An experimentally guided umbrella sampling protocol for biomolecules

Maria Mills^{a)} and Ioan Andricioaei^{b)}

Department of Chemistry, University of California, Irvine, California 92697, USA

(Received 30 January 2008; accepted 7 August 2008; published online 15 September 2008)

We present a simple method for utilizing experimental data to improve the efficiency of numerical calculations of free energy profiles from molecular dynamics simulations. The method involves umbrella sampling simulations with restraining potentials based on a known approximate estimate of the free energy profile derived solely from experimental data. The use of the experimental data results in optimal restraining potentials, guides the simulation along relevant pathways, and decreases overall computational time. In demonstration of the method, two systems are showcased. First, guided, unguided (regular) umbrella sampling simulations and exhaustive sampling simulations are compared to each other in the calculation of the free energy profile for the distance between the ends of a pentapeptide. The guided simulation use restraints based on a simulated “experimental” potential of mean force of the end-to-end distance that would be measured by fluorescence resonance energy transfer (obtained from exhaustive sampling). Statistical analysis shows a dramatic improvement in efficiency for a 5 window guided umbrella sampling over 5 and 17 window unguided umbrella sampling simulations. Moreover, the form of the potential of mean force for the guided simulations evolves, as one approaches convergence, along the same milestones as the extensive simulations, but exponentially faster. Second, the method is further validated by replicating the forced unfolding pathway of the titin I27 domain using guiding umbrella sampling potentials determined from actual single molecule pulling data. Comparison with unguided umbrella sampling reveals that the use of guided sampling encourages unfolding simulations to converge faster to a forced unfolding pathway that agrees with previous results and produces a more accurate potential of mean force. © 2008 American Institute of Physics. [DOI: 10.1063/1.2976440]

I. INTRODUCTION

The potential of mean force (PMF),¹ i.e., the free energy $W(\xi) = -k_B T \ln \langle p(\xi) \rangle$ with respect to a chosen coordinate ξ , is fundamentally related to that coordinate's distribution function, averaged over all the other degrees of freedom $\langle p(\xi) \rangle$. As such, it is a central concept in the statistical mechanical representation of molecular systems and has been employed in a number of computational applications,^{2,3} describing, for example, processes ranging from hydrophobic interactions⁴ or organic reactions in water,⁵ to proton transfer^{6–8} or ionic permeation through membrane channels,⁹ to peptide^{10,11} and protein¹² equilibria, to nucleic acid base flipping^{13,14} or more complex conformational changes¹⁵ in DNA and RNA. The PMF can be used to clearly represent the equilibrium distribution of one or a few relevant conformational variables (angles, distances, etc.) with increasing accuracy as longer sampling and improved force fields become available. Alternatively, it can also be used to describe nonequilibrium processes from the solution of a reduced Fokker–Planck equation for the time dependence of the probability distribution of the ξ variable diffusing in the potential of mean force $W(\xi)$ (describing, e.g., the transfer of an ion or a conformational transition of a biomolecule). Therefore, if the chosen coordi-

nate is a good reaction coordinate (i.e., if it is much slower than any other degree of freedom), a dynamical propagation on the PMF can simulate the kinetics of the reaction of interest. Moreover, the derivatives of the PMF along a selected set of holonomically constrained degrees of freedom^{16,17} [whose thermodynamic integration yields the PMF (Ref. 18)] can be used to locate free energy minima¹⁹ or to sample optimal reaction paths.²⁰

At the same time with the increase in the accuracy of calculating equilibrium distributions by computer simulations, advances in the experimental realm have also made it possible to determine relatively accurately *distribution functions* of measurable coordinates. For example, measurements in molecular optical spectroscopy²¹—traditionally performed in the bulk—can now be recorded at the single molecule level, thereby resolving possible differences between the behavior of a number of individual molecules versus measuring ensemble averages. This in turn can reflect for skewed probability distribution in differences between observing the most probable versus the average observables,²² and more fundamentally, can reveal new regimes for the processes at the nanoscale.^{23,24} These experimental distribution functions can be used, in principle, to obtain free energy landscapes (PMFs) for conformational dynamics of proteins and other biomolecules.^{25–27} For example, a study by Deniz *et al.*²⁵ using single molecule Förster resonance energy transfer (FRET) between sites that are fluorescently labeled in a biomolecule has enabled the characterization of protein folding

^{a)}Also in The Graduate Program in Biophysics, University of Michigan, Ann Arbor, MI 48109.

^{b)}Author to whom correspondence should be addressed. Electronic mail: andricio@uci.edu.

distributions by measuring the probability density of the distance between the N terminus and the loop of chymotrypsin inhibitor 2 at various concentrations of denaturant. FRET can also probe conformational changes for RNA molecules, which have been shown to require large equilibrium dynamical transitions as a prerequisite for their function.²⁸

Another source of experimental distribution functions comes from mechanically extending biomolecules one at a time. Single molecule pulling techniques have been employed to unfold biomolecules by using atomic force microscopy (AFM),²⁹ optical³⁰ or magnetic trapping,³¹ thereby generating extension trajectories that, when binned and processed appropriately,^{32–36} can produce a mapping of the potential of mean force and possibly of its entropic and enthalpic components³⁷ along the pulling coordinate in the absence of the pulling force or for various values of the force.³⁸ The latter aspect—a kinetic one—is important in accurately assessing the dependence of the unfolding rate constant on force,³⁹ as this has important consequences *in vivo* for instances when proteins are unfolded as part of their cellular function.^{40,41} Alternatively, in equilibrium conditions, constant-force reversible folding-unfolding measurements can also be used to map the folding energy landscape, as was showcased in recent work on DNA hairpins.⁴²

However, experiments such as the ones described above or others (including bulk experiments) can resolve general conformational changes along only a small set of degrees of freedom [for regular FRET, one or—with the emergence of three-color FRET (Refs. 43 and 44)—up to three spatial separations (but see also Ref. 45)] and they often yield only on-off signals. Theoretical work on the applicability of FRET has explored the range of conditions that interfere with the ability of FRET to reveal the true PMF, mainly poor time resolution (relative to the actual timescale of distance fluctuations) and FRET efficiency distributions that are mainly due to shot noise rather than with the actual distance fluctuations.^{46,47} Similarly, in single molecule pulling, the extension that is measurable is poorly resolved at the point of transition (e.g., when a protein unfolds). Molecular dynamics simulations,⁴⁸ on the other hand, are useful for computing dynamic and equilibrium processes of large molecules in atomic detail for any degree of freedom and with high femtosecond time resolution, but processes which involve large free energy barriers or long timescales are too computationally demanding to be simulated directly. As a consequence, this breaks any assumption of ergodicity,⁴⁹ in the sense that the computed time averages, do not equal those in the ensemble measurements.

For this reason, it is often difficult to compute the averages required for the calculation of a free energy profile from a direct molecular dynamics simulations, i.e., by directly counting the frequency of occurrence of the values of the reaction coordinate during the simulation run. The most used method for generating a PMF efficiently from molecular dynamics or Monte Carlo simulations is the venerable umbrella sampling method.⁵⁰ In this method, a reaction coordinate is chosen and restraining potentials acting on it are used over a series of windows to sample the range of the reaction coordinate. By applying the restraining potentials, the system is

encouraged to sample regions of conformational space that would not otherwise be accessible during the direct sampling. The result is a series of histograms which contain the biased distribution of the reaction coordinate from each window. These histograms are then unbiased and combined,^{51,52} usually with the aid of the weighted histogram analysis method (WHAM).⁵³ Typically, the restraining potentials are chosen through trial and error, which results in a great deal of preliminary data being discarded as the potentials are optimized. We propose here a simple method which would use umbrella sampling potentials based on an *experimentally* determined (but not necessarily accurate) PMF from measured equilibrium distributions to improve molecular dynamics simulation efficiency.

For a uniform sampling of the reaction coordinate, the ideal “restraining” potential would be the *negative* of the exact potential of mean force (that is, if the PMF would be known *a priori*). Using the negative of the PMF flattens the energy landscape, circumventing the problem of trapping by large energy barriers.^{3,53} Most techniques for improving the efficiency of umbrella sampling, such as adaptive umbrella sampling methods^{54–60} and other techniques that generate flat distributions,^{61–64} focus on this approach. In adaptive umbrella sampling, a continuously updated umbrella potential which is a function of one or a few important degrees of freedom is added iteratively to the unperturbed Hamiltonian of the system. The umbrella potential for each iteration of the simulation is chosen to be an estimate of the negative of the PMF determined from the previous iteration. As a result, one obtains a uniform sampling distribution along the important degree(s) of freedom. In the multicanonical sampling method^{61,62} or its twin sister, the entropy sampling method,⁶⁵ and in Wang–Landau sampling⁶⁴ and related methods,^{66–68} a uniform *energy* probability distribution is obtained by assigning weights that are inversely proportional to a precalculated (or a dynamically updated) density of states.

