Simulated Tempering Distributed Replica Sampling (STDR):

An Efficient Generalized Ensemble Method for Conformational Sampling

Sarah Rauscher [1,2], Chris Neale[1,2] and Régis Pomès[1,2]

1. Molecular Structure and Function, Hospital for Sick Children, 555 University Avenue, Toronto, ON, Canada M5G1X8

2. Department of Biochemistry, University of Toronto, 1 King's College Circle, Toronto, ON, Canada, M5S 1A8

Abstract

Generalized ensemble algorithms utilizing temperature space have become popular tools to enhance conformational sampling in biomolecular simulation. 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. These methods are generally very computationally intensive, and have practical limitations. The replica exchange molecular dynamics method (REX) requires a large, dedicated and homogeneous cluster of CPUs when applied to complex systems. Simulated tempering (ST) and serial replica exchange (SREM) both successfully avoid practical limitations with regard to CPU synchronization, but have the drawback of requiring extensive initial simulations, possibly adaptive, for the calculation of weight factors or potential energy distribution functions. In this paper, we introduce two novel methods: simulated tempering distributed replica (STDR) and virtual replica exchange (VREX). These methods address the practical issues inherent in the REX, ST and SREM algorithms. We perform an objective comparison of all of these algorithms in terms of both implementation issues and sampling efficiency. We use disordered peptides in explicit water as test systems, for an accumulated 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. We find that ST-based methods exhibit faster temperature diffusion, and correspondingly faster convergence of structural properties. Based on our observations, we conclude that simulated tempering is ideal for simple systems, while STDR is well-suited for complex systems. We also compare the efficiency of STDR and conventional MD for a peptide with a very complex conformational landscape and demonstrate its dramatic sampling enhancement.

Frequently Used Abbreviations:

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

---

INTRODUCTION

Achieving complete (or even adequate) conformational sampling is one of the key challenges in biomolecular simulation.[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 high, and there are often a large number of energetically-accessible states to be sampled. The timescales of many biomolecular processes, for example protein folding, are still far beyond the reach of our current computational capability, which is limited to the nanosecond timescale. For example, the timescale of beta hairpin folding and mini-protein folding occur on the 1-10μs timescale.[1] Consequently, conventional or "brute force" molecular dynamics alone is often insufficient to achieve equilibrium sampling of the important states of many biologically relevant systems. For this reason, generalized ensemble algorithms have become popular tools for conformational sampling.

A wide 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 will be most

efficient at sampling a complex conformational 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 space to enhance conformational sampling of biomolecules. We focus on the following five methods: simulated tempering (ST)[2, 3], the serial replica exchange method (SREM)[4], replica exchange (REX)[5-9] and two novel methods: virtual replica exchange (VREX) and simulated tempering distributed replica (STDR), which is a combination of simulated tempering and distributed replica sampling[10, 11]. These generalized ensemble methods all rely on the fact that the free energy surface becomes less rugged at high temperature, increasing the frequency of interconversion between conformational states.[12] Simulations performed at low temperature often require a relatively long time to cross the energetic barriers between states (depending on the barrier heights) and appear to be trapped. Transitions between regions separated by barriers may not be observed over timescales accessible to simulation. In this case, multiple simulations initiated in different conformational basins may sample different subsets of phase space, making an ergodic system appear nonergodic, a phenomenon known as quasi-nonergodicity.[13]

The sampling enhancement of generalized ensemble methods relative to single temperature molecular dynamics or Monte Carlo simulations has been demonstrated for several systems,[2, 7, 14, 15] including peptides.[6, 12, 16-22] There have also been studies which question the relative sampling efficiency of replica exchange compared to brute force molecular dynamics,[23] highlighting the importance of a rigorous definition of efficiency which accounts for the total computer time required for all temperatures.[24-26] 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 and unfolding. In general, however, the data at high temperature is not useful. Furthermore, the observed speed-up will also strongly depend on the lowest temperature.[24] It is essential to assess the convergence of both the

conventional MD simulation as well as the generalized ensemble simulation in order to perform a meaningful comparison (this crucial step is often missing[22, 26]), in addition to identifying a meaningful quantity on which to base the comparison. Any evaluation of sampling enhancement compared to MD will also be heavily system dependent (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 space. In addition to providing a comparison between generalized ensemble algorithms, we will also be providing a comparison to conventional MD.

In this paper, we begin with a brief introduction of each of the generalized ensemble methods, including the presentation of the novel methods, STDR and VREX. We then perform an objective comparison of the algorithms in terms of both practical implementation limitations and sampling efficiency. We discuss efficiency in terms of both convergence of structural properties and diffusion in temperature space, and we show that these definitions of efficiency are in fact correlated. We also compare the efficiency of STDR and conventional MD for a peptide with a very complex conformational landscape.

THEORY AND METHODS

Simulated Tempering

Simulated tempering was originally introduced to enhance sampling of a Random Field Ising Model. 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. Simulated tempering allows exchanges between these states, whereas the Monte Carlo algorithm remains trapped.[2] ST has also been shown to be effective in exploring the energy landscapes of biomolecules, which similarly have multiple energy minima separated by barriers.[27]

In the simulated tempering algorithm, temperature becomes a dynamic variable[2, 3] 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:

$$\mathbf{H}\left(X,m\right)=\beta_m H\left(X\right)+a_m \qquad \textbf{(1)}$$

where β_m is the inverse temperature, *H*(X) is the system's original Hamiltonian and a_m is a constant which depends on temperature.[2] The generalized ensemble has a corresponding generalized partition function, Z, given by:

$$\mathbf{Z}=\sum_m \int dX\left[e^{-\mathbf{H}(X,m)}\right]=\sum_m \int dX\left[e^{-\beta_m H(X)+a_m}\right]=\sum_m Z_m e^{a_m} \qquad \textbf{(2)}$$

where Z_m is the partition function corresponding to the temperature T_m.[28] The partition function of the generalized ensemble, 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".[28] The probability of sampling a certain temperature, T_m, is:

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

which depends on the generalized Hamiltonian, H, and therefore depends on the weight factor, a_m. The goal in simulated tempering 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*):

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

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)[28, 29]:

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

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.[30] However, inaccuracy in the weight factors leads to corresponding differences in the probabilities of sampling at each temperature.[3] In practice, the quality of the weight factors (that is, how close they are to accurate Helmholtz free energies) can be assessed by observing the deviation from sampling all temperatures homogeneously.

A simulated tempering simulation consists of a short molecular dynamics (or Monte Carlo) simulation in a canonical ensemble at temperature T_i followed by an exchange attempt to a neighbouring temperature, T_j. The probability of this exchange occurring is given by:

$$p\left(T_i \to T_j\right) = \min\left\{1, e^{-\left(\beta_j-\beta_i\right)E+\left(a_j-a_i\right)}\right\} \quad \textbf{(6)}$$

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.[31] The weight factors need only be accurate up to an additive constant, since only differences in weight factors are needed to determine the acceptance probability.[32] Through many repetitions of these alternating simulation and exchange steps, a random walk in temperature space is realized, corresponding to a random walk in potential energy and efficient exploration of the energy landscape.[29] In fact, simulated tempering has been shown to be as effective as the multicanonical algorithm (MUCA), which analogously employs a random walk in potential energy.[33]

The underlying challenge in the simulated tempering method 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 estimate of the weight

factors based on average energies.[28, 32] The average potential energy at each temperature, E, is obtained from initial simulations, and the differences in weight factors are calculated as follows:

$$a_{i+1} - a_i \approx \left(\beta_{i+1} - \beta_i\right)\left(\frac{E_i + E_{i+1}}{2}\right) \quad \textbf{(7)}$$

The weight factor for the lowest temperature can be set to zero (since only differences in weight factors are needed in the acceptance ratio, equation 7). These weight factors may be updated throughout the ST simulation if required.[12] In practice, calculating the dimensionless Helmholtz free energies for a complex system such as a peptide in explicit water is very computationally expensive since it requires an accurate estimate of the partition function. These calculations can require tens of nanoseconds per temperature or more (with the computational expense increasing with both system size and complexity).[12]

Replica Exchange (REX)

The replica exchange method 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 simulated tempering, and it is also known as parallel tempering[5] or multiple Markov chains.[8] In fact, parallel tempering was applied to proteins even before simulated tempering.[37] A replica exchange simulation consists of M identical copies of the system (replicas) which sample M canonical ensembles with temperatures, T_m. Exchange moves are performed between neighbouring temperatures *i* and *j*, where the probability of making this exchange depends on the potential energies, E_i and E_j, and the inverse temperatures, β_i and β_j:

$$P\left(i \leftrightarrow j\right) = \min\left\{1, e^{-\left(\beta_j - \beta_i\right)\left(E_i - E_j\right)}\right\} \quad \textbf{(8)}$$

Replica exchange is analogous to simulated tempering, but instead of weight factors in the exchange probability, the upward move of one replica is coupled to the downward move of another. Replica exchange therefore has the critical advantage of not requiring any initial simulation for the calculation of weight factors.

