Efficiency of adaptive temperature-based replica exchange for sampling large-scale protein conformational transitions

Weihong Zhang and Jianhan Chen

Published Version Information

Citation: Zhang, W., & Chen, J. (2013). Efficiency of adaptive temperature-based replica exchange for sampling large-scale protein conformational transitions. Journal of Chemical Theory and Computation, 9(6), 2849-2856.

Digital Object Identifier (DOI): doi:10.1021/ct400191b

Publisher's Link: http://pubs.acs.org/doi/full/10.1021/ct400191b

Efficiency of adaptive temperature-based replica exchange for sampling large-scale protein conformational transitions

Weihong Zhang and Jianhan Chen*

Department of Biochemistry and Molecular Biophysics

Kansas State University

Manhattan, KS 66506, USA

Submitted to *Journal of Chemical Theory and Computation* as an Article

Revised Version

Corresponding Author: Phone: (785) 532-2518; Fax: (785) 532-7278; Email: jianhanc@ksu.edu

ABSTRACT

Temperature-based replica exchange (RE) is now considered a principal technique for enhanced sampling of protein conformations. It is also recognized that existence of sharp cooperative transitions (such as protein folding/unfolding) can lead to temperature exchange bottlenecks and significantly reduce the sampling efficiency. Here, we revisit two adaptive temperature-based RE protocols, namely, exchange equalization (EE) and current maximization (CM), that were previously examined using atomistic simulations (Lee and Olson, J. Chem. Physics, 134, 24111 (2011)). Both protocols aim to overcome exchange bottlenecks by adaptively adjusting the simulation temperatures, either to achieve uniform exchange rates (in EE) or to maximize temperature diffusion (CM). By designing a realistic yet computationally tractable coarse-grained protein model, one can sample many reversible folding/unfolding transitions using conventional constant temperature molecular dynamics (MD), standard REMD, EE-REMD, and CM-REMD. This allows rigorous evaluation of the sampling efficiency, by directly comparing the rates of folding/unfolding transitions and convergence of various thermodynamic properties of interest. The results demonstrate that both EE and CM can indeed enhance temperature diffusion compared to standard RE, by ~3- and over 10-fold, respectively. Surprisingly, the rates of reversible folding/unfolding transitions are similar in all three RE protocols. The convergence rates of several key thermodynamic properties, including the folding stability and various 1D and 2D free energy surfaces, are also similar. Therefore, the efficiency of RE protocols does not appear to be limited by temperature diffusion, but by the inherent rates of spontaneous large-scale conformational re-arrangements. This is particularly true considering that virtually all RE simulations of proteins in practice involve exchange attempt frequencies ($\sim ps^{-1}$) that are several orders of magnitude faster than the slowest protein motions ($\sim \mu s^{-1}$). Our results also suggest that the efficiency of RE will not likely be improved by other protocols that aim to accelerate exchange or temperature diffusion. Instead, protocols with some types of guided tempering will likely be necessary to drive faster large-scale conformational transitions.

1. Introduction

Successful computer simulations of protein conformational equilibrium and transitions not only require accurate description of protein energetics in complex heterogeneous environments, but also require sufficient sampling of the relevant conformational space. At present, generating atomistic structure ensembles that are statistically representative of the accessible conformations of a protein under a given set of thermodynamic conditions remains a challenging problem[1]. The

difficulty arises not only because of the large and complex conformational space of biomolecules, but also due to significant energy barriers that might separate different conformational subspaces. The efficiency of conventional Monte Carlo (MC) or molecular dynamics (MD) is limited due to frequent trapping of protein in numerous local energy minima. The replica exchange (RE) method[2-5], also known as parallel tempering, has emerged as a relatively straightforward but powerful approach for enhanced conformational sampling. The basic idea is to simulate multiple replicas of the system at different temperatures independently using either MC or MD. Periodically, replicas attempt to exchange simulation temperatures according to a Metropolis criterion that preserves the detailed balance and ensures canonical distributions at all temperatures. The resulting random walk in the temperature space helps the system to escape states of local energy minima and thus facilitate conformational sampling. Replica exchange molecular dynamics (REMD) in particular has been widely applied to and shown to be successful in protein simulations[6-10]. Nonetheless, questions remain regarding the true efficiency of RE in sampling large-scale (protein) conformational transitions and its dependence on the properties of the system and key RE parameters, such as the number of replicas, exchange attempt frequency, and choice of simulation temperatures, etc.

The key parameters of REMD of protein simulations has been a subject of substantial research interest in recent years, examined both based on theoretical considerations[11-16] and through actual simulations of small peptides[8,17-19]. These studies generally confirm that RE can enhance the conformational sampling as long as the activation enthalpies are positive. In particular, recent theoretical analysis[14] and kinetic network models[13,16] of RE for two-state systems have emphasized the importance of choosing a maximum temperature slightly above the temperature where the folding rate is maximal because of the anti-Arrhenius behavior of protein folding at high temperatures. Once the maximum temperature is chosen, other REMD parameters can be set to achieve effective diffusion of replicas across the temperature ladder, arguably with the smallest possible number of replicas. Many strategies have been previously proposed for the later purpose, such as more frequent exchange attempts[18], global energy reassignment[20], and non-equilibrium switches[21]. More aggressive approaches attempt to reduce to number of replicas required for effective temperature diffusion by directly reducing the number of particles that participate in temperature exchanges[22-24], but sometimes with undesirable consequences[25].

It is typically assumed that REMD will provide the highest sampling efficiency if all the replicas spend equal amount of time at each temperature. As proposed by Sugita and Yokomoto[4] and Kofke[26], this may be achieved through optimizing the allocation of simulation temperatures to provide efficient and uniform exchanges. Nonetheless, it is not clear whether consideration of exchange efficiency alone is sufficient to optimize the sampling efficiency, which is more directly

measured by the frequency of conformational transitions, and ultimately by the convergence of various thermodynamic properties of interest. For example, REMD simulations were previously performed in the GBSW implicit solvent to examine the conformational equilibrium of a β-hairpin derived from protein GB1 domain (GB1p; residues 41-56)[27]. The exchange (acceptance) rates of both folding and control REMD simulations were very uniform and >60% for all pairs (Fig. 1A). Nonetheless, there was an apparent segregation of replicas in the temperature space: a few of replicas dominated the lowest temperature and had much lower averaged temperatures (e.g., replicas 2, 6 and 10; see Fig. 1B). Closer examination suggests that the observed temperature trapping reflects a lack of reversible folding and unfolding transitions during these simulations. The few replicas happened to fold first during the folding run (or those did not unfold at first during the control run) remained folded throughout the simulation timespan and thus had much higher tendency to visit the lower temperature windows (e.g. replicas 2 and 10; the black and red traces in Fig. 1C). Even though the folded replicas did visit higher temperatures frequently, they did not spend sufficient time to unfold spontaneously and thus quickly reversed back to lower temperatures. Vice versa was true for the unfolded replicas. As a result, there was rapid mixing of the replicas due to uniform and high exchange rates, but all replicas remained trapped in various free energy basins. Similar observations have also been made by Periole and Mark[17].

