JCTC Journal of Chemical Theory and Computation

Protein Folding Simulations Combining Self-Guided Langevin Dynamics and Temperature-Based Replica Exchange

Michael S. Lee*,†,‡,§ and Mark A. Olson§

Computational Sciences and Engineering Branch, U.S. Army Research Laboratory, Aberdeen Proving Ground, Maryland 21005, Biotechnology High Performance Computing Software Applications Institute, U.S. Army Medical Research and Materiel Command, Frederick, Maryland 21702, and Department of Cell Biology and Biochemistry, U.S. Army Medical Research Institute of Infectious Diseases, Frederick, Maryland 21702

Received February 2, 2010

Abstract: Computer simulations are increasingly being used to predict thermodynamic observables for folding small proteins. Key to continued progress in this area is the development of algorithms that accelerate conformational sampling. Temperature-based replica exchange (ReX) is a commonly used protocol whereby simulations at several temperatures are simultaneously performed and temperatures are exchanged between simulations via a Metropolis criterion. Another method, self-guided Langevin dynamics (SGLD), expedites conformational sampling by accelerating low-frequency, large-scale motions through the addition of an ad hoc momentum memory term. In this work, we combined these two complementary techniques and compared the results against conventional ReX formulations of molecular dynamics (MD) and Langevin dynamics (LD) simulations for the prediction of thermodynamic folding observables of the Trp-cage mini-protein. All simulations were performed with CHARMM using the PARAM22+CMAP force field and the generalized Born molecular volume implicit solvent model. While SGLD-ReX does not fold up the protein significantly faster than the two conventional ReX approaches, there is some evidence that the method improves sampling convergence by reducing topological folding barriers between energetically similar near-native states. Unlike MD-ReX and LD-ReX, SGLD-ReX predicts melting temperatures, heat capacity curves, and folding free energies that are closer in agreement to the experimental observations. However, this favorable result may be due to distortions of the relative free energies of the folded and unfolded conformational basins caused by the ad hoc force term in the SGLD model.

Introduction

Molecular dynamics (MD) simulations of small proteins provide insight into the mechanisms and thermodynamics of protein folding. The most traditional protocol is to simulate at a fixed temperature the folding/unfolding of a protein immersed in an explicit solvent. So far, only the smallest proteins have been folded in this way.[1] Sampling at the all-atom level is slow due to the presence of multiple small minima on the energy landscape, namely, kinetic traps, which lead to a "mountainous" and/or "pebbly" energy surface.[2] To combat this problem, several enhanced sampling methods have been developed and tested over the past decade, including replica exchange,[3] accelerated molecular dynamics,[4] self-guided molecular dynamics,[5] potential smoothing,[6] locally enhanced sampling,[7] and resolution reduction (e.g., implicit solvent models and lattice models).[8,9]

Temperature is a commonly used parameter to accelerate conformational motions in proteins.[3,10] One of the most popular approaches in recent years is temperature-based

* Corresponding author e-mail: michael.scott.lee@us.army.mil.
† U.S. Army Research Laboratory.
‡ U.S. Army Medical Research and Materiel Command.
§ U.S. Army Medical Research Institute of Infectious Diseases.

10.1021/ct100062b
Published on Web 07/26/2010

replica exchange (T-ReX),[3,11] which involves simultaneously simulating multiple fixed-temperature windows. The temperature values are assigned at exponentially spaced or adaptively spaced[12] intervals between room temperature and a user-specified higher temperature. At regular time intervals, the simulation temperatures are exchanged between neighboring pairs of thermal windows according to a Metropolis criteron.[13] The method effectively percolates conformations up and down a ladder of temperatures depending on their relative energies.

Thinking beyond sampling at various temperatures, the key to characterizing the thermodynamics of folding is to sample the multiple major conformational basins that exist between the unfolded and native states. Major conformational basins are often separated by low-frequency modes of the protein. By definition, traditional dynamics methods require a relatively long time to traverse these low-frequency modes. As a potential solution to this dilemma, the recently developed self-guided Langevin dynamics (SGLD) accelerates the lowest frequency modes of a system by the addition of an ad hoc atomic force term proportional to the running average of individual atomic momenta over a short time interval (e.g., 0.1−1 ps).[5] In this force term, high-frequency motions tend to cancel out, while low-frequency motions, which are typically unidirectional over the short averaging time, tend to be additive, and thus have the effect of adding an external boost along the low-frequency degrees freedom. Excess energy generated by the SGLD force term is removed via a fixed-energy constraint term. While the addition of an ad hoc force term may cause deviations from canonical ensemble behavior, tests to-date indicate many thermodynamic observables are not significantly altered.[5,14]

In the original paper describing SGLD, an α-helix peptide, $(AAQAA)_3$, was folded up using a simple distance-dependent dielectric electrostatic function.[5] Impressively, the SGLD simulation was estimated to be at least 65 times more efficient than standard LD in reaching the apparent lowest energy state of the α-helical conformation. Using a similar method, self-guided molecular dynamics, Wen et al. reported improved sampling versus conventional MD for the folding of a nontraditional peptide, $\beta\beta\alpha 1$, and the villin headpiece using the AMBER force field, parm94, with an analytical Poisson implicit solvation model.[15] SGLD has also been applied to various biophysical problems including ionization equilibria[16,17] and water content in the interior of proteins.[18]

In this work, we combined the merits of two disparate techniques, SGLD and temperature-based ReX, and asked if their synergy could provide further sampling enhancements versus traditional MD-ReX and LD-ReX in the prediction of thermodynamic observables. Our test system is the commonly studied 20-residue Trp-cage mini-protein "5b".[19−26] Besides its small size, the Trp-cage mini-protein is an exemplary model system because it contains many key structural elements found in larger proteins. Our evaluation criteria of MD-Rex, LD-ReX, and SGLD-ReX included how quickly the native basin was reached starting from the unfolded state and a comparison of melting temperatures, heat capacity curves, and folding free energies to recent experimental observations. In addition, we compared the free-energy landscapes generated by each method to better understand the differences in the predicted observables. Finally, we examined several criteria to determine to what extent the three ReX methods deviated from theoretical canonical ensemble behavior.

Methods

Self-Guided Langevin Dynamics. The self-guided Langevin dynamics method, developed by Wu and Brooks,[5] enhances conformational sampling by accelerating low-frequency modes through the use of an ad hoc time-averaged momentum term. The algorithm was preceded by the self-guided molecular dynamics method which, instead, used a time-averaged force term.[27] The running average of the momentum over a short-period simulation time is added back as an external force to the simulation system. This term has the effect of accelerating low-frequency motions, because the modes that are slower than the averaging time are expected to be additive. In principle, compared to MD and Langevin dynamics, SGLD should increase the rate of hopping between conformational basins which might include the lowest energy topology and/or the experimentally observed native conformation. The main drawbacks to this method, however, are that SGLD no longer provides rigorous canonical ensemble sampling and that kinetics predictions are no longer comparable to true observables.[5] Furthermore, it is not yet clear how one would recover rigorous ensemble averages by reweighting the population densities[28] resulting from a SGLD simulation.