Flattening the conformation or energy distributions to promote uniform sampling comes, however, at the substantial computational cost of predetermining either the adaptive umbrella potentials or the density of states, respectively. Moreover, as most biological molecules do not uniformly sample all conformation space, there would be no need to flatten energy barriers if the location of the barriers would be known beforehand. Instead, we propose here to add to the unperturbed Hamiltonian a set of experimentally derived harmonic potentials of different curvatures, fitted to the *positive* of an experimental PMF. This will serve as a “guide” for the reaction coordinate along the free energy path. It has been suggested previously that adding harmonic umbrella potentials minus the PMF would improve convergence for multi-window umbrella sampling techniques (as opposed to adaptive umbrella sampling, in which the negative of the PMF alone is used in a single window that is constantly updated).^{3,69} Subtracting the PMF from individual harmonic potentials gives similar potentials to our guiding technique, as long as the initial harmonic potentials are well chosen, as is depicted in the cartoon in Fig. 1 (see also discussion in Sec. III). In other words, our experimentally guided umbrella sampling should produce similar potentials to the best case

FIG. 1. Schematic examples of restraining potentials for the pedagogical case of a double well potential, showing the effect of subtracting the PMF from the restraints. (a) The original double well potential. (b) A set of uneducated-guess (poor) starting potentials (dashed lines) with the negative of the PMF (solid line) and (c) the results of subtracting the PMF from these harmonic potentials (dashed lines) overlaid with the original PMF. (d) A nearly optimal set of restraining potentials and (e) the results of subtracting the PMF from these optimal harmonic potentials, overlaid with the original PMF. The potentials in (e) resulting from the subtraction (dotted lines) produce biasing restraints similar to the type we suggest to use in our guided umbrella sampling protocol.

of this convergence method. Using a known experimental PMF should allow one to choose optimal initial restraining potentials, eliminating guesswork and wasted data.

In both the straightforward harmonic-wells-minus-PMF method and the adaptive umbrella sampling approaches, a good guess of the approximate negative of the PMF is crucial to their rapid convergence. Here we suggest how this initial guess can be provided not by preliminary simulations but directly from experimental input. This approach is similar in spirit to recent approaches that use experimental input in the form of, say, order parameter nuclear magnetic resonance (NMR) data⁷⁰ or hydrogen exchange⁷¹ to improve the description of conformational equilibrium of proteins.

To showcase our method, we report here an initial feasibility study using two systems. First, we study a test case: an unstructured pentapeptide with sequence Cys-Ala-Gly-Gln-Trp; we choose the end-to-end distance as the reaction coordinate. We selected this peptide as a test system because it has a size small enough to allow for exhaustive sampling by direct methods, and because both experimental⁷² and computational⁷³ work have used this pentapeptide as a model for contact formation in protein folding. Because no experimental potential of mean force for our chosen reaction coordinate was available, we have instead generated the PMF from extensive sampling simulations. This PMF was subsequently used as the “experimental” guiding PMF. Guided umbrella sampling was run using five windows with restraining potentials of varying curvature based on the guiding PMF. Regular, unguided umbrella sampling experiments were also run in five and seventeen windows for comparison with guided umbrella sampling simulations. Second, the method is applied to a more complex protein system, the I27 domain of the muscle protein titin, for which an experimental PMF has been determined by the forced unfolding of the I27 domain using AFM.⁷⁴ After presenting a background on umbrella sampling (in the next section) and introducing our experimentally guided umbrella sampling method (in

Sec. III), we present, in the Results sections, the outcome and the comparison of the various unguided and guided simulations for the two systems. We end with a few concluding discussions.

II. BACKGROUND ON UMBRELLA SAMPLING

In umbrella sampling,⁵⁰ a biasing potential is added to the Hamiltonian to direct the simulations toward a certain goal, for example, to induce a certain conformational transition. The biasing potential is usually in the form of a harmonic potential restraint that keeps the value of a relevant reaction coordinate fluctuating around successive positions along that coordinate. The result of this “stratification” is a series of biased histograms. In order to obtain accurate information about the free energy of the system, the raw data from individual simulations must be unbiased and recombined. The addition of the harmonic restraining potential w_i is equivalent to a multiplicative weight in the Boltzmann factor of $e^{-w_i(\xi)/k_B T}$. Thus, the biased distribution function for each window is

$$\langle p(\xi) \rangle_i^{\text{biased}} = e^{-w_i(\xi)/k_B T} \langle p(\xi) \rangle e^{-w_i(\xi)/k_B T}^{-1} \quad (1)$$

and the unbiased PMF in the i th window is

$$W_i(\xi) = -k_B T \ln \langle p(\xi) \rangle - w_i + F_i, \quad (2)$$

where the unknown free energy constant F_i is defined by the equation

$$e^{F_i/k_B T} = \langle e^{w_i/k_B T} \rangle. \quad (3)$$

In the weighted histogram analysis method (WHAM),⁵³ one uses an iterative process to determine the free energy constants, F_i . An initial guess set of F_i is used to estimate the unbiased probability distribution, where

$$\langle p(\xi) \rangle = \sum_i n_i \langle p(\xi) \rangle_i \left[\sum_j n_j e^{-(w_j(\xi) - F_j)/k_B T} \right]^{-1}. \quad (4)$$

The resulting probability is then used in the equation

$$e^{-F_i/k_B T} = \int d\xi e^{-w_i(\xi)/k_B T} \langle p(\xi) \rangle \quad (5)$$

to determine a new set of F_i values. The last two equations are solved iteratively until they are self consistent.

For regular implementations of WHAM using multiple-window umbrella sampling techniques, the strength of the harmonic biasing potentials and the number of simulation windows must be chosen to allow sufficient overlap for the data to be recombined while at the same time encouraging more rapid sampling of the reaction coordinate. Simply increasing the number of windows will not necessarily improve convergence.^{3,69} Furthermore, in order to obtain accurate information about other degrees of freedom, the chosen reaction coordinate must be the slowest one.⁶⁹ Determining optimal factors by trial and error leads to a great deal of wasted CPU time. Adaptive umbrella sampling overcomes this to a certain extent by updating the estimate of the PMF and using its negative to flatten barriers, but for regions which have not been sampled in previous simulations, the shape of the restraining potential must be set arbitrarily or

extrapolated.^{54–58} Another issue is the error in the PMF calculated by the WHAM equations, which is sensitive to the size of the bins used in constructing the histograms. Insufficient sampling leads to a large number of bins with too few configurations per bin, resulting in statistical error, while using too few bins can lead to errors in the probability density.⁷⁵

III. THE EXPERIMENTALLY GUIDED UMBRELLA SAMPLING METHOD

As discussed in the Introduction, knowing the PMF from experiment makes the choice of restraining potentials optimal, but it is important that the simulations are carried out in a way that yields information about the system of interest as a whole, rather than just the chosen reaction coordinate. This means that, in order for the experimentally guided biasing method to be practical, it is important to use experimental guiding potentials that are reported along the slow degrees of freedom of the system, and that the other degrees of freedom are sufficiently sampled so that they converge in the simulation (but see discussion later in the text regarding single molecule nonergodicity, for which departure from this convergence is worthwhile exploring as well).

The experimental PMF input in the calculation is fitted with harmonic restraining potentials of varying curvature. Assume that an experimental guiding PMF, W_{exp} is available, say from the probability distribution $p_{\text{exp}}(\xi)$ of a distance ξ measured from a FRET or micromanipulation experiment, i.e., up to a constant,

$$W_{\text{exp}}(\xi) = -k_B T \ln p_{\text{exp}}(\xi). \quad (6)$$

This guiding potential could be used in one of the following two ways (see panels B–C and D–E, respectively, in Fig. 1). The first possibility would be to choose a number of restraining potentials w_i centered each around a corresponding ξ_i value, and to run the corresponding i th simulation using the effective potential energy

$$V(\mathbf{r}) + w_i(\xi(\mathbf{r})) - W_{\text{exp}}(\xi(\mathbf{r})), \quad (7)$$

where V is the interatomic potential energy function employed describing the physical interactions in the system, with $\mathbf{r}$ the $3N$ -dimensional conformational vector. On one hand, if the simulation in the i th window would be run with a total potential consisting of the sum of just the first two terms in Eq. (7), i.e., with just $V(\mathbf{r}) + w_i(\xi(\mathbf{r}))$, this would be nothing else but regular umbrella sampling. On the other hand, in one dimensional cases, $V \approx W_{\text{exp}}$, so the subtraction of W_{exp} in Eq. (7) contributes to a flattening of the energy landscape along ξ , and the restraints w_i focus the simulation in a particular domain of ξ . In the absence of any knowledge of *a priori* details of the PMF, in regular umbrella sampling, all the w_i s are the same across all windows. For the simulation run with the potential in Eq. (7), the subtraction of W_{exp} from the sum of the first two terms is equivalent to an effective change in the individual values of the restraining potentials from window to window. In general cases, the w_i 's are to be optimized [see Figs. 1(d) and 1(e)] such that the guiding $w_i^{(g)}$ restraining potentials replace $w_i(\xi(\mathbf{r})) - W_{\text{exp}}(\xi(\mathbf{r}))$,