In light of the inadequacy if uniform exchange to promote folding/unfolding transitions, the recent proposed current maximization (CM) protocol[28] is very attractive. CM aims to systematically optimize temperature diffusion such that the number of "round trips" in the temperature space of each replica is maximized. Arguably the number of "round trips" is more directly related to the ability of RE to drive conformational transitions, and thus CM could substantially improve the sampling efficiency[28]. Lee and Olson recently compared the sampling efficiency of CM and another adaptive temperature-based REMD protocol, namely, exchange equalization (EE)[29], using the 57-residue SH3 domain of α-spectrin in implicit solvent as a model system. However, due to the system size and associated computational cost, no reversible folding/unfolding transitions were sampled. As a result, how effectively the improved temperature diffusion in CM can translate into sampling efficiency remains to be established.

Identification of appropriate protein model systems for rigorous benchmarking of sampling efficiency has actually been difficult. Due to the computational cost constraint, small peptides have been mainly used as model systems[8,17,18]. REMD simulations of the 21-residue Fs-21 peptide in implicit solvent demonstrated that REMD could enhance the sampling efficiency by 14 to 72 times at different temperatures compare to conventional MD based on examination of helicity auto-correlation functions[8]. Periole and Mark simulated a β-heptapeptide in explicit solvent to examine the convergence and sampling efficiency, which suggested REMD to be at least one order

of magnitude more efficient than MD[17]. While these studies have provided important insights into the efficacy of REMD, small model peptides lack long-range ordering that exist in actual globular proteins, and, more importantly, do not fully reflect globular proteins where folding/unfolding transitions often occur at much longer timescales compare to local conformational fluctuations. Our previous simulations of small proteins[30-32] (also see Fig. 1) have suggested that the ability to drive slow global conformational transitions is a key bottleneck in REMD sampling. Therefore, there is a need to develop realistic yet computationally tractable protein models for more rigorous benchmarking of REMD sampling efficiency.

In this work, we first construct a protein model derived from coarse-grained topology-based models[33], which are based on the principle of minimal frustration for evolved proteins and allow direct simulations of folding and unfolding transitions to characterize folding mechanisms. In particular, impressive correspondence has been demonstrated between experiment and theory for many proteins[33,34], supporting the notion that topology-based models a capture the essence of folding behaviors of globular proteins. Specifically, the original sequence-flavored Gō-like model[35] was modified to include non-specific hydrophobic interactions to mimic the presence of non-native interactions in real proteins. Inclusion of non-specific interactions also increases the complexity of the energy landscape, and thus the sampling challenge. Such a model arguably provides a reasonable balance between the level of realisticity and computational tractability for more rigorous benchmarking of how the above-discussed adaptive temperature-based REMD protocols enhance sampling of protein conformations.

2. Methods

2.1 Adaptive temperature-based REMD protocols

A typical REMD involves multiple replicas simulated simultaneously at different temperatures. For given choices of the number of replicas N, the minimum temperature, $T_{\min}$, and the maximum temperature, $T_{\max}$, the temperature distribution is usually assigned as:

$$T_1 = T_{\min}, \quad T_{i+1} = T_i \left(\frac{T_{\max}}{T_{\min}}\right)^{\left[\frac{1}{N-1}\right]}. \tag{1}$$

Exchange of simulation temperatures are attempted periodically and accepted according to a Metropolis criterion that maintains the detailed balance[4].

$$p(i \leftrightarrow j) = \min[1, e^{(\beta_i - \beta_j)(E_i - E_j)}], \tag{2}$$

where $\beta_i = 1/k_B T_i$, k_B is Boltzmann's constant, and T_i and E_i are the temperature and potential energy of replica *i*. After a successful exchange of simulation temperatures, the velocities of swapped replicas are rescaled according to,

$$P(T_j) = \sqrt{\frac{T_j}{T_i}} P(T_i), \; P(T_j) = \sqrt{\frac{T_i}{T_j}} P(T_i). \tag{3}$$

Exchange rate equalization (EE)

For systems displaying strong phase transitions such as protein folding and unfolding, the standard REMD can suffer from depressed exchange rate near the transition temperature, T_m. One approach to resolve this limitation is to dynamically adapt the temperature spacing and populate more simulation windows near T_m.. This may be achieved by equalizing the exchange rates through out the REMD temperature range. Briefly, as described by Lee and Olson[36], the simulation temperatures, $\{T_i\}$, can be updated periodically during REMD to maximize the sum of exchange acceptance rates, $s_{i \to i+1}$, raised to a properly chosen negative power, *r*,

$$\Omega = \max\left[\sum_{i=1}^{N-1} (s_{i \to i+1})^r\right]. \tag{4}$$

In the limit of $r \to -\infty$, the exchange rates would be uniform with maximized Ω. The value of *r* was assigned to be -4 in the current work as suggested by Lee and Olson[36]. Assuming that the energy distributions at all temperatures are Gaussian, the exchange rates can be estimated by integrating

the acceptance probability over the energy distributions[29]. This allows numerical optimization of $\{T_i\}$ according to Eqn. 4 to achieve approximate equalization of exchange rates between all neighboring replica pairs.

Current maximization (CM)

The CM method aims to maximize the number of round trips that the replicas travel between the lowest and highest simulation temperatures[37]. For this, a replica is marked as cold or hot depending on its last visit to either temperature extremes. The cold fraction can be then calculated for each temperature as

$$f(T) = \frac{n_{cold}(T)}{n_{cold}(T) + n_{hot}(T)}, \quad (5)$$

where $n_{cold}(T)$ and $n_{hot}(T)$ are the number of cold and hot replicas at temperature *T*. It has been shown that the current of temperature diffusion can be maximized by adjusting the temperatures, T_i, such that $f(T)$ increases linearly as a function of the temperature index, i[38]. This can be achieved by first interpolating a continuous $f(T)$ from the computed values of f at the current set of temperatures and then using it to search for new temperatures within the pre-specified temperature ranges where $f(T) = i / (N-1)$. To avoid crowding of replicas near T_m, a constraint is introduced such that no neighboring temperatures can be more than two geometric spacing units apart[29]