The main drawback of the replica exchange method is its significant computational requirements. The one-to-one correspondence between the number of replicas (M) and the number of temperatures (M) makes it necessary to use M CPUs simultaneously throughout the course of the simulation. The number of replicas needed for a replica exchange simulation is related to the number of degrees of freedom, N, as $O(N^{1/2})$. [7, 31, 38] Systems with many particles require many replicas (that is to say, a large number of simultaneously available CPUs). For all but the simplest systems, the requirement of a large and dedicated computing cluster may therefore be too prohibitive to attempt these simulations.

Inherent in the replica exchange algorithm is a synchronization of attempted moves which wastes CPU time as replicas wait to perform exchanges. Inhomogeneity of CPU speeds affects the amount of wasted time, since the speed of the calculation will actually depend on the speed of the slowest processor. Modifications to replica exchange have been developed in an effort to minimize wasted CPU time. However, both the multiplexed replica exchange method (MREM)[21] and asynchronous replica exchange[39] do not completely alleviate the need for synchronization and frequent communication between replicas.[12] This is especially important to users of distributed computing, such as the massively parallel Folding@Home project,[40] who must contend with inhomogeneity of processor speeds.[12] These methods also do not directly address the issue of the number of cores needed, which grows quickly with both system size and complexity. In particular, MREM makes use of multiplexed layers of replicas (n layers, each with M temperatures), with exchanges occurring both within and between layers.[21] Thus, MREM is even more computationally demanding, requiring n times as many

processors as REX. It is therefore not useful when there are too few 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.[39] However, the implementation of asynchronous REX requires a complex communication framework. The algorithm is also not suited to run on an arbitrary number of processors since an exchange can only occur with a certain temperature if a replica is currently there.

Serial Replica Exchange (SREM)

The serial replica exchange method (SREM)[4] was recently developed to address the main practical limitations inherent in the replica exchange method: synchronization and the large number of processors required. The exchange probability in SREM has an identical form to that of REX (equation 8) for a replica at temperature T_i attempting to move to a neighbouring temperature T_j:

$$P(i \to j) = \min \begin{cases} 1 \\ e^{-(\beta_j - \beta_i)(E_i - E_{j,PEDF})} \end{cases} \quad \textbf{(9)}$$

Unlike REX, an attempted move only involves a replica going from *i* to *j*, and no corresponding reverse move. In SREM, the potential energy, E_j, 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 conventional MD or replica exchange. 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 were required.[41] PEDFs may also need to be updated throughout the course of the SREM simulation.[12, 41] In the calculation of PEDFs, thousands to millions of energy values must be stored. This can be an especially significant drawback when dealing with complex systems, or, for example, temperature-dependent force fields.[42]

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 can only be considered to be approximately correct (in terms of obeying detailed balance) if unconverged or incorrect PEDFs are used. In contrast, the weight factors of simulated tempering can deviate from the accurate dimensionless Helmholtz free energies and still yield correct (Boltzmann-weighted) results.[2, 4, 12]

Virtual Replica Exchange (VREX)

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

$$P(i \to j) = \min\left\{ 1, e^{-(\beta_j - \beta_i)(E_i - E_{j,virtual})} \right\} \quad \textbf{(10)}$$

Here, the potential energy, $E_{j,virtual}$ for temperature T_j comes from a list of stored energy values obtained at that temperature. This is analogous to exchanging with a potential energy value from a potential energy distribution function in SREM, or the current potential energy at temperature T_j in replica exchange. Like SREM, there is only a move from temperature T_i to temperature T_j, and no reverse move. In VREX, an energy value that occurred at temperature T_j in the past is used, and following the

attempted exchange, this value is removed from the potential energy list. This constitutes a "virtual exchange".

In practice, a virtual replica exchange simulation 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 molecular dynamics simulation between exchange attempts, and values being removed as they are used in "virtual exchanges". The main advantage of VREX is that it avoids the need to calculate converged potential energy distribution functions (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 replica exchange 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 replica exchange, with the addition of a time delay between when a potential energy is produced and when it is used for an exchange.

Simulated Tempering Distributed Replica (STDR)

The second novel method in this paper is simulated tempering distributed replica (STDR). Building on the success of the simulated tempering method, we have developed a new algorithm which combines ST[2, 3] with distributed replica sampling (DR).[10, 11, 43] The combination of these two methods was originally suggested when DR was developed.[10] Distributed replica sampling, developed prior to both SREM[4] and asynchronous replica exchange[39], was designed specifically to suit shared or distributed computing platforms.[10] In the DR method, synchronization of exchange attempts is completely eliminated as replicas undergo stochastic moves independent of each other. This results in 100% CPU utilization,[10] like both ST and SREM. The algorithm also readily accommodates fluctuations in CPU availability.[10] DR was shown to sample conformational space more effectively than thermodynamic integration in the calculation of the binding free energy of benzene to T4 lysozyme, while

simultaneously optimizing the use of available computational resources.[11] In addition, DR has been combined with umbrella sampling (DRUS) to allow equilibrium exchange between different umbrella biasing potentials. When applied to alanine dipeptide, umbrella sampling alone exhibited quasi-nonergodic behaviour, while DRUS alleviated this systematic error.[43]

We briefly summarize the distributed replica algorithm, the details of which have been described previously.[10, 11] The goal of DR is to enforce uniform sampling along a reaction coordinate of interest. This may be, for example, the fourth dimension[11] or a dihedral angle[43]. In the case of STDR, we enforce homogeneous sampling of a set of temperatures, labeled by an index m=1,…,M. The probability of accepting a move from a temperature T_i to a neighbouring temperature, T_j is:

$$p\left(T_i \to T_j\right) = \min\left\{1, e^{-\left(\beta_j-\beta_i\right)E+\left(a_j-a_i\right)-\left(DRPE_j-DRPE_i\right)}\right\} \quad \textbf{(11)}$$

This is the same as the acceptance probability from simulated tempering, with the addition of the difference in distributed replica potential energy (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. The equation for the DRPE depends upon the current temperatures of all of the replicas as follows:

$$DRPE = c_1\sum_{m=1}^{M}\sum_{n=1}^{M}\left[\left(\lambda_{m,linear}-\lambda_{n,linear}\right)-\left(m-n\right)\right]^2 + c_2\left[\sum_{m=1}^{M}\lambda_{m,linear}-\sum_{m=1}^{M}m\right]^2 \quad \textbf{(12)}$$

The values of $\lambda_{m,\,linear}$ refer to a linearly-spaced temperature coordinate (for example, the lowest temperature has $\lambda_{1,\,linear}=1$, and the highest temperature has $\lambda_{M,\,linear}=M$ in this coordinate). This represents transforming the exponentially spaced temperatures into a uniformly spaced coordinate. The first term 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 towards high or low temperature. The constants c_1 and c_2 control the influence of the DRPE and can be tuned to enforce homogeneous sampling of temperatures as required.[10] Importantly, although the DRPE is an energy penalty, it is not

generally a function of system complexity[10] and its influence is completely controlled by the constants, c_1 and c_2. These constants are selected to enforce homogeneous sampling. In the case of very 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 will be required to maintain homogeneous sampling of temperature, and this will reduce the exchange probability to some degree. An example calculation of the DRPE using temperature as the reaction coordinate is provided as supplementary material.

If the weight factors, a_m, are inaccurate, simulated tempering will result in uneven sampling of the temperature coordinate. Introducing the DRPE recovers homogeneous sampling, as we will demonstrate. The STDR method is therefore more generally applicable than simulated tempering because it can make use of a poor estimate of the dimensionless Helmholtz free energies and still obtain uniform sampling of the canonical ensemble at each temperature. In this paper, we show that this is the preferred method for systems with a complex energy landscape for which limitations on computational resources would preclude obtaining sufficiently accurate estimates of Helmholtz free energies and therefore render ST practically difficult or impossible.

