Simulated Tempering Distributed Replica Sampling, Virtual Replica Exchange, and Other Generalized-Ensemble Methods for Conformational Sampling

Sarah Rauscher,[†,‡] Chris Neale,[†,‡] and Régis Pomès*,[†,‡]

Molecular Structure and Function, Hospital for Sick Children, 555 University Avenue, Toronto, ON, Canada M5G 1X8 and Department of Biochemistry, University of Toronto, 1 King's College Circle, Toronto, ON, Canada M5S 1A8

Abstract: Generalized-ensemble algorithms in temperature space have become popular tools to enhance conformational sampling in biomolecular simulations. A random walk in temperature leads to a corresponding random walk in potential energy, which can be used to cross over energetic barriers and overcome the problem of quasi-nonergodicity. In this paper, we introduce two novel methods: simulated tempering distributed replica sampling (STDR) and virtual replica exchange (VREX). These methods are designed to address the practical issues inherent in the replica exchange (RE), simulated tempering (ST), and serial replica exchange (SREM) algorithms. RE requires a large, dedicated, and homogeneous cluster of CPUs to function efficiently when applied to complex systems. ST and SREM both have the drawback of requiring extensive initial simulations, possibly adaptive, for the calculation of weight factors or potential energy distribution functions. STDR and VREX alleviate the need for lengthy initial simulations, and for synchronization and extensive communication between replicas. Both methods are therefore suitable for distributed or heterogeneous computing platforms. We perform an objective comparison of all five algorithms in terms of both implementation issues and sampling efficiency. We use disordered peptides in explicit water as test systems, for a total simulation time of over 42 μs. Efficiency is defined in terms of both structural convergence and temperature diffusion, and we show that these definitions of efficiency are in fact correlated. Importantly, we find that ST-based methods exhibit faster temperature diffusion and correspondingly faster convergence of structural properties compared to RE-based methods. Within the RE-based methods, VREX is superior to both SREM and RE. On the basis of our observations, we conclude that ST is ideal for simple systems, while STDR is well-suited for complex systems.

Introduction

Achieving complete (or even adequate) conformational sampling is one of the key challenges in biomolecular simulations.[1] The energy landscape of most biomolecules is "rugged", and the source of this ruggedness is two-fold. The energetic barriers separating accessible states are often larger than the available thermal energy, and there are typically a large number of states to be sampled. The time scales of many biomolecular processes, such as protein folding, are still far beyond the reach of our current computational capability, which is generally limited to the 10^{-8} to 10^{-7} s time scale for continuous simulations. For example, even the folding of small domains or secondary structure elements, such as β-hairpins and mini-proteins, occur on the $1-10$ μs time scale.[1] Consequently, conventional or "brute force" molecular dynamics (MD) alone is often insufficient to

* Corresponding author e-mail: pomes@sickkids.ca.
† Hospital for Sick Children.
‡ University of Toronto.

achieve complete Boltzmann sampling of the important states of many biologically relevant systems. For this reason, generalized-ensemble algorithms have become popular tools for conformational sampling.

A variety of generalized-ensemble algorithms have been developed with the common intention of overcoming energetic barriers in order to enhance sampling of conformational space. These methods use a generalized Hamiltonian for the purpose of achieving uniform sampling along a reaction coordinate of interest. Practically, one is faced with choosing the most appropriate method and reaction coordinate for a particular application. While the optimal reaction coordinate is not known a priori, it may be possible to make generalizations regarding the optimal methodology. To this end, we consider the following important question: given limited computational resources, which algorithm is most efficient at sampling a complex energy landscape? Some generalized-ensemble methods employ a random walk in potential energy, while others use different parameters which are relevant to the system of interest.[2] In this article, we compare the efficiency of a set of algorithms which make use of a random walk in temperature to enhance conformational sampling of biomolecules. We focus on the following five methods: simulated tempering (ST),[3,4] replica exchange (RE),[5−9] the serial replica exchange method (SREM),[10] and two novel methods, virtual replica exchange (VREX) and simulated tempering distributed replica sampling (STDR), which is a combination of ST and distributed replica sampling (DR).[11−13]

The generalized-ensemble algorithms compared in this paper all rely on the fact that the free energy surface becomes less rugged at high temperatures, increasing the frequency of interconversion between conformational states.[14] Simulations performed at low temperatures often require a relatively long time to cross the energetic barriers between states and appear to be trapped. Transitions between regions separated by barriers may not be observed over time scales accessible to simulation. In this case, multiple simulations initiated in different conformational basins may sample different subsets of phase space. The result is that an ergodic system appears nonergodic, a phenomenon known as quasi-nonergodicity.[15] Utilizing generalized-ensemble algorithms that induce a random walk in temperature may alleviate this source of error.

The sampling enhancement of generalized-ensemble methods relative to canonical MD or Monte Carlo (MC) simulations has been demonstrated for several systems,[3,7,16,17] including peptides.[6,14,18−24] Conversely, there have also been studies that question the relative sampling efficiency of RE compared to brute force MD,[25] highlighting the importance of a rigorous definition of efficiency which accounts for the total computer time required for all temperatures.[26−28] It is important to note that data obtained at multiple temperatures in generalized-ensemble simulations may be of interest in some studies, such as protein folding.[21,22] In general, however, the data at high temperatures are not useful. Furthermore, the observed speedup also strongly depends on the lowest temperature.[26] It is essential to assess the convergence of both the conventional MD simulations as

well as the generalized-ensemble simulations in order to perform a meaningful comparison, in addition to identifying a meaningful quantity on which to base the comparison. Any evaluation of sampling enhancement compared to single-temperature MD is also likely to depend heavily on the molecular system under study (depending on the number of basins in the landscape and the heights of barriers). It is therefore quite difficult to accurately quantify the sampling enhancement due to the introduction of a random walk in temperature.

We begin with a brief introduction of each of the generalized-ensemble methods, including the presentation of our two novel methods, STDR and VREX. We then perform a thorough comparison of the algorithms in terms of both practical implementation limitations and sampling efficiency for a disordered octapeptide in explicit water, a molecular system combining high relevance to protein folding and moderate complexity. In addition to providing a comparison between generalized-ensemble algorithms, we also provide a comparison to conventional MD. We discuss efficiency in terms of both convergence of structural properties and temperature diffusion, and we show that these definitions of efficiency are correlated. Finally, we compare the efficiency of STDR and conventional MD for a 35-residue peptide with a complex conformational landscape.

Theory and Methods

Simulated Tempering (ST). Simulated tempering was originally introduced to enhance sampling of a random field Ising model.[3] This system has a rough energy landscape for which spin-flips from the state favored by the magnetic field to the opposite state are statistically rare events. ST facilitates exchanges between these states, whereas the MC algorithm remains trapped.[3] ST has also been shown to be effective in exploring the energy landscapes of biomolecules, which similarly have multiple energy minima separated by barriers.[29]

In the ST algorithm, temperature becomes a dynamic variable[3,4] that can take on discrete values labeled by an index m ($m = 1, ..., M$). ST makes use of a generalized Hamiltonian, $H(X,m)$, which depends on all configurational degrees of freedom (X), in addition to temperature:

$$H(X, m) = \beta_m \mathrm{H}(X) - a_m \qquad (1)$$

where β_m is the inverse temperature, $\mathrm{H}(X)$ is the system's original Hamiltonian, and a_m is a constant which depends on temperature.[3] The generalized ensemble has a corresponding generalized partition function, $\mathbf{Z}$, given by:

$$\mathbf{Z} = \sum_m \int \mathrm{d}X \, [e^{-H(X,m)}] = \sum_m \int \mathrm{d}X \, [e^{-\beta_m \mathrm{H}(X)+a_m}] =$$
$$\sum_m Z_m \, e^{a_m} \quad (2)$$

where Z_m is the partition function corresponding to the temperature T_m.[30] The partition function of the generalized ensemble, $\mathbf{Z}$, is the weighted sum of the partition functions of the canonical ensembles at each temperature, Z_m. We

therefore refer to the constants, a_m, as "weight factors".[30] The probability of sampling a given temperature, T_m, is[3]

$$P(T_m) \propto e^{-H(X,m)} \equiv Z_m\, e^{a_m} \qquad (3)$$

which depends on the generalized Hamiltonian, H, and therefore depends on the weight factor, a_m. The goal in ST is to perform a random walk in temperature such that all temperatures are visited uniformly, that is, to choose weight factors such that for any two temperatures (labeled i and j)

$$Z_i\, e^{a_i} = Z_j\, e^{a_j} \qquad (4)$$

Since the partition function in the canonical ensemble, Z_m, is related to the Helmholtz free energy, A_m, the optimal weight factors are dimensionless Helmholtz free energies (the Helmholtz free energy multiplied by the inverse temperature, β):[30,31]

$$\begin{aligned} Z_m &= e^{-\beta_m A_m} = e^{-a_m} \\ a_m &= -\ln Z_m \end{aligned} \qquad (5)$$

The use of accurate dimensionless Helmholtz free energies as weight factors leads to sampling all temperatures with equal probability. In principle, the weight factors may take any value without resulting in biased, non-Boltzmann sampling at the individual temperatures. However, inaccuracy in the weight factors leads to corresponding differences in the probabilities of sampling at each temperature.[4,32]

An ST simulation consists of a short canonical MD (or MC) simulation at temperature T_i followed by an exchange attempt to a neighboring temperature, T_j. The probability of this exchange occurring is given by:

$$p(T_i \rightarrow T_j) = \min \begin{cases} 1 \\ e^{-(\beta_j - \beta_i)E + (a_j - a_i)} \end{cases} \qquad (6)$$

where E is the potential energy of the system at the end of the previous simulation at temperature T_i, and β_i and β_j are the inverse temperatures.[2] The weight factors need only be accurate up to an additive constant, since only differences in weight factors are required to determine the acceptance probability.[30] Through many repetitions of these alternating simulation and exchange steps, a random walk in temperature is realized, corresponding to a random walk in potential energy and efficient exploration of the energy landscape.[31] In fact, ST has been shown to be as effective as the multicanonical algorithm, which employs a random walk in potential energy.[33]

The underlying challenge in ST is accurately obtaining the dimensionless Helmholtz free energies, a_m. There have been two general approaches to their calculation. The first method involves making use of the weighted histogram analysis method (WHAM)[34−36] to obtain the density of states and the weight factors. The second method, which we utilize in this paper, was recently proposed as a fast and efficient scheme to obtain an accurate estimate of the weight factors based on average energies.[30,37] The average potential energy at each temperature, $\langle E \rangle$, is obtained from initial simulations, and the differences in weight factors are calculated as follows:

$$a_{i+1} - a_i \approx (\beta_{i+1} - \beta_i)\left(\frac{\langle E \rangle_i + \langle E \rangle_{i+1}}{2}\right) \qquad (7)$$

The weight factor for the lowest temperature can be set to zero since only differences in weight factors are needed in the exchange probability. The replica exchange simulated tempering method (REST) may also be used to obtain weight factors. In this method, an initial RE simulation is run for the purpose of obtaining accurate weight factors, which are then used in a ST simulation.[36,38] REST may be used with either WHAM or the method outlined in eq 7. Weight factors may be updated throughout the ST simulation if required.[14] Adaptive determination of weights using schemes like the one developed by Zhang and Ma[39] based on the Wang−Landau algorithm[40] were shown to be useful for a square lattice Ising model, a bead model of a protein and a Lennard-Jones fluid.[39] Biomolecular systems containing thousands of atoms represent a completely different level of complexity. Convergence of weights is nontrivial[14,30,31,39] and can be slow even with adaptive approaches.[30] Calculating weight factors has been the main impediment to the widespread use of the simulated tempering method.[30]

The accuracy of the weight factors (that is, how close the differences in weight factors are to accurate dimensionless Helmholtz free energy differences) can be assessed by computing the deviation from sampling all temperatures homogeneously in a sufficiently long ST simulation. In the extreme case for which all weight factors are equal and all differences in weight factors are zero, only the lowest temperature is significantly sampled. This is because the first term in the exponent of the exchange probability (eq 6) depends on the potential energy, which is generally a large, negative number for biomolecular systems. When multiplied by the difference in inverse temperatures, the resulting exchange probability dictates that moves to lower temperatures are accepted, while moves to higher temperatures are rejected. Conversely, if the differences in weight factors are equal to the differences in dimensionless Helmholtz free energies, the temperatures in the ST simulation are sampled uniformly, which is the optimal situation. In practice, weight factors obtained for ST result in temperature sampling inhomogeneity somewhere between these two extremes. Calculating the dimensionless Helmholtz free energies for a complex system such as a peptide in explicit water is computationally expensive since it requires an accurate estimate of the partition function. These calculations can require tens of nanoseconds per temperature or more, and the computational expense increases with both system size and complexity.[14]

Replica Exchange (RE). Replica exchange has been the most widely used of the methods we discuss in this paper to enhance sampling of biomolecular simulations. It can be thought of as a parallel version of ST, and it is also known as parallel tempering[5] or multiple Markov chains.[8] In fact, parallel tempering was applied to proteins even before ST.[41] An RE simulation consists of M identical copies of the system (replicas) which sample M canonical ensembles at different temperatures. Exchanges are performed between neighboring temperatures, T_i and T_j. The probability of

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2643**

making an exchange depends on the potential energies, E_i and E_j, and the inverse temperatures, β_i and β_j:

$$P(T_i \leftrightarrow T_j) = \min\left\{ \begin{array}{l} 1 \\ e^{-(\beta_j - \beta_i)(E_i - E_j)} \end{array} \right. \quad (8)$$

RE is analogous to ST, but instead of using weight factors in the exchange probability, the upward move of one replica is coupled to the downward move of another. RE therefore has the critical advantage of not requiring any initial simulation for the calculation of weight factors. Importantly, it also satisfies detailed balance.[2]

One drawback of the RE method is its significant computational requirements. There is a one-to-one correspondence between the number of replicas (M) and the number of temperatures (M). The number of replicas needed for a RE simulation is related to the number of degrees of freedom, N, as O($N^{1/2}$).[2,7,42] Systems with many particles therefore require many replicas. Although it is not a specific requirement of the RE algorithm, in its typical implementation, each replica is run on a dedicated central processing unit (CPU). This setup minimizes the amount of information that must be passed between nodes.[31] Thus, M CPUs are running simultaneously throughout the course of the RE simulation. The use of M CPUs in RE can be overcome by running multiple replicas per CPU. However, using one CPU for multiple replicas does not effectively take advantage of the parallelization inherent in the RE method.

