

# Monte Carlo vs Molecular Dynamics for All-Atom Polypeptide Folding Simulations

Jakob P. Ulmschneider,<sup>\*,†</sup> Martin B. Ulmschneider,<sup>‡</sup> and Alfredo Di Nola<sup>†</sup>

Department of Chemistry, University of Rome “La Sapienza”, Rome, Italy, and Department of Biochemistry, University of Oxford, Oxford, U.K.

Received: March 15, 2006; In Final Form: June 19, 2006

An efficient Monte Carlo (MC) algorithm including concerted rotations is directly compared to molecular dynamics (MD) in all-atom statistical mechanics folding simulations of small polypeptides. The previously reported algorithm “concerted rotations with flexible bond angles” (CRA) has been shown to successfully locate the native state of small polypeptides. In this study, the folding of three small polypeptides (trpzip2/H1/Trp-cage) is investigated using MC and MD, for a combined sampling time of  $\sim 10^{11}$  MC configurations and 8  $\mu$ s, respectively. Both methods successfully locate the experimentally determined native states of the three systems, but they do so at different speed, with 2–2.5 times faster folding of the MC runs. The comparison reveals that thermodynamic and dynamic properties can reliably be obtained by both and that results from folding simulations do not depend on the algorithm used. Similar to previous comparisons of MC and MD, it is found that one MD integration step of 2 fs corresponds to one MC scan, revealing the good sampling of MC. The simplicity and efficiency of the MC method will enable its future use in folding studies involving larger systems and the combination with replica exchange algorithms.

## Introduction

The folding of proteins into their native structure is one of the most challenging and interesting problems of molecular biology. In addition to much experimental effort, recent advances in computer simulation techniques have enabled the direct study of the folding process using all-atom representation models. Due to the high computational cost of explicit solvent representation, there has been increased use of implicit solvation models, which reduce the computational burden through a continuum treatment of the solvent. Of these, the generalized Born (GBSA) solvent model has been widely applied because it is computationally efficient and superior to earlier, simpler alternatives such as surface-area or distance-dependent dielectric models.<sup>1</sup> Originally developed by Still et al.,<sup>2</sup> the model is an extension of the Born treatment of ionic solvation to solutes containing any set of charged sites and having arbitrary molecular shape. Although GBSA can be criticized for a range of problems inherent in implicit solvation models, it has been shown to accurately reproduce relative free energies of different peptide conformations<sup>1</sup> and to identify correctly the native state of several large proteins in an extensive comparison with large decoy sets.<sup>3</sup> Many recent studies have used molecular dynamics (MD) simulations coupled with the GBSA model to fold small polypeptides in direct simulations.<sup>4–7</sup> Impressive success is achieved if this methodology is further combined with massively parallel computing,<sup>8–10</sup> or replica exchange MD.<sup>11–13</sup>

Similar results can be obtained using Monte Carlo (MC) statistical mechanics instead of MD, as demonstrated recently by folding several small polypeptides having  $\beta$ -hairpin and  $\alpha$ -helical structures.<sup>14</sup> The use of MC is motivated by the potential advantages over MD: In particular, energy derivatives are not needed including the costly ones for the GBSA free

energy. New potential functions and solvation models for use in simulations can be rapidly tested without the need to first determine the usually complicated analytical derivatives. Furthermore, MC moves can take better advantage of the implicit nature of the solvent by enabling large conformational changes to cross efficiently over energy barriers. Another advantage of MC is its simplicity: Equilibrium simulations in NVT and NPT ensembles can be studied without having to use the various thermostats and barostats necessary in MD, whose effect on simulation results is not completely clear. Also, MC simulations are run with fully flexible bonds without slowing down performance, while in MD runs bonds involving hydrogen atoms need to be constrained using methods such as LINCS<sup>15</sup> to be computationally efficient.

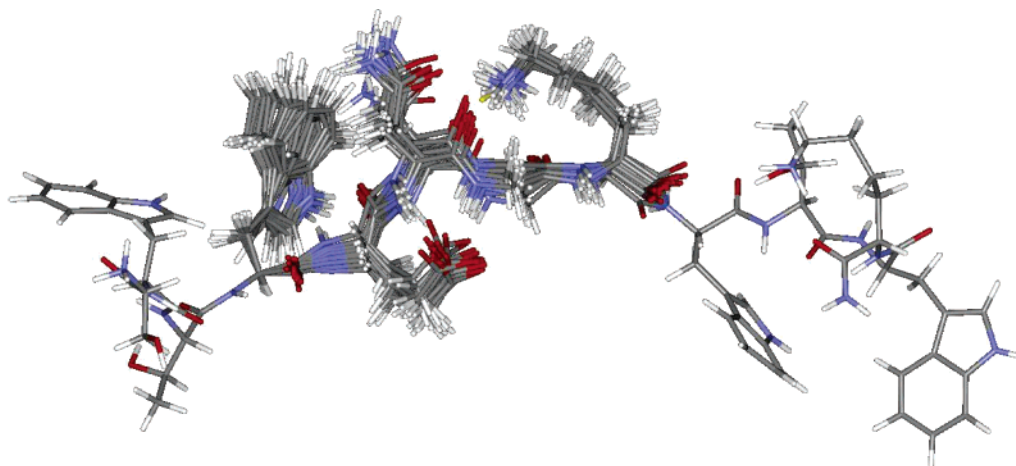
Among the prerequisites for using MC to simulate polymer dynamics are local backbone moves termed *concerted rotations*. Such moves avoid the inefficient global conformational changes of simple MC backbone moves and enable computational efficiency due to their locality.<sup>16,17</sup> A visual representation of the effect of the backbone moves is shown in Figure 1. The strength of a MC approach using concerted rotations and GBSA for folding polypeptides and identifying their native structures in aqueous solution has been demonstrated previously.<sup>14</sup>

In this work, an attempt is made to evaluate the relative strengths and weaknesses of both MC and MD in folding small polypeptides. The systems studied are the tryptophan “zipper” trpzip2, the amyloidogenic H1 peptide from the syrian hamster prion protein, and the Trp-cage. In particular, it will be of interest to compare how quickly meaningful thermodynamic properties can be obtained, how fast system observables converge, and whether the native state is reliably determined independent of the simulation method used. In addition to the assessment of equilibrium properties, also basic kinetic behavior can be studied. The multidimensional free energy surface of a polypeptide with many degrees of freedom is thought to be extremely rugged, with many stable local minima and significant barriers

\* To whom correspondence should be addressed. E-mail: Jakob@ulmschneider.com.

<sup>†</sup> University of Rome “La Sapienza”.

<sup>‡</sup> University of Oxford.