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.[44] 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 secondary structure in the form of α-helices or β-sheets).[45] Understanding the structural heterogeneity of these peptides will elucidate the structure-function relationship of elastin (for which experimental characterization is notoriously difficult due to its flexibility and insolubility). The

peptide GVGVPGVG has been studied previously, and was suggested to exhibit an 'inverse temperature transition' with an increased probability of closed conformations at higher temperatures.[46] Based on this work, the octamer seemed to be a simple yet appropriate peptide to study in the aim of understanding the temperature-dependent behaviour of elastin. Although we do not elaborate fully on the structural details of either the octamer or the 35-mer in this paper, a full characterization of the conformational landscape of these peptides will be the subject of future work. The main focus of this paper is the thorough comparison of generalized ensemble methods using these peptides as test systems. Both GVGVPGVG and $(GVPGV)_7$ are ideal test systems because of their complexity and the fact that they represent a real scientific problem in the sense that they are not well understood or characterized *a priori*. It is often the case that an overly simple test system is used for this purpose (such as alanine dipeptide[4, 43]), whereas it is expected that generalized ensemble methods will be applied to systems which are much larger and more complex. While simple test systems are useful 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 octamer 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 C and N termini are in close proximity, "hairpin"-like states, and extended structures. In figure 1B, we show the hydrogen-bonding contact map for this peptide obtained using STDR. We observe that there is only local secondary structure in the form of hydrogen-bonded turns with no α-helix or β-sheet. The most populated turn is the VPGV turn, with a hydrogen bond between the C=O group of residue 4 (valine) and the N-H group of residue 7 (valine). Several other turns form with lower population. As we will show, conventional molecular dynamics, if run for 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 assessing their relative efficiency.

The 35-mer, $(GVPGV)_7$, 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 with one generalized ensemble method because of the extensive amount of computational resources required. Of the methods we consider, STDR is the best suited to this particular application based on its performance for the octamer. It is as efficient and accurate as the other methods, while offering the most practical advantages for a large and complex system. In particular, the calculation of accurate weight factors for this system is not possible given computational resources, and the use of distributed replica sampling is necessary.

Simulation Details

For all methods (ST, STDR, SREM, REX and VREX), the same exponentially spaced temperature list was used and is provided as supplementary table 1. The simulation system consisted of the octapeptide, capped with an acetyl group at the N-terminus and an NH_2 group at the C-terminus, in a 3 x 3 x 3 nm 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,[47, 48] with the OPLS-AA/L force field[49, 50] and the TIP3P model for water.[51] Periodic boundary conditions were applied and a 1.4 nm cutoff was used for Lennard-Jones interactions. Covalent bonds involving hydrogen atoms were constrained with the SHAKE algorithm.[52] Long-range electrostatics interactions were handled using the Particle Mesh Ewald (PME) summation method[53, 54] with a Fourier spacing of 0.15 nm and a fourth-order interpolation. All 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.[55, 56] 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 (an average of approximately 144 ns per replica). This time was used because it was sufficient for all methods to achieve convergence, as demonstrated in the results. Exchanges were attempted every 25 ps, using the probabilities expressed in equations 5, 7, 8, 9 and 10, as appropriate for the method. Details of the calculation of weight factors and PEDFs are discussed in detail below. The constants c_1 and c_2 for the distributed replica potential energy in equation 11 were both 0.005. These values were found to achieve an appropriate balance between homogeneity of temperature sampling and replica mobility.[10]

The same simulation protocols were used for the simulation of the 35-mer, which was simulated in a 4.5 x 4.5 x 4.5 nm box with 2856 water molecules. Starting conformations and weight factors for each temperature were generated using conventional MD (15 ns per temperature). There were 70 temperatures used for this simulation. This list of temperatures is provided in supplementary table 1. Temperatures were spaced more closely than those of the octamer. This is because it is a much larger system, resulting in narrower potential energy distribution functions and therefore less overlap between distributions of adjacent temperatures (for the same temperature separation). This system was simulated for a total of 8.2 μs (on average 117.6 ns per replica). A conventional MD simulation of the 35-mer in the isothermal-isobaric ensemble was also performed using GROMACS version 4.0.2.[57] In this simulation a 4 fs time step was used, and constraints on bonds and angles involving hydrogen were imposed using the LINCS algorithm[58] (this was necessary to speed up the calculation for the purpose of obtaining a long trajectory). This simulation was run for 200 ns at 261 K (which corresponded to the lowest temperature in the STDR simulation).

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).[59] 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 H-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.[59] End-to-end distance is calculated as the distance between the alpha carbons of the first and last residue. Root mean square deviation (RMSD) was calculated using the g_rms program in GROMACS.[57]. All molecular visualizations in the manuscript were created using VMD.[60]

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 replica exchange algorithm specifies that the number of replicas equal the number of temperatures, and this requirement grows with system size. In general, the number of processors equals the number of replicas. If it is not possible to obtain access to the required number of processors, the replica exchange simulation simply cannot be done and an alternative method or a more advanced REX implementation must be sought. Another possible scenario is that there are extra processors available which could be utilized to speed up the calculation, but the

replica exchange algorithm does not allow the possibility of having more replicas than temperatures. In both cases, there is no general mechanism to adapt replica exchange to most efficiently use available resources. In contrast, both ST and SREM algorithms completely eliminate the need for a specific number of replicas and can therefore utilize only one processor. 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 CPUs. However, the aim of the distributed replica potential energy is to enforce homogeneous sampling of temperatures for multiple replicas. Using only one replica is therefore not optimal, and ideally one would want to have a comparable number of processors to the number of temperatures, though there is no specific requirement. Virtual replica exchange could theoretically be run on any number of processors. However, there is likely some benefit to having multiple replicas sampling different regions of the 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 replica exchange restricts the number of replicas from fluctuating during the course of the simulation. This is a major 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 REX is severely affected by inhomogeneity of CPU speeds. Each exchange step can only occur when all the replicas have completed their MD calculation. Any inhomogeneity in the computing cluster will result in a waste of computational resources as some replicas must wait for the replica with the slowest processor to finish its calculation. Since none of the other methods require any direct communication between replicas, they do not suffer from this inefficiency. Another key drawback of replica exchange is its sensitivity to CPU failure.[4, 10] If any one of the replicas is running on a processor that crashes, the entire replica exchange 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 REX will be the number of replicas times the failure rate of either SREM or ST.[4]

In contrast to replica exchange, ST, STDR, SREM and VREX all have the advantage of not requiring a fixed and synchronized cluster of CPUs. From a practical point of view, these methods are all superior to REX, except in one regard. They require initial simulations at multiple temperatures to obtain weight factors, potential energy distribution functions, or potential energy lists. In particular, ST and SREM appear to only be applicable to systems for which accurate weight factors or PEDFs can be calculated in a reasonable amount of simulation time. We will demonstrate that STDR can function with less accurate weight factors, and therefore requires less initial simulation time than ST. Similarly, VREX requires significantly less initial simulation than SREM, STDR or ST. Only short lists of potential energies at each temperature are needed to begin the VREX simulation.

An ideal temperature-based generalized ensemble method would not require a significant initial simulation (like ST and SREM), but also would not require a constant and large cluster of homogeneous CPUs (like REX). STDR and VREX address both of these issues, and are the most flexible algorithms in terms of practical concerns. These issues become 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.

Calculations of weight factors for ST and STDR, PEDFs for SREM and lists for VREX

The calculation of weight factors required initial simulations in the canonical ensemble for each of the temperatures listed in supplementary material table 1. 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 very straightforward procedure. The weight factors were computed using the average potential energy at each temperature using the method outlined in equation 7.[28] The accuracy of these weight factors was assessed by using them in a simulated tempering simulation and observing the temperature sampling uniformity (sampling uniformity will be shown in detail below). Since all temperatures were sampled with nearly equal probability, as expected from equations 3 and 5 for accurate dimensionless Helmholtz free energies, these weight factors were shown to be sufficiently converged and correct.

Using the same data from the conventional MD simulations, potential energy distribution functions were computed as described in the original SREM paper.[4] Their convergence was assessed by calculating the χ^2 measure suggested by Hagen *et al*:[4, 12]

$$\chi^2(t) = \sum_{n=1}^{Nbins} \left(P_i(t) - P_{i,reference} \right)^2 \qquad \textbf{(13)}$$

This measure computes the deviation of each bin in the current distribution, $P_i(t)$, from a reference distribution, $P_{i,reference}$. 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 non-uniform sampling of temperature space. We therefore proceeded to calculate the PEDFs using a different set of data. We used the first 25 ns at each temperature of the replica exchange 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 was used in the initial simulation. We were motivated to do this by the replica exchange simulated tempering method (REST). In this method, an initial replica exchange simulation is

run for the purpose of obtaining accurate weight factors, which are then used for a simulated tempering simulation.[36, 61] 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 replica exchange simulation. Importantly, we did not use REST to obtain the weight factors for ST to better represent the general case where it may not be possible to do so. In contrast, it was necessary to use replica exchange to obtain PEDFs for SREM in a reasonable amount of time.