$$T_{i+1} / T_i \leq (T_{\max} / T_{\min})^{2/(N-1)}. \quad (6)$$

2.2 A topology-based coarse-grained protein model with rugged energy landscape

The sequence-flavored Gō-like model[35] is an advanced topology-based model that exploits the idea that sequence can provide differing statistical weights to alternative folding pathways. By including knowledge-based pseudo-torsional potentials and using the Miyazawa-Jernigan (MJ) statistical potentials[39] for residue-specific Cα-based native interactions, the model can recapitulate subtle differences in folding mechanisms that arise from sequence differences in topologically analogous proteins[40,41]. In this work, we further extend this model to include nonspecific hydrophobic interactions. Specifically, an initial model was first generated for the B1 domain of streptococcal protein G (GB1) (PDB: 3gb1)[42] (see Fig. 2) using the Multiscale Modeling Tools for Structural Biology (MMTSB) Gō-Model Builder (http://www.mmtsb.org)[43]. We then recalibrated the model to

have $T_m \sim 350$K by uniformly adjusting the strength of all native contacts. Non-specific van der Waals (vdW) interactions were then introduced between all Cα beads with strengths similar to those of the native contacts. Specifically, the well depth was empirically set to ε = -1.1 kcal/mol, which appeared to provide a reasonable compromise between retaining fast folding rates and increasing energy landscape complexity based on pilot REMD simulations. The pilot simulations also suggested that non-specific 1-4 vdW interactions as well as those in two short loops (see Fig. 2) needed to be turned off to avoid steric restriction of rotation around Cα-Cα bonds and facilitate chain diffusion. The final model retains a sharp folding transition (Fig. 3A), but displays more complex energy landscape with multiple intermediate states (Fig. 3B; also see Section 3.4). Compared to the original sequence-flavored Gō-like model, the folding and unfolding rates are about 15 times slower at corresponding T_m's.

2.3 Simulation protocols and analysis

All the REMD simulations were performed using the MMTSB Toolset[43] with 8 replicas spanning 300 to 400K. For each replica, Langevin dynamics simulation was performed using CHARMM[44,45] with a dynamic time step of 10 fs and a friction coefficient of 0.1 ps^{-1}. The SHAKE algorithm[46] was applied to constrain all virtual bond lengths. Replica exchanges were attempted between neighboring replicas every 5000 MD steps (50 ps). The geometrically spaced distribution (Eqn. 1) was used as the initial temperature profile. Pilot runs were first performed to obtain the optimized temperature distributions according to either EE or CM protocol. Specifically, multiple cycles of 1-μs REMD were performed, each followed by optimization of the temperature distributions using the EE and CM methods until the whole profile stabilized. Additional 5-μs REMD simulations were then performed verify that the EE and CM temperature distributions were stable (for the given protein model). Finally, the converged EE and CM temperature profiles were used for independent production REMD simulations, initiated from the native structure (control run) and a fully extended structure (folding run), respectively. The length of all REMD simulations was 100 μs. The parallel control and folding runs allow additional diagnosis of convergence. For control, standard REMD and constant-temperature MD simulations (at 300 K and $T_m \approx 354$K) were also performed (initiated from the native structure). These simulations are summarized in Table 1. We note that T_m calculated from the production simulations are slightly different from the value estimated using the pilot run (~354K).

The sampling efficiencies of various simulation protocols are judged mainly based on: the number of folding/unfolding transitions N_{TS}, average time of conformational transitions, τ_{TS}, and convergence of the calculated folding free energy, ΔG, and several 1D and 2D potential of mean forces (PMFs). In the context of topology-based modeling, the fractions of native contacts provide

natural reaction coordinates for describing folding and unfolding transitions[47]. The model for protein GB1 includes 75 native contacts. Based on the 1D PMF as a function of native contact fraction (see Fig. 3B), conformations with <25 native contacts were assigned to be in the unfolded state, and those with >55 native contacts to the folded state. These definitions of the folded and unfolded states were then used to calculate the number of folding and unfolding transitions sampled in MD or by all replicas in REMD simulations (see Table 1). We further divided the native contacts into those involving the N-terminal residues 1-20 (Q-nt), the C-terminal residues 42-55 (Q-ct), or the central α-helical residues 23-36 (Q-h), and calculated 2D various PMFs. Achieving convergence on these 2D PMFs should be more challenging than converging on folding stability alone or 1D PMF. Thus, it provides a more stringent test on various REMD protocols' ability to enhance sampling. The weighted histogram analysis method (WHAM) was used to combine information from all temperatures to compute either the C_v curves or various unbiased 1D and 2D probability distributions[48]. All the analysis was performed using a combination of the MMTSB toolset, CHARMM and in-house scripts. The numbers of temperature round trips in REX simulations were calculated between the lowest two temperature and highest two temperatures.

3. Results and discussion

3.1 Convergence of the EE- and CM-optimized temperature distributions

For given temperature range and number of replicas, the EE- or CM-optimized temperature distributions should depend solely on the nature of the system (e.g., the density of state distribution). Indeed, as shown in Fig. 4, both EE- and CM-optimized temperature distributions quickly deviated from the initial geometrically spaced profile (black traces), but started to stabilize by the fourth 1-μs cycles of iterative REMD runs. In the next couple 1-μs cycles, the temperature profiles varied only slightly between cycles (<5 K), which appears to arise mainly due to the finite length of the pilot REMD cycles. Additional 5-μs pilot REMD runs further validate that the final EE- and CM-optimized temperature profiles are well converged. As previously observed[29], the EE-optimized temperature distribution sets large intervals at lower temperatures and closely places simulation windows at higher temperatures; whereas CM optimization tends to place dense windows near the melting temperature of the protein. The final EE- and CM-optimized temperature distributions (red traces in Fig. 4) were used in the 100-μs control and folding production REMD simulations.

3.2 Exchange acceptance rates and numbers of round trips in the temperature space

As a first step towards examining the efficiency of EE and CM, key REMD metrics, including exchange acceptance rates and numbers of temperature round trips, were calculated to ensure that the adaptive temperature distributions achieved the goals of EE and CM. As shown in Fig. 5, the exchange acceptance rates are indeed approximately equalized (~60%) in both control and folding EE-REMD production simulations. Note that the exchange acceptance rates between control and folding simulations are similar, indicating that simulations were well converged. In contrast, the geometrically spaced temperature distribution used in standard REMD leads to an exchange bottleneck near T_m, albeit less pronounced compared to what was observed in atomistic simulations[29]. CM-REMD yielded higher exchange acceptance rates near T_m (>80%), which is consistent of denser windows in this region. The exchange acceptance rates in CM are lower at both low and high temperature extremes, which is also similar to the previous study using atomistic model with more replicas[29]. The numbers of round trips in the temperature space are summarized in Table 1. The results show that the CM-optimized temperature distribution does enhance the diffusion in the temperature space, about 13-fold compared to the standard REMD. In comparison, equalizing the exchange rates also enhances temperature diffusion, but only about 3-fold compared to standard REMD. In the following, we further examine whether such equalized exchange rates and enhanced temperature diffusion can effectively translate into more conformational transitions and faster convergence in various thermodynamic properties.