The RE algorithm requires the synchronization of attempted moves, which results in wasted CPU time if any replica waits for other replicas to perform exchanges. Inhomogeneity of CPU speeds affects the amount of wasted time, since the speed of the calculation depends on the speed of the slowest processor. Modified versions of RE have been developed in an effort to minimize wasted CPU time, including the multiplexed replica exchange method (MREM)[23] and asynchronous replica exchange.[43] MREM makes use of multiplexed layers of replicas (n layers, each with M temperatures), with exchanges occurring both within and between layers.[23] MREM is even more computationally demanding than RE, using n times as many processors. MREM does not offer a significant advantage if there is a shortage of CPUs, but it does offer a way of using more CPUs without adding more temperatures. In asynchronous replica exchange, only the replicas undergoing exchange are synchronized, therefore increasing efficiency on heterogeneous computing platforms.[43] More complex replica management schemes have also been proposed to increase the efficiency of RE.[44,45] However, modified RE algorithms do not completely alleviate the need for synchronization and frequent communication between replicas.[14] This is especially important to users of distributed computing, such as the massively parallel Folding@Home project,[46] who must contend with inhomogeneity of processor speeds.[14]

Serial Replica Exchange (SREM). The serial replica exchange method[10] was recently developed to address the main practical limitations inherent in the RE method, namely, the need for synchronization and a large number of processors. The exchange probability in SREM has an identical form to that of RE (eq 8) for a replica at temperature T_i attempting to move to a neighboring temperature T_j:

$$P(T_i \rightarrow T_j) = \min\left\{ \begin{array}{l} 1 \\ e^{-(\beta_j - \beta_i)(E_i - E_{j,\text{PEDF}})} \end{array} \right. \quad (9)$$

Unlike RE, the attempted move from T_i to T_j does not simultaneously involve another replica moving from T_j to T_i. In SREM, the potential energy, $E_{j,\text{PEDF}}$, does not come from another replica at temperature T_j but rather is selected at random from a potential energy distribution function (PEDF) for that temperature. The PEDFs are determined through initial simulations at each temperature, which may use either constant-temperature MD or RE. These initial simulations can be very computationally demanding for biomolecular systems. For example, to obtain converged PEDFs for a small RNA hairpin, approximately 100 ns per temperature was required.[47] PEDFs may also need to be updated throughout the course of the SREM simulation.[14,47] SREM also cannot be applied to temperature-dependent force fields.[48,49]

In terms of practical implementation, SREM offers the same advantages as ST. In both methods, there is absolutely no communication required between independent simulations. Neither method requires a fixed number of CPUs, and there is no wasted CPU time in the synchronization of attempted exchanges. In principle, both ST and SREM can be run on a single CPU. SREM also presents the same critical challenge as ST: an initial simulation is needed to determine PEDFs, the length of which is highly dependent on system complexity. The significant computational cost of calculating accurate PEDFs is a key drawback of SREM, since an SREM simulation is not strictly correct if unconverged or incorrect PEDFs are used.[10,14] In contrast, the weight factors of ST can deviate from the accurate dimensionless Helmholtz free energies and still yield correct results.[3,10,14]

Virtual Replica Exchange (VREX). The first novel method we propose, virtual replica exchange, is based on the principles of both RE and SREM. A replica at temperature T_i attempts a move to temperature T_j, with the probability of exchange given by the following equation:

$$P(T_i \rightarrow T_j) = \min\left\{ \begin{array}{l} 1 \\ e^{-(\beta_j - \beta_i)(E_i - E_{j,\text{virtual}})} \end{array} \right. \quad (10)$$

Here, the potential energy, $E_{j,\text{virtual}}$, comes from a list of stored energy values obtained at temperature T_j. This is analogous to exchanging with a potential energy value selected from a PEDF in SREM, or the current potential energy of a replica at temperature T_j in RE. Like SREM, only a move from temperature T_i to temperature T_j occurs, with no simultaneous reverse move. In VREX, an energy value that occurred at temperature T_j in the past is used, and following the attempted exchange, the occurrence of this energy value is removed from the potential energy list. This constitutes a "virtual exchange".

VREX is derived to obey detailed balance in a similar manner to RE.[2,14] Consider a state A in which a replica with configuration X is at temperature T_i and a virtual replica with configuration X_{virt} is at temperature T_j. An exchange is

attempted to a state B in which the replica with configuration X is at temperature T_j and the virtual replica is at temperature T_i. The detailed balance criterion is written

$$P(A)\,P(A \rightarrow B) = P(B)\,P(B \rightarrow A) \qquad (11)$$

Transition probabilities that satisfy this criterion can then be derived as follows:

$$\frac{P(A \rightarrow B)}{P(B \rightarrow A)} = \frac{\dfrac{e^{-\beta_j H(X)}}{Z_j}\dfrac{e^{-\beta_i H(X_{\text{virt}})}}{Z_i}}{\dfrac{e^{-\beta_i H(X)}}{Z_i}\dfrac{e^{-\beta_j H(X_{\text{virt}})}}{Z_j}} = e^{-(\beta_j-\beta_i)(E(X)-E(X_{\text{virt}}))}$$

$$(12)$$

This transition probability is satisfied by the Metropolis criterion in eq 10. It also resembles the transition probability for RE (eq 8). The VREX algorithm is completely analogous to RE, except that one replica undergoes a virtual reverse move. Similarly, SREM also involves a virtual reverse move, in this case, by a replica whose energy distribution is represented by the PEDF. In contrast, the RE algorithm involves two real replicas undergoing temperature moves.

In practice, VREX requires very short initial simulations in order to generate a preliminary list of energies for each temperature. These lists are then updated as the simulation progresses, with values being added from each short MD simulation between exchange attempts and values being removed as they are used in virtual exchanges. It is possible to run out of potential energy values in the primary lists if temperatures are sampled heterogeneously. In order to address this possibility, implementations of VREX may include the use of secondary lists, to which potential energies from the primary lists are moved after a single use. Potential energies from a secondary list may be used in the rare case that the primary list for that temperature is completely used. Further, recent values can be prioritized in the primary lists, and relatively short and continually overwritten secondary lists can be maintained in order to reduce the likelihood of using pre-equilibration potential energies in post-equilibration virtual exchanges.

The main advantage of VREX is that it avoids the need to calculate converged PEDFs (like SREM) or weight factors (like ST) and only requires a short list of potential energies to begin sampling. It also addresses the main shortcoming of RE because it completely eliminates the synchronization between replicas, as well as the need for a fixed number of replicas. It is theoretically very similar to RE, with the addition of a variable time delay between the time when a potential energy is produced and when it is used for an exchange.

Distributed Replica Sampling (DR). Distributed replica sampling[11] is a general scheme for Boltzmann sampling of conformational space in which multiple replicas undergo a random walk in a reaction coordinate of interest. Individual replicas are coupled through a generalized Hamiltonian containing a potential energy term that depends on the distribution of all replicas, which acts to enforce a desired sampling distribution of the reaction coordinate. DR can therefore be used to enforce uniform sampling along a

reaction coordinate of interest. This may be, for instance, a nonphysical spatial "fourth" dimension[12] or a dihedral angle.[13] We briefly summarize the DR algorithm with temperature as the coordinate.[11] The implementation of DR in other coordinates has also been previously outlined.[11−13]

The generalized Hamiltonian of DR in temperature contains a pseudoenergy term and depends on the current inverse temperature and current configuration (q) of all replicas:

$$H(q_1,\beta_1,q_2,\beta_2,...,q_M,\beta_M) = \sum_{m=1}^{M} \beta_m\,\mathrm{E}(q_m) + \\ \mathrm{DRPE}(\beta_1,\beta_2,...,\beta_M) \quad (13)$$

where E is the potential energy. There are M replicas in total, each labeled by an index $m = 1, ..., M$. The distributed replica potential energy (DRPE) can take any functional form that depends on the distribution of replicas and fulfills the purpose of enforcing homogeneous sampling of the temperature coordinate. Importantly, although the DRPE is a pseudoenergetic penalty, it is not a function of system complexity.[11] The probability of a replica currently at temperature T_i jumping to a temperature T_j is

$$p(T_i \rightarrow T_j) = \min\left\{ \begin{array}{l} 1 \\ e^{-(\beta_j-\beta_i)E(q_i)-(\mathrm{DRPE}_j-\mathrm{DRPE}_i)} \end{array} \right. \quad (14)$$

which depends on the difference between the DRPE with the replica at temperature T_j (DRPE_j) and that at temperature T_i (DRPE_i).[11] DR can be analogously used to achieve a random walk in a parameter of the Hamiltonian, ξ, with an exchange probability:

$$p(\xi_i \rightarrow \xi_j) = \min\left\{ \begin{array}{l} 1 \\ e^{-\beta[(H(q,\xi_j)-H(q,\xi_i))+(\mathrm{DRPE}_j-\mathrm{DRPE}_i)]} \end{array} \right. \\ (15)$$

DR was designed specifically to suit shared or distributed computing platforms.[11] In contrast to RE, in which pairwise exchanges of replicas are attempted, DR considers stochastic moves of individual replicas one at a time. The stochastic move of one replica is coupled to the distribution of all other replicas through the DRPE, and no direct communication between replicas is required. In DR, synchronization of exchange attempts is therefore completely eliminated, which results in 100% CPU utilization.[11] The algorithm also readily accommodates fluctuations in CPU availability.[11] DR in combination with thermodynamic integration (TI) was shown to sample conformational space more effectively than TI alone in the calculation of the binding free energy of benzene to T4 lysozyme, while simultaneously optimizing the use of available computational resources.[12] This approach was also successfully employed to compute partial water occupancy in the pathway of proton uptake in cytochrome c oxidase.[50] In addition, DR has been combined with umbrella sampling (DRUS) to allow equilibrium exchange between different umbrella biasing potentials.[13,50] When applied to alanine dipeptide, umbrella sampling alone exhibited quasi-nonergodic behavior, while DRUS alleviated this systematic error.[13]

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2645**

Application of the DRPE restores sampling homogeneity of temperature only when the DRPE contribution is large enough to balance the preference for sampling the lowest temperature. When ST is conducted with all weight factors equal, as outlined in eq 14, all replicas migrate with a strong preference for the lowest temperature (as described above). In this case, a very strong DRPE is required to achieve sampling homogeneity. However, it has been demonstrated that, as the energetic penalty of the DRPE becomes stronger, replica mobility (as measured by acceptance ratio) decreases,[11] and therefore some modification to the DR exchange probability is necessary. This issue can be addressed by simply adding weight factors to the exchange probability, analogous to the weight factors, f, in the DRUS exchange probability:[13]

$$p(\xi_i \rightarrow \xi_j) = \min\left\{\begin{array}{l} 1 \\ e^{-\beta[(H(q,\xi_j)-H(q,\xi_i))-(f_j-f_i)+(\text{DRPE}_j-\text{DRPE}_i)]} \end{array}\right.$$

(16)

This form of the exchange probability results in good replica mobility and nearly perfect sampling homogeneity of the reaction coordinate.[13] The efficiency and practical advantages of DR in other coordinates have been well established.[11−13] It is therefore a central objective of this study to develop and test an implementation of DR which functions optimally in temperature space.

Simulated Tempering Distributed Replica Sampling (STDR). Building on the success of the both the ST method[3,4] and DR sampling,[11−13,50] we have developed a new algorithm, STDR, which combines the two approaches. STDR is essentially DR implemented rigorously in temperature. The combination of these two methods was originally suggested when DR was developed.[11] In STDR, approximately homogeneous sampling of a set of temperatures is enforced. The probability of accepting a move from a temperature T_i to a neighboring temperature T_j is

$$p(T_i \rightarrow T_j) = \min\left\{\begin{array}{l} 1 \\ e^{-(\beta_j-\beta_i)E+(a_j-a_i)-(\text{DRPE}_j-\text{DRPE}_i)} \end{array}\right.$$

(17)

This is the same as the exchange probability from ST, with the addition of the difference in DRPE between the states for which the replica is at temperature T_i (DRPE$_i$) and temperature T_j (DRPE$_j$). The calculation of the DRPE is straightforward. Its functional form depends upon the current temperatures of all replicas as follows:[12]

$$\text{DRPE} = c_1 \sum_{m=1}^{M} \sum_{n=1}^{M} [(\lambda_{m,\text{linear}} - \lambda_{n,\text{linear}}) - \omega(m-n)]^2 +$$
$$c_2 \left[\sum_{m=1}^{M} \lambda_{m,\text{linear}} - \omega \sum_{m=1}^{M} m\right]^2 \quad (18)$$

Replicas are labeled by indices m and n, where M is the number of replicas. The values of $\lambda_{m,\text{linear}}$ refer to a linearly spaced temperature coordinate. In this coordinate, the lowest temperature has $\lambda_{m,\text{linear}} = 1$, and the highest temperature has $\lambda_{m,\text{linear}}$ equal to the number of temperatures. This procedure transforms the exponentially spaced temperatures into a uniformly spaced coordinate. The factor ω, which we introduce to the DRPE in this work, is the ratio of the number of temperatures to the number of replicas. This factor allows DR to be used with an arbitrary number of replicas. The first term in eq 18 introduces an energetic penalty for two replicas sampling the same temperature, while the second term introduces a penalty for an overall drift of the replicas toward high or low temperatures. The second term is not essential when using DR in temperature. The constants c_1 and c_2 control the influence of the DRPE and can be tuned to enforce homogeneous temperature sampling as required.[11] In the case of accurate weight factors, the influence of the DRPE only needs to be small such that values of c_1 and c_2 near zero can be used. With increasingly inaccurate weight factors, larger DRPE values are required to maintain homogeneous temperature sampling, and this reduces the acceptance ratio to some degree. An example calculation of the DRPE using the temperature as the reaction coordinate is provided as Supporting Information.

If the weight factors, a_m, are inaccurate, ST results in uneven sampling of the temperature coordinate. As we will demonstrate, introducing the DRPE recovers homogeneous sampling. The STDR method is therefore more generally applicable than ST because it can make use of a poor estimate of the dimensionless Helmholtz free energies and still yield uniform sampling of the canonical ensembles at each temperature. STDR is equivalent to ST in the limit of one replica and is compatible with adaptive schemes for computing weight factors. Below, we show that STDR is the preferred method for systems with a complex energy landscape for which limitations on computational resources preclude obtaining sufficiently accurate estimates of Helmholtz free energies for a ST simulation.

Test System. For the purpose of comparing different generalized-ensemble methods, we use two related test systems, the peptides GVGVPGVG and (GVPGV)$_7$. These peptides are both based on the pentapeptide GVPGV, which is found as a repeat motif in the protein elastin.[51] In our previous study of (GVPGV)$_7$ and other related elastin-like peptides, we observed that this peptide is intrinsically disordered, having many conformations and no extended secondary structure in the form of α-helices or β-sheets.[52] Understanding the structural heterogeneity of elastin-like peptides is required to elucidate the structure−function relationship of elastin, for which experimental characterization is notoriously difficult due to its flexibility and insolubility. The peptide GVGVPGVG has also been studied previously and was suggested to exhibit an "inverse temperature transition" with an increased probability of "closed" conformations (in which the N and C termini are closer than 8 Å) at higher temperatures.[53] On the basis of this work, the octamer is a simple yet appropriate peptide to study in the aim of understanding the temperature-dependent behavior of elastin. Because the main focus of this paper is the thorough comparison of generalized-ensemble methods using these peptides as test systems, we do not elaborate fully on the structural details of either the octapeptide or the 35-residue peptide in this paper. A full characterization of the conformational landscape of these peptides will be the subject
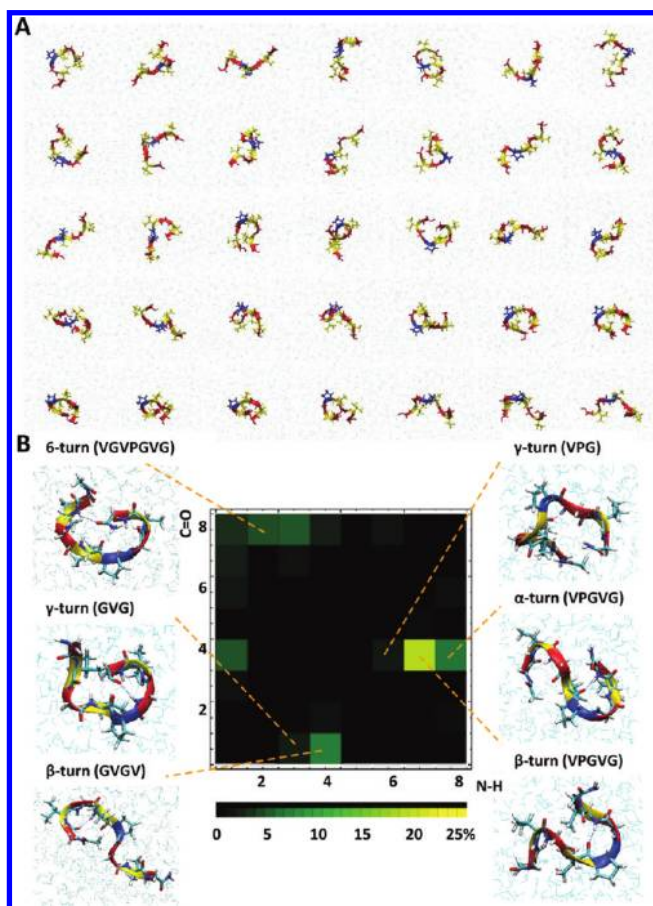
Figure 1. Conformational landscape and hydrogen-bonding contact map of GVGVPGVG. (A) A selection of 35 random conformations from the STDR simulation at 280 K, with glycine in red, valine in yellow, and proline in blue. (B) Hydrogen-bonding contact map at 280 K, with corresponding snapshots showing the presence of significantly populated contacts. N−H groups are on the horizontal axis, and C═O groups are on the vertical axis. Each square in the matrix (*i,j*) corresponds to a contact between the N−H group of residue *i* and the C═O group of residue *j*. The color scheme of the legend indicates the relationship between color and probability of contact formation.

of future work. Both GVGVPGVG and (GVPGV)$_7$ are valuable test systems because of their structural complexity and the fact that they represent a real scientific problem in the sense that they are not well understood or characterized a priori. Simple test systems are often used for comparison purposes, such as alanine dipeptide,[10,13] although generalized-ensemble methods are typically applied to systems which are much larger and more complex. While simple test systems are useful for the sake of demonstration and for the elucidation of major problems, they are less likely to detect the subtleties and practical issues experienced when studying systems of biologically relevant complexity.