The PEDFs of this system are nearly perfect Gaussian distributions, as expected due to the large number of degrees of freedom of the system, and the central limit theorem.[12] Assuming that the PEDFs are Gaussian is in general a valid assumption for biomolecular systems.[62] As an estimate of the error in the PEDFs, we considered the average deviation of the average energy of each PEDF, $\langle U_n \rangle$, from the average energy of a reference PEDF, $\langle U_n \rangle_{reference}$, as follows:

$$\sigma_{PEDFs} \approx \frac{1}{N_{temps}} \sum_{n=1}^{N_{temps}} \left| \langle U_n \rangle - \langle U_n \rangle_{reference} \right| \quad \textbf{(14)}$$

where N_{temps} is the number of temperatures. For the reference PEDFs, we used potential energy distribution functions calculated based on all of the data from the replica exchange 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 replica exchange simulation as reference data:

$$\sigma_{weight\ factors} \approx \frac{1}{N_{temps}-1} \sum_{n=1}^{N_{temps}-1} \left| (a_{n+1} - a_n) - (a_{n+1} - a_n)_{reference} \right| \quad \textbf{(15)}$$

The selection of the replica exchange 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 (equation 9) was computed as follows:

$$\sigma_{P_{ij}} = \sqrt{\left(\left(\frac{\partial P_{ij}}{\partial E_{j,PEDF}}\right)\sigma_{E_{j,PEDF}}\right)^2} = e^{-\left(\beta_j-\beta_i\right)\left(E_i-E_{j,PEDF}\right)}\left(\beta_j-\beta_i\right)\sigma_{E_{j,PEDF}} \quad \textbf{(16)}$$

We estimate this error by using the estimate for the error in the PEDFs obtained in equation 14, and the average acceptance ratio and average difference in inverse temperatures:

$$\sigma_{P_{ij},estimate} \approx \left\langle e^{-\left(\beta_j-\beta_i\right)\left(E_i-E_{j,PEDF}\right)}\right\rangle\left\langle\beta_j-\beta_i\right\rangle\sigma_{PEDFs} \quad \textbf{(17)}$$

Similarly, the error in the exchange probability for simulated tempering (given by equation 6) is:

$$\sigma_{P_{ij}} = \sqrt{\left(\left(\frac{\partial P_{ij}}{\partial\left(a_j-a_i\right)}\right)\sigma_{\left(a_j-a_i\right)}\right)^2} = e^{\left(\beta_j-\beta_i\right)E-\left(a_j-a_i\right)}\sigma_{\left(a_j-a_i\right)} \quad \textbf{(18)}$$

and this error is estimated using the average error in the weight factor differences from equation 15 and the average acceptance ratio:

$$\sigma_{P_{ij},estimate} \approx \left\langle e^{\left(\beta_j-\beta_i\right)E-\left(a_j-a_i\right)}\right\rangle\sigma_{weight\ \ factors} \quad \textbf{(19)}$$

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 weight factors of ST produce an average 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%). This explains why the weight factors 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 replica exchange. This set of data was used to calculate the PEDFs used in the SREM simulation, with an error in the exchange probability of less than 4%. The convergence of the PEDFs estimated using all of the data from replica exchange in figure 2d, it can be seen that the

error in the acceptance ratio had only decreased to less than 2% after approximately 60 ns/temperature. That is, SREM would have required preliminary simulations which were half as computationally expensive as the entire REX simulation in order to produce sampling homogeneity equivalent to that of simulated tempering. The slow convergence of PEDFs is likely why they have been updated throughout the course of the simulation in other studies.[4, 12, 41] However, an SREM simulation is strictly only correct with accurate PEDFs.[12]

It is clear from figure 2 that the error in the weight factors of ST leads to smaller error in the acceptance ratio than the error in the average energy of PEDFs. This is in qualitative agreement with a previous study comparing SREM and ST for a helical peptide. The PEDFs of SREM were observed to converge more slowly than the weight factors of ST when starting from a coil conformation. However, when both ST and SREM are started with a helical conformation, the PEDFs and weight factors were observed to converge after a similar amount of time.[12] 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 (since they are free energies).[4] In fact, we have shown for this system that the exact opposite is true. We have observed that weight factors converge significantly faster than PEDFs and lead to more homogeneous sampling of temperature space. The difference in errors is likely because the acceptance ratio in ST uses a difference in dimensionless free energies, whereas the absolute value of the potential energy is used in SREM's acceptance ratio. There is additional error in SREM, which we have not accounted for, due to using discrete distributions, as well as error in the variance of the distributions. Accuracy is affected by the choice of the number of bins and the bin width.[4] The accuracy of the value chosen from the distribution is decreased by having too few bins, whereas the convergence of the distribution is slower with a larger number of bins. These errors must then be balanced. Even if the PEDFs and weight factors had converged at the same rate, simulated tempering would have the

advantage of convenience (consider storing a short list of weight factors versus a distribution of energy values for each temperature).

We have also tested the effects of using a poor estimate of the weight factors in simulated tempering. In order to generate suboptimal weight factors, we used only the data from the first 750 ps of the replica exchange simulation. This required a total of 24.75 ns summed over all temperatures, compared to 643.5 ns needed to generate accurate weight factors. These weight factors produced inhomogeneous sampling of temperature space, confirming that they were inaccurate estimates of the dimensionless Helmholtz free energies (this will be demonstrated below). The purpose of this exercise was to emulate the more general case of a very complex system for which one may not be able to accurately calculate weight factors due to the prohibitive computational cost. Simulated tempering and STDR 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 replica exchange data. A list of 1000 energy values from the first 1ns was used for each temperature. Practically, one must make sure not to run out of potential energy values. One way to address this is to create a secondary list of all energy values that have been used for each temperature, and select from these lists in the rare case that the primary list has been completely used. We did not encounter this issue in our VREX simulation, but it is something to consider in the general implementation and 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: REX (0 ns), VREX (33 ns), SREM (825 ns), ST (643.5 ns), STDRb (24.75 ns) and STb (24.75 ns).

Methods Comparison: Characterizing Diffusion in Temperature Space

We now 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 replica exchange simulations.[63] The methods separate into two categories based on their acceptance ratios: the REX-based methods (REX, VREX and SREM), and the ST-based methods (ST, STb, STDR and STDRb). Simulated tempering has a higher acceptance ratio than replica exchange 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.[12] Zhang and Ma also observed that the rate of traversing temperatures is faster in simulated tempering, and that this becomes especially apparent if separations between adjacent temperatures are large, or if exchanges are attempted less frequently.[30] Park proved that this is generally true for a given set of temperatures[32] 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 parallel tempering (REX, VREX and SREM) do in fact lead to faster structural convergence later in the paper. 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 equation 12.[10]

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. In order to account for these unproductive moves, we calculate the replica speed as the average distance travelled after 50 exchanges, and these values 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 faster replica speeds.

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 slope of the mean squared deviation of distance versus time plot. We notice that simulated tempering with both 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 space compared to simulated tempering. Replica exchange 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 criteria is the deviation from sampling homogeneity, which indicates the amount of deviation from uniform sampling averaged over all the temperatures:

$$average\ \ deviation\ \ from\ \ homogeneity = \frac{1}{M}\sum_{m=1}^{M}\frac{\left|N_m - \langle N_m \rangle\right|}{\langle N_m \rangle} \qquad \textbf{(20)}$$

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. As supplementary information figure S1, we report the deviation from sampling homogeneity for each temperature. The coupling of upward and downward moves in the REX algorithm results in perfectly uniform sampling of all temperatures, and an average deviation of 0%. 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.[10] Even with inaccurate weight factors, the sampling of temperature is still uniform, and the diffusion coefficient is still approximately the same. This indicates that STDR in the general case with inaccurate weight factors (as one might obtain in a more complex system) still successfully produces uniform sampling and good mobility in temperature space. Simulated tempering with accurate weight factors also results in nearly uniform sampling, with an average deviation of only 3.81%, confirming the accuracy of the weight factors (based on equations 3 and 5). VREX also produces relatively uniform sampling, with an average

deviation of only 6.62%. Most of the inhomogeneous sampling in VREX occurred early in the simulation when the potential energy lists were based on a small amount of sampling, and the sampling became more 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%. Simulated tempering with inaccurate weight factors (STb) produced the least uniform sampling by design (17.4%). We intentionally calculated weight factors to produce uneven sampling to represent a more complex system for which calculating weight factors accurately would be very computationally expensive. Using these inaccurate weight factors, STb oversamples high temperatures.