FIG. 2. (a) “Experimental” potential of mean force for the end-to-end distance of pentapeptide with fitted biasing potentials. (b) Second derivative of the PMF calculated using a finite-difference scheme and by eliminating, with a running average filter of width 2.5 Å, changes in the curvature occurring on length scales smaller than the thermally accessible range for the harmonic potentials. The plot has five flat regions; correspondingly, five guiding umbrella potentials are chosen with the locations of their minima within the five regions placed according to a root-mean-square best fit to the experimental PMF.

i.e., the combination of negative PMF plus equal force-constant restraint is recast as a window-dependent restraint. In that case, this is equivalent to running simulations in the i th window on an effective potential

$$V(\mathbf{r}) + w_i^{(g)}(\xi(\mathbf{r})), \quad (8)$$

where this time the guiding potentials,

$$w_i^{(g)}(\xi) = \frac{1}{2} k_i^{(g)} (\xi - \xi_i^{(g)})^2 \quad (9)$$

have optimal $\xi_i^{(g)}$ positions of their minima and optimal $k_i^{(g)}$ force-constant values; the optimization is found from piecewise fitting of the W_{exp} PMF. This second way to perform umbrella sampling using experimental guiding data is the essence of the method we propose.

For the pentapeptide test case we showcase herein, the guiding potentials were chosen such that the center of the harmonic well was located at either a minimum of a well or at the central position of a shoulder. In other words, the piecewise fit of the above guiding potentials to the experimental PMF was done based on the plot of the second derivative $d^2 W_{\text{exp}} / d\xi^2$, i.e., based on the long-scale change of the curvature of the “experimentally” available PMF (see Fig. 2 for details). The value of each force constant $k_i^{(g)}$ was chosen such that the harmonic well approximated the curvature of the PMF around the minimum. The choice of a harmonic potential that approximates the shape of the well is an obvious one, since there would be no point in using a potential wider than the well, and using a narrower potential would require more windows in order to sample over the excluded space. In situations where this does not provide

sufficient overlap, another potential may be added with its center between those of the surrounding potentials and its $k_i^{(g)}$ value also chosen to approximate the curvature of the PMF.

The PMF for the pentapeptide test system used had no barriers; in the case when barriers in the experimental PMF would exist (see Fig. 1), a potential with the center placed at the peak of a barrier would have the $k_i^{(g)}$ value chosen to approximate the (negative) local curvature at the transition point. In the case of the pentapeptide studied here (see Fig. 2), the choice of five w_i windows was obvious (cf. Fig. 2, there are five regions in the curvature plots). For the unfolding of the titin I27 domain, potentials were fitted to the curvature of the experimental PMF. The positions of the potentials were chosen to optimize overlap between umbrella sampling windows. In general applications, an iterative optimization procedure that minimizes a piecewise cost function describing the root-mean-square difference between the experimental PMF and the fitting windows can be done over the relevant set of parameters, i.e., number of windows, positions of the restraints $\xi_i^{(g)}$, and force constants $k_i^{(g)}$.

In any case, regardless of their particular choice, the addition of the guiding potentials $w_i^{(g)}$ in Eq. (9) to the system's unperturbed Hamiltonian can be unbiased using the regular weighted histogram analysis method as presented in Sec. II.

IV. RESULTS

A. Pentapeptide test case

Molecular dynamics simulations were performed on a five amino acid peptide with sequence Cys-Ala-Gly-Gln-Trp with the polar-hydrogen parameter set 19 of the CHARMM force field,⁷⁶ using the analytical continuum electrostatics (ACE II) implicit solvent model,⁷⁷ which is based on an analytical approximation to the generalized Born solvation equation.

The starting structure of the peptide was built using typical internal coordinate values. All simulations were minimized with 1000 steps of steepest descent followed by 2000 steps of the adopted-basis Newton-Raphson procedure and heated up to 300 K in 100 ps. All simulations used Langevin dynamics at 300 K with a (water) frictional coefficient set to 91 ps⁻¹. The reaction coordinate ξ was defined as the distance in Angstroms between the sulfur of cysteine and the center of mass of the tryptophan aromatic ring. The various sets of simulations we ran were as follows:

- (a) Exhaustive sampling simulations totaling 0.5 μ s simulation time were run using constant-temperature molecular dynamics (in the absence of any restraining potential). We generated 25 initial conformations from a 10 ns long equilibration out of which originated 25 equilibrium trajectories, each 20 ns long, with different canonically distributed initial conditions; the conformational snapshots such obtained during the 0.5 μ s cumulative time were thus representative conformations drawn from the canonical ensemble and were binned by the value of the reaction coordinate ξ to generate the experimental potential of mean force. In other words,

this is, in the absence of a real experimental PMF, what we will use as the guiding PMF to which the guiding potentials will be fit.

- (b) Unguided (regular) equal-restraint umbrella sampling simulations run in
 - (i) seventeen windows; $w_i = k_i(\xi - \xi_i)^2$, $i = \overline{1, 17}$, all with the same restraining potential constant, $k_i = 1.2$ kcal/(mol $\text{\AA}^2$) were run for 15 ns each for a total of 255 ns. The windows were placed at positions ξ_i that ranged from 3 to 19 $\text{\AA}$ in 1 $\text{\AA}$ increments. The value of the reaction coordinate was recorded every 10 fs.
 - (ii) five windows; $w_i = k_i(\xi - \xi_i)^2$, $i = \overline{1, 5}$; this second set of umbrella sampling simulations was run using the same range but fewer windows and wider restraining potentials with force constant $k_i = 0.4$ kcal/(mol $\text{\AA}^2$) placed at 3, 7, 11, 15, and 19 $\text{\AA}$ for 15 ns each for a total of 75 ns.
- (c) experimentally guided umbrella sampling with five guiding potentials $w_i^{(g)} = k_i^{(g)}(\xi - \xi_i^{(g)})^2$ fitted to the simulated experimental PMF obtained from the snapshots recorded during the 0.5 μ s exhaustive sampling trajectory (Fig. 2). The guided umbrella sampling potential force constants were: $k_1^{(g)} = 1.1$ kcal/(mol $\text{\AA}^2$), $k_2^{(g)} = 0.11$ kcal/(mol $\text{\AA}^2$), $k_3^{(g)} = 0.07$ kcal/(mol $\text{\AA}^2$), $k_4^{(g)} = 0.12$ kcal/(mol $\text{\AA}^2$), $k_5^{(g)} = 0.3$ kcal/(mol $\text{\AA}^2$), placed at $\xi_1^{(g)} = 5.2$ $\text{\AA}$, $\xi_2^{(g)} = 7.7$ $\text{\AA}$, $\xi_3^{(g)} = 10.6$ $\text{\AA}$, $\xi_4^{(g)} = 13.9$ $\text{\AA}$, and $\xi_5^{(g)} = 16.5$ $\text{\AA}$. Each window was run for 15 ns for a total of 75 ns.

All umbrella sampling distributions were unbiased using a memory efficient implementation of the WHAM [Eqs. (4) and (5)] by Grossfield.⁷⁸ PMFs were also generated for the exhaustive and guided umbrella sampling simulations at various times during the simulations in order to compare their evolution in time.

Comparison of the guided umbrella sampling simulations to unguided, regular, equal-constraint umbrella sampling simulations run in 17 and 5 windows showed that the guided method is more effective and requires fewer windows than the unguided one. Comparison of the final calculated PMFs of the umbrella sampling methods with the PMF from extensive sampling on the unperturbed Hamiltonian at various simulation times shows that only the guided method produces a PMF that mimics the convergence of unbiased simulations.

Importantly, however, the PMF obtained from guided umbrella sampling continued to evolve beyond its input, i.e., beyond the one obtained from the “extensive” simulations, to a slightly (up to 0.69 kcal/mol) different PMF before converging. This implies that, even if the experimental PMF is not accurate, the guided simulation based on it will evolve towards improving the accuracy of the final PMF.

Comparison of the probability distributions and PMFs. An efficient umbrella sampling simulation typically consists of a set of as few restraining windows as possible, with as much overlap between neighboring windows as possible. The histograms for the three umbrella sampling experiments

FIG. 3. Biased histograms for the pentapeptide in the case of (a) the unguided 17 window umbrella sampling, (b) 5 window unguided umbrella sampling, and (c) 5 window guided umbrella sampling after 10 ns per window for each. Note the poor overlap for the histograms in panel (b) relative to (a), and the recovery of the overlap in (c).

are shown in Fig. 3. While the overall shape looks similar for all three, the five window unguided umbrella sampling does not have a substantial amount of overlap between the histograms. This could be improved by adding more windows or altering the restraining potentials, but either would result in the need for more computational time.