The conformational landscape of the octapeptide is complex, with many energetically accessible states that must be sampled in order to accurately compute free energies. A representative selection of these conformations is shown in Figure 1A, with "closed" states in which the N and C termini are in close proximity, "hairpin"-like states, and extended structures. Although it is a short peptide, GVGVPGVG

represents a challenging sampling problem due to the large number of thermally accessible conformations. In Figure 1B, we show the hydrogen-bonding contact map for this peptide obtained using STDR. The only secondary structure consists of hydrogen-bonded turns, with no α-helix or β-sheet. The most populated turn is the VPGV β turn, with a hydrogen bond between the C═O group of valine 4 and the N−H group of valine 7. Several other turns form with lower populations. As we will show, single-temperature MD, if run for a sufficiently long time, provides a converged description of the conformational landscape. This makes it an ideal test system because we can verify that the generalized-ensemble algorithms, given sufficient sampling, lead to correct Boltzmann-weighted sampling of conformational space, in addition to an assessment of their relative efficiency.

The 35-residue peptide, (GVPGV)$_7$, is used as a more complex test system to demonstrate the sampling enhancement provided by STDR for a landscape which not only has many populated states but also has significant energetic barriers between those states. The larger system is only simulated using constant temperature MD, ST, and STDR because of the extensive amount of computational resources required. Of the methods we consider, STDR is better suited to this particular application on the basis of its performance for the octapeptide. It is as efficient and accurate as the other methods, while offering the most practical advantages for a large and complex system (see below).

Simulation Details. For all five methods (ST, STDR, SREM, RE, and VREX), the same exponentially spaced temperature list was used. This list is provided as Supporting Information Table S1. The simulation system consists of the GVGVPGVG octapeptide, capped with an acetyl group at the N terminus and an NH$_2$ group at the C terminus, in a 3 × 3 × 3 nm^3 box with 872 water molecules. The same fully extended starting structure was used for all temperatures and all methods. Simulations were performed using the GROMACS MD simulation package, version 3.3.1,[54,55] with the OPLS-AA/L force field[56,57] for the solute and the TIP3P model for water.[58] Periodic boundary conditions were applied. The switch function of GROMACS was used for Lennard-Jones interactions, which corresponds to the usual Lennard-Jones function until 1.3 nm is reached, after which it is switched to reach zero at 1.4 nm. Covalent bonds involving hydrogen atoms were constrained with the SHAKE algorithm.[59] Calculations of electrostatic forces utilized the particle mesh Ewald summation method[60,61] with a Fourier spacing of 0.15 nm and a fourth-order interpolation. The real-space Coulombic cutoff was 1.49 nm. All MD simulations were performed in the canonical ensemble. Peptide and solvent were coupled to the same reference temperature bath with a time constant of 2 ps using the Nosé−Hoover method.[62,63] An integration step size of 2 fs was used, and coordinates were stored every 1 ps.

In order to compare the generalized-ensemble methods, the simulations were conducted as similarly as possible. To this end, the same total amount of simulation time (summed over all replicas) was performed. This amount was 4.75 μs, with an average of approximately 144 ns per replica. This time was used because it was sufficient

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2647**

for all methods to achieve statistical convergence, as shown in the results. Stochastic exchanges using the Metropolis Monte Carlo algorithm[64] were attempted every 25 ps. Exchange probabilities were calculated using eqs 6, 8, 9, 10, and 17, as appropriate for the method. Details of the calculation of weight factors and PEDFs are discussed below. The constants c_1 and c_2 for the DRPE in eq 18 were both 0.005. These values were found to achieve an appropriate balance between homogeneity of temperature sampling and replica mobility.[11] The value of the factor ω was 1.0, since the number of replicas equaled the number of temperatures. The generalized-ensemble algorithms were implemented using an in-house bash script. Software for distributed replica sampling is also available online at www.pomeslab.com.

The same simulation protocol was used for the simulation of (GVPGV)$_7$, which was simulated in a 4.5 × 4.5 × 4.5 nm^3 box with 2856 water molecules using both ST and STDR. Starting conformations and weight factors for each temperature were generated using canonical MD for 15 ns per temperature (storing 250 energy values per picosecond). A total of 70 temperatures were used for each generalized-ensemble simulation. The list of temperatures is provided in Supporting Information Table S1. Temperatures were spaced more closely than those of the octapeptide. This is because it is a larger system, resulting in less overlap between potential energy distributions of adjacent temperatures for a given temperature separation. This system was simulated for a total of 8.2 μs (117.6 ns per replica on average) using the STDR algorithm. An ST simulation using the same weight factors was also performed for 420 ns (6 ns per replica on average). ST and STDR simulations were also performed using weight factors calculated using the first 500 ps of continuous MD at each temperature for a total of 280 ns.

A conventional MD simulation of the 35-residue peptide system in the isothermal−isobaric ensemble was also performed using GROMACS, version 4.0.2.[65] In this simulation, a 4 fs time step was used, and constraints on bonds and angles involving hydrogen were imposed using the LINCS algorithm.[66] This simulation was run for 200 ns at 261 K, which corresponds to the lowest temperature in the STDR simulation. The pressure was kept constant at 1.0 bar using the Parrinello−Rahman algorithm.[67]

The analysis of the data accumulated in the trajectories was performed using an in-house script based on a modified version of the Dictionary of Secondary Structure in Proteins (DSSP).[68] For each snapshot, possible backbone hydrogen bonds were evaluated using both (a) the energetic criterion of DSSP and (b) the following geometric criteria: (i) donor−acceptor and hydrogen-acceptor distances are less than 3.5 and 2.5 Å, respectively, and (ii) the value of the acceptor−donor-hydrogen angle is less than 60°. Definitions of turns and bends are the same as those in DSSP.[68] End-to-end distance (EED) is calculated as the distance between the α carbons of the first and last residue. Root mean square deviation (rmsd) was calculated using the g_rms program in GROMACS.[65] All molecular visualizations in the manuscript were produced using VMD.[69]

Calculations of Weight Factors for ST and STDR, PEDFs for SREM, and Potential Energy Lists for VREX. The calculation of weight factors for eqs 6 and 17 required initial simulations of the octapeptide in the canonical ensemble for each of the temperatures listed in Supporting Information Table S1. These simulations were performed using conventional MD for 19.5 ns (for a total simulation time of 643.5 ns). Although obtaining these accurate weight factors was resource-intensive, it involved a straightforward procedure. The weight factors were computed using the average potential energy at each temperature according to eq 7.[30] The accuracy of these weight factors was assessed by using them in an ST simulation and observing the temperature sampling uniformity, as shown in the results. Since all temperatures were sampled with nearly equal probability, as expected from eqs 3 and 5 for accurate dimensionless Helmholtz free energies, these weight factors were deemed to be sufficiently converged and correct.

Using the same data from the conventional MD simulations, PEDFs were computed as described in the original SREM paper.[10] The convergence of the PEDFs was assessed by calculating the χ^2 measure suggested by Hagen *et al.*:[10]

$$\chi^2(t) = \sum_{n=1}^{N_{bins}} (P_i(t) - P_{i,\text{reference}})^2 \qquad (19)$$

This measure computes the deviation of each bin in the current distribution, $P_i(t)$, from a reference distribution, $P_{i,\text{reference}}$. The current distribution is cumulative, using the data up to time t. For the reference distributions, we used PEDFs computed using all of the data at each temperature. By this assessment, the PEDFs appeared to be stationary, as shown in Figure 2A. When χ^2 was plotted individually for each temperature, we also observed that each PEDF was stationary. However, an initial SREM simulation using these PEDFs resulted in nonuniform sampling of temperatures. We therefore proceeded to calculate the PEDFs using a different data set. We used the first 25 ns at each temperature of the RE simulation (for a total time of 825 ns), and these were the PEDFs used for the SREM simulation. While this procedure is similar to what would likely be done in practice with SREM, we emphasize that making this selection of PEDFs gave SREM somewhat of an advantage over ST, since more data were used in the initial simulation. The use of RE in the calculation of PEDFs is similar to REST.[36,38] Although REST results in faster convergence of the weight factors compared to conventional MD, it may be difficult or impossible to obtain access to the required number of homogeneous and dedicated CPUs for the initial RE simulation. Thus, we did not use REST to obtain the weight factors for ST to better represent the general case where it may not be convenient to do so. In contrast, it was necessary to use RE to obtain PEDFs for SREM in a reasonable amount of time.

Figure 2B shows the error in the exchange probability for both SREM and ST using the data from 19.5 ns of conventional MD at each temperature. The method for computing the error in exchange probabilities is provided in Appendix 1. The weight factors of ST produce an average
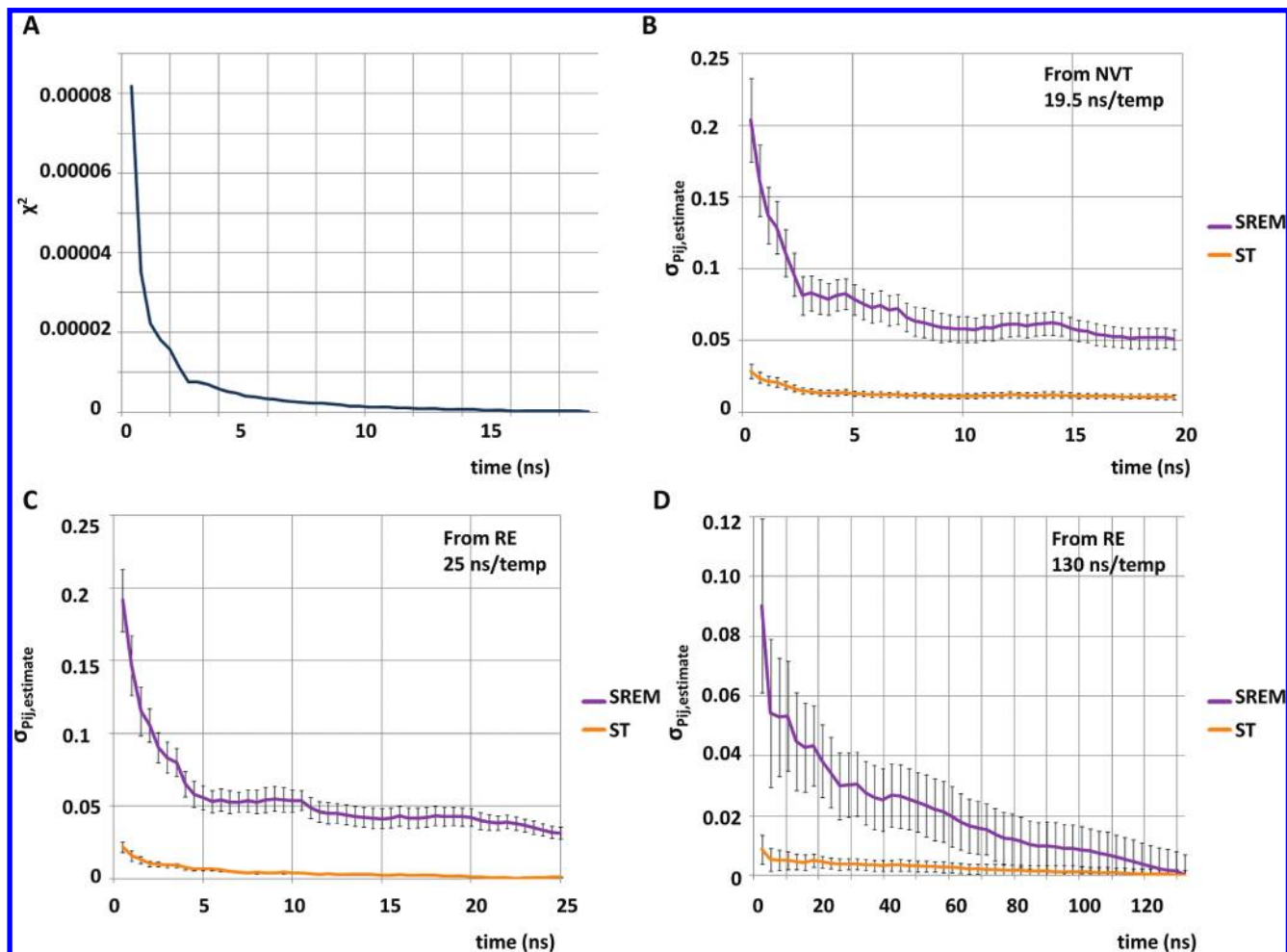
Figure 2. Assessing convergence of weight factors and PEDFs. (A) The convergence of the PEDFs for SREM is quantified using the χ^2 measure defined in eq 19. Using this measure, the PEDFs obtained using 19.5 ns of conventional MD at each temperature appear to be stationary. (B) Convergence of the PEDFs and weight factors using data from 19.5 ns of conventional MD per temperature, using the data from the complete RE simulation as a reference. (C) Similarly, data from the RE simulation, using only the first 25 ns per temperature, with the data from the complete RE simulation as a reference. (D) Similarly, using all of the data from the RE simulation (4.75 μs). Error in the acceptance ratio is shown in B, C, and D for both ST in yellow (computed using eqs 25 and 29) and SREM in purple (computed using eqs 24 and 27).

error in the exchange probability of less than 2% after 19.5 ns per temperature. Using the same amount of data, the PEDFs produce a significantly higher error in the exchange probability (more than 5%), which is why the weight factors used in ST from conventional MD produced more homogeneous sampling than the PEDFs. In Figure 2C, the error in the exchange probability for both ST and SREM is shown using the data from the first 25 ns at each temperature of RE. This data set was used to calculate the PEDFs for the SREM simulation, producing an error in the exchange probability of less than 4%. The convergence of the PEDFs estimated using all of the data from RE is shown in Figure 2D. The error in the exchange probability had only decreased to less than 2% after approximately 60 ns per temperature. That is, SREM would have required preliminary simulations which were half as computationally expensive as the entire RE simulation in order to produce error in the exchange probability equivalent to that of ST. The slow convergence of PEDFs is likely why they have been updated throughout

the course of the simulation in other studies.[10,14,47] However, an SREM simulation is strictly correct only with accurate PEDFs.[14]

Figure 2 demonstrates that the error in the weight factors of ST leads to smaller error in the exchange probability than the error in the average energy of PEDFs. This finding is in qualitative agreement with a previous study comparing SREM and ST for a helical peptide.[14] The PEDFs of SREM were observed to converge more slowly than the weight factors of ST when starting from a coil conformation, but not when both ST and SREM were started with a helical conformation.[14] In the original SREM paper, it was hypothesized, but not shown, that the calculation of PEDFs should be significantly easier than the calculation of weight factors for ST.[10] In fact, we observe that the opposite is true for this system. The weight factors converge significantly faster than PEDFs and lead to more homogeneous sampling of the temperature. The difference in errors is likely because the exchange probability in ST uses a difference in dimensionless free energies, whereas the absolute value of the potential

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2649**

energy is used in the exchange probability of SREM. Additionally, the method for computing the weight factors uses only the average potential energy at each temperature.[30] It is harder to converge a distribution of potential energies than the average of the distribution. Since weight factors are computed using average potential energies, they converge more quickly than PEDFs. The accuracy of the potential energy value selected from the discrete PEDF in SREM is also affected by the number of bins and the bin width.[10] The accuracy is decreased by having too few bins, whereas the convergence of the distribution is slower with a larger number of bins. These errors must therefore be balanced. Even if the PEDFs and weight factors converged at the same rate, ST has the advantage of convenience, since it entails storing a short list of weight factors rather than a distribution of energy values for each temperature.