As outlined and derived in the original work, the SGLD method uses the following equation of motion:

$$\dot{\mathbf{p}}_i = \mathbf{f}_i - \gamma_i \mathbf{p}_i + \mathbf{R}_i + \lambda \mathbf{g}_i \tag{1}$$

where the rate of change of the momentum of a particle i, $\dot{\mathbf{p}}_i$, is a function of the force on the particle, $\mathbf{f}_i$, a friction constant, γ_i, the momentum itself,$\mathbf{p}_i$, a random force, $\mathbf{R}_i$, and a memory function, $\mathbf{g}_i$, which is scaled by a guiding factor, λ. The memory function, $\mathbf{g}_i$, is defined by the moving average of the momentum of the system over an interval of time, L:

$$\mathbf{g}_i = \gamma_i \langle \mathbf{p}_i \rangle_L \tag{2}$$

Inevitably, the addition of a memory term $\mathbf{g}_i$ to the atomic forces will result in a lack of energy conservation, typically heating up the system. Therefore, an energy conservation term is applied to the equation of motion which leads to uniform scaling of the atomic velocities at each time step (see Wu and Brooks for details[5]). The result of this conservation mechanism is that while the diffusivities of the lower frequency modes are enhanced due to the memory function, the opposite effect occurs for the higher frequency modes.[5]

Temperature-Based Replica Exchange. The replica-exchange protocol involves performing simultaneous simulations over a range of temperatures and/or biasing potentials.[11] In this work, only a range of simulation temperatures is considered. Each simulation, a, exchanges its temperature with another simulation, b, if $\Delta_{ab} < 0$, or $\exp(-\Delta_{ab})$ is greater

than a random number with uniform distribution, $r \in (0,1)$, where Δ_{ab} is defined as

$$\Delta_{ab} = (U_b - U_a)(1/k_B T_a - 1/k_B T_b) \quad (3)$$

where U_i is a defined energy measure of the replica-exchange simulation client i corresponding to temperature T_i and k_B is Boltzmann's constant. The ReX procedure can be thought of as an autonomous heating and cooling procedure, whereby, to a large extent, lower energy conformations settle into lower temperature windows and vice versa.

Simulation Setup. In this work, we used the PARAM22[29] force field with the CMAP backbone dihedral cross-term extension[30] and the generalized Born (GB) implicit solvent model, GBMV2.[31] GBMV2 is one of the more accurate implicit solvent models currently available, as it correctly mimics the Poisson solvation energy using a molecular surface-based dielectric boundary.[32] Implicit solvent models greatly reduce the number of simulation degrees of freedom compared to explicit solvent. This has two key benefits. First, in the context of ReX simulations, a reduced magnitude of potential energies leads to a smaller number of temperature clients necessary to span the desired temperature range to ensure frequent Metropolis exchanges of among clients. Also, implicit solvent accelerates conformational sampling in its own right by eliminating the diffusive reorientation of solvent molecules upon changes in protein conformation.

An integration time step of 2 fs was used as the SHAKE algorithm[33] is applied to fix all covalent bonds with hydrogen atoms. Nonbonded electrostatics and van der Waals interactions were truncated smoothly from 12 to 14 Å. For the MD simulations, a Nose-Hoover thermostat was used with a temperature coupling constant of 50 kcal/s^2. For the Langevin dynamics and self-guided Langevin dynamics simulations, the friction constant, γ, was set to 1 ps^{-1} for all heavy atoms. The SGLD guiding factor, λ, was set to 1, while the averaging time, τ, was set to 1 ps. These two values were arrived at on the basis of a compromise between sampling efficiency and preserving the backbone $\varphi-\psi$ free energy landscape of the alanine tripeptide (results not shown). In particular, smaller values of λ lead to diminished sampling benefits of the self-guided formalism, while larger values lead to distortions in the $\varphi-\psi$ free energy map.

Replica exchange was performed using the MMTSB[34] script *aarex.pl* which is a front-end to the CHARMM molecular dynamics package (version c33b2).[35] The replica exchange protocol utilized 16 simulation clients, with temperatures exponentially spaced between 298 and 500 K. Default GBMV2 parameters were used with the exception of the β value, which was set to -12 to improve energy conservation.[36] A surface tension value of 0.00542 kcal/(mol·Å^2) was used for the solvent accessible surface area nonpolar solvation term.[37]

The original Trp-cage protein (dubbed by its inventors, "5b") was used in this work. It has the sequence "NLYIQWLKDGGPSSGRPPS" and was structurally determined by NMR (PDB ID: 1L2Y).[19] We performed six simulations in total: MD-ReX, LD-ReX, and SGLD-ReX each starting from the native (NMR conformer 1) and unfolded trans conformations. The simulations are heretofore labeled as method/starting structure, e.g., SGLD-ReX/native. The trans simulations were each run for 100 ns. The native simulations were each run for 200 ns.

Evaluation Metrics. We compared predicted structural and thermodynamic properties of protein folding between the three simulation methods and experimental observation. Our measure of model quality was root-mean-squared deviation of the α-carbon trace to the native NMR conformer 1 structure (C_α rmsd). We predicted the heat capacity as a function of temperature, melting temperature, and folding free energy of the Trp-cage, all of which can be compared directly to experiment.[38]

Heat capacity was calculated in two ways. The "instantaneous" heat capacity, C_v^1, was computed as

$$C_v^1(T) = \frac{{\sigma_U}^2}{k_B T^2} = \frac{\langle U^2 \rangle - \langle U \rangle^2}{k_B T^2} \quad (4)$$

where U is the potential energy derived from WHAM analysis (described below) and T is the temperature. The distinction of constant volume (subscript v) vs constant pressure (subscript p) is meaningless in the context of implicit solvent simulations. Useful for canonical ensemble tests (described below), an alternative heat capacity formula does not require WHAM analysis, but is only defined at the midpoints of adjacent replica-exchange client temperatures. It is computed as a finite difference of the potentials of two contiguous temperature clients:

$$C_v^2\left(\frac{T_a + T_b}{2}\right) = \frac{\langle U_b \rangle - \langle U_a \rangle}{T_b - T_a} \quad (5)$$

We estimated the melting temperature, T_m, using three approaches. First, T_m^1 is the location of the maximum value of the computed $C_v^1(T)$ function. Second, T_m^2 is computed as the point where the derivative of the average rmsd as a function of temperature is maximum, which is roughly the inflection point. Finally, T_m^3 is found by iteratively searching for the temperature at which the free-energy difference between the native and unfolded state, $\Delta\Delta G_{fold}(T)$, is approximately equal to zero. Free energy of folding as a function of temperature, $\Delta\Delta G_{fold}(T)$, was computed as

$$\begin{aligned} \Delta\Delta G_{fold}(T) &= -k_B T \ln\left(\frac{\rho_{fold}(T)}{\rho_{unfold}(T)}\right) \\ &= -k_B T \ln\left(\frac{\int_{rmsd<rmsd_{fold}} \rho(T)\, dV}{\int_{rmsd>rmsd_{fold}} \rho(T)\, dV}\right) \end{aligned} \quad (6)$$

where the population density, ρ, as a function of C_α rmsd is integrated over the rmsd-delineated "native" and "unfolded" domains and dV is the volume element. The boundary dividing the folded and unfolded regions, $rmsd_{fold}$, is somewhat arbitrary given that experimental rmsd values cannot be observed. Nonetheless, $rmsd_{fold}$ can be deduced as the point where the population density is equal in both domains (i.e., $\Delta\Delta G_{fold}(T) = 0$) when computed at a predicted melting temperature, T_m, using either the T_m^1 or T_m^2 definitions. On the basis of this analysis and inspection of the free-energy

landscapes of all simulation methods as a function of rmsd, $rmsd_{fold}$ was chosen to be 3.4 Å. This definition liberally includes not only the native-like basin (rmsd ~ 1.0 Å) but several near-native compactly folded basins.