The resulting unbiased probability distribution for the reaction coordinate [i.e., the result of unbiasing the histograms in Fig. 3 with the WHAM formula in Eqs. (4) and (5)] were quantitatively compared to the experimental distribution from extensive sampling quantitatively using a Kolmogorov–Smirnov (K-S) test. The K-S statistic⁷⁹ is defined as

$$D = \max_{-\infty < \xi < \infty} |N_1(\xi) - N_2(\xi)|, \quad (10)$$

where $N_1(\xi)$ and $N_2(\xi)$ are the cumulative probability distributions for the two sets to be compared; it has been used previously⁸⁰ as a stringent test for accurately gauging conformational sampling efficiency.

The K-S statistics are plotted as a function of simulation time in Fig. 4. A small value of the K-S statistic indicates a higher probability that the two data sets that are compared are drawn from the same underlying probability distribution, with a value of 0 indicating identical distributions. The results suggest that our guided umbrella sampling approaches the exhaustive sampling probability distribution much more quickly than the 17 window unguided umbrella sampling while the 5 window unguided umbrella sampling never comes as close to reproducing the results of the extensive sampling as the other methods, even after 75 ns. At about 30 ns, the guided umbrella sampling gives a minimum K-S value of 0.025 as compared to extensive sampling, but after 30 ns the K-S value starts to increase. However, if the guided umbrella sampling is compared to the 17 window unguided umbrella sampling after 170 ns (shown in purple in Fig. 4), the similarity of the guided umbrella sampling to the unguided begins to increase at the same time that its similarity to the exhaustive sampling begins to decrease. The K-S sta-

FIG. 4. (Color online) K-S test of similarity for the pentapeptide model. The cumulative probability distribution functions for the distance between the sulfur of cysteine and the center of mass of the aromatic ring of tryptophan of the pentapeptide for the three umbrella sampling methods at different simulation times were compared to the exhaustive simulations at 0.5 μ s. The K-S values for the 17 window umbrella sampling, 5 window umbrella sampling, and the 5 window guided umbrella sampling are shown. The K-S statistics for the guided umbrella sampling at various times as compared to the 17 window umbrella sampling at 170 ns is also shown. The lines are meant as guides to the eye. A lower K-S value indicates higher likelihood that the distributions represent the same data.

tistic for the final distributions of the guided and 17 window unguided umbrella sampling simulations is 0.018. Visual comparison of the probability curves and PMFs shows that the guided umbrella sampling is actually most like the exhaustive one at about 10 ns. The PMFs of the extensive sampling after the full simulation time and, for the guided and 17 window unguided umbrella sampling simulations, at the point in the simulation at which they most closely resembled the guiding extensive sampling, at 10 and 68 ns, respectively, are shown in Fig. 5(a). The guided umbrella sampling result is the closest to the extensive sampling one. The five window unguided umbrella sampling, also shown after the full 75 ns, never resembles the extensive strongly. The final

FIG. 5. (Color online) (a) Potentials of mean force for the end-to-end distance of the pentapeptide for the guiding extensive sampling at 0.5 μ s and for the umbrella sampling methods at the time they most resemble the extensive run: i.e., at 68 ns for the 17 window unguided run and at 10 ns for the guided run. The 5 window unguided umbrella sampling never fully resembles the extensive simulation, not even after 75 ns. (b) PMFs for the 17 window unguided umbrella sampling after the full 255 ns and the guided umbrella sampling after the full 75 ns compared to 0.5 μ s of extensive sampling.

FIG. 6. (Color online) Comparison of probability curves of pentapeptide end-to-end distance at various simulation times. (a) Exhaustive sampling at 0.5 μ s, guided umbrella sampling at 10 ns (2 ns per window), and the 17 window unguided umbrella sampling at 68 ns (4 ns per window). (b) Exhaustive sampling at 0.5 μ s, guided umbrella sampling at 30 ns (6 ns per window), and 17 window unguided umbrella sampling at 102 ns (6 ns per window). It can be seen that the umbrella sampling probabilities are beginning to shift to the right. (c) Probability distribution for the unguided 17 window umbrella sampling is shown at 255 ns (15 ns per window) with the guided umbrella sampling at 50 and 75 ns. There is very little difference in the guided umbrella sampling between 50 and 75 ns.

PMFs after convergence of the umbrella sampling techniques are compared to the extensive sampling in Fig. 5(b). At this point the umbrella sampling techniques yield PMFs that are more similar to each other than to the one obtained from the extensive sampling.

In other words, after reaching a profile very close to that of the extensive sampling, the probability distributions of the guided and 17 window unguided umbrella sampling techniques begin to shift toward the right and eventually resemble each other more than the extensive. These relationships are shown qualitatively in Fig. 6, which shows the probability distributions of extensive sampling with the guided umbrella sampling at 10 ns and the 17 window unguided sampling at 68 ns (a) and the guided sampling at 30 ns and unguided at 102 ns (b). Figure 6(c) shows the final 17 window unguided umbrella sampling probability distribution with those for the guided at 50 and 75 ns. A possible explanation for this drift away from the exhaustive sampling distribution is that there is a transition which is not seen after the 0.5 μ s of extensive sampling, i.e., that the later did not have sufficient time to exhaustively sample to convergence of the right weight of the ξ variable. This possibility seems most likely given the large timescales involved in the relaxation dynamics of many biomolecules of similar size and the fact that the umbrella sampling distributions start out looking similar to the exhaustive then gradually shift to a different distribution.

Time scaling of the sampling efficiency: A timing comparison. In order to test the possibility that there is a transition, the PMFs for the guided umbrella sampling were compared with the guiding extensive simulations at various simulation times using percent similarities and visual comparison. The PMFs appear to line up at several points. Figure 7(a) shows a plot of the extensive simulation times versus the guided umbrella sampling simulation times. An exponential of the form $t_{\text{ex}} = 1.707 \exp(0.485 t_{\text{gu}})$, where t_{ex} and t_{gu} are the total extensive sampling time and the total (cumulative) guided umbrella sampling time, respectively, was the result of an exponential fit to the plot and was used to transform the guided umbrella sampling simulation with respect to time. A plot of the guided umbrella sampling PMF curve as a function of cumulative simulation time is shown in Fig. 7(b). The PMF begins to shift after about 12 ns. Figure 7(c) shows the extensive PMF as a function of cumulative simulation time (purple) overlaid with the first 10 ns of the guided umbrella sampling PMF transformed with the exponential time relationship (red). The two plots line up extremely well with the exception of a slight decrease in the exhaustive around 12 Å from 50 to 200 ns. The plots do not match before 25 ns because the guided umbrella sampling simulations were started with the peptide in a conformation near the median reaction coordinate for each potential. There is no such relationship between the PMFs with respect to time for 17 window unguided umbrella sampling simulations and the extensive simulations.

Sampling of the “perpendicular” coordinates: Dihedral angle distributions. To determine if this method, given a chosen reaction coordinate to restrain, would yield accurate information for degrees of freedom to which no restraining potential was explicitly applied, we also compared the probability distributions of the peptide backbone dihedral angles Φ and Ψ .

Dihedral angle probability curves for all simulations were calculated using a modification of the WHAM equation which addressed the need to unbiased the probabilities of reaction coordinates on which the restraining potential entering the perturbed Hamiltonian did not depend explicitly. The unrestrained reaction coordinate, say, a dihedral angle Φ , was unbiased using the equation

$$\langle p(\Phi) \rangle = \sum_{\xi} \sum_i n_i \langle p(\Phi) \rangle_i \left[\sum_j n_j e^{-(w_j(\xi) - F_j)/k_B T} \right]^{-1}, \quad (11)$$

where ξ is the restrained reaction coordinate and Φ is the unrestrained coordinate. Iterative calculations of p and ξ were done using Eqs. (4) and (5).^{56,81,82}

Since the peptide chosen contains only five amino acids, it was feasible to calculate the unbiased PMFs of all the Φ and Ψ dihedral angles. The individual angles from the 0.5 μ s long extensive sampling simulation were compared with those from the umbrella sampling methods over 68 ns for the 17 window unguided and 10 ns for the guided umbrella sampling protocols (Fig. 8). The five window unguided umbrella sampling simulation was not compared as it failed to produce a valid distribution for the reaction coordinate.