Temperature sampling efficiency is characterized by an overall score. The five measures of efficiency defined in this section are combined by averaging their normalized values. Normalization was performed by dividing each value by the maximum value of that measure. The overall score for each method is reported in the last line of table 2. Simulated tempering with accurate weight factors performs the best overall, and all of the ST-based methods perform better than the REX-based methods (especially SREM, which had the lowest overall score).

Methods Comparison: Characterizing 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). Shown in figure 3 is the probability distribution of the end-to-end distance obtained using each of the generalized ensemble methods at 280K (the lowest temperature). Also shown is the average distribution, which is obtained by taking the average of all of the methods. There is no discernable trend or better agreement between REX-based methods compared to ST-based methods. For example, for the peak at 5Å, REX, STDRb and SREM are above the average while STDR, VREX and STb are below. This suggests that there is no inherent bias of either ST-based or REX-based methods towards

sampling particular conformations. Based on this observation and on 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 end-to-end distance distribution of each method ($P_{eed}(n)$) from the "gold standard" end-to-end distance distribution ($P_{eed,\,reference}(n)$) by computing:

$$\sigma_{eed} = \sum_{n=1}^{Nbins} \left(P_{eed}(n) - P_{eed,reference}(n) \right)^2 \qquad \textbf{(21)}$$

where the index n labels bins, and there are N_{bins} in total. These values for σ are reported in figure 3. STDR exhibits the best agreement with the average distribution. In general, the ST-based methods have lower values for σ, corresponding to more accurate end-to-end distance distributions than the REX-based methods.

In order to confirm that the ST-based methods produce more accurate EED distributions when compared to the REX-based methods, EED distributions for the lowest ten temperatures for each generalized ensemble method are also computed and compared to the reference using equation 21. The EED distributions for each method and each temperature are displayed in figure 4, along with the σ value which is the average of the ten temperatures. The ST-based methods produce EED distributions which are quantitatively more accurate for all temperatures compared to the REX-based methods. STDR shows the best overall agreement with the reference data set, with a σ value of 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 are considered. A useful ergodic measure is the 1,4 pair distance metric,[64, 65] which quantifies the convergence of the distance between 1,4 residue pairs

(residues 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:

$$d_{dC\alpha\ matrix}(t) = \frac{1}{N_{residues}^2} \sum_{i=1}^{N_{residues}} \sum_{j=1}^{N_{residues}} \left(r_{ij}(t) - r_{ij,reference} \right)^2 \quad \textbf{(22)}$$

This involves computing the difference between each average pairwise distance (r_{ij}) and the same pairwise distance from the reference α-carbon distance matrix ($r_{ij,reference}$). In this equation, time, t, refers to simulation time accumulated at the temperature considered, and ($r_{ij}(t)$) is a cumulative average. As with end-to-end distance, the average of all seven generalized ensemble methods is used as the reference. We compute an analogous measure of convergence for the hydrogen bonding contact map (depicted in figure 1):

$$d_{contact\ map}(t) = \frac{1}{N_{residues}^2} \sum_{i=1}^{N_{residues}} \sum_{j=1}^{N_{residues}} \left(P_{ij}(t) - P_{ij,reference} \right)^2 \quad \textbf{(23)}$$

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 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" turn (shown in figure 1, the most probable turn). In addition, the average probability of forming a hydrogen bond and a "bend" (as defined in the DSSP algorithm[59]) 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 octamer's structure.

A representative example of how these structural properties measure structural convergence is shown in figure 5. In figure 5a, the convergence of the α-carbon distance matrix, the hydrogen bonding contact map, and the end-to-end distance distribution are displayed. 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 is converged is ambiguous. Each structural property appears to be converged at a slightly different time. This highlights the importance of considering multiple metrics when discussing the convergence of a simulation. Certain properties converge more quickly than others. In order to define convergence quantitatively, we consider the time taken to reach the reference value of the structural property of interest. As mentioned, the reference value is taken from the combined data set of all seven methods. Figure 5c depicts the time each structural property takes to reach the reference value, and remain within both one and two standard deviations. 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 simulation to reach the reference value within one standard deviation. The structural convergence times are provided in figure 5d for each of the generalized ensemble methods at 280K. At this temperature, STDR converges fastest to the reference data, closely followed by ST and STDRb.

To be systematic in ranking the structural convergence rates of the generalized ensemble methods, we also calculate t_{sc} for the lowest seven temperatures. These times are provided as supplementary material, figure S2. While STDR converges faster than the other methods at 280K, this is not a general trend for all temperatures. Each temperature has a different t_{sc} for each method. This highlights the importance of evaluating more than the lowest temperature when comparing the performance the methods, in addition to considering several structural metrics. It also suggests a way of quantifying the error in the measure 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 $<t_{sc}>$ is then the standard error of these measurements. Figure 6a shows a two-dimensional plot of $<t_{sc}>$ to within two standard deviations versus $<t_{sc}>$ to within one standard deviation. Lower values for $<t_{sc}>$ indicate faster structural convergence. A clear trend emerges: ST-based methods reach structural convergence more quickly than REX-based methods. The method that reaches convergence the fastest is simulated tempering with accurate weight factors (ST), while the method slowest to converge is SREM. It is not possible to unambiguously rank the other methods due to error in $<t_{sc}>$. However, it is important to note that both VREX and REX converge faster than SREM.

We can now answer a key question: does faster diffusion in temperature lead to a corresponding speed up in efficiency of conformational sampling? Figure 6b demonstrates that this is in fact the case. The combined average structural convergence time (obtained by taking the sum of $<t_{sc}>$ 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 method based on simulated tempering because it provides enhanced efficiency in terms of conformational sampling. Simulated tempering with accurate weight factors is the clear winner in both temperature diffusion and structural convergence, while serial replica exchange 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, simulated tempering 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 STDRb and STb here). Using this octapeptide, it is not possible to conclude which of these options is preferable. Inaccurate weight factors for this system yield

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

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 should still yield correct statistics, only with suboptimal sampling of temperature space.[30] We have explored the effect of suboptimal Helmholtz free energies on the accuracy of the data and we have demonstrated that the resulting conformational populations are not biased by the use of inaccurate free energies. This is demonstrated in figure 6a. Both simulated tempering and STDR with inaccurate weight factors (STb and STDRb) converge to the reference data set, which indicates that they achieve accurate conformational sampling. It has been stated that SREM is not exactly correct with inaccurate PEDFs (due to only approximately preserving detailed balance).[4, 12] We have shown in figure 6a that SREM, even with inaccurate PEDFs, still leads to Boltzmann-weighted sampling of conformational space (within one standard deviation) for this system. However, it converges more slowly than REX, and all the other generalized ensemble algorithms considered here.

Comparison of STDR and Conventional MD

The relative sampling enhancement of replica exchange compared to conventional MD has been the subject of significant controversy.[18] For example, one study found that replica exchange produced a speed up of 71.5 times at 275K for a 21 residue helical peptide with implicit solvent, based on the auto-correlation function of helicity.[17] In another work, replica exchange simulations of met-enkephalin in explicit solvent sampled five times more conformational space than a conventional MD simulation of the same duration (considering principal components based on the REX trajectory).[16] It has also been shown analytically that the expected speed up of replica exchange is directly related to the activation enthalpy for two-state protein folding. The efficiency of replica exchange is optimal when the maximum

temperature is chosen just slightly above the temperature at which the folding activation enthalpy is zero. [18] There are several key issues that emerge when reviewing comparative studies of replica exchange and conventional MD. First, the observed sampling enhancement (or lack thereof) is always heavily system dependent, as well as dependent on the structural or thermodynamic parameter on which the comparison is based. Second, examination of convergence for either the replica exchange simulation or the MD simulation is often neglected.[22, 26]

A comparison between STDR and conventional MD is shown in figure 7. Figure 7a and 7b show a superposition of 200 structures obtained using STDR and MD respectively. The amount of simulation time is the same for both methods (144ns for conventional MD, and 144ns in total for all temperatures for STDR, corresponding to 4.4ns at 280K). The root mean square deviations (RMSD) of these two collections of structures, 3.52Å for STDR and 3.88Å 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 7c and 7e) and conventional MD (in figures 7d and 7f). STDR converges more quickly, approximately by a factor of 2-3. However, given that STDR requires sampling 33 temperatures for the same amount of time, it is much less computationally efficient. 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 temperature because of the predicted temperature transition of the octapeptide GVGVPGVG.[46, 66, 67] 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. Pseudo-convergence can be observed for a structural ensemble generated by conventional MD which is energetically trapped.[43] 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, should they exist. 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.