The heat capacity, C_v^1, folding free energies, and free-energy landscapes were derived from the multidimensional temperature-based-weighted histogram method WHAM[13] algorithm applied to the ReX simulation data.[39] The following dimensions were binned in the WHAM algorithm: potential energy, C_α rmsd, and radius of gyration, R_g. Structural representatives on the rmsd vs R_g landscape were selected visually and verified by rmsd analysis against the entire set of structures at 270 K.

Because of the ad hoc force term, SGLD may not strictly adhere to rigorous statistical mechanics and thus fail to produce a canonical ensemble.[5] We wanted to ascertain to what extent, if any, does SGLD deviate from canonical behavior, especially in combination with T-ReX. There are several measures one can use to assess departure from theoretical canonical ensemble behavior.[40] We evaluated four criteria in this work. First, we compared the actual average temperatures of the replica-exchange client simulations vs the temperatures specified in the input. Second, the ratio of heat capacity values, g, derived from a single temperature vs two temperatures was calculated as

$$g(T) = C_v^2(T)/C_v^1(T) \tag{7}$$

In a hypothetical canonical ensemble of a single state, the value of g should be 1.[40] The skew, S, of the potential energy distribution for a fixed temperature client is defined as

$$S = \frac{\langle (U - \langle U \rangle)^3 \rangle}{\langle (U - \langle U \rangle)^2 \rangle^{3/2}} \tag{8}$$

The skew for a perfect canonical ensemble of a single state should be zero.[40] In other words, the canonical energy distribution far away from a transition temperature should be strictly Gaussian. Our final metric was the kinetic energies or "temperatures" of the individual normal modes of the system. This measure, as far we are aware, has not been reported elsewhere. The question addressed in this metric is whether SGLD-ReX overheats low-frequency modes and cools high-frequency modes to retain the correct total macroscopic kinetic energy (i.e., temperature). If we define A_{ni} to be the ith Cartesian degree of freedom for the nth normal mode vector, and v_i is the velocity of the ith degree of freedom, then the temperature of the nth normal mode, T_n is

$$T_n = \frac{1}{2k_B}\left(\sum_i A_{ni}\sqrt{m_i}v_i\right)^2 \tag{9}$$

Velocities were obtained at 0.1 ps intervals from independent 1 ns simulations using the MD, LD, and SGLD protocols. The Cartesian normal mode matrix, **A**, was computed with the *vibran* module in CHARMM after optimizing NMR conformer 1 with 2000 steps of adopted-basis Newton−Raphson minimization.

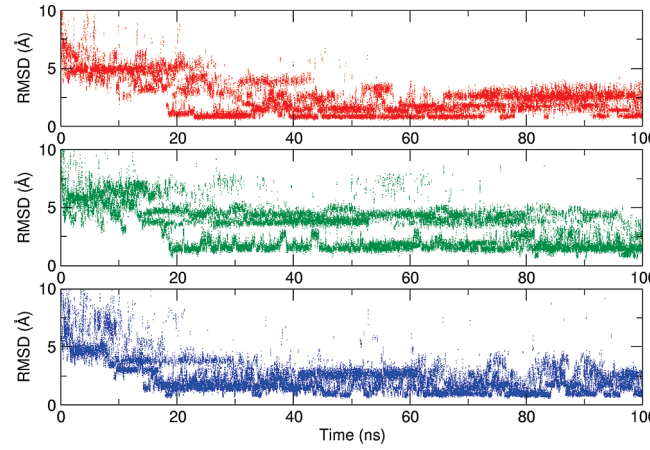

Figure 1. C_α rmsd of the lowest temperature ReX window (270 K) as a function of simulation time (smoothed by a 1 ns running average) for the folding simulations starting from the unfolded trans structure. Legend: red, MD-ReX; green, LD-ReX; blue, SGLD-ReX.

We also looked at three model systems to investigate possible deviations between SGLD and LD using one-dimensional potentials with two wells. The first model potential was a symmetric double well:

$$U_1(x) = x^2(x - 2)^2 \tag{10}$$

The second potential was a double well with one minimum higher than the other:

$$U_2(x) = x^2(x - 2)^2 + 2x \tag{11}$$

The third potential had two wells of different curvatures, narrow and broad:

$$U_3(x) = 0.499 + 0.7372x + 12.51x^2 - 23.883x^3 + 16.659x^4 - 5.1411x^5 + 0.59399x^6 \tag{12}$$

A fictitious particle was propagated along the model potential in CHARMM with a mass of 2171 amu. The mass was chosen to reproduce that of the Trp-cage. The y and z dimensions were restrained by a 500 (kcal/mol)/Å^2 harmonic potential. LD and SGLD parameters were the same as those used in the Trp-cage simulations. The time step was set to 10 fs, and the total simulation times were 10 μs. Trajectories of the x-coordinates were saved every picosecond and placed into histograms that were binned in 0.01 Å increments.

Results

First, we assessed how quickly the three simulation models fold up the Trp-cage protein starting with the trans conformation. In Figure 1, the C_α rmsd of the lowest temperature ReX window (270 K) is plotted as a function of simulation time. Because of the Metropolis criterion, low-temperature selection is a reasonable measure of energy-based structure detection. The SGLD-ReX, LD-ReX, and MD-ReX simulations first detect structures below 1.5 Å rmsd at 14.1, 18.5, and 18.5 ns, respectively. These results suggest only a modest speedup for SGLD-ReX in approaching the native basin starting from the unfolded state. This result is in stark

Table 1. Predicted Melting Temperatures and Folding Free Energies[a]

simulation	simulation time (ns)	T_m^1 (K) (max C_p)	T_m^2 (K) (rmsd)	T_m^3 (K) ($\Delta\Delta G_{fold} = 0$)	$\Delta\Delta G_{fold}$ ($T = 298$ K)
MD-ReX/trans	50−100	360	362	351	−1.8
MD-ReX/native	50−100	366	367	355	−1.6
MD-ReX/native	150−200	369	369	348	−1.7
LD-ReX/trans	50−100	375	352	290	0.2
LD-ReX/native	50−100	339	345	335	−1.9
LD-ReX/native	150−200	372	370	354	−1.4
SGLD-ReX/trans	50−100	315	315	311	−0.8
SGLD-ReX/native	50−100	339	338	331	−1.4
SGLD-ReX/native	150−200	324	320	306	−0.4
experiment	n/a	324	316[b]	317	−0.76 (0.05)

[a] Experimental results are taken from Streicher and Makhatadze.[38] [b] Derived from midpoint of "fraction unfolded" graph.[38]

comparison to the 65-fold speedup of helix formation reported in the original SGLD paper.[5] The likely explanation is that ReX already provides sufficient sampling enhancement for MD and LD to overcome the unfolded/folded transition barrier to fold up the Trp-cage. For all methods, the lowest C_α-rmsd basin appears to reside around 0.9 Å from the NMR conformer 1. This result compares favorably to the fact that the 37 other NMR conformers in PDB entry 1L2Y[19] are also, on average, 0.9 Å (rmsd) from conformer 1.