In the particular case of backbone dihedrals for this system, a good amount of similarity is observed for the (unre-

FIG. 7. (Color) Comparison of the exhaustive and guided umbrella sampling PMFs for the pentapeptide end-to-end distance as a function of simulation time. (a) Plot of exhaustive sampling times vs the guided umbrella sampling times with a similar PMF; exponential fit yields $t_{\text{ex}} = 1.707 \exp(0.485 t_{\text{gu}})$. (b) Guided umbrella sampling PMF as a function of simulation time. The time evolution of the potential of mean force of the reaction coordinate shows that the free energy well begins to narrow after about 25 ns and does not converge until about 45 ns. (c) Overlay of the PMFs from the exhaustive sampling (purple) and the guided umbrella sampling (red). The time of the guided umbrella sampling has been transformed using the exponential from panel (a) in order to make it line up with the exhaustive run results. The two simulations follow the same relative time path with respect to the reaction coordinate.

strained) dihedral distribution when comparing exhaustive sampling with umbrella sampling. However, there were exceptions, as detailed below. In general cases, the quality of the match will depend on the amount of coupling between the restrained and unrestrained coordinates. A suitable measure of the degree of relationship between two coordinates has been shown to be provided, when it can be calculated to convergence, by the mutual entropy,⁸² a high value for it indicating that the coordinates are correlated. This correlation is unlikely to be the case for the present two coordinate manifolds (end-to-end distance and backbone dihedrals) because of the pliability of protein backbones. This means that guiding along the end-to-end distance will not enhance significantly the sampling of relevant backbone dihedral values.

FIG. 8. (Color) Free energy profiles for “perpendicular” degrees of freedom, i.e., the pentapeptide dihedral angles. Exhaustive sampling distribution at 0.5 μs is shown in green, 17 window umbrella sampling (68 ns) in blue, and 5 window guided umbrella sampling (10 ns) in red.

As described below, very rare events in the perpendicular degrees of freedom not coupled to the reaction coordinate may still not be seen with this method, and if properties of the perpendicular manifold are desired, then multidimensional umbrella sampling (i.e., sampling along two or more reaction coordinates) should be employed.⁸³ In this regard, it would also be interesting to apply multidimensional umbrella sampling using guiding data from three-color FRET experiments^{43,44} that can report up to three interfluorophore distances.

Despite the general agreement, there were angles which differed in terms of their sampled distributions between the three numerical experiments significantly: the Φ and Ψ angles of alanine, the Ψ angle of glycine, and the Ψ angle of glutamine. The differences between the Ψ angle of alanine for the three techniques are relatively small and may be due in part to the larger range of flexibility of this angle due to the small size of the side chain. The same is true of the glycine Ψ angle, although in this case the guided umbrella sampling yielded results that were more like those from extensive sampling than from the unguided one. The most significant differences are in the Φ dihedral of alanine and Ψ dihedral of glutamine. The extensive simulation gives a small well in the PMF of alanine Φ which is not seen in the other sampling methods. This well is the result of a single event during the 0.5 μs of simulations. The other difference which appears to be significant is in the Ψ angle of glutamine (see Fig. 8). The well at about $-\pi/4$ is overpopulated in the 17 window unguided umbrella sampling and not

seen in the guided umbrella sampling. The guided umbrella sampling protocol appears to have partially brought about crossing of the barrier, but then returned to the more favorable angle. This well also represents occurrences of a rare event, which took place only four times during the extensive simulations. The fact that nonconvergence of the Φ dihedral of alanine and Ψ dihedral of glutamine did not affect the convergence of the distribution of ξ further emphasizes the fact that neither of these dihedrals is expected to be coupled to the reaction coordinate (i.e., that integration over the Boltzmann factor of the dihedral energy term comes out as a constant in the expression for the PMF).

Ideally, in general cases, any efficient umbrella sampling simulations would visit every possible configuration in the directions perpendicular to the reaction coordinate. Unfortunately, given the limited simulation time, this may not always be possible. For example, as seen here, for singular events such as those which contribute to the differences in the PMFs for the alanine Φ and glutamine Ψ dihedrals, it may not be possible to obtain a completely accurate profile without enhancing sampling in those directions. However, in the particular case of weak coupling between the perpendicular coordinates and the reaction coordinate, this lack of accuracy might not affect significantly the quality of the PMF along the reaction coordinate. A related example has to do with the calculation of order parameters that report on the motion of bond vectors in proteins as measured by solution NMR relaxation experiments.⁸⁴ Computer simulation of these experiments have shown that there is good agreement with experiment for backbone order parameters (which can be calculated well with a simple harmonic vibrational approximation^{85–87} or by an even simpler contact model⁸⁸). By contrast, for order parameters of side chains (which are the “perpendicular coordinates” in the language of our previous discussion) agreement to experiment is relatively poorer^{89,90} and a more thorough dynamical sampling is needed.⁹¹ The relatively larger difficulty in converging side-chain versus backbone parameters is because of the more complex motion of the side chains, but the lack of convergence for side chains does not preclude the ability to calculate accurate backbone parameters, which would indicate that there is relatively limited coupling between the backbone and the side chains, at least as far as the order parameter estimator is concerned.

B. Application to titin protein: Unfolding free energy profile

To further test our method to a more complex system, we applied it to the computation of the free energy profile for the forced unfolding of the I27 domain of the muscle protein titin. Because the “force field” used by nature is different from the empirical force field used in the simulations, it was important to test our method with a guiding PMF that was obtained from a different source than the simulation used to compute the final PMF (as was the case for the *simulated* experimental PMF of the pentapeptide in the previous section). To this end, we have used an *actual* experimental PMF: recently, the free energy profile for forced unfolding of the I27 domain of titin has been determined from single mol-

ecule pulling experiments⁷⁴ by using the Jarzynski equality.⁹²

The I27 domain of titin was chosen as a test system for our experimentally guided umbrella sampling method because, in addition to the recently determined PMF, its forced unfolding pathway had previously been studied extensively by both experimental^{74,93–95} and numerical^{96–100} methods. Taken together, the combination of AFM, protein engineering (by ϕ value mutational analysis), molecular dynamics pulling simulations, and NMR structural considerations present a detailed picture of the unfolding pathway of this domain that involves breaking a cluster of hydrogen bonds between parallel beta sheets experiencing shear.

The experimentally derived PMF (Ref. 74) revealed a free energy change of 11.4 kcal/mol between the equilibrium state and a transition state, the latter assumed at a previously determined value of 6 Å extension.⁹⁵ Using this PMF, we ran experimentally guided umbrella sampling simulations on the titin I27 domain using the CHARMM parameters and the implicit solvent model as described in the previous section. We compared the final PMF such obtained with that obtained directly from regular, unguided umbrella sampling simulations. Each umbrella sampling window was equilibrated for 500 ps followed by a production run of 5 ns. The reaction coordinate used was the difference of the distance between the N and C termini from the equilibrium distance, i.e., the equilibrium distance, 46 Å is designated $r=0$, 47 Å is $r=1$, and so on up to $r=8.5$ Å. For guided umbrella sampling, potentials were assigned every 1 or 0.5 Å, with a force constant based on fitting to the experimental PMF. Distance between adjacent windows was determined based on overlap between the restraining potentials. For unguided umbrella sampling, potentials were assigned every 0.5 Å with a force constant of 5 kcal/(mol Å²). The total simulation time was 75 ns for guided umbrella sampling and 90 ns for the unguided umbrella sampling. The starting configuration for each window was taken from the previous window.

Figure 9(a) shows the experimentally derived PMF with a sample of the restraining potentials we used. In order to ensure adequate sampling, potentials were fit to the experimental curve every 1 Å initially and every 0.5 Å in the instances where 1 Å did not provide sufficient overlap between reaction coordinate distributions in adjacent windows. For the fitting, we have used the assumption that the equilibrium native state of titin I27 is 6 Å away from the transition state,⁹⁵ the value also used by Harris *et al.*⁷⁴ to estimate the height of the barrier relative to the folded state. As can be seen from the figure, at the low energy equilibrium state the force constant of the umbrella potential is low. As the free energy increases, so do the force constants. This makes intuitive sense, as stronger potentials will be needed to drive the protein away from its native state. The area of the curve immediately after the transition state cannot be determined directly from experiment, due to snapping of the cantilever after the domain begins to unfold. In this region a constant k value of 5 kcal/(mol Å²) was used.

Figure 9(b) shows the PMFs calculated from the guided umbrella sampling simulation and the unguided umbrella sampling with a constant k value of 5 kcal/(mol Å²) every

FIG. 9. (Color online) (A) Experimental PMF and a selection of guiding umbrella sampling potentials used in the experimentally guided umbrella sampling simulations of unfolding of titin I27. As the simulations progress toward higher energy states, the force constant increases. (B) PMF for guided and unguided umbrella sampling simulations. The two methods produce substantially different curves; in contrast to the unguided runs, the guided sampling converges significantly faster, and to a profile consistent with experimental data (see text).