The Importance of STDR for More Complex Systems

For relatively small and simple systems, such the octamer in this study or a short poly-alanine peptide in water,[28] the calculation of dimensionless Helmholtz free energies is a relatively straightforward procedure. For these cases, simulated tempering 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 prevent 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. Thus, even with near optimal weight factors which are updated throughout the simulation, Park and Pande still observed an average deviation from sampling homogeneity of 4.9% for a short peptide.[28] If the weight factors deviate severely from the true Helmholtz free energies, then sampling may be so far from uniform that it may not be possible to perform simulated tempering even with updates of the weight factors throughout the simulation. That is, there will be too little sampling at

certain temperatures to obtain a reasonable estimate of the free energies. We expect this to 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. We now describe a complex system for which this is the case.

In addition to studying the octapeptide, GVGVPGVG, we also simulated 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 (15ns per temperature for 70 temperatures, for a total of 1.05μs). Even with this large amount of data, the sampling of temperature space for simulated tempering using these weight factors is heterogeneous. The average deviation from sampling homogeneity is 21.3% (computed using equation 20), and the sampling at each temperature is shown in supplementary material, figure S5. Sampling at temperatures in the middle of the temperature range is less than both low and high temperatures, which impedes temperature diffusion from high to low temperature. Therefore, a simulated tempering simulation using these weight factors would be ineffective. A replica exchange simulation requiring 70 homogeneous and constantly available CPUs was not an option due to limited computational resources. Based on the simulations of the octapeptide, we concluded that STDR would be the most suitable method for this application. For comparison, we also performed a conventional MD simulation at the lowest temperature (261K). In figure 8a, we show a superposition of 200 structures, obtained every 1ns from a 200ns trajectory generated using constant-temperature MD. These structures have an average RMSD of 1.66Å, clearly indicating that the peptide is trapped in one conformational basin and undergoes only small conformational changes. This set of structures is contrasted to the set of 200 randomly-selected structures from the complete STDR simulation, which clearly represents completely different conformations (having an average RMSD of 8.40Å). For clarity, we also show six example structures in figure 8d to demonstrate the variety of conformations observed in the STDR simulation. To make a more direct comparison between

conventional MD and STDR, figure 8c shows 200 structures from STDR using the same amount of simulation time as the conventional MD simulation (200ns summed over all the temperatures, corresponding to approximately 3ns at 261K). The radius of gyration distributions for conventional MD, STDR and the first 3ns of STDR are shown in figure 8e. The distribution observed in STDR (both with all the data and with only the first 3ns) show several conformational states, while the conventional MD simulation is clearly trapped in one state. Even using the same amount of computational resources, STDR clearly produces a more heterogeneous ensemble of conformations.

Figure 9 displays hydrogen bonding contact maps for STDR, STDR with 3ns of sampling and conventional MD. STDR produces a conformational ensemble with many contacts formed with low probability. In contrast, conventional MD generates a contact map with only a few 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. Conventional MD severely underestimates the heterogeneity of the conformational landscape and exhibits both pseudo-convergence and quasi-nonergodicity. Even 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 for this system to quantitatively measure the speed-up of STDR versus conventional MD because limited computational resources preclude conventional MD simulations for the time that it might require to achieve structural convergence. We observe that conventional MD was trapped in one conformational basin for 200ns. 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 space does

in fact correspond to enhanced sampling, establishing the efficacy of the STDR method for a complex polypeptide.

Discussion and Conclusions

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 (should they exist). However, if one is not simultaneously interested in the behaviour 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. That is to say, the cost of the enhanced sampling simulation is the computer time needed to obtain the data at all temperatures, not only the temperature of interest. In order to truly "enhance sampling" relative to molecular dynamics 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 shows that it is possible to observe pseudo-convergence using molecular dynamics (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). This was the case in the simulation of $(GVPGV)_7$ using conventional MD. Long-time molecular dynamics does not yield the appropriate conformational distribution and the system remains trapped in a local minimum of the energy landscape. In contrast, we observed that conventional molecular dynamics was able to satisfactorily reproduce the conformational ensemble of GVGVPGVG at a significantly reduced computational cost compared to using a generalized ensemble method. How are the averages of quantities obtained using molecular dynamics to be interpreted in light of this issue? Based on 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 (a trade-off for increased confidence in the convergence and accuracy of the data). There have been several other examples where the enhanced sampling provided by generalized ensemble methods provides convergence that would not be feasible with molecular dynamics alone. [6, 12, 16-22] These observations underscore the need not only 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 and ergodicity of simulations. 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" molecular dynamics approach. It may be advisable to use a generalized ensemble approach (especially if the research question only demands conformational sampling and not dynamic information).

We now proceed to formulate a general statement regarding the choice of an appropriate generalized ensemble method. We cannot think of any system for which SREM would be the method of choice, since its PEDFs converge more slowly than weight factors, and it exhibits slower structural convergence and slower temperature diffusion than ST-based methods. It is also more inconvenient to implement compared to simulated tempering, and is only correct with accurate PEDFs, which we have observed to be difficult to obtain. Inaccurate PEDFs lead to both sampling inhomogeneity as well as simulations which do not preserve detailed balance.[4, 12] 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, replica exchange is not well suited to complex systems. A REX simulation typically requires the synchronization of a number of replicas (and processors) that grows with the number of degrees of freedom as $O(N^{1/2})$.[7, 31, 38] Although there is no theoretical limit on the number of replicas that one can use for a replica exchange simulation, it is in practice generally very difficult to have access to a very 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. Virtual replica exchange represents an attractive alternative to replica exchange since it completely eliminates synchronization and communication between replicas. It produces more homogeneous sampling of temperature space compared to SREM with much less initial simulation time. It therefore is preferable to both SREM and REX. We have also shown that in general REX-based methods suffer from slower structural convergence and temperature diffusion compared to ST-based methods. It is therefore preferable to use an ST-based method in temperature space. However, VREX may be a more suitable method in another reaction coordinate other than temperature for which weight factors are much more difficult to obtain.

If one has a relatively simple system, such as a small peptide, for which weight factors can be accurately calculated, simulated tempering is the most appropriate method. We have shown that ST exhibits the fastest temperature diffusion, and correspondingly, the fastest structural convergence. It requires no communication between CPUs and no fixed number of CPUs, resulting in no waste of computational resources. As one moves to more complex biomolecular systems, simulated tempering distributed replica becomes more suitable than simulated tempering. Even in the limit of 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 STDR method, combining the attributes of both simulated tempering and distributed replica sampling, offers increased computational efficiency and flexibility. We have demonstrated that STDR can make use of inaccurate weight factors to achieve homogeneous sampling of temperature space and consequently structural convergence. Because it is an ST-based method, it exhibits both faster temperature diffusion and structural convergence than REX-based methods. STDR is suitable for any computing cluster or distributed computing environment, and is easily implemented. It requires no fixed number of CPUs and there is no wasted CPU time in synchronization of exchanges. It is ideal for very complex systems. Based

on our thorough comparison of the available temperature-based generalized ensemble methods, STDR is the current method of choice for efficient conformational sampling of biomolecules in temperature space.

References:

1. Gnanakaran, S.; Nymeyer, H.; Portman, J.; Sanbonmatsu, K. Y.; Garcia, A. E., Peptide folding simulations. *Curr. Opin. Struct. Biol.* **2003,** *13* (2), 168-174.

2. Marinari, E.; Parisi, G., Simulated Tempering - A New Monte-Carlo Scheme. *Europhys. Lett.* **1992,** *19* (6), 451-458.

3. Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsovvelyaminov, P. N., New Approach to Monte-Carlo Calculation of the Free-Energy - Method of Expanded Ensembles. *J. Chem. Phys.* **1992,** *96* (3), 1776-1783.

4. Hagen, M.; Kim, B.; Liu, P.; Friesner, R. A.; Berne, B. J., Serial replica exchange. *J. Phys. Chem. B* **2007,** *111* (6), 1416-1423.

5. Hansmann, U. H. E., Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.* **1997,** *281* (1-3), 140-150.

6. Sugita, Y.; Okamoto, Y., Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999,** *314* (1-2), 141-151.

7. Hukushima, K.; Nemoto, K., Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan* **1996,** *65* (6), 1604-1608.

8. Tesi, M. C.; vanRensburg, E. J. J.; Orlandini, E.; Whittington, S. G., Monte Carlo study of the interacting self-avoiding walk model in three dimensions. *J. Stat. Phys.* **1996,** *82* (1-2), 155-181.