3.3 Number of folding and unfolding transitions

The sampling efficiency of REMD is arguably more directly related to the number of major conformational transitions sampled, such as between folded and unfolded states. As summarized in Table 1, 28 and 24 folding-unfolding transitions were observed during the control and folding EE-REMD runs, respectively. Interestingly, despite enhanced temperature diffusion, the numbers of conformational transitions sampled by CM-REMD simulations are actually smaller, 23 in the folding run and only 18 in the control run. Using the differences between the control and folding runs as rough estimates of the uncertainties, EE- and CM-REMD generated 26 ± 4 and 20.5 ± 5 reversible folding/unfolding transitions in an aggregated length of 800 μs. These numbers are not significantly higher than a total of 21 reversible folding/unfolding events sampled in the standard REMD simulation. Therefore, enhanced temperature diffusion, using either EE or CM, does not appear to accelerate large-scale conformational transitions. The implication is that temperature diffusion is not limiting in driving large-scale conformational transitions even with the standard REMD. We note that similar observations have also been made by Kouza and Hansmann in their recent study of so-called rejection free replica exchange simulations[49].

Theoretical studies based on two-state model systems have predicted that REMD cannot drive transitions faster than the maximal rates at all the temperatures sampled (which should occur at slightly above T_m for actual proteins)[13]. Indeed, the average time per conformational transition, τ_{TS}, is ~30 μs or more in all REMD simulations, which over twice longer than τ_{TS} of ~14.3 μs observed in regular MD at T_m (see Table 1). We note that a key advantage of REMD, however, is the ability to generate correct thermodynamic dynamic ensembles at all temperatures, such that conformations sampled at all temperatures can be combined together using WHAM[50] to calculate thermodynamic properties any temperature of interest (e.g., the lowest temperature sampled). In contrast, recovering thermodynamic properties at the room temperature using a single simulation at T_m is generally unreliable. Therefore, slower average τ_{TS} in REMD compared to MD at T_m does not suggest that REMD is less efficient than regular MD (the contrary has been shown to be true extensively by previous works as discussed in Introduction). We also note that at 300 K no reversible folding and unfolding transition was sampled in multiple 100-μs MD runs and REMD is thus necessary for efficient generation of converged ensembles.

3.4 Convergence of thermodynamic properties

Arguably, the ultimate goal of REMD simulations is to generate statistically representative structure ensembles such that well-converged thermodynamic properties can be derived. Here, we focus on several typical thermodynamic properties frequently involved in protein folding studies, including T_m, C_V, folding stability (ΔG_{fold}, "zero-dimensional" free energy), 1D PMF, and several 2D PMFs. Importantly, both C_V and 1D PMF as a function of the total native contact fraction (Q) are well-converged among all REMD simulations (e.g., see Fig. 3). ΔG_{fold} values at $T_m \approx 354$K derived from all REMD and MD simulations are within 0.25 kT from each other (except the CM-REMD control run, which deviates by ~0.5 kT from other runs). Sample 2D PMFs, calculated from the control EE-REMD run, are shown in Fig. 6. We note that inclusion of non-specific hydrophobic interactions does not change the fundamental features of these free energy surfaces compared to the original sequence-flavored model[35]. For example, the folding of the C-terminal hairpin precedes that of the N-terminal hairpin (see Fig. 6D). However, all the free energy surfaces are significantly more rugged with several local minima that are not present in the surfaces derived from the original model (e.g., see Fig. 6B).

Fig. 7 compares the self-convergence of 1D PMF and a representative 2D free energy (Q vs. Q-nt) as a function of simulation time, measured using the root mean square deviation (RMSD) from the final profiles calculated using all 100 μs data. A few observations can be made. First, among all three simulations initiated from the folded state (EE control, CM control and standard REMD), both EE- and CM-REMD runs appear to converge faster compared to the standard REMD. For example, the

1D PMF quickly converged within about 5 μs in both EE- and CM-REMD control runs while it took ~15 μs to achieve a similar level of convergence for the standard REMD (Fig. 7A). Similar observations were also made in 2D PMFs (e.g., see Fig. 7B). Importantly, the apparent faster convergence of EE- and CM-REMD control simulations within short timescales does not appear to arise from more efficient conformational sampling. Instead, it is mainly attributed to faster mixing of conformation states sampled at different temperatures due to more efficient temperature diffusion in EE- and CM-REMD. Second, comparing the results of the control and folding runs suggests that the true convergence of thermodynamic properties is much slower than the apparent rates of self-convergence in the control runs. Interestingly, EE-REMD appears to be slightly more efficient than CM-REMD despite less efficient temperature diffusion, which is consistent with the larger number of folding/unfolding transitions sampled. Third, the convergence of higher dimensional PMFs expectedly is much slower. All three REMD protocols perform very similarly, with the RMSD from the final profiles gradually decreasing throughout the 100 μs simulation span.

To further investigate how the number of actual conformational transitions sampled affects the accuracy of calculated thermodynamic properties, we divided both EE and CM control simulations into uniform fragments of various lengths and calculated the average numbers of conformational transitions and standard deviations of calculated ΔG_{fold}. The results are summarized in Fig. 8. It shows a clear inverse correlation of the standard deviation of ΔG_{fold} and the number of conformational transitions sampled. Importantly, due to rapid mixing of conformations sampled at different temperatures (as noted above), reasonable estimates of ΔG_{fold}, with ~1 kT uncertainty, appear feasible even with "ultra-short" REMD simulations on the order of 100 ns (which is actually the typical length of atomistic REMD simulations), even though virtually no reversible transitions could be sampled in such a short simulation timespan. In addition, the uncertainty in the calculated ΔG_{fold} does not decrease significantly with longer simulations, until significant numbers of reversible folding/unfolding transitions are sampled. The later requires simulation timescales a few fold of the inherent folding/unfolding time, which is >20 μs for the current protein model. Along this line, it appears that equalized exchange achieved by EE-REMD is slightly advantageous (e.g., compare the black vs. red traces in Fig. 8), in allowing more rapid mixing of different temperatures to achieve better convergence with simulations much shorter than the folding/unfolding timescales (which is the case for most atomistic simulations).