0.5 Å. The guided umbrella sampling PMF shows an inflection point at 6 Å. While the guided umbrella sampling generates the correct intermediate structure reported to exist at 6 Å, the free energy value at $r=6$ is 26.8 kcal/mol, considerably larger than the experimental value of 11.4 kcal/mol.⁷⁴ This is not entirely surprising, since our unfolding simulations use an implicit solvent model optimized for folded states, and similar differences between free energies calculated by simulation and experiment have been reported previously.^{101,102} The lack of explicit-water hydrogen bonding might additionally limit accurate energy estimation in the particular case of I27, as it has been shown that water molecules can bridge the mechanical unfolding transition state of this protein.¹⁰³

In contrast with the guided PMF, the PMF of the unguided umbrella sampling has a significantly different shape, with an inflection point at 4 Å. By 6 Å, in the unguided PMF calculation, the protein had not yet unfolded, even if the computing time spent (90 ns) exceeded that for the guided simulations (75 ns). To check that the PMF obtained by guiding is indeed the converged free energy profile, three additional simulation sets were run. The first one involved doubling the simulation time of the original 75 ns guided umbrella sampling windows: no noticeable change in the PMF was observed after a total of 150 ns. Second, we increased the simulation time for the unguided sampling and included starting structures in the unguided windows from the original 75 ns guided sampling: in the absence of any guiding potential, the unguided PMF started to converge towards the 75 ns guided PMF after about 165 ns total simulation time [see Fig. 9(b)]. Third, doubling the simulation time of the unguided umbrella sampling with no input from

FIG. 10. (Color) (A) Equilibrated structure of the titin I27 domain. (B) Guided umbrella sampling titin structure for $r=6$ Å. (C) Unguided umbrella sampling structure for $r=6$ Å. The dotted lines mark the distance between C α carbons for residues 3 to 26 and 5 to 24, which are hydrogen bonded in the equilibrium structure. The dissociation of the A strand (in red), residues 3 to 7, is the first step in the forced unfolding pathway, which is correctly reproduced by the guided sampling, but not by the unguided one, even for twice the simulation time (see text).

the guided sampling had no significant convergence effect towards the experimental PMF (data not shown).

The equilibrium structure and the snapshots for the two sets of umbrella sampling simulations [Figs. 10(a)–10(c), respectively] reveal the structural basis for the differences in the PMFs. In the guided umbrella sampling simulations at $r=6$ Å, the N-terminal strand of the domain is pulled away from the C terminus. This is in agreement with the unfolding pathway as determined by experimental ϕ values and forced unfolding simulations,^{97,99} in which titin I27 initially unfolds to an intermediate state with the N-terminal beta sheet partially separated from the rest of the domain before reaching the transition state, in which the N- and C-terminal strands are completely separated from each other. The unguided umbrella sampling simulations, on the other hand, did not replicate this pathway. At 6 Å, the N-terminal beta strand was stretched but not separated from the domain [see Fig. 10(b)]. It was only at 13 Å (results not shown) that the N-terminal strand had begun to pull away, but the overall structure looked more like the guided umbrella sampling intermediate at 6 Å than a more extended transition state. Moreover, a H-bond analysis of the purely unguided simulation showed that the unguided restraining windows by themselves did not follow the unfolding pathway indicated by the above described experimental and forced unfolding studies. In contrast, despite the fact that the guiding PMF was incomplete beyond the transition state, using the experimental PMF as a guide for simulations resulted in an improvement in the accuracy of umbrella sampling in the region of 0–6 Å, accuracy which included the confirmation of the 6 Å transition intermediate and of the putative structural mechanism underlying unfolding.

The application of our guided PMF calculation to I27 unfolding served the purpose to show that it can be applied with experimentally derived data to more complex systems. While a full description of the entire folding free energy surface of the molecule is beyond the scope of the present methodology-oriented paper, the application did reveal that using the experimentally guided PMF yields data on other properties that were not “fitted” in the PMF that is consistent with what is known about the molecule: (i) an inflection point with a free energy change of 2.4 kcal/mol at 1.5 Å, exactly where the experimental studies of Williams *et al.* find the location of a pretransition state from the native to an on-path intermediate (cf. Fig. 2 in Ref. 95), and (ii) structural data on the transition state consistent with NMR, mutational

(ϕ -value analysis), and single molecule experiments. Further simulations of the system of extensive duration are needed to compute the conformational entropy component of the free energy for the unfolded chain beyond the transition point. As well, accurate explicit solvent representation will probably be of relevance to account for the role of water bridging and stabilization of the intermediates identified, in consistency with solvent substitution studies.¹⁰³

V. CONCLUDING DISCUSSION

We have presented a theoretical analysis of the feasibility of an experimentally guided umbrella sampling calculation protocol. Using guiding restraining potentials based on an experimental free energy profile derived from measured equilibrium distributions of a conformational variable, we suggest a method that is able to “guide” umbrella sampling simulations to a converged potential of mean force along that variable. Using the distribution of the end-to-end distance of a pentapeptide model system, the guided umbrella sampling was more efficient than regular, unguided umbrella sampling, requiring fewer windows and giving the same results in less than one sixth the total simulation time. This did not count the simulation time of discarded unguided umbrella sampling runs with restraining potentials which were not effective. The new method also outperformed an extensive sampling done with regular molecular dynamics of 0.5 μ s total time. Our guided umbrella sampling also provided potentials of mean force for unbiased reaction coordinates, the dihedral angles of the peptide, which were similar to those of the extensive sampling, with a few notable exceptions. Given the fact that they did not affect the convergence of the PMF along the constrained coordinate, the exceptions were attributed to negligible coupling between the reaction coordinate values and the respective dihedral angle values.

Our results for the pentapeptide test case also indicated that, with the guided umbrella sampling procedure, there was a transition of the converged free energy profile which went beyond that obtained from the 0.5 μ s of direct sampling, but which is seen with regular umbrella sampling. This is consistent with explicit water simulations of the same peptide and others,⁷³ which needed microsecond scale simulations to converge the results of contact formation observables in peptide loops such that they matched the corresponding experimental values.⁷² Interestingly, although the guided umbrella simulations on the peptide were based on the PMF from the extensive simulations, the distribution of the reaction coordinate began to look more like the 17 window regular umbrella sampling simulations after about 30 ns cumulative time. An implication is that the experimentally guided umbrella sampling protocol we propose will be effective even when the experimental PMF used to determine the restraining potentials is not an exactly accurate PMF. This holds true also for the titin test case, in which experimentally guided umbrella sampling produced better results than the unguided version, despite errors inherent in both the experimental and simulation methods.

The comparison of the guided umbrella sampling and direct exhaustive molecular dynamics of the pentapeptide as

a function of simulation time showed that the guided umbrella sampling simulations roughly follow the same time course as the exhaustive simulation result, but did so exponentially faster. The time plots also showed how the guided PMF evolves in time past the profile reached by exhaustive sampling. Similarly, for the forced unfolding of titin I27, guided umbrella sampling simulations followed an unfolding pathway in agreement with previous simulations and experiment,^{97,99} whereas umbrella sampling with arbitrarily chosen force constants did not.

Although the usefulness of our proposed technique obviously depends on the availability of reliable experimental data, we believe that the potential applications are promising. We intend to extend this method from its use for equilibrium processes to using it for kinetic processes, i.e., for the calculation of time correlation functions. We are currently applying our guiding method to study the mechanism of folding of CI2, for which the PMF for the distance between two residues in various concentrations of denaturant has been determined using single molecule FRET by Weiss and co-workers.^{25,26} While errors in the FRET experiments can lead to inaccuracies in the resulting PMFs,^{46,47} our work shows that umbrella sampling is improved even when the guiding potential is not the exact PMF (as was the case for the pentapeptide), or incomplete (as in the case of titin I27).

Having the experimental guiding potential for large complex biomolecules might prove useful because it can guide sampling in regions of configuration space that would not even be visited by the other methods (because of long time sampling problems). As such, properties of perpendicular degrees of freedom might be revealed that would not have been gauged before. For example, FRET experiments can report on the distance between two residues of CI2 at different folded states, but cannot directly reveal the sequence of folding in the manifold perpendicular to the interfluorophore distance.^{25,26} In addition, in comparison to techniques such as steered,¹⁰⁴ biased,¹⁰⁵ or targeted molecular dynamics,¹⁰⁶ which use information about the final state of a system to direct simulations that reveal information about a pathway, this technique can use experimental information about a reaction coordinate to direct simulations in a way that can reveal more detailed information about a system in which details of the final state may not be known. This may be important, as studies on several proteins including titin¹⁰⁷ indicate, to discern between unfolding by pulling at one end through a pore versus at both end,⁴¹ or to assess the dependence on the direction of pulling.¹⁰⁸

The insights derived from the PMF curve, an inherent equilibrium calculation, can be further enhanced by advanced dynamical sampling techniques such as transition path sampling.^{109–111} This powerful method for calculating kinetics—whose use in biomolecular applications has been reported^{112–114}—can be applied with transition points from the structures generated from the guided simulations to obtain dynamical trajectories that span the length between the initial and final states of the PMF. These trajectories can probe if the dynamics progresses as expected from the equilibrium assumption inherent to the PMF calculations (i.e., to probe whether the choice of the reaction coordinate as the

slowest order parameter was appropriate,¹¹⁵ or if coupling to other degrees of freedom is significant^{116,117}). Because transition path sampling is a method that does not require the definition of a reaction coordinate, the ensembles of dynamical transition trajectories generated by it (or by other long-time dynamical methods^{118–123}) can be grouped together and further analyzed to obtain information about distinct reaction pathways.

