPTI, it is reasonable to interpret this transition as one in which nonequilibrium fluctuations from the initial structure cause the KWW parameters to fluctuate until the protein begins to sample equilibrium phase space.

V. CONCLUSIONS

We have demonstrated important limitations of existing equilibration methods and proposed an equilibration method based on the stretched exponential function. Because the stretched exponential is based on a summation of contributions from independent harmonic oscillators, our method implicitly assumes multiple, decoupled relaxation processes in

order to separate nonequilibrium and equilibrium fluctuations. The initial nonequilibrium fluctuations could potentially play a spurious role in calculating quantities related to rare events. Thus, computational models and simulations that require dissipation of the initial nonequilibrium fluctuations will benefit from this approach. In particular, this method will benefit computational determination of quantities that are highly sensitive to protein structure, such as steered MD simulations of ligand-receptor unbinding forces or free energies of protein docking.

ACKNOWLEDGMENTS

We have benefited from computational resources funded by the National Science Foundation (Grant No. IMR-0414849). E.B.W. gratefully acknowledges financial support from the Massachusetts Institute of Technology Provost's Office and Ida M. Green.

APPENDIX

To ensure that the KWW approximation [Eq. (3)] is valid for our simulation of the solvated protein BPTI, we examined the RMSD on a logarithmic time scale to clarify the short-time behavior as suggested by Apitz *et al.* [22]. There was no indication that the time constant distribution consisted of multiple distinct peaks. Rather, the data corresponded with a time constant distribution described by broad, overlapping peaks (see Fig. 10), indicating that the KWW approximation is acceptable in our test system. This conclusion is supported by similar results in the RMSD for biotin-streptavidin.

[1] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, 1987).
[2] M. Karplus and J. A. McCammon, Nat. Struct. Biol. **9**, 646 (2002).
[3] W. F. van Gunsteren and A. E. Mark, Eur. J. Biochem. **204**, 947 (1992).
[4] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[5] J. N. S. Evans, *Biomolecular NMR Spectroscopy* (Oxford University Press, New York, 1995).
[6] L. J. Smith, X. Daura, and W. F. van Gunsteren, Proteins: Struct., Funct., Bioinf. **48**, 487 (2002).
[7] L. S. Caves, J. D. Evanseck, and M. Karplus, Protein Sci. **7**, 649 (1998).
[8] B. Hess, Phys. Rev. E **65**, 031910 (2002).
[9] J. B. Clarage, T. Romo, B. K. Andrews, B. M. Pettitt, and J. George N. Phillips, Proc. Natl. Acad. Sci. U.S.A. **92**, 3288 (1995).
[10] E. B. Walton and K. J. VanVliet (unpublished).
[11] V. Daggett and M. Levitt, Annu. Rev. Biophys. Biomol. Struct. **22**, 353 (1993).
[12] L. Stella and S. Melchionna, J. Chem. Phys. **109**, 10115 (1998).
[13] D. van der Spoel, E. Lindahl, B. Hess, A. R. van Buren, E. Apol, P. J. Meulenhoff, D. P. Tieleman, A. L. T. M. Sijbers, K. A. Feenstra, and R. van Drunen *et al.*, *GROMACS User Manual version 3.2* (2004); URL www.gromacs.org
[14] C. Baysal and A. R. Atilgan, Biophys. J. **83**, 699 (2002).
[15] B. R. Brooks and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **80**, 6571 (1983).
[16] N. Go, T. Noguti, and T. Nishikawa, Proc. Natl. Acad. Sci. U.S.A. **80**, 3696 (1983).
[17] M. Levitt, C. Sander, and P. S. Stern, J. Mol. Biol. **181**, 423 (1985).
[18] T. Noguti and N. Go, Nature (London) **296**, 776 (1982).
[19] S. Hayward and N. Go, Annu. Rev. Phys. Chem. **46**, 223 (1995).
[20] K. Kohlrausch, Pogg. Ann. Phys. Chem. **91**, 56 (1854).
[21] G. Williams and D. Watts, Trans. Faraday Soc. **66**, 80 (1970).
[22] D. Apitz and P. M. Johansen, J. Appl. Phys. **97**, 063507 (2005).
[23] S. Hayward, A. Kitao, F. Hirata, and N. Go, J. Mol. Biol. **234**, 1207 (1993).
[24] A. Kitao, F. Hirata, and N. Go, Chem. Phys. **158**, 447 (1991).
[25] P. Richard, M. Nicodemi, R. Delannay, P. Ribiere, and D. Bideau, Nat. Mater. **4**, 121 (2005).
[26] E. Zaccarelli, G. Foffi, F. Sciortino, P. Tartaglia, and K. A. Dawson, Europhys. Lett. **55**, 157 (2001).
[27] K. Kremer and G. S. Grest, J. Chem. Phys. **92**, 5057 (1990).
[28] G. V. Vijayadamodar and A. Nitzan, J. Chem. Phys. **103**, 2169 (1995).
[29] It has been brought to our attention that a preprint using the KWW model to analyze the temperature dependence of relaxation of carbonmonoxymyoglobin is available on the arXiv preprint server at http://mentor.lanl.gov/PS_cache/physics/pdf/0411/0411017.pdf
[30] H. Czapinska, J. Otlewski, S. Krzywda, G. Sheldrick, and M. Jaskolski, J. Mol. Biol. **295**, 1237 (1999).
[31] H. Berendsen, D. van der Spoel, and R. van Drunen, Comput. Phys. Commun. **91**, 43 (1995).
[32] E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. **7**, 306 (2001).
[33] H. J. C. Berendsen, J. P. M. Postma, A. DiNole, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).
[34] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, J. Comput. Chem. **18**, 1463 (1997).
[35] S. Miyamoto and P. A. Kollman, J. Comput. Chem. **13**, 952 (1992).
[36] E. Pettersen, T. Goddard, C. Huang, G. Couch, D. Greenblatt, E. Meng, and T. Ferrin, J. Comput. Chem. **25**, 1605 (2004).
[37] P. C. Weber, D. Ohlendorf, J. Wendoloski, and F. Salemme, Science **243**, 85 (1989).