We also tested the effects of using a poor estimate of the weight factors in ST. In order to generate suboptimal weight factors, we used the data from the first 750 ps of the RE simulation. This required a total of 24.75 ns summed over all temperatures, compared to 643.5 ns used to generate accurate weight factors. These weight factors produced inhomogeneous sampling of the temperature, confirming that they were inaccurate estimates of the dimensionless Helmholtz free energies (as is demonstrated in the results below). The purpose of this exercise was to emulate the more general case of a complex system for which one may not be able to accurately calculate weight factors due to the prohibitive computational cost. ST and STDR simulations carried out with these inaccurate weight factors will hereafter be referred to as STb and STDRb, respectively.

Potential energy lists for the VREX simulation were also generated using the RE data. A list of 1000 energy values from the first 1 ns was used for each temperature. We did not run out of potential energy values in the primary lists and therefore no secondary lists were used. However, secondary potential energy lists may be necessary in the application of this method to other systems. In summary, we highlight the varying costs of the initial simulations for each of the methods in terms of the simulations times: RE (0 ns), VREX (33 ns), SREM (825 ns), STDR (643.5 ns), STDRb (24.75 ns), ST (643.5 ns), and STb (24.75 ns).

Results and Discussion

Practical Implementation Issues. Before we begin a detailed comparison of the efficiency of the temperature-based generalized-ensemble methods, we briefly compare them with regard to the practical issues encountered in their implementation. A summary of this comparative discussion is provided in Table 1. Prior to beginning an enhanced sampling simulation, it is necessary to assess the available computational resources, including the number of processors available, the heterogeneity of their speeds, and their failure rate (frequency of "crashes").

In terms of the number of CPUs required, the RE algorithm specifies that the number of replicas equal the number of temperatures, which grows with system size. In the typical implementation of RE, the number of processors equals the

Table 1. Practical Advantages and Disadvantages of Generalized-Ensemble Algorithms[a]

Implementation Issue	RE	SREM	VREX	STDR	ST
Scalable to any number of CPU's (even one CPU)?					
Algorithm readily accommodates a fluctuating number of CPU's?					
Efficiency impaired by inhomogeneity of CPU's?					
Performance severely affected by CPU failure?					
Initial calculation of weight factors, potential energy distribution functions or potential energy lists required?					
Initial simulation times (ns)	0	825	33	643.5 (STDR) 24.75 (STDRb)	643.5 (ST) 24.75 (STb)

[a] If a method is not affected by an implementation issue, the corresponding square is colored in green. Yellow indicates that the issue is somewhat of a concern, and red indicates that it is potentially a major pitfall. The only major issues for SREM and ST are the calculation of PEDFs and weight factors, respectively. STDR and VREX are not severely affected by any implementation issue. However, they do require very short initial simulations to obtain weight factors and potential energy lists and, ideally, would not be run with only one replica. RE, in its typical implementation in which each replica is run on a dedicated CPU, is hindered by all of the issues listed, except that it does not require any initial simulation.

number of replicas. If it is not possible to obtain access to the required number of processors, an alternative method or a more advanced RE implementation must be sought. Another possible scenario is that extra processors are available, which could be utilized to speed up the calculation, but the RE algorithm does not allow the possibility of having more replicas than temperatures. This particular issue is addressed by the MREM algorithm, which utilizes multiple layers of replicas.[23] However, there is no general mechanism to adapt RE to use available resources most efficiently. In contrast, both ST and SREM completely eliminate the need for a specific number of replicas. Multiple ST or SREM simulations can be run independently to take advantage of a computing cluster or distributed computing. The benefit of utilizing several processors simultaneously, each running an independent ST or SREM simulation, is simply reaching convergence more quickly in terms of wall clock time. Similarly, STDR and VREX algorithms do not require a fixed number of replicas. However, the aim of the DRPE is to enforce homogeneous sampling of temperatures for multiple replicas. Using only one replica is therefore not optimal, and ideally one would use a number of replicas comparable to the number of temperatures, though there is no specific requirement. A VREX simulation can in principle have any number of replicas. However, there is likely some benefit to having multiple replicas sampling different regions of conformational space in the updating of the potential energy lists (that is, running more than one replica at a time).

Of the generalized-ensemble methods we consider, only RE prevents the number of replicas from fluctuating during the course of the simulation. This may be a drawback in distributed

computing platforms and shared computing clusters, for which there is no way to predict the number of available processors in advance. Furthermore, the efficiency of RE is significantly affected by inhomogeneity of CPU speeds. Each exchange step can only occur when all of the replicas have completed their MD calculation. Any inhomogeneity in the computing environment results in a waste of computational resources, as some replicas must wait for the replica with the slowest processor to finish its calculation. This issue has been partly addressed by the asynchronous replica exchange method, although some degree of synchronization is still required for the replicas undergoing exchange.[43] Since none of the other methods require any direct communication between replicas, they do not suffer from this inefficiency. Another key drawback of the typical implementation of RE is its sensitivity to CPU failure.[10,11] If one of the replicas is running on a processor that crashes, the entire RE simulation is stalled until this replica can be restarted on a functioning processor. The time wasted due to CPU failure depends on the failure rate of the cluster and can be quite significant. Failure rates also rise with the number of replicas, and therefore the failure rate of RE is equal to the number of replicas times the failure rate of either SREM or ST.[10]

In contrast to RE, the other four methods all have the advantage of not requiring a fixed and synchronized cluster of CPUs to function optimally. From a practical point of view, these methods are all superior to RE, except in one regard. Only RE does not require initial simulations at multiple temperatures to obtain weight factors, PEDFs, or potential energy lists. In particular, ST and SREM appear to only be suited to systems for which accurate weight factors or PEDFs can be calculated in a reasonable amount of simulation time. For the test system in the present study, accurate weight factors for ST were computed using single-temperature MD, whereas SREM required more simulation time and the use of RE in order to obtain sufficiently accurate PEDFs (see methods section). We will demonstrate that STDR can function with less accurate weight factors and therefore requires less initial simulation time than ST. Finally, VREX requires significantly less initial simulation than SREM, STDR, or ST. Only short lists of potential energies at each temperature are needed to begin a VREX simulation.

An ideal temperature-based generalized-ensemble method would not require a significant initial simulation (as do ST and SREM) but also would not involve the use of a large cluster of homogeneous CPUs (as is common for RE). STDR and VREX address both of these issues, and are the most flexible algorithms in terms of practical concerns. These issues are particularly important if one is using a distributed computing platform with fluctuating numbers of heterogeneous CPUs in many different locations, or a shared computing cluster, which may present similar limitations.

Diffusion in Temperature. We characterize the efficiency of the temperature diffusion of each method using several different metrics, which are summarized in Table 2. First, we calculate the average acceptance ratio, which is a metric commonly reported for RE simulations.[70] The methods separate into two categories based on their acceptance ratios: the RE-based methods (RE, VREX, and SREM) and the ST-based methods (ST, STb, STDR, and STDRb). ST has a

Table 2. Evaluating Temperature Diffusion

Property	RE	VREX	SREM	STDR	STDRb	ST	STb
Acceptance Ratio	0.237	0.241	0.248	0.378	0.376	0.463	0.404
	7	6	5	3	4	1	2
Replica Speed (distance/time)	0.058	0.050	0.051	0.058	0.059	0.065	0.065
	4	7	6	4	3	1	1
Mean Free Path	0.322	0.245	0.255	0.402	0.402	0.431	0.445
	5	7	6	3	3	2	1
Diffusion Coefficient	0.208	0.181	0.156	0.195	0.196	0.246	0.249
	3	6	7	5	4	2	1
Average Deviation from Sampling Homogeneity (%)	0	6.62	12.61	2.50	2.98	3.81	17.40
	1	5	6	2	3	4	7
Composite Score (a normalized linear combination)	0.79	0.70	0.68	0.87	0.88	0.98	0.94
	5	6	7	4	3	1	2

The quality of the random walk in temperature for each generalized-ensemble method is assessed using five criteria, and an overall score is obtained by taking the normalized linear combination. Normalization is performed by dividing each value by the maximum value of that measure. The fifth measure was included in the normalized linear combination as $1.0 - $ (average deviation from sampling homogeneity)/100% for consistency with the other measures of temperature diffusion. The overall scores for each property are ranked from 1 to 7, representing decreasing performance. Scores are colored as follows: 1 and 2 (green), 3−5 (yellow), and 6 and 7 (red).

higher acceptance ratio than RE for the same set of temperatures, in agreement with a previous comparison of the methods.[36] Similarly, ST has a higher acceptance ratio than SREM.[14] Zhang and Ma also observed that the rate of traversing temperatures is faster in ST, and that this effect becomes especially apparent if separations between adjacent temperatures are large, or if exchanges are attempted less frequently.[32] Park proved that this is generally true for a given set of temperatures[37] and concluded with a question as to whether the enhanced acceptance ratio affects the rate of sampling different microstates, and therefore structural convergence. We investigate whether the higher acceptance ratios in serial tempering algorithms (both ST and STDR) compared to those in parallel tempering (RE, VREX, and SREM) do in fact lead to faster structural convergence in the next section. It should be noted that the DRPE in STDR decreases the acceptance ratio relative to ST, since it increases the probability of rejecting moves that result in inhomogeneous temperature sampling. The extent of this effect depends on the constants c_1 and c_2 in eq 18.[11]

Next, we consider a quantity which we call "replica speed". Back exchanges can occur in which a replica accepts a move to an adjacent temperature and at the next exchange returns to its previous position. These back exchanges contribute to the acceptance ratio, but they result in no net change in temperature, and typically no crossing of significant energetic barriers. In order to account for these "unproductive" moves, we calculate the replica speed as the average distance traveled after 50 exchange attempts. Values of replica speed are reported in Table 2. All of the methods have similar values for the replica speed, with SREM and VREX being slightly slower. The higher acceptance ratios of the ST-based methods do not correspond to significantly

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2651**

faster replica speeds. That is to say, the higher acceptance ratios for the ST-based methods are partly due to an increased frequency of unproductive moves.

Making an analogy with the replicas traveling in temperature space as a type of diffusion in a one-dimensional coordinate, we calculate the mean free path and diffusion coefficient for each method. Mean free path is defined as the average distance traveled between successive rejected moves ("collisions"). The diffusion coefficient is defined as the rate of change of the mean squared deviation of distance over time. We notice that ST, both with accurate and inaccurate weight factors, has the highest mean free path and diffusion coefficient. Both STDR simulations behave remarkably similarly and are slightly slower at diffusion in temperature compared to ST. RE has a higher diffusion coefficient than STDR, but a lower mean free path. It is also slightly more efficient at temperature diffusion than VREX or SREM.

Another important criterion is the deviation from sampling homogeneity, which indicates the amount of deviation from uniform sampling averaged over all of the temperatures:

$$\text{average deviation from homogeneity} = \frac{1}{M} \sum_{m=1}^{M} \frac{|N_m - \langle N_m \rangle|}{\langle N_m \rangle} \quad (20)$$

where the number of samples at temperature m is N_m, the average number of samples per temperature is $\langle N_m \rangle$, and M is the number of temperatures. We report the deviation from sampling homogeneity for each temperature in Supporting Information Figure S1. The coupling of upward and downward moves in the RE algorithm results in perfectly uniform sampling of all temperatures. STDR produces nearly uniform sampling, with deviations from uniformity of 2.50% and 2.98% for accurate (STDR) and inaccurate (STDRb) weight factors, respectively. This is expected because the application of the DRPE favors uniform sampling of the temperature coordinate.[11] Even with inaccurate weight factors, the temperature sampling is still uniform, and the diffusion coefficient is still approximately the same. This indicates that STDR in the general case (i.e., with inaccurate weight factors) still successfully produces uniform sampling and good mobility in temperature. Our ST simulation also results in nearly uniform sampling, confirming the accuracy of the weight factors. ST with inaccurate weight factors (STb) produces the least uniform sampling, by design (17.40%). We intentionally selected weight factors to produce uneven sampling to represent a more complex system for which calculating weight factors accurately would be computationally expensive. VREX produces relatively uniform sampling, with an average deviation of 6.62%. Most of the inhomogeneous sampling in VREX occurs early in the simulation when the potential energy lists were based on a small amount of sampling, and the sampling is increasingly homogeneous with time. In contrast, SREM does not produce uniform sampling, with less sampling at the lowest temperatures and an average deviation of 12.61%.

Temperature sampling efficiency is characterized by an overall score. The five measures of efficiency defined in this
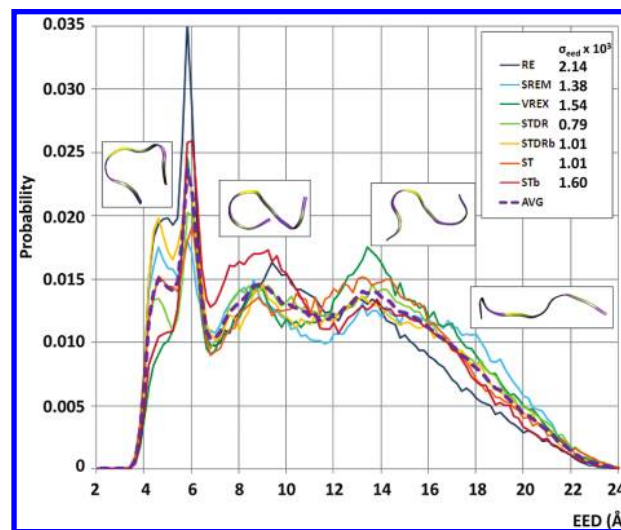

Figure 3. Assessing the accuracy of the EED distribution. The EED probability distribution is shown for each method with colors indicated in the legend using data from 280 K. The average distribution is computed as the average of all seven methods and is shown in purple (dashed line). The error of the distribution of each generalized-ensemble algorithm, σ_{eed}, is shown next to the legend and was computed using eq 21 with the average distribution as the reference.

section are combined by averaging their normalized values. The overall score for each method is reported in the last row of Table 2. ST with accurate weight factors performs the best overall, and all of the ST-based methods perform better than the RE-based methods (especially SREM, which has the lowest overall temperature diffusion score).