**Figure 1.** Visual demonstration of the concerted rotation Monte Carlo algorithm with flexible torsion and bond angles. The graph shows the sampling of a segment of a polypeptide, with 50 overlaid structures from a short MC run of  $10^5$  steps. Note that the conformational change is restricted to a local window of the protein backbone. In a full Monte Carlo simulation, the position of the window is varied randomly along the chain, and three side chain moves are attempted for every backbone move. For this figure, the side chain moves were switched off in the simulation to reveal the sampling effect of the backbone moves only.

that delay the folding of the system into the global free energy minimum of the native state.<sup>18</sup> On a short simulation time scale MC runs will yield a Markov chain of conformations that does not incorporate a time element. On a much longer time scale, determined by the lag time necessary to escape from main local minima, it is expected that MC and MD will show similar behavior, since both methods are expected to require a comparable effort to escape from the same large conformational trap or cross the same large kinetic barrier. Thus, the comparison of both methods will enable an assessment of how many MC moves/scans on average correspond to an MD time step, a quantity that is unique to the system studied. Finally an efficiency comparison can be made taking into account the CPU time used.

### Simulation Methods

The MC simulations were run with a MC program developed by the authors especially for the simulation of protein folding and includes the concerted rotations, as detailed in the original report.<sup>16</sup> A newer version of the concerted rotation algorithm was used in which the root search of the chain closure was replaced by an analytical solution, resulting in slightly better performance. Each simulated system consisted of just a single copy of the polypeptide. Normal protonation states were adopted for pH 7, i.e., deprotonated carboxylic acids and protonated amines and guanidines; the termini were treated to reproduce the experimental conditions: acetylated N-terminus and amidated C-terminus for the H1 peptide,<sup>19</sup> charged N-terminus and amidated C-terminus for trpzip2,<sup>20</sup> and charged termini for the Trp-cage.<sup>21</sup> The potential energy was evaluated with the OPLS-AA force field,<sup>22</sup> and the simulations used sampling at a temperature of 300 K for the H1 peptide and 323 K for the trpzip2 and Trp-cage. The full potential energy was evaluated with no cutoffs for the nonbonded interactions and with a dielectric constant of 1 for the Coulombic interactions. The utilized GBSA method was the fast asymptotic pairwise summation model developed by Qiu, Still, and co-workers,<sup>23</sup> which has been demonstrated to yield excellent results in predicting experimental free energies of solvation as well as hydration effects on conformational equilibria.<sup>24</sup> The electrostatic energy and, therefore, the Born radii are recomputed for every MC configuration; the constituent atomic radii are taken from the OPLS-AA force field ( $r = 0.5\sigma$ ) except in the case of

hydrogens for which radii of 1.15 Å are assigned, as in the original study.<sup>23</sup> For the MC simulations, the GB energy was only updated for the part of the molecule close to the move site, significantly increasing the performance with only a minimum loss of accuracy. The nonpolar contribution to the solvation free energy was calculated as in the original method by Still et al.<sup>2</sup> to be proportional to the total solvent accessible surface area (SASA) with an effective surface tension of 4.9 cal/(mol Å<sup>2</sup>). The SASA was computed using a probe radius of 1.4 Å. Since exact SASA calculations are usually time-consuming, SASA is slowly varying, and the contributions of the SASA term to the free energies are relatively small, a SASA mimic based on the Born radii was used, which has been shown to be very accurate, but much faster.<sup>25</sup>

For each polypeptide a series of eight MC runs was performed starting from completely stretched conformations. Attempted MC backbone moves were made every fourth MC step; the remainder were single side chain moves, which are rapid. Each of the MC runs was 4 billion configurations of length, for a total simulation “time” of  $3.2 \times 10^{10}$  configurations. Instead of MC steps, we have used the more meaningful quantity of a MC “scan” or “sweep”, which is defined as the number of MC steps required to move—on average—each residue of the system once. Thus, 1 MC scan = 12 MC steps for trpzip2, 14 MC steps for H1, and 20 MC steps for Trp-cage.

The MD simulations, with fixed bond lengths<sup>15</sup> and a time step of 2 fs for numerical integration were performed with the GROMACS software package,<sup>26</sup> modified by us to include the GBSA implicit solvation model described above. The setup was identical to the MC runs, and the eight simulations per system were run for  $250 \times 10^6$  time steps (500 ns) each, resulting in an aggregate simulation time of  $2 \times 10^9$  time steps (4 μs). The systems were coupled to a heat bath at the respective simulation temperature (see below) and a time constant of  $\tau = 0.1$  ps using a Berendsen thermostat.<sup>27</sup>

Considerable effort was spent to ensure that both the MC and MD simulations sample the polypeptide systems using identical potential functions. A fit of the total molecular mechanics system energy plus the GBSA solvation energy for 1000 evenly spaced conformations along a trpzip2 trajectory, as calculated by MD and MC, gave a correlation coefficient of  $r^2 > 0.99$  and a root mean square error of  $\sim 0.1\%$ , which was the accuracy of the input data.

For both the trpzip2 and H1 runs, a cluster analysis was carried out to assess the main secondary structural motifs populated during the simulations. The pairwise method of Daura et al.<sup>28</sup> was employed. Since clustering becomes computationally costly for large coordinate sets, the structures were taken every millionth MC step/every 200 ps and were superimposed using main-chain least-squares fitting (RMSD), with a similarity cutoff (the maximum value of the RMSD of a cluster member to the cluster center) of 1.8 Å. Since most clusters are sparsely populated and do not represent the main features of the simulations, only the most populated clusters were further analyzed. In the case of H1, clusters with similar secondary structural motifs were grouped together as a result of clustering.

The principal experimental data for comparison with the present simulation results are the structures of the polypeptides as obtained from detailed NMR studies. For the H1 peptide, a low-level X-ray structure is available.<sup>29</sup> To quantify the similarity to the native state, we aligned each conformation to the C<sub>α</sub> positions of the relevant experimental structure and calculated the root-mean-square C<sub>α</sub> deviation (RMSD). The reference NMR structure for trpzip2 was the most representative conformer, 1, of the 20 submitted structures (PDB code 1LE1).<sup>20</sup> For H1, a previously determined β-hairpin structure derived from the low-level X-ray conformer was used.<sup>30</sup> The reference structure for Trp-cage (PDB code 1L2Y) was conformer 1 of the 38 refined NMR structures.<sup>21</sup> Since the conformational space of even small polypeptides has many degrees of freedom, it is helpful to choose order parameters and project the ensemble into two dimensions. We constructed a free energy function in terms of surfaces with the axis of RMSD to the experimental structure and the radius of gyration R<sub>g</sub>. For a system in thermodynamic equilibrium, the change in free energy on going from one state of the system to another is given by

$$\Delta G = -RT \ln \frac{p_1}{p_2}$$

where  $R$  is the ideal gas constant,  $T$  is the temperature, and  $p_i$  is the probability of finding the system in state  $i$ . A two-dimensional space of peptide conformations was divided into a grid with a spacing of 0.2 Å in both dimensions, and the free energy (the negative logarithm of the population) was calculated for each bin. In addition, the average total energy  $E$  was calculated for each cell, which is the sum of peptide internal energy plus the interactions with the solvent given by the GBSA solvation free energy. As such it includes the solvation entropy. Finally the solute entropy was obtained using

$$\Delta S = \frac{\Delta E - \Delta G}{T}$$

for each bin. All values for  $G$  and  $E$  have been shifted such that the lowest value of the free energy surface and the potential energy surface is zero. Thus, the reported surface values  $\Delta G$  are the transfer free energies with respect to the bin that has been set to zero. Similarly, the  $\Delta E$  value of each bin is the relative potential energy to the bin with  $E = 0$ .