Comparison of calculated and experimental folding free energies requires defining the native basin. The resolution limit of our force-field/implicit solvent model energy function compelled us to use a liberal definition of the native basin. By evaluating the population distribution as a function of rmsd at the melting temperature of each simulation, we found that there were three compact basins that resided below 3.4 Å. Therefore, the dividing line between native/nonnative was set to 3.4 Å rmsd. Using this definition and data from the last 50 ns of the 200 ns native simulations, the free energies of folding predicted by SGLD-ReX, LD-ReX, and MD-ReX are −0.4, −1.4, and −1.7 kcal/mol, respectively, as reported in Table 1. The SGLD-ReX result is closest to the experimental folding free energy of −0.76 kcal/mol.

Also, in Table 1, predicted melting temperatures are compared. Melting temperatures derived from heat capacity (T_m^1) and rmsd (T_m^2) are nearly the same except for LD-ReX (see the PMF analysis below.) In contrast, melting temperatures derived from the transition point ($\Delta\Delta G_{fold} = 0$), T_m^3, are 10−20 K lower than T_m^1, which suggests the rmsd = 3.4 Å unfolded/folded dividing line is somewhat imprecise. SGLD-ReX simulations predict a melting temperature coinciding with experimental results. However, MD-ReX and LD-ReX predict melting temperatures ∼40−50 K higher than experiment. The large discrepancies in predicted melting temperatures between SGLD-ReX and traditional ReX approaches cannot be explained by the uncertainty of the calculations, which is ∼10 K.

To better understand the melting transition, the predicted heat capacity curves were compared against the experimental results in Figure 2. SGLD-ReX, as discussed earlier, predicts the closest temperature peak to experiment. Predictions from MD and LD are fairly consistent between the two simulation models in the temperature peak and profiles. The magnitudes of the heat capacities were not expected to correspond to experiment because implicit solvent models incorporate the free energy associated with the solvent degrees of freedom as part of the total potential energy, U, or "enthalpy" of the system.[41] This fact can distort calculations of observables on the basis of the enthalpy (e.g., heat capacity). However, the width of the heat capacity profile as a function of temperature was not expected to be affected to the same degree.

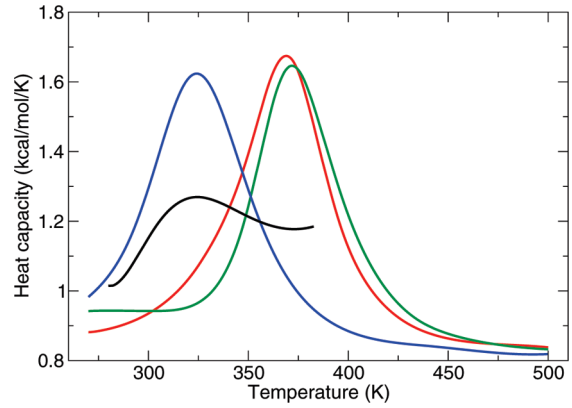

Figure 2. Predicted heat capacity as a function of temperature for the three simulation models (starting from the native structure) using the last 50 ns of 200 ns simulation data compared to experiment.[38] Legend: red, MD-ReX; green, LD-ReX; blue, SGLD-ReX; black, experiment.

To comprehend why the three different simulation models predict dissimilar observables even though the underlying potential energy function is identical, we investigated the potentials of mean force (PMFs) along various coordinate dimensions and performed several tests of canonical ensemble behavior. The PMFs along C_α rmsd to native and R_g coordinates in Figure 3 provide insight into the various conformational basins explored by the simulations. For the simulations starting from the unfolded state, the LD-ReX simulations generated three distinct intermediate folded states and only one near-native state when the PMF is mapped onto rmsd and R_g coordinates. In contrast, the SGLD-ReX simulations populated three near-native basins, while the MD-ReX simulations obtained similar results, further splitting the middle basin.

The nearest-to-native basin incorporates conformations with most of the correct features compared to the NMR structure as seen in a structure representative of that basin (Figure 3, model R1). The only defects are that the tryptophan side chain has a slight twist compared to the native, and the elusive 3_{10}-helix (as gauged by DSSP[42]) is

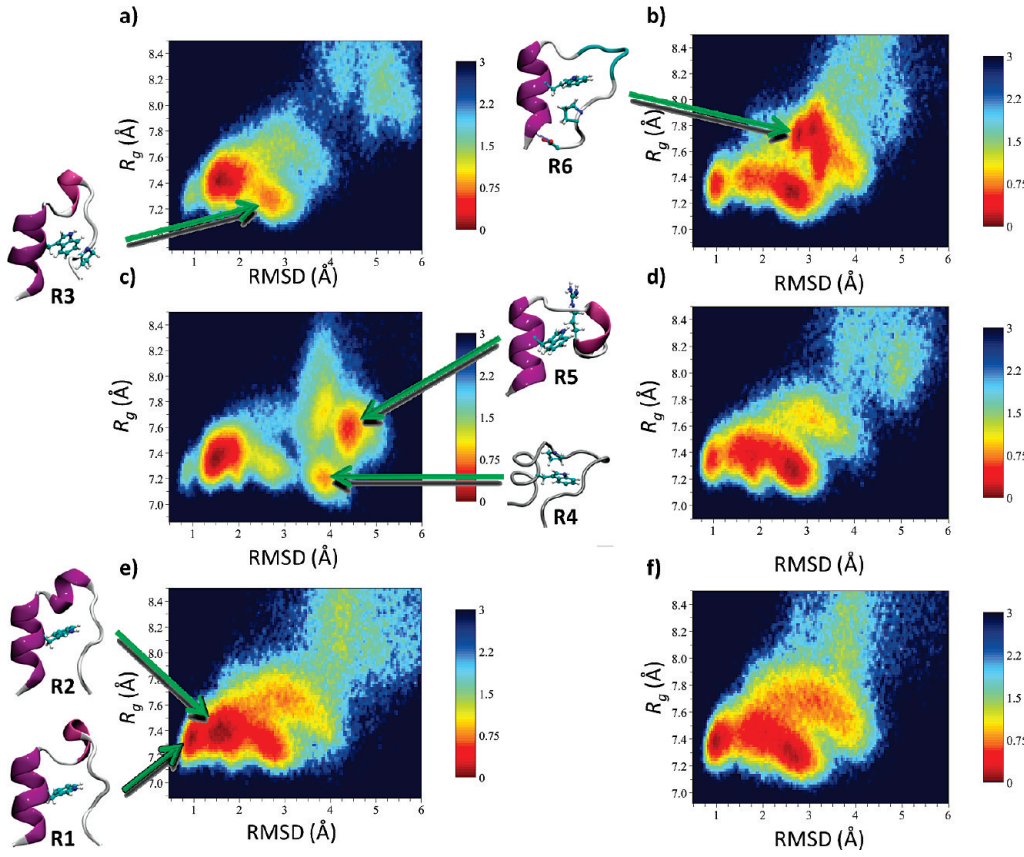

Figure 3. Free-energy landscapes at $T = 298$ K as a function of C_α rmsd to native and radius of gyration, R_g, using the last 50 ns of data from the 100 ns simulations starting from the folded conformation (right plots) and the unfolded trans conformation (left plots): (a, b) MD-REX; (c, d) LD-REX, and (e, f) SGLD-REX. The free energies are represented by a range of colors and are arbitrarily capped at 3 kcal/mol.