Convergence of Structural Properties. The octamer GVGVPGVG is a disordered peptide with many thermally accessible conformations, as shown in Figure 1. A useful descriptor of the conformation of such a short and flexible peptide is the end-to-end distance, EED. The EED probability distributions obtained using each of the generalized-ensemble methods at 280 K (the lowest temperature) are shown in Figure 3. Also shown is the average distribution, which is obtained by taking the average of all of the methods. There is no systematic bias of either ST-based or RE-based methods toward sampling particular conformations. For example, for the peak at 5 Å, RE, STDRb, and SREM are above the average while STDR, VREX, and STb are below. On the basis of this observation and of the large amount of sampling in the combined data set of all seven methods (nearly 35 μs), we take the average to be the "gold standard" for comparison throughout the analysis of structural convergence (it is hereafter referred to as the "reference"). We quantify the deviation σ_{eed}, of the EED distribution of each method $P_{\text{eed}}(n)$, from the reference EED distribution $P_{\text{eed,reference}}(n)$, by computing

$$\sigma_{\text{eed}} = \sum_{n=1}^{N_{\text{bins}}} (P_{\text{eed}}(n) - P_{\text{eed,reference}}(n))^2 \quad (21)$$

where the index n labels bins, and there are N_{bins} in total. The values of σ_{eed} are reported in Figure 3. STDR exhibits the best agreement with the average distribution. In
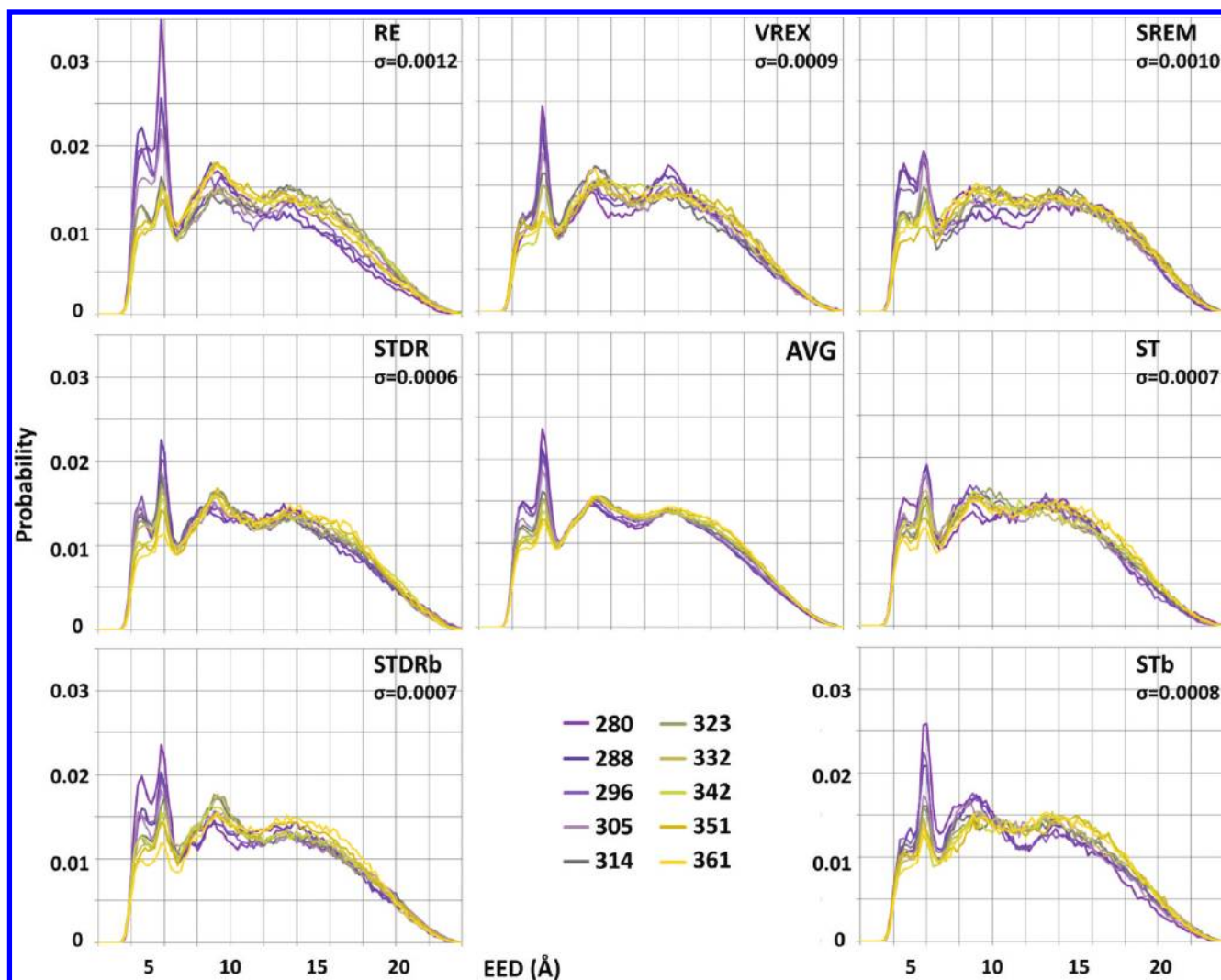
Figure 4. EED distributions at different temperatures. The EED probability distributions are shown for the lowest 10 temperatures for each generalized-ensemble method, as well as the average of all seven methods. The average error of the distributions of each generalized-ensemble algorithm, σ, is also shown. This was computed for each of the 10 temperatures using eq 21 with the average distribution as the reference, and the average of these errors is shown on each plot. The RE-based methods are shown in the top row and have larger errors than the ST-based methods, shown in the second and third rows. The average over all methods is shown in the central plot.

general, the ST-based methods have lower values for σ_{eed}, corresponding to more accurate EED distributions than the RE-based methods.

In order to confirm that the ST-based methods produce more accurate EED distributions when compared to the RE-based methods, EED distributions for the lowest 10 temperatures for each generalized-ensemble method are also computed and compared to the reference using eq 21. The EED distributions for each method and each temperature are displayed in Figure 4, along with the σ_{eed} value which is the average of the 10 temperatures. The ST-based methods produce EED distributions which are quantitatively more accurate than the RE-based methods at all temperatures. STDR shows the best overall agreement with the reference data set, with an average σ_{eed} value of only 0.006, and distributions which clearly show the same temperature trend as the reference distributions.

For a systematic comparison of the generalized-ensemble methods, the convergence of several structural properties in

addition to the EED distribution is considered. A useful ergodic measure is the 1,4 pair distance metric,[71,72] which quantifies the convergence of the distance between 1,4 residue pairs (residues with indices i and $i + 3$) over time. We extend this metric to include all residue pairs and therefore quantify the convergence of the α-carbon distance matrix as follows:

$$d_{\text{dC}\alpha\text{matrix}}(t) = \frac{1}{(N_{\text{residues}} - 1)!} \sum_{i=1}^{N_{\text{residues}}} \sum_{j=1}^{N_{\text{residues}}} (\bar{r}_{ij}(t) - \bar{r}_{ij,\text{reference}})^2$$

(22)

where the indices i and j correspond to residue number and the number of residues is N_{residues}. The difference between each average pairwise distance $\bar{r}_{ij}$ and the same average pairwise distance from the reference α-carbon distance matrix ($\bar{r}_{ij,\text{reference}}$) is computed. In this equation, t refers to simulation time accumulated at the temperature considered, and $\bar{r}_{ij}(t)$ is a cumulative average. As with EED, the average of all seven
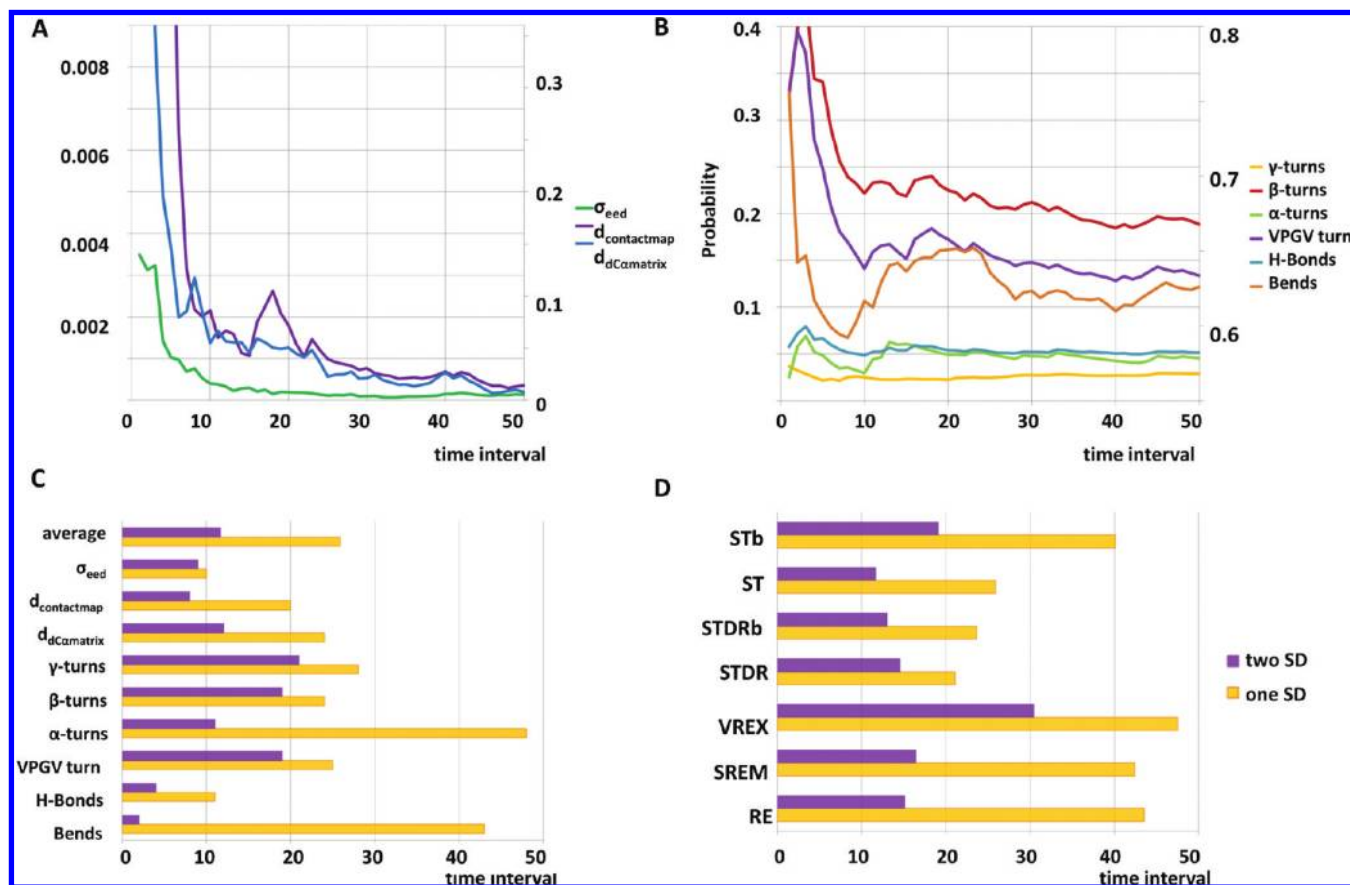
Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2653**

Figure 5. Assessing structural convergence using multiple criteria. The data for A, B, and C are from ST at 280 K. The trajectory is separated into 50 time intervals, and the quantities reported are calculated cumulatively. Time intervals are used to compare all methods fairly, since each method results in a different amount of sampling time at the lowest temperature. (A) Structural convergence is assessed using σ_{eed} (eq 21), $d_{contactmap}$ (eq 23), and $d_{dC\alpha matrix}$ (eq 22, plotted on the secondary axis). (B) The probability per residue of a γ-turn, β-turn, and α-turn are shown, as well as the population of the VPGV β-turn. The probability of a hydrogen bond per residue and a bend per residue (plotted on the secondary axis) are also shown. (C) For each of the structural properties shown in A and B, the time intervals at which they reached and remained within one and two standard deviations of the reference data set are shown. The average of these times is also shown, corresponding to the average structural convergence time, $\langle t_{sc} \rangle$. One standard deviation is calculated on the basis of the values of each of the seven generalized-ensemble methods at the end of the simulation and their standard deviation from the reference value. (D) The average structural convergence times for one and two standard deviations are shown for all seven methods at 280 K in yellow and purple, respectively. These times are provided for temperatures 288, 296, 305, 314, 323, and 332 K in Supporting Information Figure S2.

generalized-ensemble methods is used as the reference. We compute an analogous measure of convergence for the hydrogen-bonding contact map, depicted in Figure 1b:

$$d_{contactmap}(t) = \frac{1}{N_{residues}^2} \sum_{i=1}^{N_{residues}} \sum_{j=1}^{N_{residues}} (P_{ij}(t) - P_{ij,reference})^2 \tag{23}$$

where P_{ij} is the probability of a hydrogen bond forming between the C=O group of residue i and the N−H group of residue j, and $P_{ij}(t)$ is a cumulative average of all of the data. The elements of the reference contact map, $P_{ij,reference}$, are computed using the data from all seven methods. We also directly compute the probability of forming certain turns (γ, β, and α turns, defined by hydrogen bonds between residues i and $i + 2$, $i + 3$, and $i + 4$, respectively) as well as the VPGV β-turn (shown in Figure 1, the most probable turn). In addition, the average probabilities of forming a hydrogen

bond and a "bend" (as defined in the DSSP algorithm[68]) on a per-residue basis are computed. The convergence of each of these structural properties is considered individually and compared to the reference data. Taken as a set, these structural properties provide a detailed description of the octapeptide's complex conformational ensemble.

A representative example of how these structural properties measure convergence is shown in Figure 5 for ST at 280 K. The convergence of the α-carbon distance matrix, the hydrogen-bonding contact map, and the EED distribution are displayed in Figure 5A. The cumulative averages for the different types of turns, as well as hydrogen bonds and bends, are shown in Figure 5B. It is apparent from both of these plots that selecting a particular time at which the simulation has converged is ambiguous. Each structural property appears converged at a slightly different time. This ambiguity highlights the importance of considering multiple metrics when discussing the convergence of a simulation. In order

to define convergence quantitatively, we consider the time taken to reach the reference value of the structural property of interest and remain within both one and two standard deviations, shown in Figure 5C. Taking the average of these times provides a composite measure of when structural convergence is reached, and this average is a "structural convergence time", t_{sc}. By comparing to the reference data, both convergence and accuracy are simultaneously assessed. The time at which each structural metric reaches the reference value is significantly different. For example, the EED distribution reaches the reference distribution faster than any of the other structural metrics, while the population of α-turns requires nearly the entire ST simulation to reach the reference value to within one standard deviation. The structural convergence times are provided in Figure 5D for each of the generalized-ensemble methods at 280 K. At this temperature, STDR converges fastest to the reference data, closely followed by ST and STDRb.