To check the effect of a different grid size on the thermodynamic properties, different grid spacings of 0.1–0.5 Å were used to construct the same free energy surfaces. Surfaces constructed on smaller bin sizes tended to be rougher as fewer points are available per bin, but all were similar, with the same overall shape and spread of the free energy. Finally, the folding time (the average time required to reach the native state) and dwell time (the average time the system remains in the native

state) of folded states were estimated by averaging over the small number of folding events encountered during the simulations.

## Results

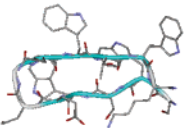
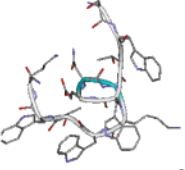
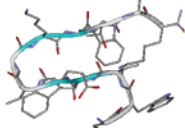
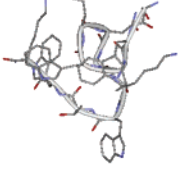
**Structural Comparison of trpzip2.** The first test system is the tryptophan zipper trpzip2. Tryptophan zippers are a series of small peptides recently synthesized by Cochran et al.<sup>20</sup> Despite their size of only 12–16 residues, they form remarkably stable β-hairpins in aqueous solution characterized by a structural motif of tryptophan–tryptophan cross-strand pairs. The system trpzip2 (sequence SWTWENGKWTWK) was chosen for simulation due to its high stability. CD spectroscopy and NMR experiments reveal a β-hairpin with a type I' β-turn at the Asn-Gly junction.<sup>20</sup> To speed up conformational sampling, the temperature was set to 323 K, close to the experimental melting temperature of 345 K.<sup>20</sup>

First, we compared the various different conformations sampled during both the MC and MD simulations. All runs show an immediate relaxation of the system from the extended state. This initial fast collapse is much faster in terms of CPU time for the MD simulations (<0.05 × 10<sup>6</sup> time steps, or 100 ps) than for MC (<1.7 × 10<sup>6</sup> MC scans). Straight downhill folding from a high-energy unfolded conformation to a compact equilibrium-like structure is akin to a minimization for the MD algorithm, enhanced in this case by the use of implicit solvent. On the other hand, in MC simulations the energy can only be slowly lowered in a rattle-like fashion converging on the minimum only after many accepted and rejected moves.

Once collapsed conformations are reached, the initial difference in the dynamics of MC and MD vanishes, and both show frequent transitions between compact folded states. Multiple folding/unfolding cycles are observed, and the system can be trapped in compact coil or β-hairpin conformations for considerable time, usually with several backbone hydrogen bonds formed. A cluster analysis of the MC simulations reveals the main secondary structure motifs sampled (Table 1): The native state is the most populated cluster, with 24.2% of all structures. A type I' β-turn with the Gly at position 3, Glu-Asn-Gly-Lys is found as observed in the NMR measurements. All native interstrand hydrogen bonds are formed, and the average backbone RMSD to the NMR structure is 0.8 ± 0.2 Å. A partially folded state is populated 4.8% of the simulation time and is characterized by the core of the hairpin formed but with the last two interstrand hydrogen bonds broken and a salt bridge between the charged N-terminus and the COO<sup>-</sup> of Glu5. It has a rather short lifetime of about 11 × 10<sup>6</sup> MC scans. A main misfolded structure encountered in the simulation (20.9%) is represented as cluster 2: This coiled conformation is stabilized by three interstrand hydrogen bonds, a salt bridge between the NH<sub>3</sub><sup>+</sup> of Lys12 and the COO<sup>-</sup> of Glu5, and a favorable stacking of three of the tryptophan rings. It has a long lifetime of 167 × 10<sup>6</sup> MC scans. The remaining ~50% of all structures fall into a large number of clusters that can be mainly characterized as random coil and misfolded hairpins, with very low occupancy and short lifetime.

The result of the corresponding cluster analysis for the MD simulations is shown in Table 2. The native state is not the most populated cluster and has only an occupancy of 12.6%, about half that of the MC runs. The hairpin is identical to the one sampled with MC, and the average backbone RMSD to the NMR structure is an even lower 0.6 ± 0.2 Å. Interestingly, both MC and MD runs located the native state nearly as often (three to four times). The most populated cluster in the MD

TABLE 1: Cluster Analysis of All MC Runs for trpzip2<sup>a</sup>

Cluster	Structure	Occupancy	$N_{\text{Visit}}$	$t_{\text{Visit}}$ [ $10^6$ MC scans]
1		24.2 %	3	95 ± 97
2		20.9 %	3	169 ± 98
3		4.8 %	1	110
4		4.2 %	1	85
5 - 259	Coils, helices, misfolded hairpins	45.9 %	~	< 33

<sup>a</sup> A total of 259 clusters were found. The first four most populated clusters are represented, classified according to secondary structure, the time the clusters are visited during the simulations ( $N_{\text{Visit}}$ ), and their average lifetime in million MC scans ( $t_{\text{Visit}}$ ).

runs is the misfolded coiled conformation found as cluster two in the MC study, involving the salt bridge between the  $\text{NH}_3^+$  of Lys12 and the  $\text{COO}^-$  of Glu5. The occupancy (22.6%) is similar to the MC runs. This cluster has a larger  $N_{\text{Visit}}$  and the time the simulations stay in it is shorter than the corresponding cluster in the MC runs because of the frequent transition to the almost similar clusters 3 and 4, leading to a combined population of  $\sim 34\%$  for this misfold. The stability of this structure in both MC and MD sampling could be due to the noted overstabilization of salt bridges in GBSA models.<sup>12,31</sup> Recently, corrections have been proposed to the generalized Born equation to increase dielectric screening for side chains involved in erroneous salt bridges.<sup>32</sup>

**Thermodynamics of trpzip2.** To obtain insights into the folding mechanism, we project the many-dimensional system onto one or two structural coordinates. Figure 2 shows the free energy surface using the backbone RMSD to the NMR structure as descriptor of the system. A two-state folding pattern is evident, with a deep well for the folded states at  $\sim 0.5\text{--}1.1$  Å and a shallow broad basin of the unfolded structures ranging from 4 to 6 Å. For both the MC and MD runs, the magnitude of the folding barrier in this landscape is about  $\sim 3$  kcal/mol, and the native state is only marginally stable with  $\Delta G \sim 0$  kcal/mol compared to the unfolded basin. These results are almost identical to a previous study, although a different force field was used there.<sup>33</sup> Although MD and MC are entirely different simulation techniques, the results are remarkably similar, confirming thermodynamic properties can be reliably obtained by both.