4. Conclusions

Existence of sharp cooperative transitions in proteins can lead to temperature exchange bottleneck and subsequently limit the sampling efficiency of the popular REMD method. Several approaches

have been proposed to address this bottleneck, including two adaptive temperature REMD protocols previously examined by Lee and Olson[29]. In this work, we constructed a reasonably realistic yet computationally tractable protein model to re-evaluate how effective equalizing the exchange rates or maximizing the current of temperature diffusion can enhance the sampling of large-scale protein conformational transitions. The results demonstrate that, despite substantially enhanced temperature diffusion, neither EE- nor CM-REMD could generate significantly more folding/unfolding transitions. As the result, the convergence of key thermodynamic properties is similar with EE-, CM- or standard REMD, except for very short simulations. In the later case, the apparent convergence can be mainly attributed to mixing of conformations sampled at different temperatures and thus can benefit from enhanced temperature diffusion from either EE- or CM-REMD. We note that typical atomistic REMD simulations are on the oder of 100 ns and thus do fall into the "very short" category. A key lesson from the current work is also that, with exchange attempt frequency ($\sim ps^{-1}$) several orders of magnitude faster than the inherent timescale of slowest protein motions ($\sim \mu s^{-1}$), temperature diffusion does not appear to be limiting in the ability of REMD to drive large-scale conformational transitions. Therefore, it is unlikely that any protocol that aims to accelerate temperature exchange or temperature diffusion will lead to substantial enhancement in true sampling efficiency of REMD protein simulations. One will likely need to explore protocols that involve some types of guided tempering to allow the protein to take advantage of faster unfolding at high temperatures simultaneously with rapid folding near the transition temperature.

Acknowledgement

We greatly appreciate Dr. Michael S. Lee for sharing his Perl scripts for generating EE- and CM-REMD temperature profiles. This work was supported by the National Science Foundation (MCB 0952514) and the KSU Johnson Center for Basic Cancer Research. This work is contribution number 13-153-J from the Kansas Agricultural Experiment Station.

Supporting Information Available

Representative time series of temperature, energy, RMSD and the numbers of native and total contacts for a single replica. This information is available free of charge via the Internet at http://pubs.acs.org

References

(1) Liwo, A.; Czaplewski, C.; Oldziej, S.; Scheraga, H. A. *Curr. Opin. Struct. Biol.* **2008**, *18*, 134.
(2) Swendsen, R. H.; Wang, J. S. *Phys. Rev. Lett.* **1986**, *57*, 2607.
(3) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Japan* **1996**, *65*, 1604.
(4) Sugita, Y.; Okamoto, Y. *Chem Phys Lett* **1999**, *314*, 141.

(5) Hansmann, U. H.; Okamoto, Y. *Curr. Opin. Struct. Biol.* **1999**, *9*, 177.

(6) Nymeyer, H.; Gnanakaran, S.; Garcia, A. E. In *Numerical Computer Methods, Pt D* 2004; Vol. 383, p 119.

(7) Earl, D. J.; Deem, M. W. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910.

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

(9) Gallicchio, E.; Levy, R. M.; Parashar, M. *J. Comput. Chem.* **2008**, *29*, 788.

(10) Roitberg, A. E.; Okur, A.; Simmerling, C. *J. Phys. Chem. B* **2007**, *111*, 2415.

(11) Predescu, C.; Predescu, M.; Ciobanu, C. V. *J. Phys. Chem. B* **2005**, *109*, 4189.

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

(13) Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 15340.

(14) Nymeyer, H. *J. Chem. Theory Comput.* **2008**, *4*, 626.

(15) Denschlag, R.; Lingenheil, M.; Tavan, P. *Chem. Phys. Lett.* **2008**, *458*, 244.

(16) Zheng, W. H.; Andrec, M.; Gallicchio, E.; Levy, R. M. *J. Phys. Chem. B* **2008**, *112*, 6083.

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

(18) Sindhikara, D.; Meng, Y. L.; Roitberg, A. E. *J. Chem. Phys.* **2008**, *128*.

(19) Rao, F.; Caflisch, A. *J. Chem. Phys.* **2003**, *119*, 4035.

(20) Li, X. F.; O'Brien, C. P.; Collier, G.; Vellore, N. A.; Wang, F.; Latour, R. A.; Bruce, D. A.; Stuart, S. J. *J. Chem. Phys.* **2007**, *127*.

(21) Ballard, A. J.; Jarzynski, C. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 12224.

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

(23) Cheng, X.; Cui, G.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2005**, *109*, 8220.

(24) Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2006**, *2*, 420.

(25) Huang, X. H.; Hagen, M.; Kim, B.; Friesner, R. A.; Zhou, R. H.; Berne, B. J. *J. Phys. Chem. B* **2007**, *111*, 5405.

(26) Kofke, D. A. *J. Chem. Phys.* **2004**, *121*, 1167.

(27) Chen, J. H.; Im, W. P.; Brooks, C. L. *J. Am. Chem. Soc.* **2006**, *128*, 3728.

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

(29) Lee, M. S.; Olson, M. A. *J. Chem. Phys.* **2011**, *134*, 244111.

(30) Chen, J. H.; Brooks, C. L. *Phys. Chem. Chem. Phys.* **2008**, *10*, 471.

(31) Click, T. H.; Ganguly, D.; Chen, J. *Int. J. Mol. Sci.* **2010**, *11*, 5292.

(32) Chen, J. *Arch. Biochem. Biophys.* **2012**, *524*, 123.

(33) Wolynes, P. G. *Q. Rev. Biophys.* **2005**, *38*, 405.

(34) Hills, R. D.; Brooks, C. L. *Int. J. Mol. Sci.* **2009**, *10*, 889.

(35) Karanicolas, J.; Brooks, C. L. *Protein Sci.* **2002**, *11*, 2351.

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

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

(38) Katzgraber, H. G.; Trebst, S.; Huse, D. A.; Troyer, M. *J. Stat. Mech.-Theory E.* **2006**, P03018.

(39) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623.

(40) Karanicolas, J.; Brooks, C. L. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 3954.

(41) Hills, R. D.; Brooks, C. L. *J. Mol. Biol.* **2008**, *382*, 485.

(42) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Am. Chem. Soc.* **1999**, *121*, 2337.

(43) Feig, M.; Karanicolas, J.; Brooks, C. L. *J. Mol. Graph. Modell.* **2004**, *22*, 377.

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

(45) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545.

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

(47) Cho, S. S.; Levy, Y.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 586.

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

(49) Kouza, M.; Hansmann, U. H. *J. Chem. Phys.* **2011**, *134*, 044124.

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

Tables

Table 1. Summary of production REMD and MD simulations.

Method	Initial State	Length (μs)	T_m (K)	N_{TR}	N_{TS}	τ_{TS} (μs)
EE	Folded	100	351	4423	28	28.6
	Extended	100	352	4003	24	33.3
CM	Folded	100	354	18533	18	44.4
	Extended	100	352	19570	23	34.8
Standard	Folded	100	352	1411	21	38.1
MD @ 354K	Folded	200	–	–	14	14.3
MD @ 300K	Folded	100	–	–	0	Undef

T_m values shown were estimated based on the peak of the final calculated C_V curves. N_{TR} is the total number of round trips that all replicas travel between temperature extremes.