For a systematic ranking of the structural convergence times, t_{sc} is also calculated for the lowest seven temperatures. These times are provided in Supporting Information Figure S2. While STDR converges faster than the other methods at 280 K, this is not a general trend for all temperatures. Each temperature has a different t_{sc} for each method. The ranking of the methods varies between temperatures. For example, at 305 K, RE reaches convergence faster than STDR. This highlights the importance of evaluating more than the lowest temperature when comparing the performance of the methods, in addition to considering several structural metrics. It also suggests a way of quantifying the error in the measurement of t_{sc}. An average structural convergence time, $\langle t_{sc} \rangle$, for each method is obtained by averaging t_{sc} for the lowest seven temperatures, for both one and two standard deviations. The error in $\langle t_{sc} \rangle$ is then the standard error of these measurements. Figure 6A shows a two-dimensional plot of $\langle t_{sc} \rangle$ to within two standard deviations versus $\langle t_{sc} \rangle$ to within one standard deviation. Lower values for $\langle t_{sc} \rangle$ indicate faster structural convergence. A clear trend emerges: ST-based methods reach structural convergence more quickly than RE-based methods. The method that reaches convergence the fastest is ST with accurate weight factors, while the method slowest to converge is SREM. It is not possible to conclusively rank the other methods due to error in $\langle t_{sc} \rangle$. However, it is important to note that both VREX and RE converge faster than SREM. VREX is therefore not only a more convenient method for removing the synchronization from the RE algorithm than SREM, but is also faster at conformational sampling.

We can now answer a key question: does faster diffusion in temperature lead to a corresponding speedup in conformational sampling? Figure 6B demonstrates that this is in fact the case. The combined average structural convergence time, obtained by taking the sum of $\langle t_{sc} \rangle$ for one and two standard deviations, is plotted versus the composite temperature diffusion score from Table 2. The ST-based methods, which have higher acceptance ratios and diffusion coefficients, also exhibit faster structural convergence. This key observation indicates that, in general, it is preferable to use a ST-based method because it provides enhanced efficiency
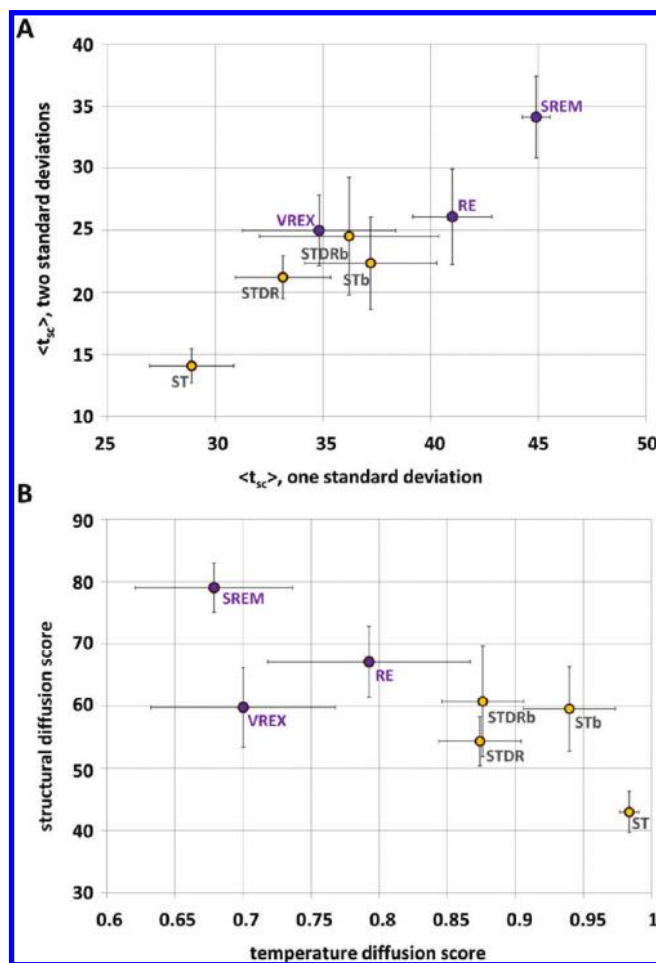

Figure 6. Correlation between structural convergence and temperature diffusion. (A) Average structural convergence times, $\langle t_{sc} \rangle$, obtained using the lowest seven temperatures are shown. The $\langle t_{sc} \rangle$ to reach two standard deviations is plotted against the $\langle t_{sc} \rangle$ to reach one standard deviation for each method. Error bars represent the standard error of $\langle t_{sc} \rangle$ for the seven temperatures. Another version of this plot is provided as Supporting Information Figure S3, with the $\langle t_{sc} \rangle$ for each temperature shown. (B) The $\langle t_{sc} \rangle$ times for one and two standard deviations from A are added together to create a structural convergence score, which is plotted against the temperature diffusion score from Table 2 for each method. A correlation is observed between structural convergence and temperature diffusion. ST-based methods (in yellow) have superior temperature diffusion, which leads to faster structural convergence compared to RE-based methods (in purple).

in terms of conformational sampling. ST with accurate weight factors is clearly superior in both temperature diffusion and structural convergence, while SREM is the least efficient method in terms of both of these metrics. In the case of a simple system for which weight factors can be obtained accurately with relatively little computational expense, ST is the method of choice. In the case of a more complex system for which sufficiently accurate weight factors might be expensive to obtain, the best choice would be to compute an initial estimate for the weight factors and use ST or STDR (corresponding to STb and STDRb here). Using the octapeptide as a test system, it is not possible to conclude which of these options is preferable. Inaccurate weight factors for this

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2655**

system yield comparable temperature diffusion and structural convergence for both STb and STDRb. To investigate this issue further, a more complex system, (GVPGV)$_7$, is also studied below.

Finally, another important question is whether inaccurate weight factors or PEDFs still lead to accurate, Boltzmann-weighted sampling at each temperature. It has been suggested that simulations with incorrect weight factors still yield correct statistics, only with suboptimal sampling of the temperature.[4] Analysis of the effect of suboptimal Helmholtz free energies on the accuracy of the data demonstrates that the resulting conformational populations are not biased by the use of inaccurate free energies (Figures 3 and 6a). Both ST and STDR with inaccurate weight factors (STb and STDRb) converge to the reference data set, which indicates that they achieve accurate, Boltzmann-weighted conformational sampling. Figure 6A shows that, even with inaccurate PEDFs, SREM still leads to Boltzmann-weighted sampling of conformational space (within one standard deviation) for this system. However, it converges more slowly than RE and all of the other generalized-ensemble algorithms considered in this study.

It has been pointed out that SREM is not rigorously correct when employing PEDFs that are not representative of the potential energies sampled during the simulation.[10,14] In other words, SREM is correct only when stationary potential energy distributions are used, which may in the general case require adaptation of the PEDFs. In contrast, VREX is inherently adaptive due to the update of the potential energy lists. All RE-based methods, including VREX and SREM, require an initial equilibration phase to reach stationary potential energy distributions. Quasi-nonergodicity due to finite run length is a potential issue for any of the generalized-ensemble algorithms but is most significant for SREM. Systematic error associated with selecting potential energies from unrepresentative conformations can be minimized in VREX by maintaining short, frequently updated potential energy lists. Reducing the size of the potential energy list in the VREX approach decreases the equilibration lag. Both VREX and RE use recent potential energy values for attempted moves. In addition, another inherent advantage of VREX over SREM is that, in VREX, potential energy values utilized in the virtual temperature transitions are drawn from the same ensemble as those utilized in the evaluation of actual transitions, albeit with a stochastic time delay.

Comparison of STDR and Conventional MD. The relative sampling enhancement of RE compared to conventional MD has been the subject of significant controversy.[20] For example, one study found that RE produced a speedup of 71.5 times at 275 K for a 21-residue helical peptide with implicit solvent, based on the autocorrelation function of helicity.[19] In another work, an RE simulation of met-enkephalin in explicit solvent sampled 5 times more conformational space than a conventional MD simulation of the same duration.[18] It has also been shown analytically that the expected speedup of RE is directly related to the activation enthalpy for two-state protein folding. The efficiency of RE is optimal when the maximum temperature is chosen just slightly above the temperature at which the folding activation

enthalpy is zero.[20] There are several key issues that emerge when reviewing comparative studies of RE and conventional MD. First, the observed sampling enhancement, or lack thereof, is heavily system-dependent, as well as dependent on the structural or thermodynamic parameter on which the comparison is based. Second, an evaluation of convergence for either the RE simulation or the MD simulation is often neglected. Comparisons of other generalized-ensemble methods, including ST and SREM, and conventional MD have also been performed.[14]

Here, we attempt to provide a rigorous comparison between STDR and conventional MD for the octapeptide (Figure 7). Figure 7A and B show a superposition of 200 structures obtained using STDR and MD, respectively, at 280 K. The amount of simulation time is the same for both methods (144 ns for conventional MD and 144 ns in total for all temperatures for STDR, corresponding to 4.4 ns at 280 K). The root-mean-square deviations (rmsd) of these two collections of structures, 3.52 Å for STDR and 3.88 Å for conventional MD, are comparable. By this measure, both STDR and conventional MD produce a similar amount of conformational sampling using the same amount of CPU time.

We also show the convergence of the structural properties described in the previous section for both STDR (in Figure 8C and E) and conventional MD (in Figure 7D and F). STDR converges more quickly, approximately by a factor of 2−3 at 280 K. However, given that STDR requires sampling 33 temperatures for the same amount of time, it is much less computationally efficient. Specifically, the STDR simulation was 4.75 μs, compared to 200 ns for conventional MD. Overall, for this particular system, there is no computational advantage in using STDR over conventional MD when the total cost of simulating all temperatures is considered.

However, in the present case, we are interested in the conformational ensemble at both low and high temperatures because of the predicted temperature transition of the octapeptide GVGVPGVG.[53,73,74] It is therefore still beneficial to use STDR because it enhances sampling at the individual temperatures. It is of key importance to note that we only know that conventional MD is able to satisfactorily reproduce the conformational ensemble of the octamer by simultaneously using generalized-ensemble algorithms. It is only by comparing to STDR, as well as the combined data set of all the generalized-ensemble methods, that we are able to verify the convergence of the conventional MD simulation. Pseudoconvergence can be observed for a structural ensemble generated by conventional MD which is energetically trapped.[13] In this way, it is possible to achieve convergence without simultaneously achieving accuracy. Using a generalized-ensemble method and allowing a random walk in temperature allows the system to overcome energetic barriers. Without knowledge of the energy landscape of the system of interest, it is hard to predict the expected sampling enhancement of a generalized-ensemble method. Similarly, it is hard to assess the accuracy of an apparently converged value, which is also expected to depend on the topology of the energy landscape.
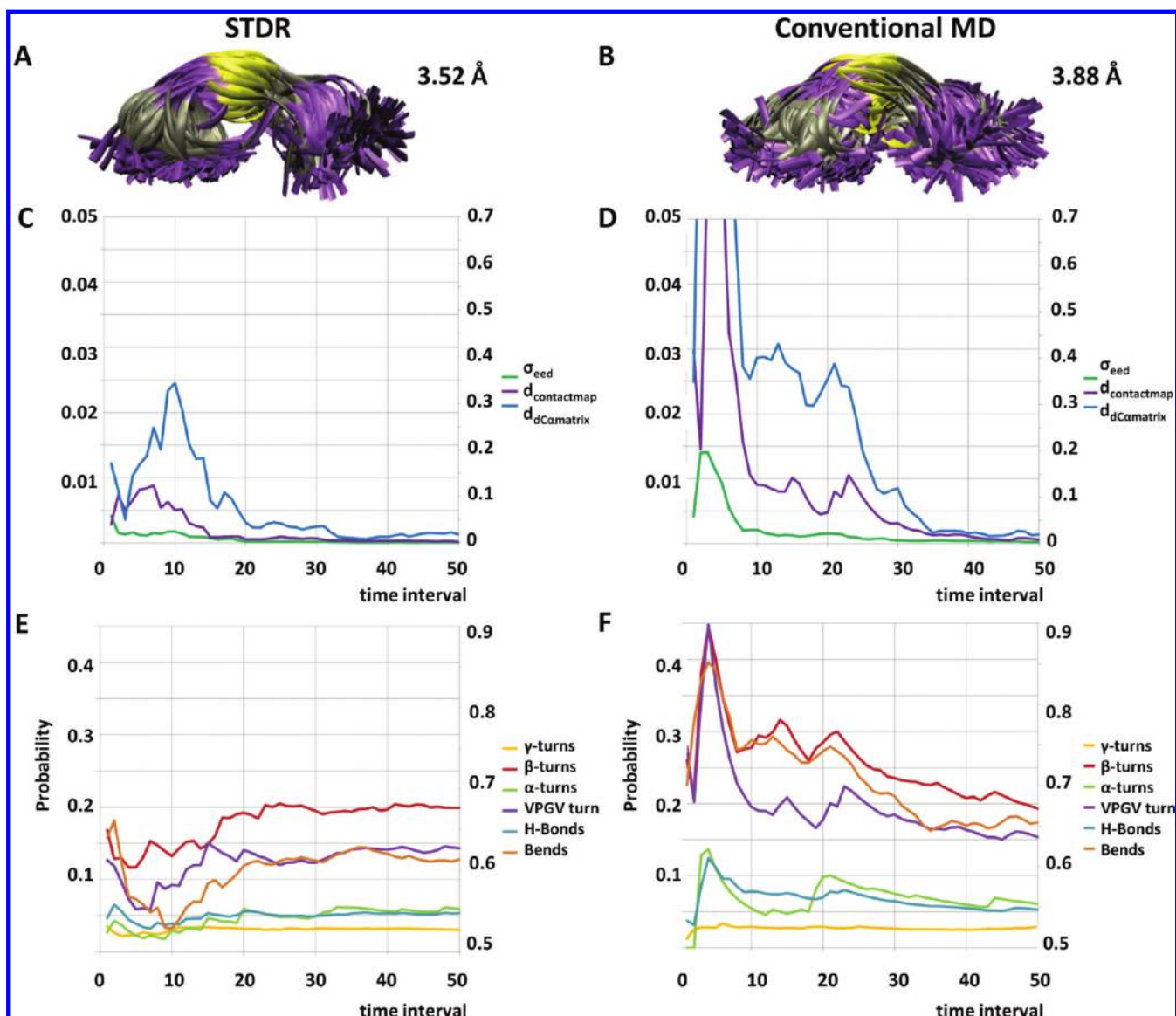
Figure 7. Comparing STDR and conventional MD for GVGVPGVG. (A) A total of 200 structures in ribbon representation obtained using the first 4.4 ns at 280 K for STDR are shown, and (B) for 144 ns of conventional MD, along with the corresponding rmsd. Glycine is in purple, proline is in yellow, and valine is in gray. C and D show σ_{eed} (eq 21), $d_{contactmap}$ (eq 23), and $d_{dC\alpha matrix}$ (eq 22, plotted on the secondary axis) for STDR and conventional MD, respectively. The trajectories are separated into 50 time intervals, and the quantities reported are calculated cumulatively, as in Figure 5. E and F show the probability per residue of a γ-turn, β-turn, and α-turn, as well as the population of the VPGV β-turn. The probabilities of observing an intramolecular hydrogen bond per residue and a bend per residue (plotted on the secondary axis) are also shown.