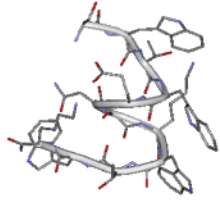
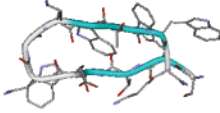
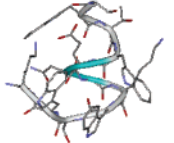
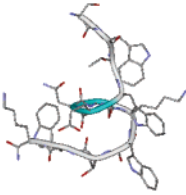
To determine the factors that stabilize the two minima, we also show the components of the free energy in the same figure. The effective energy and entropy contributions span a range of  $\sim 15$  kcal/mol, and significant compensation is seen.  $\Delta E$  is more

poorly converged than  $\Delta G$ , with a standard deviation of  $\sigma_E \approx 10$  kcal/mol for all histogram points. This is because many conformations with diverse energy contribute to every point. Nevertheless, the energetic stabilization of the native state and the competing misfolded structures can clearly be seen. The native state has the lowest solute entropy.

To evaluate the convergence of the free energy profile, we calculated the root-mean-square deviation of the free energy at each bin averaged over the eight trajectories considering as the expectation (reference) profile the one calculated using all trajectories. Since the various trajectories do not individually sample all of the available phase space, there is an average error of 0.5–1.6 kcal/mol (smaller around the native state since averaging is limited to the trajectories that sampled there), indicating that longer simulations are necessary to obtain more accurate free energy surfaces. However, since only three runs located the native state during the MC and four during the MD simulations, the relative populations of the states are not converged, and the well depths, the barrier heights, and the folding free energy are expected to deviate significantly from converged values. Figure 4 shows the development over time of the average root-mean-square error over all bins and trajectories, revealing the convergence of the free energy values along both the MC and MD simulations.

A second system descriptor can be introduced to project the free energy surface onto two structural coordinates. Figure 3 shows the free energy of trpzip2 as a function of the RMSD from the native structure and the radius of gyration, which is a measure of the overall compactness of a conformation. The two-state nature of the landscape is evident with the deep native well and the broader basin of the unfolded structures. Extended structures ( $R_g > 9$  Å) all occupy regions of large free energy and have a short lifetime. Previous calculations of trpzip2 have

TABLE 2: Cluster Analysis of All MD Runs for trpzip2<sup>a</sup>

Cluster	Structure	Occupancy	$N_{\text{visit}}$	$t_{\text{visit}}$ [ $10^6$ MD steps]
1		22.6 %	18	$36.5 \pm 35$ (73 ns)
2		12.6 %	4	$79 \pm 71$ (158 ns)
3		6.1 %	22	$25.5 \pm 25.5$ (51 ns)
4		5.0 %	14	$34 \pm 38.5$ (68 ns)
5 - 518	Coils, helices, misfolded hairpins	53.7 %	~	< 25 (< 50 ns)

<sup>a</sup> A total of 518 clusters were found. The first four most populated clusters are represented, classified according to secondary structure, the time the clusters are visited during the simulations ( $N_{\text{visit}}$ ), and their average lifetime in million MD steps and nanoseconds ( $t_{\text{visit}}$ ).

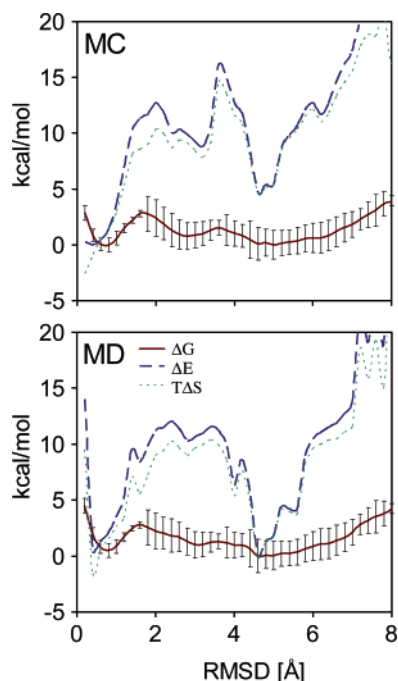
revealed very similar landscapes.<sup>10</sup> The same shape of the landscape of MC and MD demonstrates the independence of the thermodynamic results on the chosen sampling method.

**H1.** The second system studied is the 14-residue amyloidogenic H1 peptide MKHMAGAAAAGAVV from the syrian hamster prion protein (residues 109–122). This peptide is considered to be important for the  $\alpha$ -to- $\beta$  conformational transition that leads to amyloid formation and is responsible for prion diseases. An experimental structure is not available at high resolution, since in aqueous solution it aggregates very rapidly to form  $\beta$ -sheet-rich fibrils,<sup>19,29</sup> while in 2,2,2-trifluoroethanol or membrane-mimicking environments it adopts an  $\alpha$ -helical conformation.<sup>34–36</sup> Previous work has revealed the folding of H1 into a  $\beta$ -hairpin in simulations with explicit solvent representation.<sup>30,36,37</sup>

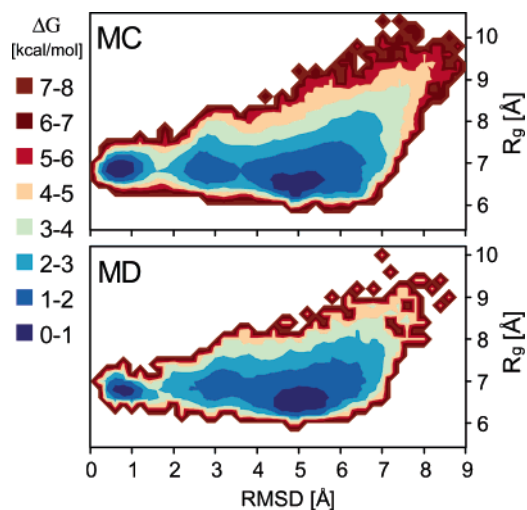
Tables 3 and 4 show the conformations sampled during the MC and MD runs. The H1 peptide exhibits much more flexibility in the simulations, as can be inferred from the higher number of clusters found and the broader distribution of the cluster population as compared with the trpzip2 system. The increased flexibility is due to the large number of alanines in the H1 peptide, facilitating conformational transitions compared to the trpzip2 system with more bulky side chains. Since only the first two clusters have a population of more than 5%, but many clusters represent similar conformations, the results of the cluster analysis were grouped according to secondary structure. The simulations reveal the previously reported beta hairpin with type II'  $\beta$ -turn involving the residues A113-G114-A115-A116, with a population of 13.6% in the MC and 7.3% in MD. This is less than in simulations with explicit solvent, where the hairpin showed up to 30% occupancy.<sup>30</sup> The chain ends are mostly frayed, with only the middle four hydrogen

bonds of the hairpin significantly formed. Although the MC simulations show this structure to be quite stable, the average lifetime in the MD runs is somewhat shorter. A second hairpin with a type IV  $\beta$ -turn involving G114-A115-A116-A117 encountered in the explicit solvent simulations has a very low occupancy of  $\sim 1.2\%$  in the MD runs and  $< 0.5\%$  in the MC simulations. Helical structures are also poorly populated, with  $\sim 1.1\%$  for both MC and MD, which is a little less than the 5% encountered in the in explicit solvent study.<sup>30</sup> The MC and MD results agree well with each other. However, compared to the previous explicit solvation simulations in which the hairpin was found to be more stable, the GBSA simulations show considerable flexibility, with only the inner part of the hairpin showing fully formed hydrogen bonds while the chain ends tended to be frayed.