only sampled a small percentage of the time. In comparison, the main flaw of the representative structure of the 1.5 Å basin (Figure 3, R2) is incorrect folding in the turn region even though the Trp indole ring is stacked parallel to the correct proline (residue 18). The ∼2.7 Å rmsd basin representative (Figure 3, R3) has even larger flaws in the turn region despite the fact that the tryptophan is once again packed against Pro18. Because this basin is slightly more compact than the basins that are nearer to native, nonspecific hydrophobic collapse from the implicit solvent surface area term is a likely culprit. Representative compact structures of the LD-ReX/trans simulation that are even farther from the native (Figure 3, R4 and R5) show nonnative packing of the indole ring against residues Arg16 (R4) and Pro12 (R5). The MD simulation starting from the native produced a stable basin of structures (R6) in which the carboxyl terminus of Ser20 hydrogen bonds to the amide nitrogen of N-terminus residue Leu2, while Pro18 stacks perpendicularly to the tryptophan ring. In contrast, SGLD sees that same part of conformational space in rmsd/R_g dimensions but does not dwell there for any significant length of time, probably due to the ad hoc force term which favors increased global sampling vs local sampling.

From PMFs in the dimensions of potential energy and C_α rmsd to native at the transition temperature ($\Delta G_{fold} = 0$), further insights into the differences among the protocols can be gleaned as seen in Figure 4. All three methods sample the nearest-to-native basin (∼1 Å) at their respective transition temperatures, with SGLD-ReX having the most density there. Since the nearest-to-native basin does not appear to be the lowest in free energy, this could be due to the fact that SGLD-ReX performs the most excursions among basins in a given simulation time. Another positive feature of SGLD-ReX (Figures 4g−i) is how similar the PMFs are among the different starting conformations and data collection times. This suggests that, of the three methods, SGLD-ReX is the most self-consistent and arguably the most converged, at least in the conformational space of compact folds. The 150−200 ns data windows of MD-ReX and LD-ReX do have qualitative agreement with SGLD-ReX, suggesting that longer equilibration times can bring these three methods into better agreement. In particular, the sampled rmsd basins are similar among three methods at 150−200 ns. The main drawback with SGLD-ReX is that the computed transition temperature and resultant basin energies are significantly lower, suggesting a distortion of the relative free energy among folded and unfolded states. Finally, note that the rmsd in the range of 3.2−3.5 Å appears to be a sensible dividing line between folded/unfolded states, as, in many cases, the next basin above this line is higher in energy.

In Figure 5, we computed several properties of the simulation protocols to determine if there were any major deviations from theoretical canonical ensemble behavior. In Figure 5a, only MD-ReX produced temperatures precisely in line with those specified by the user, thanks to strict temperature control with a Nose-Hoover thermostat. The

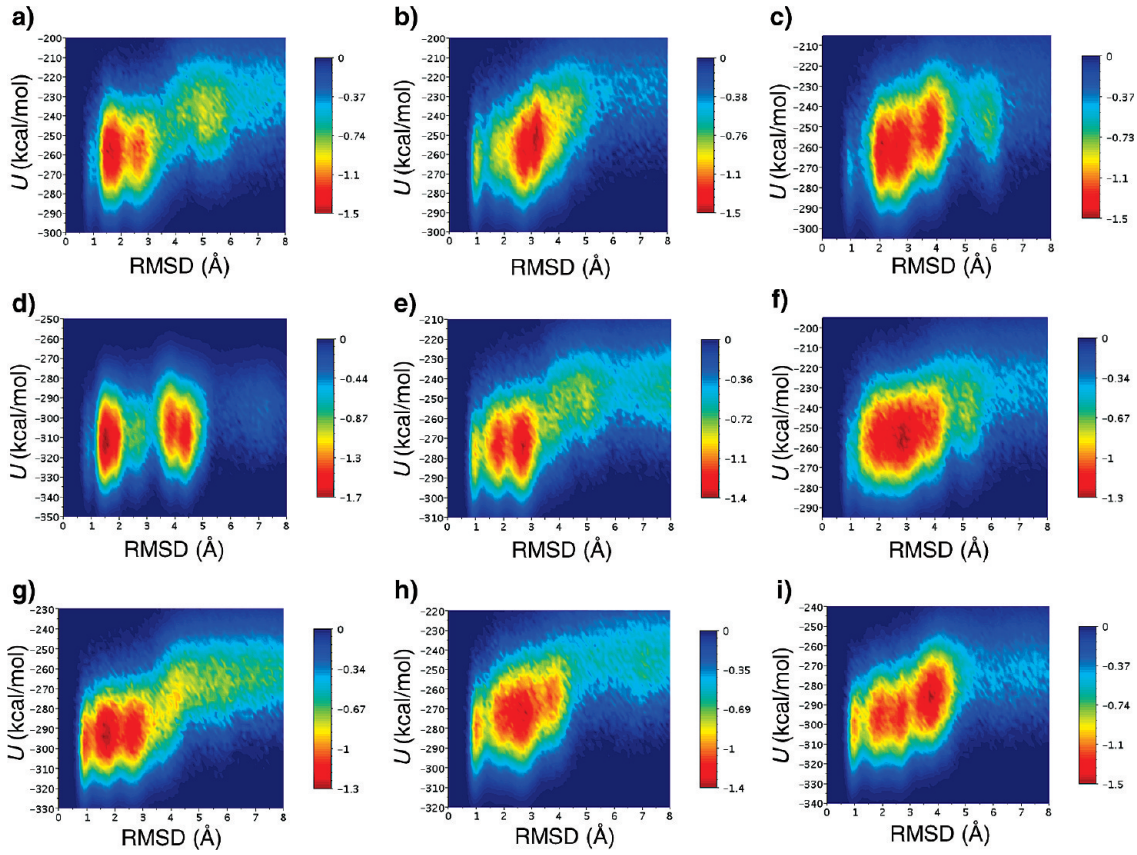

Figure 4. Free-energy landscapes at respective melting temperatures ($\Delta G_{fold} = 0$) of individual simulations ⟨*method/starting structure/simulation data*⟩ in the coordinates of potential energy, U, and C_α rmsd to native: (a) MD-ReX/trans/50−100 ns (T = 351.3 K), (b) MD-ReX/native/50−100 ns (T = 354.6 K), (c) MD-ReX/native/150−200 ns (T = 348.2 K), (d) LD-ReX/trans/50−100 ns (T = 290.1 K), (e) LD-ReX/native/50−100 ns (T = 335.1 K), (f) LD-ReX/native/150−200 ns (T = 353.9 K), (g) SGLD-ReX/trans/50−100 ns (T = 311.2 K), (h) SGLD-ReX/native/50−100 ns (T = 331.5 K), and (i) SGLD-ReX/native/150−200 ns (T = 306.4 K).

deviations for SGLD-ReX are the most substantial, up to 3 K too low for the highest temperature window. Fortunately, in Figure 5b, the Metropolis exchange factors implied by the simulation temperatures for SGLD-ReX are still consistent with the ones actually used for exchange. In Figure 5c, the ratio of heat capacities calculated by instantaneous and finite difference methods are all close to the theoretical value of 1 away from the transition state. In all cases, the ratio deviates from 1, as expected for a two-state superposition around the predicted melting temperatures of the respective methods. Figure 5d paints a similar story, whereby the potential energy histograms of all three methods have zero skew except for the transition temperature where two Gaussians are expected.