Choosing Between ST and STDR. For relatively small and simple systems, such as the octapeptide used in this study or a short polyalanine peptide in water,[30] the calculation of dimensionless Helmholtz free energies is possible, although computationally intensive. For these cases, ST is an ideal method, since it alleviates the need for communication between processors in parallel tempering and the subsequent waste of computational resources. However, calculation of the Helmholtz free energies increases in difficulty as system size and complexity increase. When the system is sufficiently large and complex, as is often the case for biomolecular systems of interest, limited computational resources may preclude the calculation of sufficiently accurate weight factors. That is, it is only possible to obtain dimensionless Helmholtz free energies which result in an acceptable level

of sampling uniformity with very extensive initial simulations. Even with near optimal weight factors updated throughout the simulation, Park and Pande still observed an average deviation from sampling homogeneity of 4.9% for a short peptide (calculated on the basis of the data in Table 1 of ref 30).[30] With very inaccurate weight factors, sampling of temperatures may be far from uniform. That is, there may be too little sampling at certain temperatures to obtain a reasonable estimate of the weight factors to facilitate adaptation. ST may therefore not be an appropriate method, even with adaptation of the weight factors throughout the simulation. This may be the case for many biomolecular systems of interest, which are larger than the small peptides or peptides in implicit solvent commonly used to test
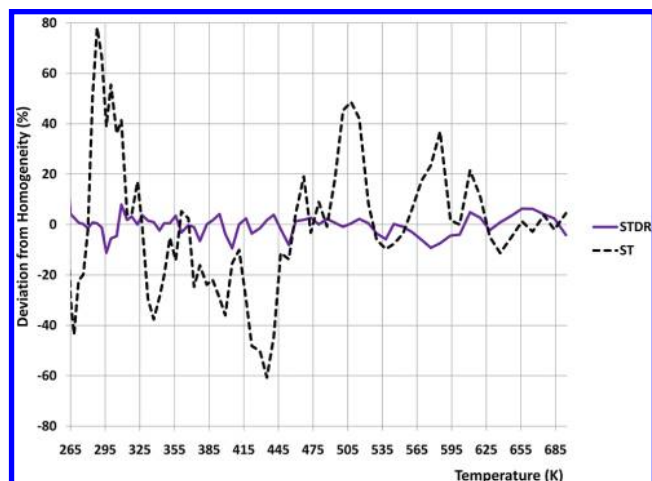
Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2657**

Figure 8. Deviation from sampling homogeneity for the ST simulation of $(GVPGV)_7$. For each temperature, the deviation from sampling homogeneity is computed as %deviation from homogeneity $= (N_m - \langle N_m \rangle)/\langle N_m \rangle \times 100\%$, where N_m is the number of samples at temperature T_m and $\langle N_m \rangle$ is the average number of samples per temperature. The ST simulation used the same weight factors as the STDR simulation, with 6 ns of sampling per replica (each started at a different temperature), for a total of 420 ns of simulation. A decrease in sampling between 410 and 460 K results in poor mobility of the replicas between high and low temperatures in the ST simulation.

generalized-ensemble methods. We now describe a complex system for which STDR is better suited than ST.

In addition to studying the octapeptide, GVGVPGVG, we also studied a longer peptide based on the same motif, $(GVPGV)_7$. Accurate weight factors for this system could not be obtained using a reasonable investment of computational resources (15 ns per temperature for 70 temperatures, for a total of 1.05 μs). Even with this large amount of data, the sampling of temperature in a ST simulation using these weight factors is heterogeneous. The average deviation from sampling homogeneity is 21.3% (computed using eq 20). In contrast, when STDR and the same weight factors are used, the average deviation from sampling homogeneity is only 3.4%. In Figure 8, the deviation from sampling homogeneity at each temperature is shown for both ST and STDR. In the ST simulation, sampling in the middle of the temperature range is less than that at both low and high temperatures, deviating from homogeneity by more than 60%. Since it is unfavorable in this case to sample intermediate temperatures, diffusion from high to low temperatures is impeded. In particular, the ST simulation using these weight factors experiences 56% fewer transitions between 417 and 454 K compared to the STDR simulation using the same weight factors. The sampling barrier in the intermediate temperature range impedes the random walk. The application of the DRPE results in a slight decrease of the acceptance ratio from 0.43 in the ST simulation to 0.38 in the STDR simulation. Importantly, the average replica speed in STDR is higher than that of the ST simulation (0.058 and 0.053 for STDR and ST, respectively). This indicates that replicas are able to efficiently explore temperature in the STDR simulation. In this case, the addition of the DRPE does not

significantly impair replica mobility and allows nearly uniform sampling of all temperatures. It is therefore a more suitable method than ST for this particular system.

We performed another comparison between ST and STDR using weight factors obtained with only 500 ps of canonical MD at each temperature. These weight factors are more inaccurate than those based on 15 ns of simulation at each temperature, leading to an average sampling inhomogeneity of 96.88% in a ST simulation. By applying the DRPE in a STDR simulation, the average sampling inhomogeneity is reduced to 8.84%. The acceptance ratio for the STDR simulation with these inaccurate weight factors is 0.38. This is exactly the same as the acceptance ratio of the STDR simulation using the weight factors based on 15 ns at each temperature. This observation is in agreement with the results of the STDR and STDRb simulations of the octapeptide. As shown in Table 2, both STDR and STDRb simulations exhibit nearly identical temperature diffusion coefficients, acceptance ratios, and mean free paths. Importantly, the results for both the octapeptide and the 35-residue peptide demonstrate that replica mobility in STDR simulations is not significantly affected by inaccuracy of the weight factors. Further, more accurate weight factors do not improve temperature diffusion, or corresponding structural convergence. Thus, we recommend that, when the STDR algorithm is used, the computational investment for the initial calculation of weight factors should be minimized.

Performance of STDR for More Complex Systems. To compare STDR with conventional MD, we also performed a MD simulation at the lowest temperature of the STDR simulation (261 K). A superposition of 200 structures, obtained every 1 ns from a 200 ns trajectory generated using conventional MD at 261 K, is shown in Figure 9A. These structures have an average rmsd of 1.66 Å, indicating that the peptide is trapped in one conformational basin and undergoes only small conformational changes. This set of structures contrasts with the set of 200 randomly selected structures from the complete STDR simulation at 261 K (Figure 9B), which represents completely different conformations with an average rmsd of 8.40 Å. For clarity, we also show six example structures in Figure 9D to demonstrate the variety of conformations observed in the STDR simulation. To make a more direct comparison between conventional MD and STDR, Figure 9C shows 200 structures from STDR using the same amount of simulation time as the conventional MD simulation (200 ns summed over all of the temperatures, corresponding to approximately 3 ns at 261 K). Finally, the radius of gyration distributions for conventional MD, STDR, and the first 3 ns of STDR are shown in Figure 9E. Both distributions from STDR show several conformational states, while the conventional MD simulation is trapped in one state. Even when using the same amount of computational resources, STDR produces a more heterogeneous ensemble of conformations.

Figure 10 displays hydrogen-bonding contact maps at 261 K for STDR, STDR with 3 ns of sampling, and conventional MD. STDR produces a conformational ensemble in which many contacts are formed with low probability. In contrast, conventional MD generates a contact map with only a few
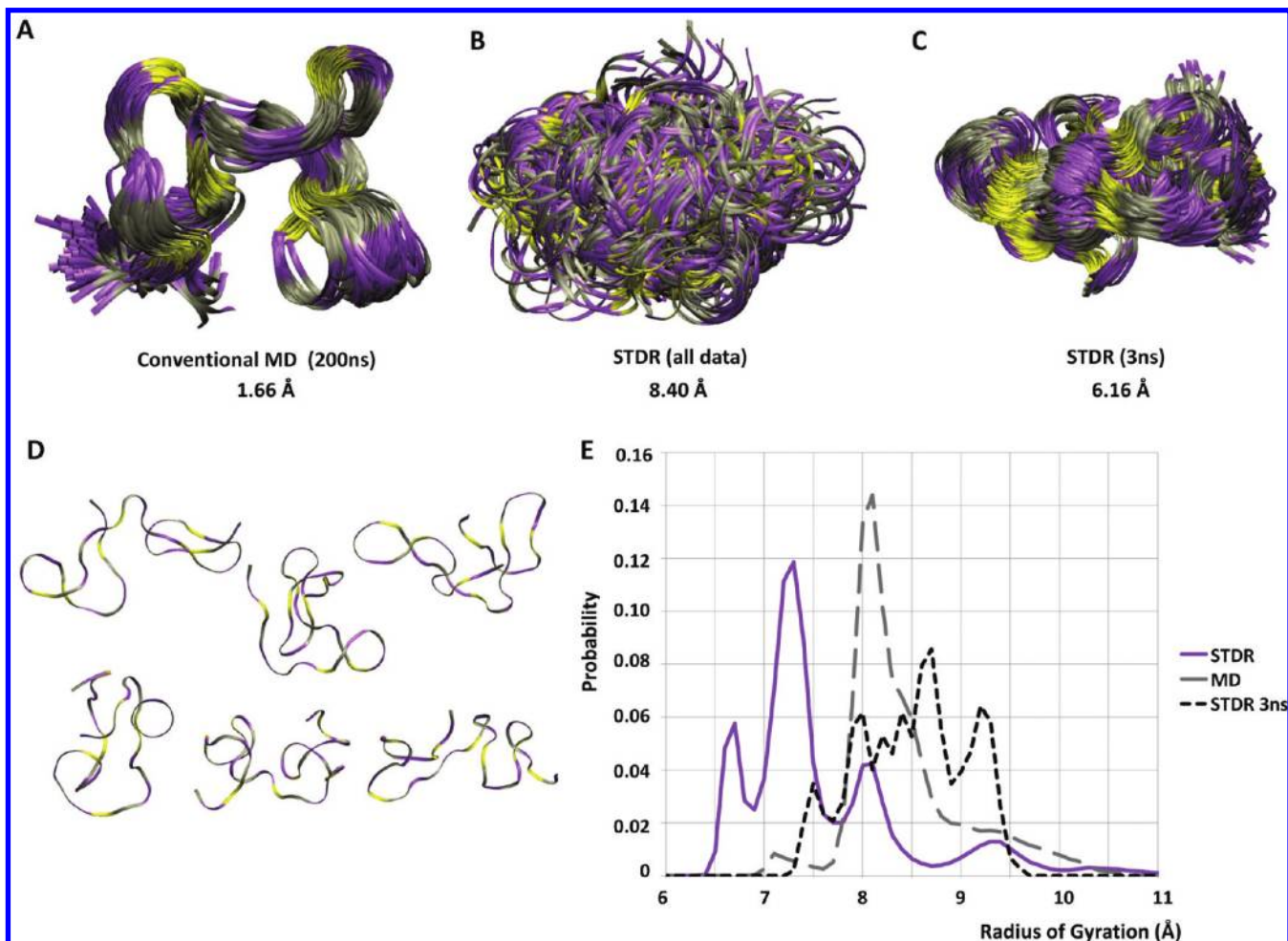
Figure 9. Comparing STDR and conventional MD for $(GVPGV)_7$. A total of 200 structures in ribbon representation along with their rmsd are shown for (A) the conventional MD simulation of length 200 ns at 261 K, (B) for the STDR simulation at 261 K using all of the data (120 ns at this temperature), (C) and for the STDR simulation at 261 K using the first 3 ns (this is the same simulation time summed over all replicas as A). Glycine is in purple, proline is in yellow, and valine is in gray. (D) A selection of six example structures is shown from the structures in B to illustrate the structural diversity obtained using STDR. (E) Probability distributions of the radius of gyration.

contacts, some of which are formed for nearly the entire simulation. The contact maps are shown with two different vertical scales to emphasize this point. If only the conventional MD simulation had been performed, a completely different understanding of the conformational landscape would have emerged. Single-temperature MD severely underestimates the heterogeneity of the conformational landscape and exhibits both pseudoconvergence and quasi-nonergodicity. Even when using the same amount of simulation time as MD, the contact map from STDR has more contacts, none of which has a probability of more than 30%.

It is not possible to quantitatively measure the speedup of STDR versus conventional MD for this system because limited computational resources preclude performing conventional MD simulations for the time required to achieve structural convergence. We observe that conventional MD is trapped in one conformational basin for 200 ns. It is not possible to accurately predict how long it would take to sample all relevant states and reach convergence. Qualitatively, we observe a dramatic sampling enhancement due to STDR. Using the same amount of computational resources, STDR generates more unique conformations for this peptide.

This indicates that the random walk in temperature does in fact lead to enhanced sampling, establishing the efficacy of the STDR method for a complex polypeptide.

Before sampling the complete energy landscape of a system of interest, there is no way to confidently predict the height of the energy barriers, or the number of energetically stable conformations (local minima of the energy surface). By coupling to simulations at higher temperatures, high energetic barriers can be overcome. However, if one is not simultaneously interested in the behavior of the system at multiple temperatures, it may be less computationally expensive to run very long simulations, or a collection of simulations, at a single temperature.[75] In the present study, conventional MD successfully produced the conformational ensemble of the octapeptide but resulted in quasi-nonergodicity for the 35-residue peptide. In order to truly "enhance sampling" relative to single-temperature MD simulations, an enhanced sampling method must achieve convergence at a rate which is greater than the product of the number of replicas and the computer time for each replica.

The present study of $(GVPGV)_7$ shows that it is possible to observe pseudoconvergence using single-temperature MD

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2659**
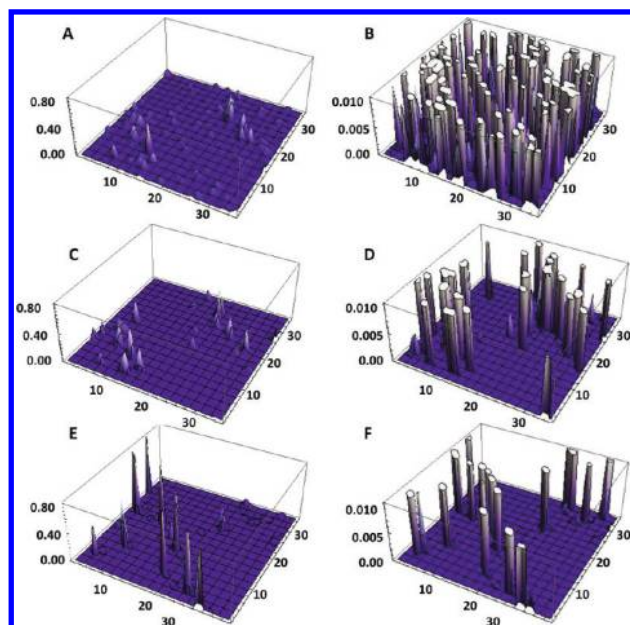

Figure 10. Hydrogen-bonding contact maps from STDR and conventional MD. Hydrogen-bonding contact maps are depicted as three-dimensional maps, where peak height represents the probability of contact formation. These plots are shown on two scales. On the left, the scale has a maximum of 0.8, and on the right, 0.01, for clarity in showing the contacts formed with low probability. A and B, the STDR simulation at 261 K using all of the data (120 ns at this temperature). C and D, the STDR simulation using the first 3 ns (this is the same simulation time summed over all replicas as E and F). E and F, 200 ns of conventional MD. Some contacts are formed over 80% of the time.

(that is, to observe convergence of a quantity of interest without observing the true value of that quantity, Boltzmann-weighted by the populations of all possible conformations). Long-time MD simulations do not yield the appropriate conformational distribution, and the system remains trapped in a local minimum of the energy landscape. In contrast, we observe that conventional MD is able to satisfactorily reproduce the conformational ensemble of the octapeptide at a significantly reduced computational cost compared to using a generalized-ensemble method. In light of this apparent contradiction, how are the averages of quantities obtained using MD simulations to be interpreted? On the basis of this work, it appears that using generalized-ensemble algorithms is a more prudent approach, even if in some cases it may be less efficient overall to do so (for increased confidence in the accuracy of the data). Several other examples have shown that the enhanced sampling provided by generalized-ensemble methods provides convergence that would not be feasible with single-temperature MD.[6,14,18−24] These observations underscore not only the need for enhanced sampling methods but also the shortcomings of techniques such as block averaging over simulations initiated in a single conformational basin in estimating the convergence of results. The challenge in simulating complex systems is that a priori one does not know the efficiency of the generalized-ensemble approach relative to the "brute force" MD approach. It may be advisable to use a general-

ized-ensemble algorithm, especially if conformational sampling, and not dynamic information, is sought.