The free energy surfaces of the H1 peptide as a function of RMSD to the experimental structure are shown in Figure 5. A hairpin conformation derived from the low-resolution X-ray structure was used as the reference conformation, neglecting the first two and last two residues of the chain in the comparison due to their large conformational flexibility. A broad barrier-free basin is seen, with remarkably similar shape for both the MC and MD runs. In the previous study with explicit solvation, a similar featureless smooth surface was found, with a maximum spread of 3.3 kcal/mol of the free energy.<sup>30</sup> A similar spread of  $< 2$  kcal/mol is seen here for the main sampled region. The  $\beta$ -hairpin, although clearly lowest in energy, is higher in free energy by  $\sim 1$  kcal/mol compared to the unfolded and misfolded structures that populate the main basin at  $\sim 4$  Å RMSD. Again, the magnitude of the errors indicates that the free energy surfaces are not fully converged, despite the similar shape for MC and



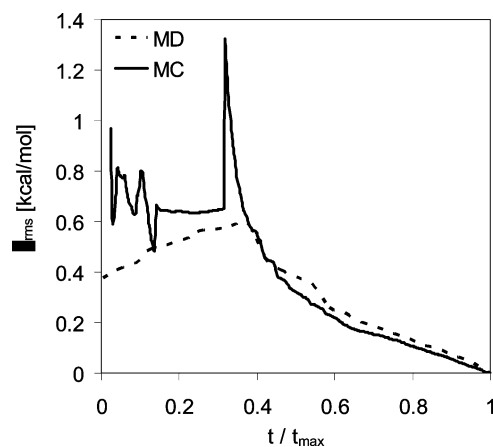
**Figure 2.** Free energy profile of trpzip2 as a function of RMSD from the NMR structure for the MC runs (top panel) and the MD runs (bottom panel). The plotted value of  $\Delta G$  is the free energy relative to the lowest bin which was set to  $G = 0$ . The same holds for  $\Delta E$ . A narrow valley of the native state and the broad basin of misfolded structures can be seen. Also shown are the energetic and solute entropic contribution to the free energy, with low average energy for both the native state and the misfolded conformations. The errors of the free energy are calculated as the root-mean-square deviation of  $\Delta G$  averaged over the eight trajectories with the complete profile as the reference. The standard deviation of  $\Delta E$  is a much larger  $\sigma \approx 10$  kcal/mol for all histogram bins, revealing the conformational flexibility.



**Figure 3.** Two-dimensional free energy profile of trpzip2 as a function of RMSD from the NMR structure and the gyration radius for the MC runs (top panel) and the MD runs (bottom panel). The free energy contours are in units of kcal/mol.

MD. The instability of the hairpin was also reported in the previous study, where a value of  $+0.6$  kcal/mol was determined.<sup>30</sup>

Figure 6 shows the two-dimensional free energy surfaces as a function of RMSD to the experimental structure and  $R_g$ . The large featureless basin is even more apparent, and the difference to the corresponding landscape of trpzip2 striking. MD and MC are seen to yield almost identical results.

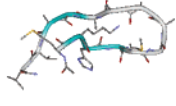
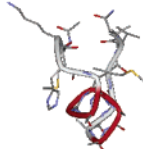


**Figure 4.** Convergence of the free energy profile of trpzip2 for both the MC and MD runs. The graph shows the change of the root-mean-square deviation of the free energy at each bin averaged over the eight trajectories considering as the expectation (reference) profile the one calculated using all trajectories. The relative progress of MC and MD is shown with  $t_{\max} = 250 \times 10^6$  MD steps and  $333 \times 10^6$  MC scans. Both methods show a steady convergence to the final profile. Large changes in the initial phase correspond to transitions where previously unsampled phase space regions are entered.

**Folding Times.** A comparison of the folding times of the studied polypeptides gives insight into the folding kinetics of both the MC and MD simulations. Experimentally, trpzip2 has a folding time of  $1.8\text{--}2.5 \mu\text{s}$  at 300 K, as determined by laser T-jump spectroscopy,<sup>10</sup> and previous computational estimates are  $3\text{--}6 \mu\text{s}$ .<sup>10</sup> Although we study this system at a combined length of  $4 \mu\text{s}$ , the dynamics is expected to be much faster due to the neglect of the viscous drag of water in our MD simulations, which is necessary to measure the unbiased sampling performance of MD as compared to MC. In addition, the temperature is slightly higher, 323 K. For all folding events, the mechanism of  $\beta$ -hairpin formation was found to be a “zipper”, with the turn region forming first before the subsequent backbone hydrogen bonds, in agreement with previous work.<sup>10</sup> The growth of the population of the native state over the course of the simulations (the cumulative ratio of nativelike structures to all structures) is shown in Figure 7. In sufficiently long simulations this curve is expected to level out at the equilibrium population, indicating again that the 500 ns/4 billion MC steps chosen are not long enough for full convergence. Here, as noted above, the MC runs exhibit a  $\sim 2$  times higher population at the end of the simulations with respect to MD. A quantitative assessment of the performance of both MC and MD runs for trpzip2 is given in Table 5. The average of the folding time  $t_{\text{av}}$  of the observed folding events is  $128 \times 10^6$  MC scans and  $130 \times 10^6$  time steps (259 ns) for MD. If the relative CPU time is factored in, it is visible that the MC runs fold  $\sim 2.55$  faster than MD, a factor strongly dependent on the level of optimization of the programs used.

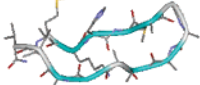
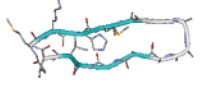

The low number of folding events leads to errors of the folding times that are quite large. For example, in a previous, much shorter MC study of trpzip2, folding times in the range of  $(29\text{--}46) \times 10^6$  MC scans were encountered using a similar setup,<sup>14</sup> significantly lower than the values found here. In addition, several trajectories show an initial fast collapse into the native state: One of the MD runs of trpzip2 folds in just  $2.5 \times 10^6$  time steps (5 ns) (visible in Figure 7), and two of the MC runs of the H1 system collapse straight into the hairpin in only  $3.2 \times 10^6$  and  $8.3 \times 10^6$  MC scans. Such fast “lucky collapses” reveal a problem with starting simulations from extended conformations rather than the equilibrated compact

TABLE 3: Cluster Analysis of All MC Runs for H1<sup>a</sup>

Structure	$N_{\text{cluster}}$	Type	Occupancy	$N_{\text{visit}}$	$t_{\text{visit}}$ [10 <sup>6</sup> MC scans]
	8	II' $\beta$ turn - hairpin	13.6 %	2-5	from $1.8 \pm 1.1$ to $108 \pm 98$
	2	Partial $\alpha$ -helix	1.1 %	3	$10 \pm 1.1$
Coils, unfolded structures	520	-	85.7 %	-	-

<sup>a</sup> A total of 530 clusters were found. The most populated clusters are shown, grouped together according to secondary structure, the number of clusters  $N_{\text{cluster}}$  of each group, the time the clusters are visited during the simulations ( $N_{\text{visit}}$ ), and their average lifetime in million MC scans ( $t_{\text{visit}}$ ).