Next, the average potential energies of replica-exchange clients over the last 50 ns of simulation data elucidates how the client simulations sampled the potential energy surface starting from the unfolded structure (Figure 5e). The profile of the average energy over different client temperatures mirrors the heat capacity curve in that different transition temperatures (seen here as deviations from linearity) can be observed for the various protocols. In addition, among the three protocols at the lowest temperature, MD-ReX digs deepest into the potential energy surface, providing the lowest energy structures by several kilocalories per mole. This small discrepancy could partially explain the free energy differences between the three methods.

Another test stems from a concern that SGLD may overheat low-frequency modes and cool high-frequency modes to compensate for total energy conservation and maintenance of the user-specified system temperature. Fortunately, as seen in Figure 5f, all methods have roughly the same profile of normal mode temperature vs index of normal mode (going from slowest to fastest). Interestingly, the individual mode kinetic energies of all three methods tend to start out high and go lower with increasing mode frequency. This trend can be attributed as an artifact of SHAKE. When SHAKE is not run, the kinetic energy profiles are virtually flat along the entire range of modes with the correct value associated with the simulation temperature (results not shown).

Finally, we compare LD and SGLD for generating free-energy profiles of three double-well potentials (Figure 6). In the first potential model, Figure 6a, the potential is a symmetric double well. The deviation from ideal behavior is quite small for both LD and SGLD ($\tau = 0.1$ ps), namely, ∼0.005 kcal/mol. However, SGLD accelerates sampling by crossing between wells 4883 times, while LD only crosses 1117 times in the 10 μs simulation. The increased crossing frequency results in a lowered free-energy barrier between

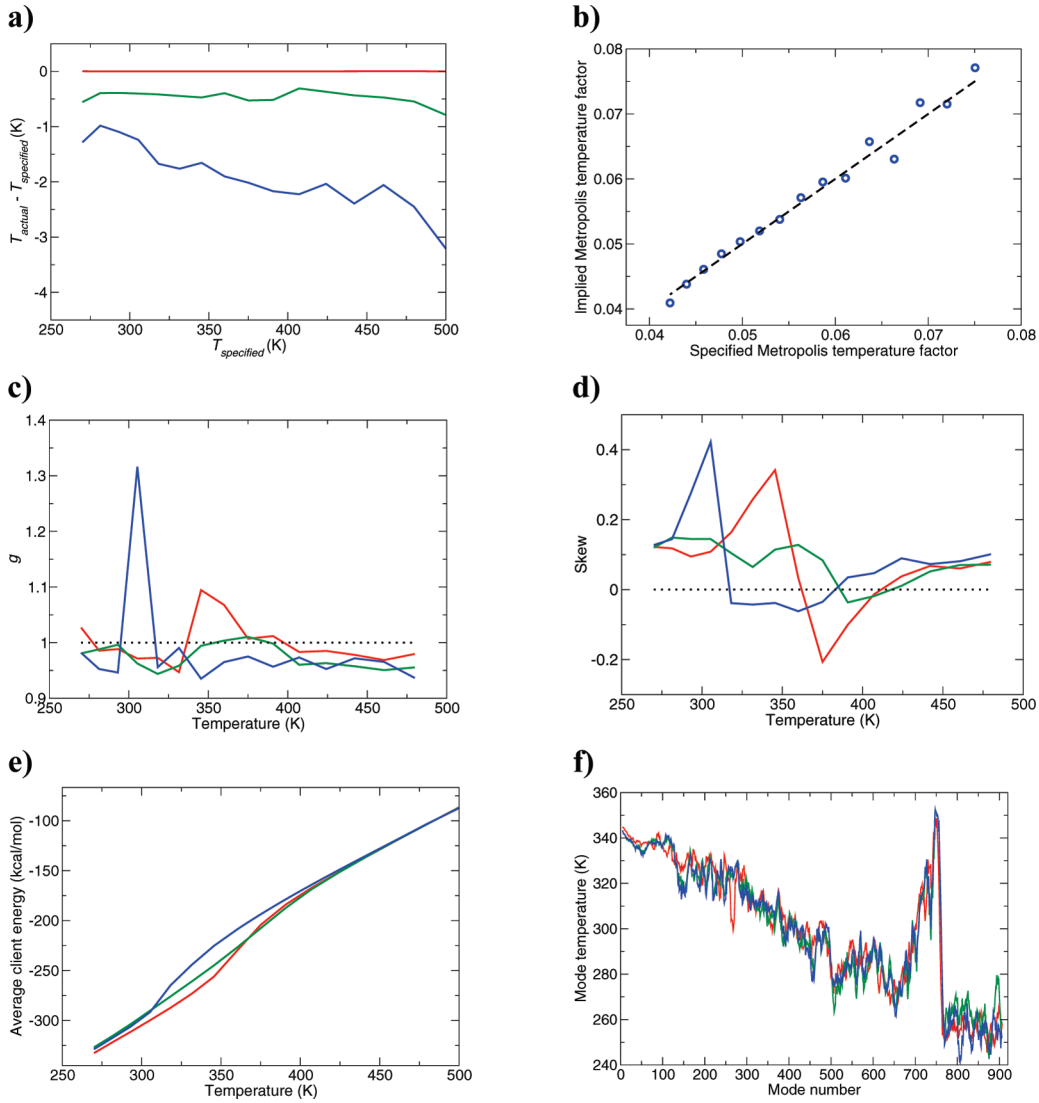

Figure 5. Various metrics of the ensembles generated at each replica-exchange temperature window for the three simulation protocols: (a) deviations from specified temperature (trans starting structures), (b) Metropolis temperature factors implied by simulation temperatures vs factors actually used for exchange, (c) *g*, ratio of heat capacities calculated by fluctuations and finite difference, (d) skew of potential energy distribution, (e) average potential energy of temperature windows over the last 50 ns of simulations starting from the unfolded trans conformation, and (f) temperatures of normal modes for each simulation protocol. Legend: red, MD-ReX; green, LD-ReX; blue, SGLD-ReX. Dashed lines represent optimal values.

the two wells of about 1 kcal/mol. Next, for a heterogeneous model system with one well higher in energy than the other as in Figure 6b, SGLD shifts the higher energy well down by ~0.3 kcal/mol. Finally, for a different heterogeneous system where the curvature of the higher well is broader than the lower well (Figure 6c) mimicking a folded/unfolded peptide landscape, the higher energy well is actually stabilized by ~0.3 kcal/mol. In both heterogeneous model systems, the free-energy barriers are reduced as expected. Unfortunately, the heterogeneous model systems indicate that SGLD distorts the relative free energy of minima with different energies and curvatures. How this result directly translates to a real protein system such as Trp-cage with a complex topological landscape is unclear.