Figures

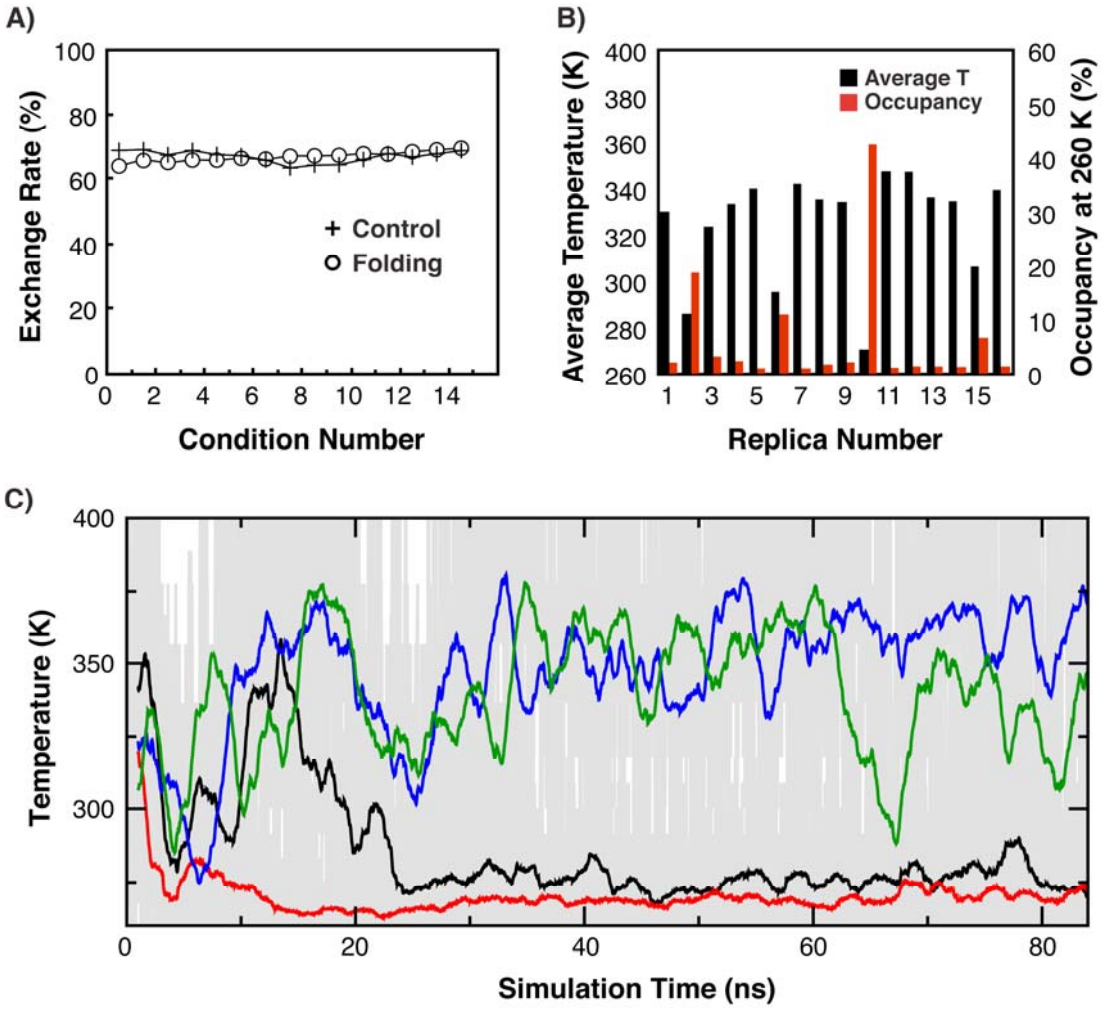

Figure 1 REMD simulations of GB1p β-hairpin. A). Exchange acceptance rates between neighboring temperatures for the control (initiated from the native structure) and folding (from a fully extended structure) REMD simulations. Condition numbers correspond to different simulation temperatures. B). Average temperatures and percentages of occupancy at the lowest temperature (260 K) of all replicas in the folding REMD simulation. C). 50-ps running averages of the temperature history of four representative replicas (2: black, 10: red, 12: blue, 16: green) during the folding REMD simulation. The grey traces in the background are the raw time traces. Details of these simulations were described in Reference [27].

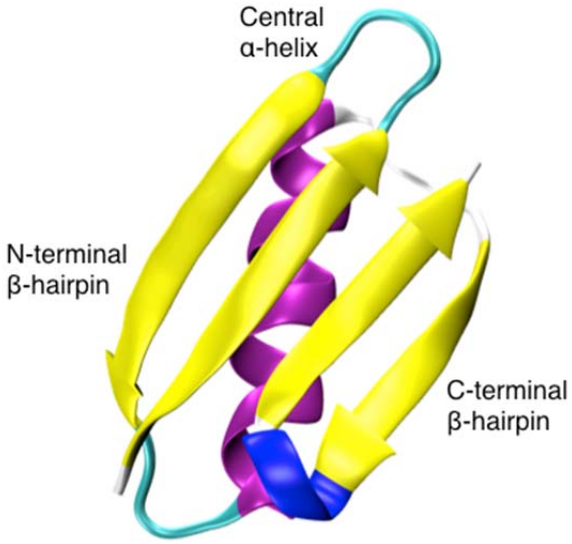

Figure 2. The structure of B1 domain of streptococcal protein G (PDB ID: 3GB1)[42]. The center helix is colored in violet, β-strands in yellow, and loops in cyan.

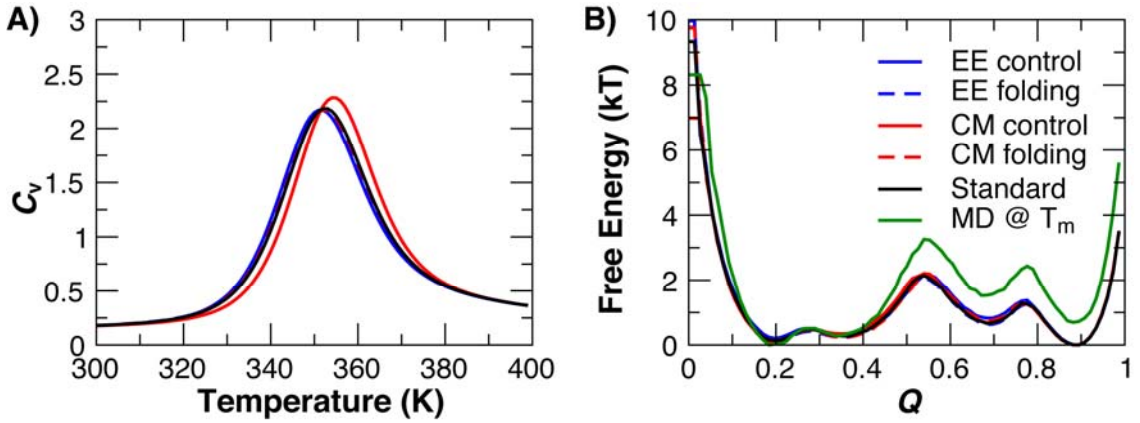

Figure 3. A) The heat capacity, C_V, as a function of temperature and B) free energy as a function of the native contact fraction (Q) at T_m = 354 K, calculated from various 100-μs EE- and CM-REMD and standard REMD simulations. The PMF derived from a 200-μs constant temperature MD at T_m = 354K is also shown (green trace in panel B).