In the same vein, there may be exciting applications that have to do with heterogeneity and the apparent nonergodicity¹²⁴ observed in single molecule studies of complex molecules. For example, in the case of the hairpin ribozyme, single molecule FRET revealed complex structural dynamics that would have been difficult to detect in ensemble measurements. This ribozyme, which undergoes a conformational change from an undocked to a docked conformation has several distinct docked/undocked substates with different rates of undocking, and the ribozyme exhibited a memory effect, where switching between different docked substates was rarely seen.¹²⁵ This apparent non-Markovian, memory effect in the hairpin ribozyme is presumably due to different conformations of two loops within the molecule.¹²⁵ These different conformations cannot be directly detected experimentally, but could be probed during a molecular dynamics simulation that is guided along a distinct experimental (single molecule) PMF.

ACKNOWLEDGMENTS

M.M. was partially supported by an NIH Molecular Biophysics Predoctoral Research Training Program. I.A. acknowledges support from the NSF CAREER award program (CHE-0548047) and the donors of the ACS Petroleum Research Foundation.

- ¹J. G. Kirkwood, *J. Chem. Phys.* **3**, 300 (1935).
- ²B. Roux and M. Karplus, *Biophys. J.* **59**, 961 (1991).
- ³B. Roux, *Comput. Phys. Commun.* **91**, 275 (1995).
- ⁴L. R. Pratt and D. Chandler, *J. Chem. Phys.* **67**, 3683 (1977).
- ⁵J. Gao, *J. Am. Chem. Soc.* **113**, 7796 (1991).
- ⁶K. Hinsen and B. Roux, *J. Chem. Inf. Comput. Sci.* **37**, 1018 (1997).
- ⁷A. Smondyrev and G. Voth, *Biophys. J.* **82**, 1460 (2002).
- ⁸H. Chen, Y. Wu, and G. Voth, *Biophys. J.* **93**, 3470 (2007).
- ⁹B. Roux, T. Allen, S. Berneche, and W. Im, *Q. Rev. Biophys.* **37**, 15 (2004).
- ¹⁰B. M. Pettitt and M. Karplus, *Chem. Phys. Lett.* **121**, 194 (1985).
- ¹¹D. J. Tobias and C. L. Brooks III, *J. Phys. Chem.* **96**, 3864 (1992).
- ¹²E. Boczek and C. Brooks, *Science* **269**, 393 (1995).
- ¹³N. Huang, N. Banavali, and J. MacKerell, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 68 (2003).
- ¹⁴K. Hart, B. Nystrom, M. Ohman, and L. Nilsson, *RNA* **11**, 609 (2005).
- ¹⁵J. Wereszczynski and I. Andricioaei, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 16200 (2006).
- ¹⁶E. Carter, G. Ciccotti, J. Haynes, and R. Kapral, *Chem. Phys. Lett.* **156**, 472 (1989).
- ¹⁷M. Sprik and G. Ciccotti, *J. Chem. Phys.* **109**, 7737 (1998).
- ¹⁸K. Kuczera, *J. Comput. Chem.* **17**, 1726 (1996).
- ¹⁹Y. Wang and K. Kuczera, *J. Phys. Chem. B* **101**, 5205 (1997).
- ²⁰G. Li and Q. Cui, *J. Mol. Graphics Modell.* **24**, 82 (2005).
- ²¹D. Andrews and A. Demidov, *Resonance Energy Transfer* (Wiley, New York, 1999).
- ²²A. Cooper, *Proc. Natl. Acad. Sci. U.S.A.* **73**, 2740 (1976).
- ²³T. L. Hill, *Thermodynamics of Small Systems* (Dover, New York, 1994).
- ²⁴C. Bustamante, J. Liphardt, and F. Ritort, *Phys. Today* **58**, 43 (2005).
- ²⁵A. A. Deniz, T. A. Laurence, G. S. Beligere, M. Dahan, A. B. Martin, D. S. Chemla, P. E. Dawson, P. G. Schultz, and S. Weiss, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5179 (2000).
- ²⁶S. Weiss, *Nat. Struct. Biol.* **7**, 724 (2000).
- ²⁷B. Schuler and W. Eaton, *Curr. Opin. Struct. Biol.* **18**, 16 (2008).
- ²⁸N. G. Walter, D. A. Harris, M. J. B. Pereira, and D. Rueda, *Biopolymers* **61**, 224 (2001).
- ²⁹M. Carrion-Vazquez, A. F. Oberhauser, T. E. Fisher, P. E. Marszalek, H. Li, and J. M. Fernandez, *Prog. Biophys. Mol. Biol.* **74**, 63 (2000).
- ³⁰C. Bustamante, J. C. Macosko, and G. J. L. Wuite, *Nat. Rev. Mol. Cell Biol.* **1**, 130 (2000).
- ³¹T. R. Strick, M. N. Dessinges, G. Charvin, N. H. Dekker, J. F. Allemand, D. Bensimon, and V. Croquette, *Rep. Prog. Phys.* **66**, 1 (2003).
- ³²G. Hummer and A. Szabo, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 3659 (2001).
- ³³J. MacFadyen and I. Andricioaei, *J. Chem. Phys.* **123**, 074107 (2005).
- ³⁴M. Schlierf and M. Rief, *Biophys. J.* **90**, L33 (2006).
- ³⁵O. K. Dudko, G. Hummer, and A. Szabo, *Phys. Rev. Lett.* **96**, 108101 (2006).
- ³⁶D. West, P. Olmsted, and E. Paci, *J. Chem. Phys.* **125**, 204910 (2006).
- ³⁷J. Nummela, F. Yassin, and I. Andricioaei, *J. Chem. Phys.* **128**, 024104 (2008).
- ³⁸J. Nummela and I. Andricioaei, *Biophys. J.* **93**, 3373 (2007).
- ³⁹S. Kirmizialtin, L. Huang, and D. E. Makarov, *J. Chem. Phys.* **122**, 234915 (2005).
- ⁴⁰A. Matouschek, *Curr. Opin. Struct. Biol.* **13**, 98 (2003).
- ⁴¹P. Tian and I. Andricioaei, *J. Mol. Biol.* **350**, 1017 (2005).
- ⁴²M. T. Woodside, P. C. Anthony, W. M. Behnke-Parks, K. Larizadeh, D. Herschlag, and S. M. Block, *Science* **314**, 1001 (2006).
- ⁴³S. Hohng, C. Joo, and T. Ha, *Biophys. J.* **87**, 1328 (2004).
- ⁴⁴J. P. Clamme and A. Deniz, *ChemPhysChem* **6**, 74 (2005).
- ⁴⁵C. L. Ting and D. E. Makarov, *J. Chem. Phys.* **128**, 115102 (2008).
- ⁴⁶B. Schuler, E. Lipman, and W. Eaton, *Nature (London)* **419**, 743 (2002).
- ⁴⁷I. Gopich and A. Szabo, *J. Phys. Chem. B* **107**, 16111 (2003).
- ⁴⁸M. Karplus and J. A. McCammon, *Nat. Struct. Biol.* **9**, 646 (2002).
- ⁴⁹R. Palmer, *Adv. Phys.* **31**, 669 (1982).
- ⁵⁰G. M. Torrie and J. P. Valleau, *J. Comput. Phys.* **23**, 187 (1977).
- ⁵¹A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.* **63**, 1195 (1989).
- ⁵²A. M. Ferrenberg, *Phys. Rev. Lett.* **63**, 1658 (1989).
- ⁵³S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, *J. Comput. Chem.* **13**, 1011 (1992).
- ⁵⁴M. Mezei, *J. Comput. Phys.* **68**, 237 (1987).
- ⁵⁵R. W. W. Hooft, B. P. Vaneijck, and J. Kroon, *J. Chem. Phys.* **97**, 6690 (1992).
- ⁵⁶C. Bartels and M. Karplus, *J. Comput. Chem.* **18**, 1450 (1997).
- ⁵⁷C. Bartels and M. Karplus, *J. Phys. Chem. B* **102**, 865 (1998).
- ⁵⁸M. Schaefer, C. Bartels, and M. Karplus, *J. Mol. Biol.* **284**, 835 (1998).
- ⁵⁹E. Darve, M. A. Wilson, and A. Pohorille, *Mol. Simul.* **28**, 113 (2002).
- ⁶⁰A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12562 (2002).
- ⁶¹B. A. Berg and T. Neuhaus, *Phys. Lett. B* **267**, 249 (1991).
- ⁶²B. A. Berg and T. Neuhaus, *Phys. Rev. Lett.* **68**, 9 (1992).
- ⁶³U. H. E. Hansmann and Y. Okamoto, *J. Comput. Chem.* **14**, 1333 (1993).
- ⁶⁴F. Wang and D. Landau, *Phys. Rev. Lett.* **86**, 2050 (2001).
- ⁶⁵J. Lee, *Phys. Rev. Lett.* **71**, 211 (1993).
- ⁶⁶N. Rathore, T. Knotts, and J. de Pablo, *J. Chem. Phys.* **118**, 4285 (2003).
- ⁶⁷M. Shell, P. Debenedetti, and A. Panagiotopoulos, *J. Chem. Phys.* **119**, 9406 (2003).
- ⁶⁸J. Kim, J. Straub, and T. Keyes, *Phys. Rev. Lett.* **97**, 050601 (2006).
- ⁶⁹T. C. Beutler and W. F. van Gunsteren, *J. Chem. Phys.* **100**, 1492 (1994).
- ⁷⁰R. B. Best and M. Vendruscolo, *J. Am. Chem. Soc.* **126**, 8090 (2004).
- ⁷¹M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus, *J. Am. Chem. Soc.* **125**, 15686 (2003).
- ⁷²L. J. Lapidus, W. A. Eaton, and J. Hofrichter, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 7220 (2000).
- ⁷³I. C. Yeh and G. Hummer, *J. Am. Chem. Soc.* **124**, 6563 (2002).
- ⁷⁴N. C. Harris, Y. Song, and C.-H. Kiang, *Phys. Rev. Lett.* **99**, 068101 (2007).
- ⁷⁵M. N. Kobra, *J. Comput. Chem.* **24**, 1437 (2003).
- ⁷⁶B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *J. Comput. Chem.* **4**, 187 (1983).
- ⁷⁷M. Schaefer and M. Karplus, *J. Phys. Chem.* **100**, 1578 (1996).
- ⁷⁸A. Grossfield, source code from <http://membrane.urmc.rochester.edu/index.html>.
- ⁷⁹W. Press, B. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in FORTRAN 77: The Art of Scientific Computing* (Cambridge