Discussion

Self-guided Langevin dynamics was originally devised to accelerate low-frequency motions in order to enhance sampling. Our results for simulating folding−unfolding of the Trp-cage indicate that, compared to MD and LD, the topological free-energy barriers among major conformational basins were effectively reduced, but overall folding times were not significantly improved. The melting temperature predicted by SGLD is noticeably lower than those produced by the MD and LD approaches. This result could be a sign of actual changes in the effective population density sampled by the SGLD simulations or errors in canonical ensemble behavior that might be expected from the use of an ad hoc force term. Investigating the latter possibility, several canonical ensemble tests indicate that any error introduced by the SGLD force term can be ruled out. At worst, the actual simulation temperatures do drift moderately (~3 K) from their specifications, which could partly be due to the use of a high-energy derivative formalism such as GBMV[36] which, in turn, may compound errors associated with the total energy correction term in SGLD. As shown previously,[5] the diffu-

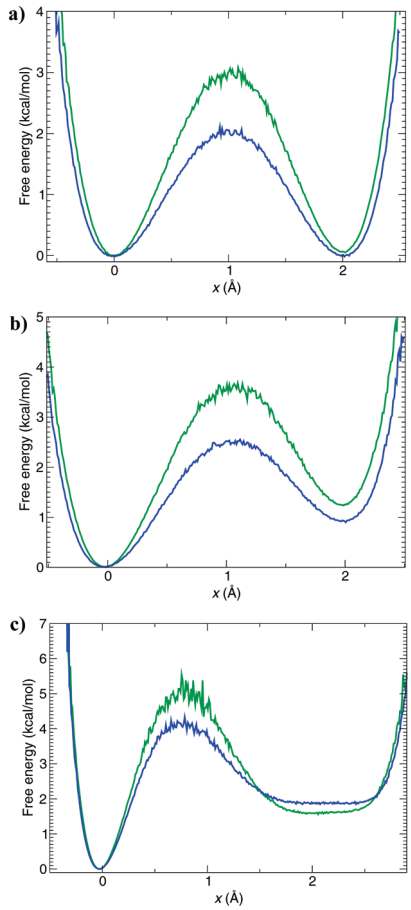

Figure 6. Potentials of mean force for three one-dimensional model systems (eqs 10, 11, and 12) using LD and SGLD integrators: (a) symmetric double well, (b) double well with different energies, and (c) narrow well/broad well combination. Legend: green, LD; blue, SGLD.

sivity of an SGLD simulation is greatly increased compared to LD. However, an analysis of the kinetic energies of the normal modes ruled out the possibility that the low-frequency modes in SGLD were actually "hotter" compared to MD and LD.

While our measured deviations from canonical ensemble behavior were not significant, predicted thermodynamic observables and the PMF landscapes in Figures 3 and 4 do differ substantially. Therefore, we cannot rule out the possibility that SGLD may have lowered the melting temperature by smoothing the basins of folded and unfolded states, thereby raising the relative free energy of the folded state. This could happen if the ad hoc force term in SGLD somehow propels the system out of narrow basins such as the native state more often than the broad unfolded regions. Another possibility is that SGLD slightly scales down the entire free-energy surface, reducing not just the free-energy barriers but the energetic difference between free-energy minima. To definitively test this prospect, we ran three model one-dimensional potentials. The SGLD simulation of the symmetric double well agreed with LD for the relative free energy of the minima but had a reduced barrier that increased crossing between states by a factor of 4. However, the SGLD simulation of the asymmetric narrow well potential yielded a free energy difference too small by 0.3 kcal/mol. Interestingly, SGLD simulation of narrow/broad well combination yielded the opposite effect: the magnitude of ΔG was 0.3 kcal/mol too large. This result confirms that SGLD does distort relative free energies of minima, albeit less than its favorable reduction of transition barriers. In any case, these test systems cannot be quantifiably translated to a complex system such as Trp-cage. Thermodynamic distortions could be definitively resolved by reweighting the simulation data. Unfortunately, such a reweighting scheme has not been devised yet.

While SGLD induces distortions in the simulations of simple model systems, molecular dynamics with a strict thermostat such as Nose-Hoover is not without fault. It should be granted that MD is better overall at locally sampling protein conformations compared to SGLD and LD. The evidence for this assertion is that in both the native and trans simulations, MD samples the lowest average energy of conformations at room temperature. In addition, the conformational basins as viewed by the PMF landscapes of MD are tighter than in LD and SGLD. The drawback of local sampling is that MD simulations can get trapped in certain conformational basins while insufficiently sampling others. In fact, it has been shown that in the limit of systems with a small number of degrees of freedom, Nose-Hoover MD simulations with only a single fictitious coordinate can become nonergodic, i.e., repeat the same trajectory ad infinitum.[43] Therefore, MD sampling may be incomplete or at least slower than SGLD at visiting all of the relevant compact conformational basins.

Thermodynamic issues aside, SGLD-ReX has a few salient properties. First, SGLD-ReX seems to show improved sampling convergence as the PMF native results at 50−100 ns visually match up quite well to the 150−200 ns data. This could be due to the ad hoc force term which increases transition probabilities among neighboring conformational basins. Better global sampling could partially explain the improved agreement of SGLD to experiment as previous studies have shown that all dominant basins need to be sampled to obtain accurate free-energy estimates.[41] Next, of the three approaches, SGLD-ReX has the highest sampling density of the nearest-native basin (<1 Å rmsd). This result may be due to the fact that SGLD-ReX does a better job of skimming the potential surface, revisiting the nearest-native basin more often. Finally, unlike coarse-grained methods, which smooth the potential energy surface by reducing spatial resolution, the self-guided formalism does not introduce any distortions into the generated conformations. This opens up the possibility of applying SGLD-ReX to the protein structure refinement problem.[44]

There are several caveats in this study. First and foremost, an implicit solvent model was used rather than explicit water molecules.[45] This choice enabled straightforward application of temperature-based replica exchange because of reduced degrees of freedom. Nonetheless, implicit solvent models have reduced conformational resolution due to artifacts such as too strong (or too weak) salt bridges[46] and hydrogen bonds, and missing noncovalent attractions between protein and solvent.[47] Another issue in this work is that we used a fixed-charge force field rather than a flexible-charge one

which can be problematic in studies such as this one where large conformational changes are expected[48] (i.e., unfolded to folded.) Also, while the simulation time of 100 ns was sufficient to permit ab initio folding of the Trp-cage protein followed by roughly 50 ns of production data, sampling convergence was not achieved in some cases. For example, the MD-ReX and LD-ReX simulations, starting from either native or unfolded trans conformations, produced quite different rmsd vs R_g PMF landscapes. Moreover, to fold up proteins larger than the Trp-cage, longer simulation times may be required. Finally, while fixed simulation temperatures for replica exchange was sufficient for this work, adaptive changes to the temperature set[12] may be necessary for larger proteins, especially ones with a sharper peak in their heat capacity profile (i.e., sharper energetic transition between the folded and unfolded state).[39]