TABLE 4: Cluster Analysis of All MD Runs for H1<sup>a</sup>

Structure	$N_{\text{cluster}}$	Type	Occupancy	$N_{\text{visit}}$	$t_{\text{visit}}$ [10 <sup>6</sup> MD steps]
	6	II' $\beta$ turn - hairpin	7.3 %	4 - 13	1 - 6 (2-12 ns)
	2	IV $\beta$ turn - hairpin	1.7 %	7	$2 \pm 1$ (4 ns)
	2	Partial $\alpha$ -helix	1.2 %	5	$2 \pm 1.5$ (4 ns)
Coils, unfolded structures	693	-	89.8 %	-	-

<sup>a</sup> A total of 703 clusters were found. The most populated clusters are shown grouped together according to secondary structure, the number of clusters  $N_{\text{cluster}}$  of each group, the time the clusters are visited during the simulations ( $N_{\text{visit}}$ ), and their average lifetime in million MD steps and nanoseconds ( $t_{\text{visit}}$ ).

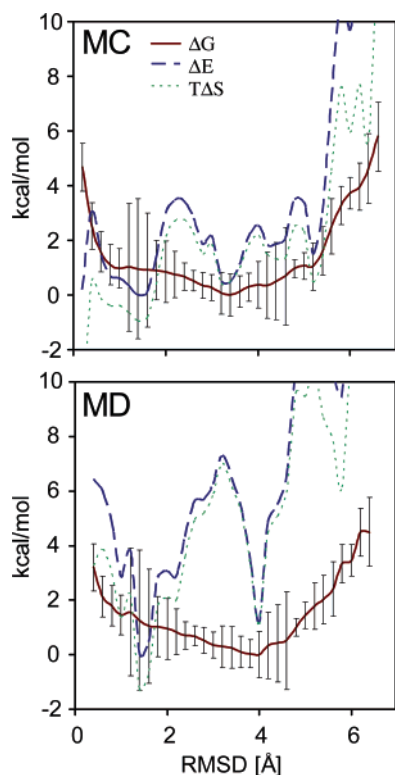
ensemble of the denatured state. This is done to avoid biasing the results, but it should be noted that there will still be a bias due to all simulations beginning from the same structure. In these few fast collapses, the folding from high-energy extended unfolded conformations proceeds without crossing any barrier straight to the folded state, which is a special case that is unlikely to occur when initiating folding from low-energy compact structures. It has been noted before that fast folding events are not representative of the major folding pathways.<sup>10,38</sup> Including such runs in the overall statistics increases the uncertainties. If, e.g., the fast 5 ns folding event of trpzip2 would be discarded, the  $t_{\text{av}}$  would increase to  $343 \pm 146$  ns. A large spread of values is also seen with the average lifetime  $\tau$  in the folded state.

Results for the H1 peptide are shown in the same table. The errors are larger here, since H1 exhibits a significantly broader range of both  $t_{\text{av}}$  and  $\tau$ . Due to a large number of extremely short folding events in the MD simulations ( $\tau < 5$  ns), the average lifetime of the hairpin is very short, as noted above. The MC runs show a larger stability of the hairpin, and again MC locates this state faster than MD by a factor of  $\sim 2$ .

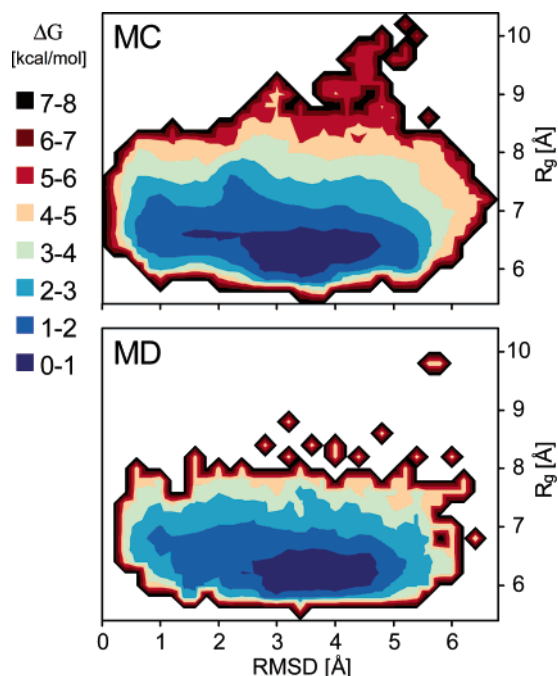
The folding time  $t_f$  can also be determined from a fit to a single-exponential kinetics model,<sup>10</sup> in which the probability that a molecule has folded is expected to follow  $P_{\text{folded}} = 1 - \exp$

$(-t/t_f)$ . In the limit of short simulations with  $t \ll t_f$ ,  $P_{\text{folded}} \approx t/t_f$ . A linear fit of the cumulative number of folding events versus time yields  $t_f$ , and these numbers are shown in the same table. The MD folding time is lower than both the experiment and previous studies due to the neglect of solvent viscosity, and the inaccuracy is high since there are only three to four folding events. Calculating  $t_f$  for the H1 system is not possible due to the poor stability of the  $\beta$ -hairpin.

**MC vs MD Comparison.** How does MC compare to MD? The folding times enable an estimate of the ratio of MC moves to MD time in the simulations. For trpzip2,  $\sim 6 \times 10^6$  MC steps correspond to  $0.5 \times 10^6$  time steps MD, which is a ratio of  $\sim 12$  MC moves for 1 MD integration step of 2 fs. The H1 system yields a ratio of  $\sim 7 \times 10^6$  MC steps per  $0.5 \times 10^6$  time steps, or  $\sim 14$  moves/time step. Interestingly, these numbers are exactly proportional to the amount of residues of the peptides (12 for trpzip2, 14 for H1). If the ratio is expressed in the number of MC scans, the simulations both independently reach the result of 1 MC scan  $\approx$  1 MD time step. This ratio reveals the efficiency of MC (in which half of all moves are usually rejected) to sample the conformational space using all-atom force fields and the latest solvation models. Interestingly, the result is almost identical to a previous comparison of MC and

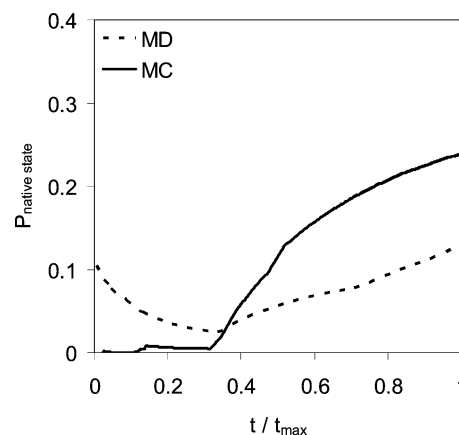


**Figure 5.** Free energy profile of the H1 peptide as a function of RMSD from the  $\beta$ -hairpin structure for the MC runs (top) and the MD runs (bottom) (see caption of Figure 2 for details).  $\Delta G$  is almost similar and the graphs reveal that the  $\beta$ -hairpin, despite its low conformational energy, is about  $\sim 1$  kcal/mol higher in free energy than competing compact conformations.