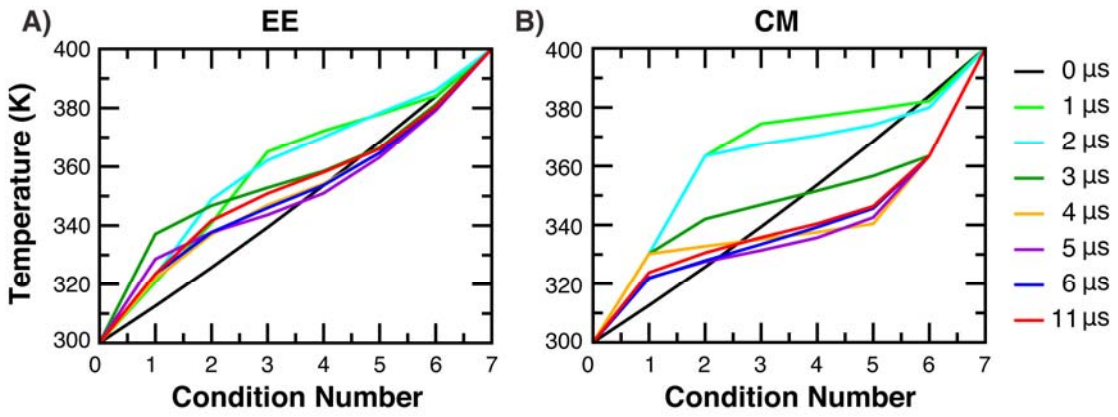

Figure 4. Temperature distributions optimized by **A)** EE and **B)** CM at the end of multiple cycles of 1-μs and 5-μs pilot REMD runs (see Methods).

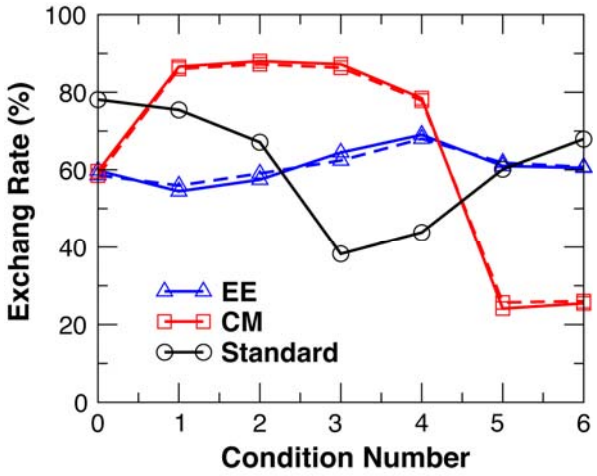

Figure 5. Exchange acceptance rates extracted from the 100-μs production REMD simulations. Solid and dash lines correspond to results from the control and folding simulations, respectively.

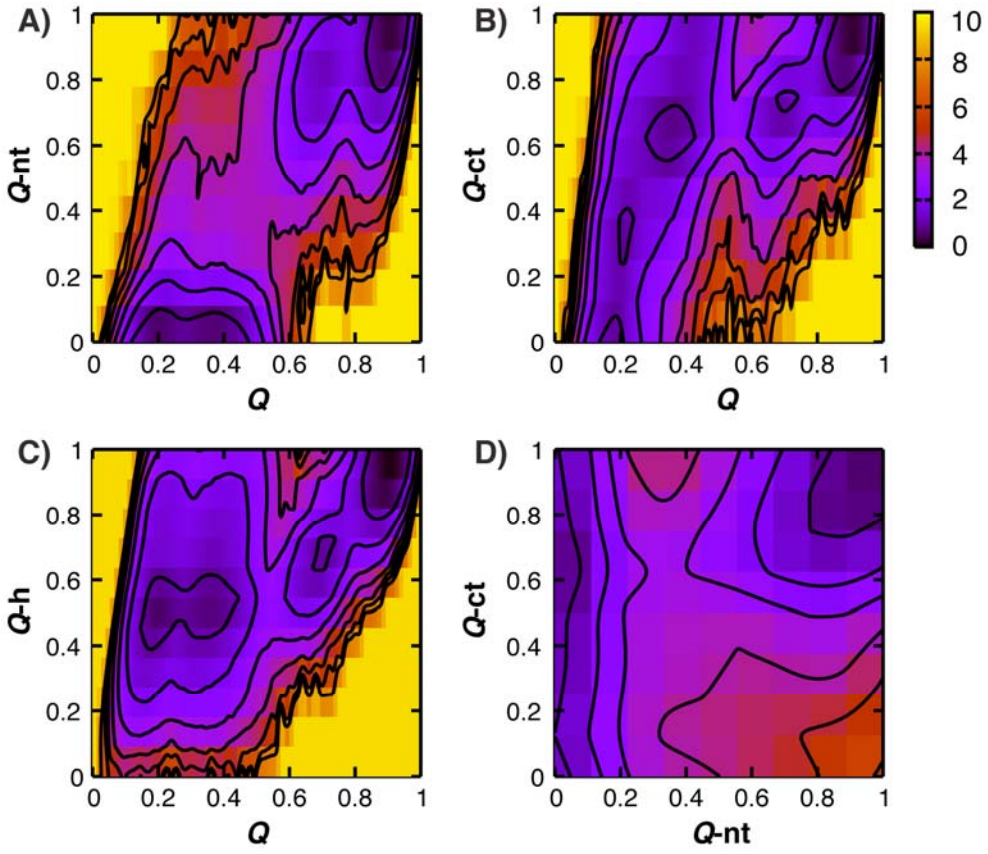

Figure 6. 2D free energy surfaces of folding of protein GB1 along various order parameters at T_m calculated from control EE-REMD simulations. Q, Q-nt, Q-ct and Q-h are fraction of native contacts of the whole protein, the N-terminal hairpin, the C-terminal hairpin and central helix respectively.