9. Ferrenberg, A. M.; Swendsen, R. H., New Monte Carlo Technique for Studying Phase Transitions. *Phys. Rev. Lett.* **1988,** *61* (23), 2635-2638.

10. Rodinger, T.; Howell, P. L.; Pomes, R., Distributed replica sampling. *J. Chem. Theory Comput.* **2006,** *2* (3), 725-731.

11. Rodinger, T.; Howell, P. L.; Pomes, R., Calculation of absolute protein-ligand binding free energy using distributed replica sampling. *J. Chem. Phys.* **2008,** *129* (15), 12.

12. Huang, X. H.; Bowman, G. R.; Pande, V. S., Convergence of folding free energy landscapes via application of enhanced sampling methods in a distributed computing environment. *J. Chem. Phys.* **2008,** *128* (20), 15.

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

14. Bedrov, D.; Smith, G. D., Exploration of conformational phase space in polymer melts: A comparison of parallel tempering and conventional molecular dynamics simulations. *J. Chem. Phys.* **2001,** *115* (3), 1121-1124.

15. Yamamoto, R.; Kob, W., Replica-exchange molecular dynamics simulation for supercooled liquids. *Phys. Rev. E* **2000,** *61* (5), 5473-5476.

16. Sanbonmatsu, K. Y.; Garcia, A. E., Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins* **2002,** *46* (2), 225-234.

17. Zhang, W.; Wu, C.; Duan, Y., Convergence of replica exchange molecular dynamics. *J. Chem. Phys.* **2005,** *123* (15), 9.

18. Nymeyer, H., How efficient is replica exchange molecular dynamics? An analytic approach. *J. Chem. Theory Comput.* **2008,** *4* (4), 626-636.

19. Periole, X.; Mark, A. E., Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent. *J. Chem. Phys.* **2007,** *126* (1), 11.

20. Rao, F.; Caflisch, A., Replica exchange molecular dynamics simulations of reversible folding. *J. Chem. Phys.* **2003,** *119* (7), 4035-4042.

21. Rhee, Y. M.; Pande, V. S., Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J.* **2003,** *84* (2), 775-786.

22. Tsai, H. H.; Reches, M.; Tsai, C. J.; Gunasekaran, K.; Gazit, E.; Nussinov, R., Energy landscape of amyloidogenic peptide oligomerization by parallel-tempering molecular dynamics simulation: Significant role of Asn ladder. *Proc. Natl. Acad. Sci. U. S. A.* **2005,** *102* (23), 8174-8179.

23. Denschlag, R.; Lingenheil, M.; Tavan, P., Efficiency reduction and pseudo-convergence in replica exchange sampling of peptide folding-unfolding equilibria. *Chem. Phys. Lett.* **2008,** *458* (1-3), 244-248.

24. Zuckerman, D. M.; Lyman, E., A second look at canonical sampling of biomolecules using replica exchange simulation. *J. Chem. Theory Comput.* **2006,** *2* (4), 1200-1202.

25. Zuckerman, D. M.; Lyman, E., A second look at canonical sampling of biomolecules using replica exchange simulation. (vol 2, pg 1200, 2006). *J. Chem. Theory Comput.* **2006,** *2* (6), 1693-1693.

26. Beck, D. A. C.; White, G. W. N.; Daggett, V., Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations. *J. Struct. Biol.* **2007,** *157* (3), 514-523.

27. Hansmann, U. H. E.; Okamoto, Y., Monte Carlo simulations in generalized ensemble: Multicanonical algorithm versus simulated tempering. *Phys. Rev. E* **1996,** *54* (5), 5863-5865.

28. Park, S.; Pande, V. S., Choosing weights for simulated tempering. *Phys. Rev. E* **2007,** *76* (1), 5.

29. Okamoto, Y., Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *Journal of Molecular Graphics and Modelling* **2004,** *22*, 425-439.

30. Zhang, C.; Ma, J. P., Comparison of sampling efficiency between simulated tempering and replica exchange. *J. Chem. Phys.* **2008,** *129* (13), 7.

31. Mitsutake, A.; Sugita, Y.; Okamoto, Y., Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers* **2001,** *60* (2), 96-123.

32. Park, S., Comparison of the serial and parallel algorithms of generalized ensemble simulations: An analytical approach. *Phys. Rev. E* **2008,** *77* (1), 6.

33. Hansmann, U. H. E.; Okamoto, Y., Numerical comparisons of three recently proposed algorithms in the protein folding problem. *J. Comput. Chem.* **1997,** *18* (7), 920-933.

34. Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M., The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules .1. The Method. *J. Comput. Chem.* **1992,** *13* (8), 1011-1021.

35. Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A., Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. *J. Chem. Theory Comput.* **2007,** *3* (1), 26-41.

36. Mitsutake, A.; Okamoto, Y., Replica-exchange simulated tempering method for simulations of frustrated systems. *Chem. Phys. Lett.* **2000,** *332* (1-2), 131-138.

37. Hansmann, U. H. E.; Okamoto, Y., Prediction of Peptide Conformation by Multicanonical Algorithm - New Approach to the Multiple-Minima Problem. *J. Comput. Chem.* **1993,** *14* (11), 1333-1338.

38. Fukunishi, H.; Watanabe, O.; Takada, S., On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.* **2002,** *116* (20), 9058-9067.

39. Gallicchio, E.; Levy, R. M.; Parashar, M., Asynchronous replica exchange for molecular simulations. *J. Comput. Chem.* **2008,** *29* (5), 788-794.

40. Shirts, M.; Pande, V. S., Computing - Screen savers of the world unite! *Science* **2000,** *290* (5498), 1903-1904.

41. Bowman, G. R.; Huang, X. H.; Yao, Y.; Sun, J.; Carlsson, G.; Guibas, L. J.; Pande, V. S., Structural insight into RNA hairpin folding intermediates. *J. Am. Chem. Soc.* **2008,** *130* (30), 9676-+.

42. Shen, H. J.; Czaplewski, C.; Liwo, A.; Scheraga, H. A., Implementation of a serial Replica Exchange Method in a physics-based united-residue (UNRES) force field. *J. Chem. Theory Comput.* **2008,** *4* (8), 1386-1400.

43. Neale, C.; Rodinger, T.; Pomes, R., Equilibrium exchange enhances the convergence rate of umbrella sampling. *Chem. Phys. Lett.* **2008,** *460* (1-3), 375-381.

44. Miao, M.; Bellingham, C. M.; Stahl, R. J.; Sitarz, E. E.; Lane, C. J.; Keeley, F. W., Sequence and structure determinants for the self-aggregation of recombinant polypeptides modeled after human elastin. *J. Biol. Chem.* **2003,** *278* (49), 48553-48562.

45. Rauscher, S.; Baud, S.; Miao, M.; Keeley, F. W.; Pomes, R., Proline and glycine control protein self-organization into elastomeric or amyloid fibrils. *Structure* **2006,** *14* (11), 1667-1676.

46. Baer, M.; Schreiner, E.; Kohlmeyer, A.; Rousseau, R.; Marx, D., Inverse temperature transition of a biomimetic elastin model: Reactive flux analysis of folding/unfolding and its coupling to solvent dielectric relaxation. *J. Phys. Chem. B* **2006,** *110* (8), 3576-3587.

47. Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C., GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005,** *26* (16), 1701-1718.

48. Lindahl, E.; Hess, B.; van der Spoel, D., GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001,** *7* (8), 306-317.

49. Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996,** *118* (45), 11225-11236.

50. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. In *Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum*

chemical calculations on peptides, Symposium on Molecular Dynamics - The Next Millennium, New York, New York, Jun 02-03; Amer Chemical Soc: New York, New York, 2000; pp 6474-6487.

51. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983,** *79* (2), 926-935.

52. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical Integration of Cartesian Equations of Motion of a System With Constraints - Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* **1977,** *23* (3), 327-341.

53. Darden, T.; York, D.; Pedersen, L., Particle Mesh Ewald - An NLOG(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993,** *98* (12), 10089-10092.

54. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G., A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995,** *103* (19), 8577-8593.

55. Nose, S., A Molecular-Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* **1984,** *52* (2), 255-268.

56. Hoover, W. G., Canonical Dynamics - Equilibrium Phase-Space Distributions *Phys. Rev. A* **1985,** *31* (3), 1695-1697.

57. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E., GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008,** *4* (3), 435-447.

58. Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J., LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997,** *18* (12), 1463-1472.