**Figure 6.** Two-dimensional free energy profile of the H1 peptide as a function of RMSD from  $\beta$ -hairpin structure and the gyration radius for the MC runs (top panel) and the MD runs (bottom panel). The free energy contours are in units of kcal/mol.

MD, in which the conformational equilibration of a box of liquid hexane molecules was studied:<sup>39</sup> The ratio found was 1.3 MC scans  $\approx$  1 MD time step. The MC runs were found to be 1.6–3.8 times faster than those of MD. In our simulations, the MC runs were  $\sim 2$ –2.5 times more efficient than those of MD. Of



**Figure 7.** Growth of the population of the native state of trpzip2 (the cumulative ratio of natively like structures to all structures). The relative progress is compared along the MC and MD runs with  $t_{\max} = 250 \times 10^6$  MD steps and  $333 \times 10^6$  MC scans. The high starting population of the MD runs is due to a single fast  $\sim 5$  ns collapse event from an extended initial structure, remaining folded for only 23 ns.

course, this ratio is highly dependent on the level of performance of the individual MC and MD programs.

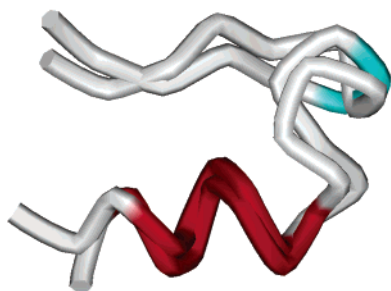
**Trp-Cage.** With the encouraging performance of the MC folding simulations of trpzip2 and H1, we tested this method on a larger and more complex system, the “Trp-cage” mini-protein. This 20 residue polypeptide with the sequence NLYIQWLKDGGPSSGRPPPS has been optimized by Neidigh et al.<sup>21</sup> and is one of the smallest proteins displaying two-state folding properties. Due to the larger number of MD studies available of this system,<sup>5,6,8,40,41</sup> we only ran the MC simulations. A setup identical to the trpzip2 runs was used, with eight runs at 4 billion MC steps ( $200 \times 10^6$  MC scans) each at a temperature of 323 K. The experimental reference is the NMR structure (PDB code 1L2Y), which shows a short  $\alpha$ -helix, and a triplet polyproline II helix. Due to the larger size and complexity of the system, the sampling is found to be less exhaustive than with either trpzip2 or H1. Only one folding event is detected, after  $67 \times 10^6$  MC scans, and the system remains firmly folded for  $104 \times 10^6$  MC scans. The average backbone RMSD to the NMR structure over the folded state is  $\sim 2.7$  Å, a value somewhat higher than obtained in previous studies due to the flexibility of the chain ends. The chain ends are not well defined, as can be seen by the structural variation of the submitted NMR conformers in the PDB file 1L2Y. A representative overlay of the folded phase conformation with the NMR structure (conformer 1) is shown in Figure 8. Although uncertain, the folding time expressed in CPU days gives a similar result to the other two polypeptides, with  $t_{\text{av}} = 5.9$  days, and a performance of  $12 \times 10^6$  MC scans/day. Since the native state was only encountered once, the thermodynamic properties are too unreliable to report—much longer simulations or probably replica exchange runs are necessary to yield these quantities. The folding time of Trp-cage is experimentally estimated to be  $\sim 4$   $\mu\text{s}$ .<sup>42</sup> Many MD folding studies report much faster folding events, in the range of 10–30 ns,<sup>5,6,8</sup> and these probably have to be considered extreme cases<sup>8</sup> or, as elaborated above, lucky collapses. Using the analysis of trpzip2 and H1, the reported folding event of  $67 \times 10^6$  MC scans would correspond to a larger MD time of 133 ns. The seven other MC trajectories that do not locate the native state spend most of their time in compact and stable misfolded conformations.



**TABLE 5: Average Folding Time and Performance for trpzip2 and H1<sup>a</sup>**

	trpzip2		H1	
	MC	MD	MC	MD
MC scans/MD steps	$333 \times 10^6$	$250 \times 10^6$	$286 \times 10^6$	$250 \times 10^6$
CPU time (days)	13.6	25.6	10.6	18.8
MC scans/MD steps per day	$24.6 \times 10^6$	$9.75 \times 10^6$	$27.0 \times 10^6$	$13.3 \times 10^6$
$t_{av}$	$(128 \pm 27) \times 10^6$	$(130 \pm 103) \times 10^6$ (259 ± 207 ns)	$(72 \pm 115) \times 10^6$	$(70 \pm 53) \times 10^6$ (140 ± 107 ns)
$t_{av}$ (days)	5.2	13.3	2.7	5.3
$t_f$	$225 \times 10^6$	1.3 $\mu$ s		
$\tau$	$(153 \pm 87) \times 10^6$	$(60 \pm 83) \times 10^6$ (121 ± 167 ns)	$(96 \pm 102) \times 10^6$	$(7.4 \pm 9.8) \times 10^6$ (14.8 ± 19.5 ns)

<sup>a</sup> All calculations were performed on 3-GHz Intel Xeon CPUs. The total simulation effort, the used CPU time, and the progress per day for each system are given. Units are given in MC scans (1 MC scan = 12 MC steps for trpzip2, 14 MC steps for H1), and in time steps for MD (1 MD time step = 2 fs). The average folding time  $t_{av}$  is the average of the folding events observed by both MC and MD. It is also shown in CPU time, as a measure of performance.  $t_f$  is the folding time estimated from a fit to a single-exponential kinetics model (see main text). The last row gives the average lifetime  $\tau$  of the folded state.



**Figure 8.** Overlay of the NMR structure (conformer 1) and a representative structure from the folded phase of the MC simulation that located the native state of the Trp-cage miniprotein. The quality of the match is lower at the chain ends, which were found to be quite flexible at the chosen temperature of 323 K.