Our study shows that combining SGLD and ReX produces a sampling method that has both advantages and disadvantages compared to Nose-Hoover-based MD-ReX and LD-ReX. Enhanced sampling convergence for SGLD-ReX is seen in the free-energy landscapes and is likely due to reducing the transition barriers between unfolded and folded states. Nonetheless, SGLD-ReX lacks direct application to studying kinetics of protein folding due to possibly modifying the folding pathway and its degrees of free-energy frustration. In addition, SGLD-ReX may produce lower predicted free energies of folding and melting temperatures by shifting the relative heights of free energy minima in the same way that it reduces transition barriers between conformational basins. If a proper reweighting scheme were devised, these problems would be alleviated. All in all, with appropriate selection of parameters and acknowledgment of some distortions due its ad hoc nature, SGLD-ReX should find application in the calculation of thermodynamics of protein folding−unfolding and protein−ligand association. The method is also currently being evaluated in the emerging field of comparative protein model refinement.

Acknowledgment. We would like to thank Dr. I.-C. Yeh for supplying his WHAM code. Funding for this research was provided by the U.S. Department of Defense Threat Reduction Agency Grants 3.10010_06_RD_B and TMTI0004_09_BH_T, and the Department of Defense Biotechnology High Performance Computing Software Applications Institute. Computational time was provided, in part, by the U.S. Army Research Laboratory Major Shared Resource Center and Maui High Performance Computing Center. The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department of Defense. This paper has been approved for public release with unlimited distribution.

References

(1) Freddolino, P. L.; Schulten, K. *Biophys. J.* **2009**, *97*, 2338.

(2) Zuckerman, D. M.; Lyman, E. *J. Chem. Theory Comput.* **2006**, *2*, 1200.

(3) Ishikawa, Y.; Sugita, Y.; Nishikawa, T.; Okamoto, Y. *Chem. Phys. Lett.* **2001**, *333*, 199.

(4) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919.

(5) Wu, X.; Brooks, B. R. *Chem. Phys. Lett.* **2003**, *381*, 512.

(6) Pappu, R. V.; Marshall, G. R.; Ponder, J. W. *Nat. Struct. Biol.* **1999**, *6*, 50.

(7) Hornak, V.; Simmerling, C. *Proteins* **2003**, *51*, 577.

(8) Skolnick, J.; Zhang, Y.; Arakaki, A. K.; Kolinski, A.; Boniecki, M.; Szilagyi, A.; Kihara, D. *Proteins* **2003**, *53* (Suppl. 6), 469.

(9) Zhang, Y.; Arakaki, A. K.; Skolnick, J. *Proteins* **2005**, *61*, Suppl. 7, 91−8.

(10) Kirkpatrick, S.; Gelatt, C. D., Jr.; Vecchi, M. P. *Science* **1983**, *220*, 671.

(11) Paschek, D.; Nymeyer, H.; Garcia, A. E. *J. Struct. Biol.* **2007**, *157*, 524.

(12) Trebst, S.; Troyer, M.; Hansmann, U. H. *J. Chem. Phys.* **2006**, *124*, 174903.

(13) Gallicchio, E.; Andrec, M.; Felts, A. K.; Levy, R. M. *J. Phys. Chem. B* **2005**, *109*, 6722.

(14) Wu, X.; Brooks, B. R. *Biophys. J.* **2004**, *86*, 1946.

(15) Wen, E. Z.; Hsieh, M. J.; Kollman, P. A.; Luo, R. *J. Mol. Graphics Modell.* **2004**, *22*, 415.

(16) Damjanovic, A.; Garcia-Moreno, E. B.; Brooks, B. R. *Proteins* **2009**, *76*, 1007.

(17) Damjanovic, A.; Wu, X.; Garcia-Moreno, E. B.; Brooks, B. R. *Biophys. J.* **2008**, *95*, 4091.

(18) Damjanovic, A.; Miller, B. T.; Wenaus, T. J.; Maksimovic, P.; Garcia-Moreno, E. B.; Brooks, B. R. *J. Chem. Inf. Model.* **2008**, *48*, 2021.

(19) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. *Nat. Struct. Biol.* **2002**, *9*, 425.

(20) Paschek, D.; Hempel, S.; Garcia, A. E. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 17754.

(21) Kannan, S.; Zacharias, M. *Proteins* **2009**, *76*, 448.

(22) Snow, C. D.; Zagrovic, B.; Pande, V. S. *J. Am. Chem. Soc.* **2002**, *124*, 14548.

(23) Zhou, R. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 13280.

(24) Steinbach, P. J. *Proteins* **2004**, *57*, 665.

(25) Juraszek, J.; Bolhuis, P. G. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 15859.

(26) Gattin, Z.; Riniker, S.; Hore, P. J.; Mok, K. H.; van Gunsteren, W. F. *Protein Sci.* **2009**, *18*, 2090.

(27) Wu, X.; Wang, S. *J. Chem. Phys.* **1999**, *110*, 9401.

(28) Andricioaei, I.; Dinner, A. R.; Karplus, M. *J. Chem. Phys.* **2003**, *118*, 1074.

(29) Mackerell, A. D., Jr.; Bashford, D.; Bellott, D. M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, I. W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.

(30) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., 3rd. *J. Am. Chem. Soc.* **2004**, *126*, 698.

(31) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., 3rd. *J. Comput. Chem.* **2003**, *24*, 1348.

(32) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L., 3rd. *J. Comput. Chem.* **2004**, *25*, 265.

(33) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.

(34) Feig, M.; Karanicolas, J.; Brooks, C. L., 3rd. *J. Mol. Graphics Modell.* **2004**, *22*, 377.

(35) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminatham, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.

(36) Chocholousova, J.; Feig, M. *J. Comput. Chem.* **2006**, *27*, 719.

(37) Feig, M.; Brooks, C. L., 3rd. *Proteins* **2002**, *49*, 232.

(38) Streicher, W. W.; Makhatadze, G. I. *Biochemistry* **2007**, *46*, 2876.

(39) Yeh, I. C.; Lee, M. S.; Olson, M. A. *J. Phys. Chem. B* **2008**, *112*, 15064.

(40) Rosta, E.; Buchete, N.-V.; Hummer, G. *J. Chem. Theory Comput.* **2009**, *5*, 1393.

(41) Chang, C. E.; Gilson, M. K. *J. Am. Chem. Soc.* **2004**, *126*, 13156.

(42) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577.

(43) Cooke, B.; Schmidler, S. C. *J. Chem. Phys.* **2008**, *129*, 164112.

(44) Lee, M. S.; Olson, M. A. *J. Chem. Theory Comput.* **2007**, *3*, 312.

(45) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740.

(46) Zhou, R.; Berne, B. J. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12777.

(47) Swanson, J. M.; Mongan, J.; McCammon, J. A. *J. Phys. Chem. B* **2005**, *109*, 14769.

(48) Patel, S.; Mackerell, A. D., Jr.; Brooks, C. L., 3rd. *J. Comput. Chem.* **2004**, *25*, 1504.

CT100062B