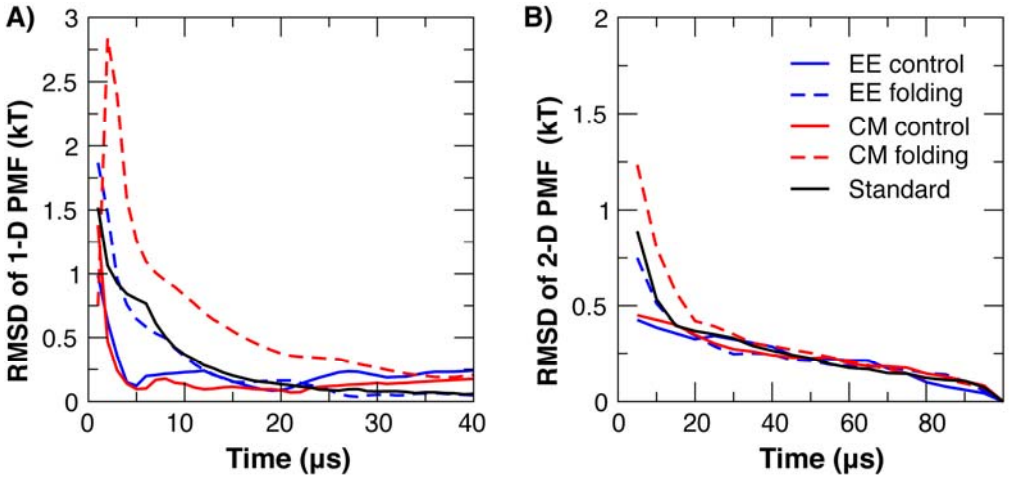

Figure 7. Self-convergence of A) 1D PMF (as a function of *Q*), and B) a representative 2D free energy surface (as a function of *Q* and *Q*-nt) during various 100- μs REMD simulations. The RMSD values shown were calculated with reference to the final 1D or 2D free energy profiles extracted using all 100 μs data of corresponding REMD runs.

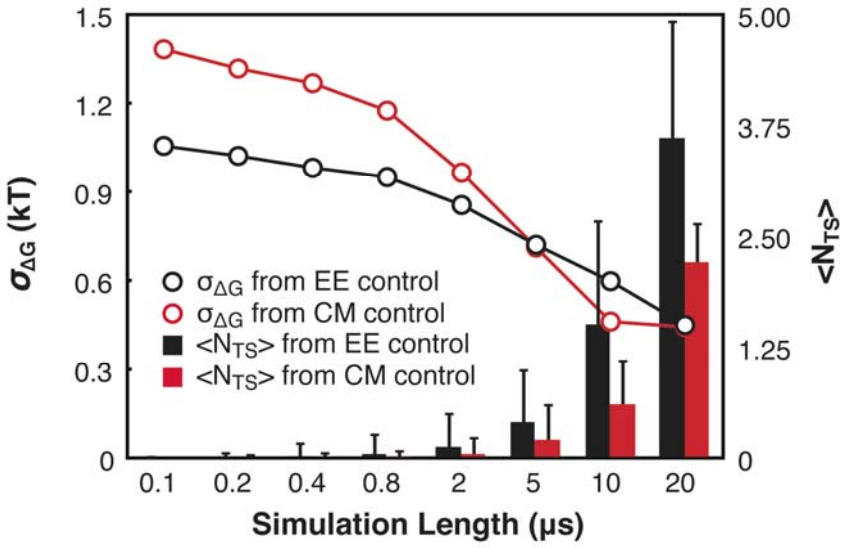

Figure 8. Standard deviation of calculated folding free energy ($\sigma_{\Delta G}$) and average number of conformational transitions ($<N_{TS}>$) calculated using various length of EE- and CM-REMD control simulations.

Supporting Materials

"Efficiency of adaptive temperature-based replica exchange for sampling large-scale protein conformational transitions", Weihong Zhang and Jianhan Chen

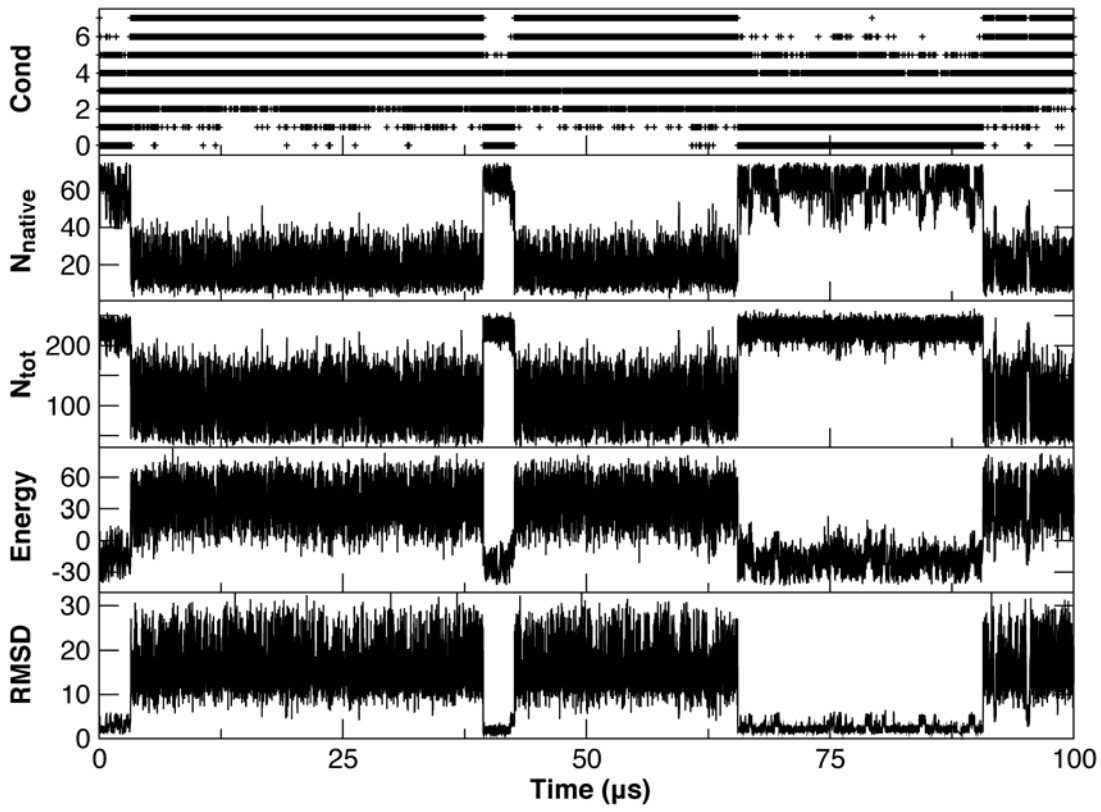

Figure S1: Time series of the temperature (condition number), total energy, RMSD and the numbers of native and total (nonspecific) contacts for a single replica (replica 1 from the EE control run; see main text). Nonspecific contacts are considered formed whenever the CA-CA distance is below 10 Å.