- University Press, Cambridge, 1992).
- ⁸⁰ I. Andricioaei and J. E. Straub, *J. Chem. Phys.* **107**, 9117 (1997).
- ⁸¹ E. Gallicchio, M. Andrec, A. K. Felts, and R. M. Levy, *J. Phys. Chem. B* **109**, 6722 (2005).
- ⁸² D. D. L. Minh, *J. Phys. Chem. B* **111**, 4137 (2007).
- ⁸³ S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, *J. Comput. Chem.* **16**, 1339 (1995).
- ⁸⁴ G. Lipari, A. Szabo, and R. M. Levy, *Nature (London)* **300**, 197 (1982).
- ⁸⁵ E. R. Henry and A. Szabo, *J. Chem. Phys.* **82**, 4753 (1985).
- ⁸⁶ R. Brüschweiler and D. A. Case, *Phys. Rev. Lett.* **72**, 940 (1994).
- ⁸⁷ S. Sunada, N. Go, and P. Koehl, *J. Chem. Phys.* **104**, 4768 (1996).
- ⁸⁸ F. Zhang and R. Brüschweiler, *J. Am. Chem. Soc.* **124**, 12654 (2002).
- ⁸⁹ L. J. Smith, A. E. Mark, C. M. Dobson, and W. F. van Gunsteren, *Biochemistry* **34**, 10918 (1995).
- ⁹⁰ D. Ming and R. Brüschweiler, *J. Biomol. NMR* **29**, 363 (2004).
- ⁹¹ R. Best, J. Clarke, and M. Karplus, *J. Mol. Biol.* **349**, 185 (2005).
- ⁹² C. Jarzynski, *Phys. Rev. Lett.* **78**, 2690 (1997).
- ⁹³ P. E. Marszalek, H. Lu, H. B. Li, M. Carrion-Vazquez, A. F. Oberhauser, K. Schulten, and J. M. Fernandez, *Nature (London)* **402**, 100 (1999).
- ⁹⁴ H. Li, A. F. Oberhauser, S. B. Fowler, J. Clarke, and J. M. Fernandez, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6527 (2000).
- ⁹⁵ P. M. Williams, S. B. Fowler, R. B. Best, J. L. Toca-Herrera, K. A. Scott, A. Steward, and J. Clarke, *Nature (London)* **422**, 446 (2003).
- ⁹⁶ H. Lu and K. Schulten, *Biophys. J.* **79**, 51 (2000).
- ⁹⁷ S. B. Fowler, R. B. Best, J. L. T. Herrera, T. J. Rutherford, A. Steward, E. Paci, M. Karplus, and J. Clarke, *J. Mol. Biol.* **322**, 841 (2002).
- ⁹⁸ P. C. Li and D. E. Makarov, *J. Chem. Phys.* **119**, 9260 (2003).
- ⁹⁹ R. B. Best, S. B. Fowler, J. L. T. Herrera, A. Steward, E. Paci, and J. Clarke, *J. Mol. Biol.* **330**, 867 (2003).
- ¹⁰⁰ G. Hummer and A. Szabo, *Biophys. J.* **85**, 5 (2003).
- ¹⁰¹ R. H. Zhou and B. J. Berne, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12777 (2002).
- ¹⁰² H. Nymeyer and A. E. Garcia, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 13934 (2003).
- ¹⁰³ L. Dougan, G. Feng, H. Lu, and J. Fernandez, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 3185 (2008).
- ¹⁰⁴ B. Isralewitz, M. Gao, and K. Schulten, *Curr. Opin. Struct. Biol.* **11**, 224 (2001).
- ¹⁰⁵ E. Paci and M. Karplus, *J. Mol. Biol.* **288**, 441 (1999).
- ¹⁰⁶ J. Schlitter, M. Engels, and P. Kruger, *J. Mol. Graphics* **12**, 84 (1994).
- ¹⁰⁷ T. Sato, M. Esaki, J. M. Fernandez, and T. Endo, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 17999 (2005).
- ¹⁰⁸ D. K. West, D. J. Brockwell, and E. Paci, *Biophys. J.* **91**, L51 (2006).
- ¹⁰⁹ C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, *J. Chem. Phys.* **108**, 1964 (1998).
- ¹¹⁰ C. Dellago, P. G. Bolhuis, and D. Chandler, *J. Chem. Phys.* **110**, 6617 (1999).
- ¹¹¹ C. Dellago, P. G. Bolhuis, and P. L. Geissler, *Adv. Chem. Phys.* **123**, 1 (2002).
- ¹¹² P. G. Bolhuis, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12129 (2003).
- ¹¹³ J. Marti and F. Csajka, *Phys. Rev. E* **69**, 061918 (2004).
- ¹¹⁴ R. Radhakrishnan and T. Schlick, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 5970 (2004).
- ¹¹⁵ P. G. Bolhuis, C. Dellago, and D. Chandler, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5877 (2000).
- ¹¹⁶ A. Ma, A. Nag, and A. R. Dinner, *J. Chem. Phys.* **124**, 144911 (2006).
- ¹¹⁷ R. Best, E. Paci, G. Hummer, and O. Dudko, *J. Phys. Chem. B* **112**, 5968 (2008).
- ¹¹⁸ R. Olender and R. Elber, *J. Chem. Phys.* **105**, 9299 (1996).
- ¹¹⁹ R. Elber, J. Meller, and R. Olender, *J. Phys. Chem. B* **103**, 899 (1999).
- ¹²⁰ A. K. Faradjian and R. Elber, *J. Chem. Phys.* **120**, 10880 (2004).
- ¹²¹ D. Moroni, P. G. Bolhuis, and T. S. van Erp, *J. Chem. Phys.* **120**, 4055 (2004).
- ¹²² A. F. Voter, *Phys. Rev. B* **57**, R13985 (1998).
- ¹²³ M. R. Shirts and V. S. Pande, *Phys. Rev. Lett.* **86**, 4983 (2001).
- ¹²⁴ Z. Xie, N. Srividya, T. R. Sosnick, T. Pan, and N. F. Scherer, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 534 (2004).
- ¹²⁵ X. W. Zhuang, H. Kim, M. J. B. Pereira, H. P. Babcock, N. G. Walter, and S. Chu, *Science* **296**, 1473 (2002).