59. Kabsch, W.; Sander, C., Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983,** *22* (12), 2577-2637.

60. Humphrey, W.; Dalke, A.; Schulten, K., VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996,** *14* (1), 33-&.

61. Mitsutake, A.; Okamoto, Y., Replica-exchange extensions of simulated tempering method. *J. Chem. Phys.* **2004,** *121*, 2491-2504.

62. Rathore, N.; Chopra, M.; de Pablo, J. J., Optimal allocation of replicas in parallel tempering simulations. *J. Chem. Phys.* **2005,** *122* (2), 8.

63. Abraham, M. J.; Gready, J. E., Ensuring mixing efficiency of replica-exchange molecular dynamics simulations. *J. Chem. Theory Comput.* **2008,** *4* (7), 1119-1128.

64. Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J., Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proc. Natl. Acad. Sci. U. S. A.* **2005,** *102* (39), 13749-13754.

65. Thirumalai, D.; Mountain, R. D.; Kirkpatrick, T. R., Ergodic Behavior in Supercooled Liquids and in Glasses. *Phys. Rev. A* **1989,** *39* (7), 3563-3574.

66. Rousseau, R.; Schreiner, E.; Kohlmeyer, A.; Marx, D., Temperature-dependent conformational transitions and hydrogen-bond dynamics of the elastin-like octapeptide GVG(VPGVG): A molecular-dynamics study. *Biophys. J.* **2004,** *86* (3), 1393-1407.

67. Schreiner, E.; Nicolini, C.; Ludolph, B.; Ravindra, R.; Otte, N.; Kohlmeyer, A.; Rousseau, R.; Winter, R.; Marx, D., Folding and unfolding of an elastinlike oligopeptide: "Inverse temperature transition," reentrance, and hydrogen-bond dynamics. *Phys. Rev. Lett.* **2004,** *92* (14), 4.

FIGURE CAPTIONS

TABLE 1: Practical Advantages and Disadvantages of Generalized Ensemble Algorithms | 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 on only one CPU. Replica exchange is severely hindered by all of the issues listed, except that it does not require any initial simulation.

TABLE 2: Evaluating Diffusion in Temperature Space | 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.

FIGURE 1: The Conformational Landscape and Hydrogen-Bonding Contact Map of GVGVPGVG | (A) A selection of 35 random conformations from the STDR simulation at 280K of GVGVPGVG, with glycine in red, valine in yellow and proline in blue. **(B)** The hydrogen-bonding contact map of GVGVPGVG, 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.

FIGURE 2: Assessing Convergence of Weight Factors and PEDFs | (A) Convergence of PEDFs for SREM was quantified using the χ^2 measure defined in equation 13. Using this measure, the PEDFs obtained using 19.5ns of conventional MD at each temperature appeared to be stationary. **(B)** Data based on 19.5ns of conventional MD per temperature. **(C)** Data from the replica exchange simulation, using only the first 25ns per temperature. **(D)** All of the data from the REX simulation (4.75μs). Error in the acceptance ratio is shown in (B), (C) and (D) for both ST in yellow (computed using equations 15 and 19) and SREM in purple (computed using equations 14 and 17).

FIGURE 3: Assessing the Accuracy of the End-to-End Distance Distribution | The EED probability distribution is shown for each method with colors indicated using data from 280K. 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, is shown next to the legend, and was computed using equation 15 with the average distribution as the reference.

FIGURE 4: End-to-End Distance Distributions at Different Temperatures| The EED probability distributions are shown for the lowest ten 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 ten temperatures using equation 20 with the average distribution as the reference, and the average of these errors is shown on each plot. The REX-based methods are shown in the top row, and have larger errors than the ST-based methods, shown in the second and third rows.

FIGURE 5: Assessing Structural Convergence Using Multiple Criteria| The data for (A), (B) and (C) are from simulated tempering at 280K. The trajectory is separated into fifty 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} (equation 21), $d_{contact\ map}$ (equation 22) and $d_{dCa\ matrix}$ (equation 23, plotted on the secondary axis). **(B)** The probability per residue of a gamma turn, beta turn and alpha turn are shown, as well as the population of the VPGV turn (the most populated turn, depicted in figure 1). 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 interval at which they reached and remained within one and two standard deviations of the reference data set are shown as a bar graph. The average of these times is also shown, corresponding to the average structural convergence time, $<t_{sc}>$. One standard deviation is calculated based on 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 280K in yellow and purple respectively. These times are provided for temperatures 288K, 296K, 305K, 314K, 323K and 332K as supplementary material figure S2.

FIGURE 6: Structural Convergence is Correlated to Temperature Diffusion| (A) Average structural convergence times, $<t_{sc}>$, obtained using the lowest seven temperatures are shown. The $<t_{sc}>$ to reach two standard deviations is plotted against the $<t_{sc}>$ to reach one standard deviation for each method. Error bars represent the standard error of the $<t_{sc}>$ for the seven temperatures. Another version of this plot is provided as supplementary material figure S3, with the $<t_{sc}>$ for each temperature shown. **(B)** The $<t_{sc}>$ 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 clear 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 REX-based methods (in purple).

FIGURE 7: Comparing STDR and Conventional MD for GVGVPGVG| (A) Two hundred structures in ribbon representation obtained using the first 4.4ns at 280K for STDR are shown, and **(B)** for 144ns of conventional MD, along with the corresponding RMSD. Glycine is in purple, proline is in yellow, and valine is in grey. **(C)** and **(D)** show σ_{eed} (equation 15), $d_{contact\ map}$ (equation 17) and $d_{dCa\ matrix}$ (equation 16,

plotted on the secondary axis) for STDR and conventional MD respectively. The trajectories are separated into fifty time intervals, and the quantities reported are calculated cumulatively, as in figure 5. **(E)** and **(F)** show the probability per residue of a gamma turn, beta turn and alpha turn as well as the population of the VPGV turn. The probability of a hydrogen bond per residue and a bend per residue (plotted on the secondary axis) are also shown.

FIGURE 8: Comparing STDR and Conventional MD for $(GVPGV)_7$| Two hundred structures in ribbon representation along with their RMSD are shown for **(A)** the conventional MD simulation of length 200ns at 261K, **(B)** for the STDR simulation at 261K using all of the data (120ns at this temperature), **(C)** and for the STDR simulation at 261K using the first 3ns (this is the same simulation time summed over all replicas as **(A)**). Glycine is in purple, proline is in yellow, and valine is in grey. A selection of six example structures is shown from the structures in **(B)** to illustrate the structural diversity obtained using STDR. The probability distributions of radius of gyration for the data sets described are shown in **(E)**.

FIGURE 9: Hydrogen Bonding Contact Maps from STDR and Conventional MD | In this figure, we depict hydrogen-bonding contact maps as three dimensional maps, where peak height represents the probability of contact formation. We show these plots on two scales. On the left, the scale goes up to 0.8, and on the right, to 0.01, for clarity in showing the contacts formed with low probability. **(A)** and **(B)** are the contact maps for the STDR simulation at 261K using all of the data (120ns at this temperature). **(C)** and **(D)** are the contact maps for the STDR simulation using the first 3ns (this is the same simulation time summed over all replicas as **(E)** and **(F)**). **(E)** and **(F)** are the contact maps for 200ns of conventional MD. Some contacts are formed over 80% of the time.

Supplementary Material

TABLE S1: List of Temperatures | A list of the 33 exponentially spaced temperatures for generalized ensemble simulations of GVGVPGVG, and the 70 temperatures used in the STDR simulation of $(GVPGV)_7$.

FIGURE S1: Deviation from Sampling Homogeneity | The deviation from perfectly homogeneous sampling of temperature for each method. REX is not shown since it samples temperatures uniformly by definition. Deviation from homogeneity is computed as described in equation 14. The average deviations from sampling homogeneity reported in table 2 were computed as averages over temperature using the data in this figure. STb was created to have inhomogeneous sampling, and it clearly achieves this. SREM also exhibits significant deviation from homogeneity.

FIGURE S2: Structural Convergence at Multiple Temperatures, Bar Graphs | The average structural convergence times for one and two standard deviations are shown for all seven methods in yellow and purple respectively. These times are provided for temperatures 280K, 288K, 296K, 305K, 314K, 323K and 332K. Below each plot, a ranking of the methods is provided, from the fastest to converge to the slowest.

FIGURE S3: Structural Convergence at Multiple Temperatures, 2d Plot | 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 the $\langle t_{sc} \rangle$ for the seven temperatures. In **(A)** the methods are colored red for ST-based and blue for REX-based, while in **(B)** they are colored individually.