## Conclusions

The presented work demonstrates that MC in combination with GBSA solvation and all-atom force fields can reliably predict the native state of small proteins and polypeptides. The total simulation effort involved  $\sim 10^{11}$  MC steps and 8  $\mu$ s MD time. In all studied systems, trpzip2, H1, and Trp-cage, results were virtually indistinguishable from equivalent MD simulations using the same force field and solvation method: free energy surfaces were almost identical; the population of major clusters was very similar; and folding times as well as the lifetime of the folded state were proportional in MC and MD. Overall, MC was found to yield these results at a  $\sim 2$ – $2.5$  times smaller computational effort. Given the strong differences of MC and MD, such comparisons are very helpful in determining the reliability of the obtained results—thermodynamic and kinetic—and to enable potential algorithmic artifacts to be detected. Although simulations were run for an accumulated time of  $3.2 \times 10^{10}$  MC steps and 4  $\mu$ s for MD per peptide—comparable to the experimental folding time scales for the studied systems in the  $\sim 1$ – $4 \mu$ s range—only a few folding events were encountered. Thus, the uncertainties in both the thermodynamic and kinetic results are still significant. By use of replica exchange algorithms, the number of barrier crossing events could probably be substantially increased to yield free energy surfaces that are better converged in a shorter time.<sup>11</sup> However, such simulations do not reveal the interesting temporal behavior of folding, and the method is not guaranteed to increase efficiency in all situations.<sup>43</sup> Such data will ultimately be accessible from longer individual simulations, which are now possible on modern hardware.

Although the Markov chain of states generated by the Metropolis MC algorithm does not incorporate a time element,

the statistical nature of both MC and MD (MD runs are coupled to a heat bath) leads to very similar dynamics on the longer time scale of folding, misfolding, and unfolding of protein conformations. The similarity of the folding results has enabled us to estimate the ratio of MC steps (or MC scans) and MD time steps for the case of folding simulations in implicit solvent. We find a correspondence of 1 MC scan  $\approx$  1 MD time step. This is in agreement with a previous comparison of MC and MD for liquid hexane, in which a ratio of  $\sim 1.3$  was found.<sup>39</sup> In that study, MC was found to be more efficient than MD by a factor of 1.6–3.8. Our simulations find MC to be 2–2.5 times more efficient than MD, but this ratio is expected to be strongly dependent on the level of optimization of the respective MD/MC program. It will be of interest to apply the efficient MC algorithm with concerted rotations to larger systems and to investigate its performance when replica exchange moves are included.

**Acknowledgment.** We thank Dr. I. Daidone for helpful discussions. This project has been supported by the Wellcome Trust and an Emmy Noether fellowship of the Deutsche Forschungsgemeinschaft.

## References and Notes

- (1) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129.
- (2) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. J. *Am. Chem. Soc.* **1990**, *112*, 6127.
- (3) Felts, A. K.; Gallicchio, E.; Wallqvist, A.; Levy, R. M. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 404.
- (4) Jang, S.; Shin, S.; Pak, Y. J. *Am. Chem. Soc.* **2002**, *124*, 4976.
- (5) Simmerling, C.; Strockbine, B.; Roitberg, A. E. *J. Am. Chem. Soc.* **2002**, *124*, 11258.
- (6) Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Duan, Y. *J. Mol. Biol.* **2003**, *327*, 711.
- (7) Zagrovic, B.; Snow, C. D.; Shirts, M. R.; Pande, V. S. *J. Mol. Biol.* **2002**, *323*, 927.
- (8) Snow, C. D.; Zagrovic, B.; Pande, V. S. *J. Am. Chem. Soc.* **2002**, *124*, 14548.
- (9) Snow, C. D.; Nguyen, N.; Pande, V. S.; Gruebele, M. *Nature* **2002**, *420*, 102.
- (10) Snow, C. D.; Qiu, L. L.; Du, D. G.; Gai, F.; Hagen, S. J.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4077.
- (11) Zhou, R. H.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931.
- (12) Zhou, R. H.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12777.
- (13) Snow, C. D.; Sorin, E. J.; Rhee, Y. M.; Pande, V. S. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 43.
- (14) Ulmschneider, J. P.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2004**, *126*, 1849.
- (15) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. J. *Comput. Chem.* **1997**, *18*, 1463.
- (16) Ulmschneider, J. P.; Jorgensen, W. L. *J. Chem. Phys.* **2003**, *118*, 4261.

- (17) Ulmschneider, J. P.; Jorgensen, W. L. *J. Phys. Chem. B* **2004**, *108*, 16883.
- (18) Onuchic, J. N.; Wolynes, P. G. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70.
- (19) Nguyen, J.; Baldwin, M. A.; Cohen, F. E.; Prusiner, S. B. *Biochemistry* **1995**, *34*, 4186.
- (20) Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5578.
- (21) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. *Nat. Struct. Biol.* **2002**, *9*, 425.
- (22) Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. *J. Am. Chem. Society* **1996**, *118*, 11225.
- (23) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005.
- (24) Jorgensen, W. L.; Ulmschneider, J. P.; Tirado-Rives, J. *J. Phys. Chem. B* **2004**, *108*, 16264.
- (25) Schaefer, M.; Bartels, C.; Karplus, M. *J. Mol. Biol.* **1998**, *284*, 835.
- (26) van der Spoel, D.; van Drunen, R.; Berendsen, H. J. C. *GRONINGEN MACHINE FOR CHEMICAL SIMULATION*; Department of Biophysical Chemistry, BIOSON Research Institute: Nijenborgh, Groningen, 1994.
- (27) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
- (28) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angew. Chem., Int. Ed.* **1999**, *38*, 236.
- (29) Inouye, H.; Kirschner, D. A. *J. Neurochem.* **1998**, *70*, S7.
- (30) Daidone, I.; Amadei, A.; Di Nola, A. *Proteins: Struct. Funct., Bioinf.* **2005**, *59*, 510.
- (31) Zhou, R. H. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 148.
- (32) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 310.
- (33) Yang, W. Y.; Pitera, J. W.; Swope, W. C.; Gruebele, M. *J. Mol. Biol.* **2004**, *336*, 241.
- (34) Heller, J.; Kolbert, A. C.; Larsen, R.; Ernst, M.; Bekker, T.; Baldwin, M.; Prusiner, S. B.; Pines, A.; Wemmer, D. E. *Protein Sci.* **1996**, *5*, 1655.
- (35) Jayawickrama, D.; Zink, S.; Vandervelde, D.; Effiong, R. I.; Larive, C. K. *J. Biomol. Struct. Dynam.* **1995**, *13*, 229.
- (36) Daidone, I.; Simona, F.; Roccatano, D.; Broglia, R. A.; Tiana, G.; Colombo, G.; Di Nola, A. *Proteins: Struct. Funct. Bioinf.* **2004**, *57*, 198.
- (37) Daidone, I.; D'Abramo, M.; Di Nola, A.; Amadei, A. *J. Am. Chem. Soc.* **2005**, *127*, 14825.
- (38) Paci, E.; Cavalli, A.; Vendruscolo, M.; Caflisch, A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 8217.
- (39) Jorgensen, W. L.; TiradoRives, J. *J. Phys. Chem.* **1996**, *100*, 14508.
- (40) Pitera, J. W.; Swope, W. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7587.
- (41) Zhou, R. H. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13280.
- (42) Qiu, L. L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J. *J. Am. Chem. Soc.* **2002**, *124*, 12952.
- (43) Brown, S.; Head-Gordon, T. *J. Comput. Chem.* **2003**, *24*, 68.