Conclusions

We now return to the original question: given limited computational resources, which generalized-ensemble algorithm is most efficient at sampling a complex conformational landscape? The first important distinction between methods is the separation between those based on ST and those based on RE. In this paper, we demonstrate that ST-based methods result in both faster temperature diffusion and faster structural convergence. They are therefore preferable to RE-based methods. This is the most general conclusion of the paper.

Within the family of RE-based methods, the computational efficiencies of the various algorithms are not equivalent. SREM should only be applied to systems for which PEDFs can be accurately obtained. Therefore, due to limited computational resources, SREM can only be applied to simple systems. Like SREM, RE is not well-suited to complex systems because of the need to synchronize simulations of a large number of replicas (and typically, a large number of processors). Although there is no theoretical limit on the number of replicas that one can use for a RE simulation, it is generally difficult in practice to obtain access to a large, dedicated, and homogeneous computing cluster. Even if one does have access to such a computational resource, the wasted CPU time may also increase sharply with the number of replicas due to both CPU failure and inhomogeneity in CPU speeds.

VREX represents an attractive alternative to RE since it completely eliminates synchronization and communication between replicas. It produces more homogeneous sampling of temperature compared to SREM, with much less initial simulation time. It is therefore preferable to both SREM and RE. Since RE-based methods suffer from slower structural convergence and temperature diffusion compared to ST-based methods, it is preferable to use a ST-based method in temperature. This disadvantage may become less significant for complex systems. VREX may also be a more suitable method in another reaction coordinate other than temperature, for which weight factors are much more difficult to obtain. Moreover, initial simulations for weight factor calculations may also benefit from the use of VREX.

In the case of relatively simple systems for which weight factors can be accurately calculated using minimal computational resources, ST is the most appropriate method. We have shown that ST with accurate weight factors exhibits the fastest temperature diffusion and, correspondingly, the fastest structural convergence. However, for more complex systems, for which weight factors are costly to compute, STDR becomes the preferred method. Even with infinite resources, a long initial simulation to compute weight factors accurate enough to yield homogeneous sampling in ST may not be the most efficient use of computational resources. The feasibility of a ST simulation is determined by the accuracy of the weight factors, which can only be assessed by actually performing a ST simulation. Importantly, we have demonstrated that STDR can make use of inaccurate weight factors to achieve homogeneous sampling of temperature and

consequently structural convergence. Replica mobility is only slightly impeded by the DRPE. By contrast, ST with inaccurate weight factors produces heterogeneous sampling of temperatures, which is also an impediment to the random walk in temperature. STDR is suitable for any computing cluster or distributed computing environment, since it requires no fixed number of CPUs or synchronization of exchanges. Complex systems can therefore benefit from the STDR method, which is increasingly advantageous as system size and complexity grow.

Frequently Used Abbreviations

CPU, central processing unit; DR, distributed replica sampling; DRPE, distributed replica potential energy; EED, end-to-end distance; MD, molecular dynamics; RE, replica exchange; SREM, serial replica exchange method; ST, simulated tempering; STDR, simulated tempering distributed replica sampling; VREX, virtual replica exchange.

Appendix 1. Computing Error in Exchange Probabilities for ST and SREM

The PEDFs of the octapeptide are nearly perfect Gaussian distributions, as expected due to the large number of degrees of freedom of the system and the central limit theorem.[14] Assuming that the PEDFs are Gaussian is in general a valid assumption for biomolecular systems.[76] As an estimate of the error in the PEDFs, we considered the average deviation of the average energy of each PEDF, $\langle E_n \rangle$, from the average energy of a reference PEDF, $\langle E_n \rangle_{\text{reference}}$, as follows:

$$\sigma_{\text{PEDFs}} \approx \frac{1}{N_{\text{temps}}} \sum_{n=1}^{N_{\text{temps}}} |\langle E_n \rangle - \langle E_n \rangle_{\text{reference}}| \quad (24)$$

where N_{temps} is the number of temperatures. For the reference PEDFs, we used potential energy distribution functions calculated on the basis of all of the data from the RE simulation (a total of 4.75 μs for all temperatures). We computed the average error in the differences of weight factors in an analogous way, also using the RE simulation as reference data:

$$\sigma_{\text{weightfactors}} \approx \frac{1}{N_{\text{temps}}-1} \sum_{n=1}^{N_{\text{temps}}-1} |(a_{n+1} - a_n) - (a_{n+1} - a_n)_{\text{reference}}| \quad (25)$$

The selection of the RE simulation as a reference was made because it was the only generalized-ensemble method that we tested that did not make use of any initial simulation.

In order to make a fair comparison between the errors in the weight factors used in ST and the PEDFs used in SREM, it is important to consider the error in not only the potential energy distribution functions and dimensionless Helmholtz free energies but also the error in the resulting exchange probabilities. The error in the exchange probability of SREM (eq 9) was computed as follows:

$$\sigma_{P_{ij}} = \sqrt{\left[\left(\frac{\partial P_{ij}}{\partial E_{j,\text{PEDF}}}\right)\sigma_{E_{j,\text{PEDF}}}\right]^2} = e^{-(\beta_j-\beta_i)(E_i-E_{j,\text{PEDF}})}(\beta_j - \beta_i)\sigma_{E_{j,\text{PEDF}}} \quad (26)$$

We estimate this error by using the estimate for the error in the PEDFs obtained in eq 24 and the average acceptance ratio and average difference in inverse temperatures:

$$\sigma_{P_{ij},\text{estimate}} \approx \langle e^{-(\beta_j-\beta_i)(E_i-E_{j,\text{PEDF}})}\rangle\langle\beta_j - \beta_i\rangle\sigma_{\text{PEDFs}} \quad (27)$$

Similarly, the error in the exchange probability for ST (given by eq 6) is

$$\sigma_{P_{ij}} = \sqrt{\left[\left(\frac{\partial P_{ij}}{\partial(a_j - a_i)}\right)\sigma_{(a_j-a_i)}\right]^2} = e^{(\beta_j-\beta_i)E-(a_j-a_i)}\sigma_{(a_j-a_i)} \quad (28)$$

and this error is estimated using the average error in the weight factor differences from eq 25 and the average acceptance ratio:

$$\sigma_{P_{ij},\text{estimate}} \approx \langle e^{(\beta_j-\beta_i)E-(a_j-a_i)}\rangle\sigma_{\text{weightfactors}} \quad (29)$$

Supporting Information Available: Supplementary table (S1), supplementary figures (S1−S3), and an example calculation of the DRPE are provided. This material is available free of charge via the Internet at http://pubs.acs.org. Distributed Replica Sampling software is available online at www.pomeslab.com.

References

(1) Gnanakaran, S.; Nymeyer, H.; Portman, J.; Sanbonmatsu, K. Y.; García, A. E. *Curr. Opin. Struct. Biol.* **2003**, *13* (2), 168–174.

(2) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60* (2), 96–123.

(3) Marinari, E.; Parisi, G. *Europhys. Lett.* **1992**, *19* (6), 451–458.

(4) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. *J. Chem. Phys.* **1992**, *96* (3), 1776–1783.

(5) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281* (1−3), 140–150.

(6) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314* (1−2), 141–151.

(7) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65* (6), 1604–1608.

(8) Tesi, M. C.; van Rensburg, E. J. J.; Orlandini, E.; Whittington, S. G. *J. Stat. Phys.* **1996**, *82* (1−2), 155–181.

Generalized-Ensemble Methods

J. Chem. Theory Comput., Vol. 5, No. 10, 2009 **2661**

(9) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1988**, *61* (23), 2635–2638.

(10) Hagen, M.; Kim, B.; Liu, P.; Friesner, R. A.; Berne, B. J. *J. Phys. Chem. B* **2007**, *111* (6), 1416–1423.

(11) Rodinger, T.; Howell, P. L.; Pomès, R. *J. Chem. Theory Comput.* **2006**, *2* (3), 725–731.

(12) Rodinger, T.; Howell, P. L.; Pomès, R. *J. Chem. Phys.* **2008**, *129* (15), 155102.

(13) Neale, C.; Rodinger, T.; Pomès, R. *Chem. Phys. Lett.* **2008**, *460* (1−3), 375–381.

(14) Huang, X.; Bowman, G. R.; Pande, V. S. *J. Chem. Phys.* **2008**, *128* (20), 205106.

(15) Chipot, C.; Pohorille, A. *Free Energy Calculations: Theory and Applications in Chemistry and Biology*; Springer: Berlin, 2007.

(16) Bedrov, D.; Smith, G. D. *J. Chem. Phys.* **2001**, *115* (3), 1121–1124.

(17) Yamamoto, R.; Kob, W. *Phys. Rev. E* **2000**, *61* (5), 5473–5476.

(18) Sanbonmatsu, K. Y.; García, A. E. *Proteins* **2002**, *46* (2), 225–234.

(19) Zhang, W.; Wu, C.; Duan, Y. *J. Chem. Phys.* **2005**, *123* (15), 154105.

(20) Nymeyer, H. *J. Chem. Theory Comput.* **2008**, *4* (4), 626–636.

(21) Periole, X.; Mark, A. E. *J. Chem. Phys.* **2007**, *126* (1), 014903.

(22) Rao, F.; Caflisch, A. *J. Chem. Phys.* **2003**, *119* (7), 4035–4042.

(23) Rhee, Y. M.; Pande, V. S. *Biophys. J.* **2003**, *84* (2), 775–786.

(24) Tsai, H. H.; Reches, M.; Tsai, C. J.; Gunasekaran, K.; Gazit, E.; Nussinov, R. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (23), 8174–8179.

(25) Denschlag, R.; Lingenheil, M.; Tavan, P. *Chem. Phys. Lett.* **2008**, *458* (1−3), 244–248.

(26) Zuckerman, D. M.; Lyman, E. *J. Chem. Theory Comput.* **2006**, *2* (4), 1200–1202.

(27) Zuckerman, D. M.; Lyman, E. *J. Chem. Theory Comput.* **2006**, *2* (6), 1693–1693.

(28) Beck, D. A. C.; White, G. W. N.; Daggett, V. *J. Struct. Biol.* **2007**, *157* (3), 514–523.

(29) Hansmann, U. H. E.; Okamoto, Y. *Phys. Rev. E* **1996**, *54* (5), 5863–5865.

(30) Park, S.; Pande, V. S. *Phys. Rev. E* **2007**, *76* (1), 016703.

(31) Okamoto, Y. *J. Mol. Graphics Modell.* **2004**, *22*, 425–439.

(32) Zhang, C.; Ma, J. P. *J. Chem. Phys.* **2008**, *129* (13), 7.

(33) Hansmann, U. H. E.; Okamoto, Y. *J. Comput. Chem.* **1997**, *18* (7), 920–933.

(34) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13* (8), 1011–1021.

(35) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3* (1), 26–41.

(36) Mitsutake, A.; Okamoto, Y. *Chem. Phys. Lett.* **2000**, *332* (1−2), 131–138.

(37) Park, S. *Phys. Rev. E* **2008**, *77* (1), 016709.

(38) Mitsutake, A.; Okamoto, Y. *J. Chem. Phys.* **2004**, *121* (6), 2491–2504.

(39) Zhang, C.; Ma, J. P. *Phys. Rev. E* **2007**, *76* (3), 036708.

(40) Wang, F. G.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86* (10), 2050–2053.

(41) Hansmann, U. H. E.; Okamoto, Y. *J. Comput. Chem.* **1993**, *14* (11), 1333–1338.

(42) Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, *116* (20), 9058–9067.

(43) Gallicchio, E.; Levy, R. M.; Parashar, M. *J. Comput. Chem.* **2008**, *29* (5), 788–794.

(44) Luckow, A.; Jha, S.; Kim, J.; Merzky, A.; Schnor, B. *Phil. Trans. R. Soc. A* **2009**, *367* (1897), 2595–2606.

(45) Woods, C. J.; Ng, M. H.; Johnston, S.; Murdock, S. E.; Wu, B.; Tai, K.; Fangohr, H.; Jeffreys, P.; Cox, S.; Frey, J. G.; Sansom, M. S. P.; Essex, J. W. *Phil. Trans. R. Soc. A* **2005**, *363* (1833), 2017–2035.

(46) Shirts, M.; Pande, V. S. *Science* **2000**, *290* (5498), 1903–1904.

(47) Bowman, G. R.; Huang, X.; Yao, Y.; Sun, J.; Carlsson, G.; Guibas, L. J.; Pande, V. S. *J. Am. Chem. Soc.* **2008**, *130* (30), 9676–9678.

(48) Shen, H.; Czaplewski, C.; Liwo, A.; Scheraga, H. A. *J. Chem. Theory Comput.* **2008**, *4* (8), 1386–1400.

(49) Czaplewski, C.; Kalinowski, S.; Liwo, A.; Scheraga, H. A. *J. Chem. Theory Comput.* **2009**, *5* (3), 627–640.

(50) Henry, R. M.; Yu, C. H.; Rodinger, T.; Pomès, R. *J. Mol. Biol.* **2009**, *387* (5), 1165–1185.

(51) Miao, M.; Bellingham, C. M.; Stahl, R. J.; Sitarz, E. E.; Lane, C. J.; Keeley, F. W. *J. Biol. Chem.* **2003**, *278* (49), 48553–48562.

(52) Rauscher, S.; Baud, S.; Miao, M.; Keeley, F. W.; Pomès, R. *Structure* **2006**, *14* (11), 1667–1676.

(53) Baer, M.; Schreiner, E.; Kohlmeyer, A.; Rousseau, R.; Marx, D. *J. Phys. Chem. B* **2006**, *110* (8), 3576–3587.

(54) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26* (16), 1701–1718.

(55) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7* (8), 306–317.

(56) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236.

(57) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105* (28), 6474–6487.

(58) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79* (2), 926–935.

(59) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23* (3), 327–341.

(60) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092.

(61) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593.

(62) Nosé, S. *Mol. Phys.* **1984**, *52* (2), 255–268.

(63) Hoover, W. G. *Phys. Rev. A* **1985**, *31* (3), 1695–1697.

2662 *J. Chem. Theory Comput., Vol. 5, No. 10, 2009*

Rauscher et al.

(64) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21* (6), 1087–1092.

(65) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4* (3), 435–447.

(66) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. *J. Comput. Chem.* **1997**, *18* (12), 1463–1472.

(67) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52* (12), 7182–7190.

(68) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22* (12), 2577–2637.

(69) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14* (1), 33–38.

(70) Abraham, M. J.; Gready, J. E. *J. Chem. Theory Comput.* **2008**, *4* (7), 1119–1128.

(71) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (39), 13749–13754.

(72) Thirumalai, D.; Mountain, R. D.; Kirkpatrick, T. R. *Phys. Rev. A* **1989**, *39* (7), 3563–3574.

(73) Rousseau, R.; Schreiner, E.; Kohlmeyer, A.; Marx, D. *Biophys. J.* **2004**, *86* (3), 1393–1407.

(74) Schreiner, E.; Nicolini, C.; Ludolph, B.; Ravindra, R.; Otte, N.; Kohlmeyer, A.; Rousseau, R.; Winter, R.; Marx, D. *Phys. Rev. Lett.* **2004**, *92* (14), 148101.

(75) Monticelli, L.; Sorin, E. J.; Tieleman, D. P.; Pande, V. S.; Colombo, G. *J. Comput. Chem.* **2008**, *29* (11), 1740–1752.

(76) Rathore, N.; Chopra, M.; de Pablo, J. J. *J. Chem. Phys.* **2005**, *122* (2), 024111.